Stochastic Mirror Descent on Overparameterized Nonlinear Models

Navid Azizan, Sahin Lale, and Babak Hassibi

***Abstract*—Most modern learning problems are highly overparameterized, i.e., have many more model parameters than the number of training data points. As a result, the training loss may have infinitely many global minima (parameter vectors that perfectly "interpolate" the training data). It is therefore imperative to understand which interpolating solutions we converge to, how they depend on the initialization and learning algorithm, and whether they yield different test errors. Here we study these questions for the family of *stochastic mirror descent* (SMD) algorithms, of which stochastic gradient descent (SGD) is a special case. Recently, it has been shown that for overparameterized *linear* models, SMD converges to the closest global minimum to the initialization point, where closeness is in terms of the Bregman divergence corresponding to the potential function of the mirror descent. With appropriate initialization, this yields convergence to the minimum-potential interpolating solution, a phenomenon referred to as *implicit regularization*. On the theory side, we show that for sufficiently overparameterized *nonlinear* models, SMD with a (small enough) *fixed* step size converges to a global minimum that is "very close" (in Bregman divergence) to the minimum-potential interpolating solution, thus attaining *approximate implicit regularization*. On the empirical side, our experiments on the MNIST and CIFAR-10 datasets consistently confirm that the above phenomenon occurs in practical scenarios. They further indicate a clear difference in the generalization performances of different SMD algorithms: experiments on the CIFAR-10 dataset with different regularizers, ℓ_1 to encourage sparsity, ℓ_2 (SGD) to encourage small Euclidean norm, and ℓ_∞ to discourage large components, surprisingly show that the ℓ_∞ norm consistently yields better generalization performance than SGD, which in turn generalizes better than the ℓ_1 norm.**

***Index Terms*—Mirror descent, stochastic gradient descent, overparameterization, implicit regularization.**

I. Introduction

DEEP learning has demonstrably enjoyed a great deal of success in a wide variety of tasks [1]–[7]. 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 [8]. In other words, these highly overparameterized models can "interpolate" the training data, so much so that this regime has been called the "interpolating regime" [9]. 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 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 [8], [10]. More remarkably, even in the absence of any explicit regularization, dropout, or early stopping [8], the global minima obtained by these algorithms seem to generalize quite well (contrary to some other "bad" global minima). It has been also observed that even among different optimization algorithms, i.e., SGD and its adaptive variants, there is a discrepancy in the solutions achieved by different algorithms and how they generalize [11].

In this paper, we propose training deep neural networks with the family of *stochastic mirror descent* (SMD) algorithms, which is a generalization of the popular SGD. For any choice of potential function, there is a corresponding mirror descent algorithm. In particular, to see whether these algorithms lead to different minima and generalize differently, we train a standard ResNet-18 architecture on the popular CIFAR-10 dataset using SMD with a few different potential functions: ℓ_1 norm, ℓ_2 norm (SGD), and ℓ_∞ norm.[1] In all the cases, we train the network with a sufficiently-small fixed step size until we converge to an interpolating solution (global minimum). Comparisons between the histograms of these different global minima show that they are vastly different. In particular, the solutions obtained by ℓ_1-SMD are much sparser, and, on

This work was supported in part by the National Science Foundation under grant ECCS-1509977, by a grant from Qualcomm Inc., by NASA's Jet Propulsion Laboratory through the President and Director's Fund, and by fellowships from Amazon Web Services Inc. and PIMCO, LLC. This paper was presented in part at the 2019 International Conference on Machine Learning (ICML) Generalization Workshop, Long Beach, CA, USA.

N. Azizan was with the Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA, 91125 USA. He is now with the Department of Mechanical Engineering and the Institute for Data, Systems, and Society (IDSS), Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: azizan@mit.edu)

S. Lale and B. Hassibi are with the Department of Electrical Engineering, California Institute of Technology, Pasadena, CA, 91125 USA (e-mail: alale@caltech.edu; hassibi@caltech.edu)

Manuscript received June 8, 2020; revised December 7, 2020 and May 10, 2021; accepted May 24, 2021.

Digital Object Identifier 10.1109/TNNLS.2021.3087480

[1]Since the potential function needs to be differentiable and strictly convex, and ℓ_1 and ℓ_∞ norms are not, we use $\ell_{1+\epsilon}$ and ℓ_N norms for a sufficiently small ϵ and a sufficiently large N instead (See Section III).

the contrary, the solutions obtained by ℓ_∞ have virtually no zero components while having a smaller maximum. More importantly, there is a clear gap in the generalization performance of these algorithms. In fact, surprisingly and somewhat counterintuitively, the solutions obtained by ℓ_∞-norm SMD (which uses all the parameters in the already-highly-overparameterized network) consistently generalize better than the one obtained by SGD, which in turn outperform the sparser one obtained by ℓ_1-norm SMD. This begs the question:

Which global minima do these algorithms converge to, and what properties do they have?

On the theory side, we show that, for overparameterized nonlinear models, if the model is sufficiently overparameterized so that the random initialization point is close to the manifold of interpolating solutions (something that is occasionally referred to as the "blessing of dimensionality"), then the SMD algorithm for any particular potential function converges to a global minimum that is approximately ***the closest one to the initialization, in Bregman divergence corresponding to the potential.*** 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 corroborated by extensive experiments, further implies that, when initialized around zero, SGD converges to a solution that has almost the smallest Euclidean norm, thus acting as an approximate ℓ_2-norm regularizer. More generally, ***when initialized at the minimizer of the potential, SMD with any potential function ψ converges to a solution that has almost the smallest potential ψ.*** For instance, when initilized around zero, the solution obtained by SMD with ℓ_1-norm potential is approximately the minimum ℓ_1-norm one, which explains why its weights are much sparser. Similarly, the solution obtained by SMD with the ℓ_∞-norm potential has an ℓ_∞-norm regularization, which explains why the maximum of the weights is much smaller in this case.

II. Background

A. Preliminaries

Let us denote the training dataset by $\{(x_i, y_i) : i = 1, \dots, n\}$, 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 $f(x_i, w) = f_i(w)$ 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

$$\mathcal{W} = \{w \in \mathbb{R}^p \mid f(x_i, w) = y_i, i = 1, \dots, n\}. \tag{1}$$

This is a high-dimensional set (e.g., a $(p-n)$-dimensional manifold) in $\mathbb{R}^p$ and depends only on the training data $\{(x_i, y_i) : i = 1, \dots, n\}$ 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 $\ell(y_i, f(x_i, w)) = \ell(y_i - f(x_i, w))$, with $\ell(\cdot)$ convex and having a global minimum at zero (such as square loss, Huber loss, etc.). In this case, $L(w) = \sum_{i=1}^n \ell(y_i - f(x_i, w))$. For example, the conventional gradient descent (GD) algorithm can be used as an attempt to minimize $L(\cdot)$ over w.

B. Stochastic Mirror Descent

An important generalization of GD is the *mirror descent* (MD) algorithm, which was first introduced by Nemirovski and Yudin [12] and has been widely used since then [13]–[16]. Consider a strictly-convex differentiable function $\psi(\cdot)$, called the *potential function*. Then MD is given by the following recursion

$$\nabla\psi(w_i) = \nabla\psi(w_{i-1}) - \eta\nabla L(w_{i-1}), \quad w_0 \tag{2}$$

where $\eta > 0$ is known as the step size or learning rate. Note that, due to the strict convexity of $\psi(\cdot)$, the gradient $\nabla\psi(\cdot)$ defines an invertible map so that the recursion in (2) yields a unique w_i at each iteration, i.e., $w_i = \nabla\psi^{-1}\left(\nabla\psi(w_{i-1}) - \eta\nabla L(w_{i-1})\right)$. Compared to classical GD, 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 MD reduces to GD when $\psi(w) = \frac{1}{2}\|w\|^2$, since the gradient is simply the identity map.

Alternatively, the update rule (2) can be expressed as

$$w_i = \arg\min_w \ \eta w^T \nabla L(w_{i-1}) + D_\psi(w, w_{i-1}), \tag{3}$$

where

$$D_\psi(w, w_{i-1}) := \psi(w) - \psi(w_{i-1}) - \nabla\psi(w_{i-1})^T(w - w_{i-1}) \tag{4}$$

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 [17].

p-norms Algorithm. For any q-norm squared potential function $\psi(w) = \frac{1}{2}\|w\|_q^2$, with $\frac{1}{p} + \frac{1}{q} = 1$, the algorithm will reduce to the so-called p-norms algorithm [18], [19].

When n is large, computation of the entire gradient may be cumbersome. Alternatively, in online scenarios, the entire loss function $L(\cdot)$ may not be available, and only the local loss functions may be provided at each iteration. In such settings, a stochastic version of MD has been introduced, aptly called *stochastic mirror descent* (SMD), which can be considered the straightforward generalization of stochastic gradient descent (SGD):

$$\nabla\psi(w_i) = \nabla\psi(w_{i-1}) - \eta\nabla L_i(w_{i-1}), \quad w_0. \tag{5}$$

The instantaneous loss functions $L_i(\cdot)$ can be either drawn at random or cycled through periodically.

III. Training Deep Neural Networks with SMD

As mentioned earlier, the heavy overparameterization in typical deep neural networks means that the loss function for such architectures typically has infinitely many global minima, and these different minima can have very different properties and generalization performances. Motivated by this fact, we propose training deep neural networks with stochastic mirror descent algorithms, to see if they lead to different global minima and different generalization performances.

In particular, we propose training deep neural networks with SMD with potential function $\psi(w) = \frac{1}{q}\|w\|_q^q$, which can be expressed as:

$$w_i[j] = \left||w_{i-1}[j]|^{q-1}\text{sign}(w_{i-1}[j]) - \eta\nabla L_i(w_{i-1})[j]\right|^{\frac{1}{q-1}} \times \text{sign}\left(|w_{i-1}[j]|^{q-1}\text{sign}(w_{i-1}[j]) - \eta\nabla L_i(w_{i-1})[j]\right), \tag{6}$$

where $w_i[j]$ denotes the j-th element of the w_i vector.

Note that, for this particular choice of potential function, the update rule is *separable*, i.e., the j-th element of the new weight vector can be computed using only the j-th element of the weight and gradient vectors. This allows for efficient, parallel and distributed implementation of the algorithm, which is highly desirable for large-scale learning tasks.

We should also remark that the computational complexity of the ℓ_q-norm SMD is of the same order as that of the usual SGD. In other words, it is linear in the number of weights, which, again, can also be parallelized in the same way as SGD.

In addition, the storage complexity of the algorithm is exactly the same as the usual SGD. All that is stored are the weights.

A. An Experiment

We take the popular CIFAR-10 dataset and the standard ResNet-18 architecture, commonly used for this dataset. We initialize the network with random weights around zero, as usual, and train it with the ℓ_q-norm SMD for a few different values of k. In particular, we use: $\ell_{1+\epsilon}$ norm, ℓ_2 norm (SGD), ℓ_3 norm, ℓ_8 norm, ℓ_{10} norm, and ℓ_{14} norm, where $\ell_{1+\epsilon}$ is a surrogate for ℓ_1 norm, and the higher norms are surrogates for the ℓ_∞ norm. In all the cases, we choose the step size to be sufficiently small and train for a sufficiently large number of steps until we converge to an interpolating solution (global minimum).

We compare the generalization performance of these different solutions on the test set. Fig. 1 shows the test errors of the solutions. As can be seen, there is a clear gap in the generalization performance of the algorithms: SMD with higher-norms consistently outperforms SGD, which in turn performs better than the SMD with ℓ_1-norm. In fact, perhaps surprisingly, by virtue of changing the optimizer from SGD to these high-norm SMDs, without any additional tricks, we outperform the state of the art for ResNet-18 on CIFAR-10. This is particularly remarkable given that this very architecture had been designed with training with SGD in mind.

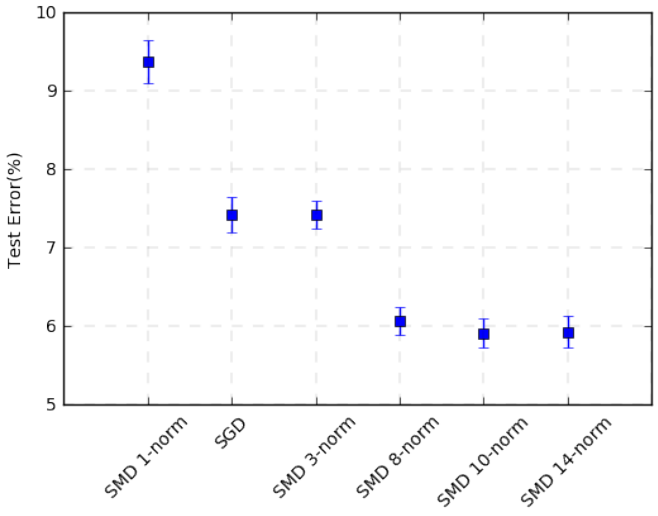


Fig. 1. Generalization performance of different SMD algorithms on the CIFAR-10 dataset using the ResNet-18 neural network. SMDs with higher norms (which are surragotes for ℓ_∞ norm) tend to achieve better generalization performance (lower test error) than the ones with smaller norms. In particular, ℓ_{14} consistently outperforms SGD (state-of-the-art), while ℓ_1-SMD performs worse than both.

One may be curious to see how different the weights obtained by different algorithms look. Fig. 2 shows the histogram of the absolute value of the weights for four different SMDs, initialized by the exact *same* set of weights. The histograms of the final weights look substantially different, and, since they all started from the same initial weights and they all interpolate the same dataset, this difference is fully attributable to the mirrors used. Remarkably, the histogram of the ℓ_1-SMD has more weights at and around zero, i.e., it is very sparse. The histogram of the ℓ_2-SMD (SGD) looks almost perfectly

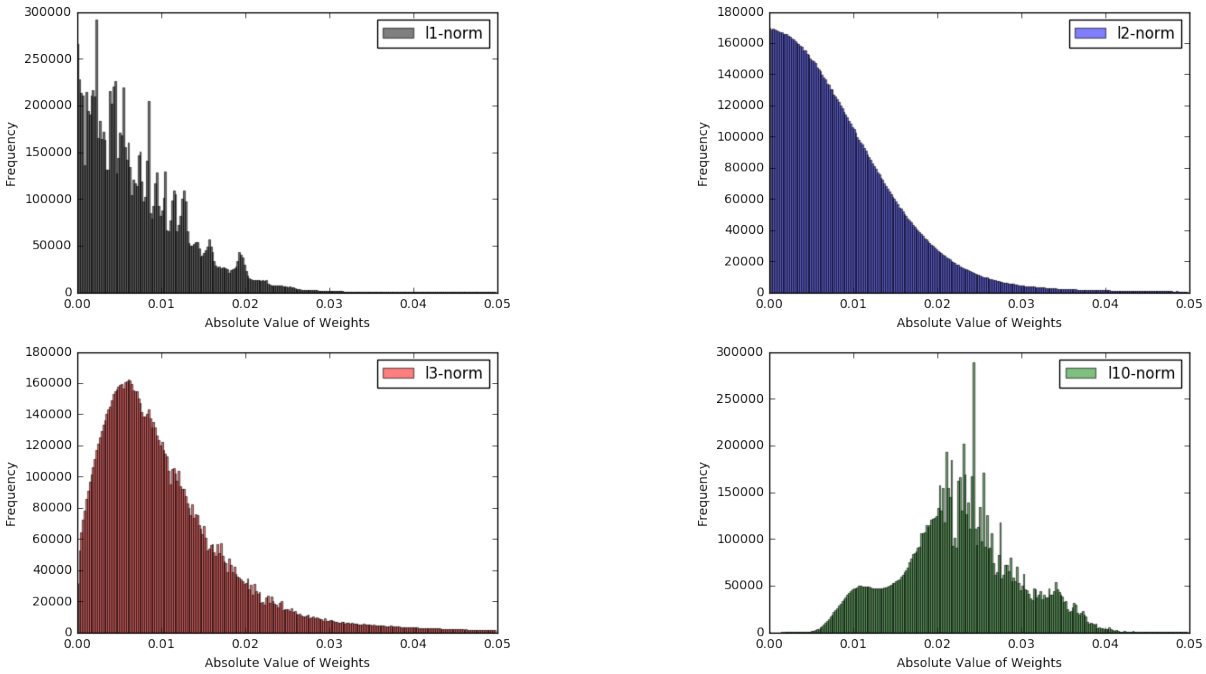


Fig. 2. Histogram of the absolute value of the final weights in the network for different SMD algorithm with different potentials. Note that each of the four histograms corresponds to an 11×10^6-dimensional weight vector that perfectly interpolates the data. Even though the weights remain quite small, the histograms are drastically different. ℓ_1-SMD induces sparsity on the weights. SGD appears to lead to a Gaussian distribution on the weights. ℓ_3-SMD starts to reduce the sparsity, and ℓ_{10} shifts the distribution of the weights significantly, so much so that almost all the weights are non-zero.

Gaussian. The one corresponding to ℓ_3 has somewhat shifted to the right, and the ℓ_∞ has completely moved away from zero (i.e., all the components are non-zero) while having no "tail." The fact that the ℓ_∞ solution, which uses all the parameters in the already-highly-overparameterized network, generalizes better than the sparser ones is quite remarkable.

IV. Theoretical Results

In this section, we provide a theoretical analysis of what different SMD algorithms converge to. In particular, we show that for highly overparameterized models, under certain assumptions: (1) SMD converges to a global minimum and (2) the global minimum obtained by SMD is approximately the closest one to the initialization in Bregman divergence corresponding to the potential.

A. Warm-up: 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 that 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 $\mathcal{W} = \{w \mid y_i = x_i^T w,\ i = 1, \dots, n\}$, and the loss is $L_i(w) = \ell(y_i - x_i^T w)$. The following result characterizes the solution that SMD converges to [20], [21].

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*

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

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 $w_0 = \arg\min_{w \in \mathbb{R}^p} \psi(w)$, under the conditions of Proposition 1, the SMD iterates converge to*

$$w_\infty = \arg\min_{w \in \mathcal{W}} \psi(w). \tag{7}$$

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 ℓ_2-norm regularizer. In this section, we show that these results continue to hold for highly overparameterized nonlinear models in an approximate sense.

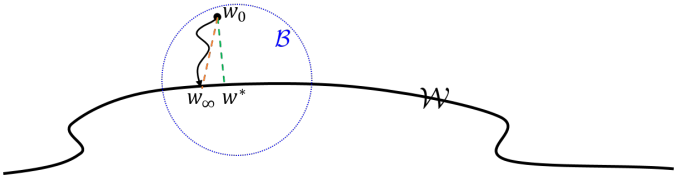


Fig. 3. An illustration of the parameter space. $\mathcal{W}$ represents the set of global minima, w_0 is the initialization, $\mathcal{B}$ is the local neighborhood, w^* is the closest global minimum to w_0 (in Bregman divergence), and w_∞ is the minimum that SMD converges to.

B. Main Results

Let us define

$$D_{L_i}(w, w') := L_i(w) - L_i(w') - \nabla L_i(w')^T (w - w'), \quad (8)$$

which is defined in a similar way to a Bregman divergence for the loss function. The difference, though, is that, due to the nonlinearity of $f(\cdot, \cdot)$, 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).

It has been argued in several recent papers that in highly overparameterized neural networks, because $\mathcal{W}$ is very high-dimensional, any random initialization w_0 is close to it, with high probability [20], [22]–[25] (see the discussion in Appendix A (Section A-B)). In such settings, it is reasonable to make the following assumption about the manifold.

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. 4 for an illustration). What we are requiring is that the initialization w_0 be inside the intersection of all such neighborhoods for $i = 1, \ldots, n$. In other words, we require a w_0 close enough to $\mathcal{W}$. The ϵ in Assumption 1 characterizes the closeness.

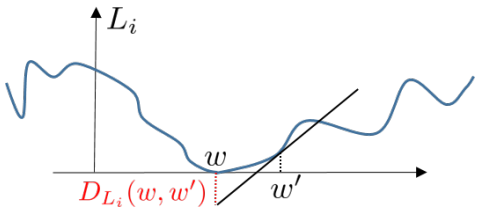


Fig. 4. An illustration of $D_{L_i}(w, w') \geq 0$ in a local region in Assumption 1.

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 $w' \in \operatorname{conv} \mathcal{B}$.*

This is a mild assumption, which is assumed in other related work such as [26] as well. Note that we do *not* require α 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_∞).*
3) *$w_\infty \in \mathcal{W}$.*

In other words, we converge to a global minimum (interpolating solution). The convergence is "local" in the sense that Assumption 1 has to be met. However, as argued earlier, that is not an unreasonable assumption in highly overparameterized settings. 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, one 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 just small enough to guarantee the strict convexity of $\psi(\cdot) - \eta L_i(\cdot)$ inside $\mathcal{B}$, and not globally.

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

$$w^* = \arg\min_{w \in \mathcal{W}} D_\psi(w, w_0). \quad (9)$$

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_∞ which is "very close" to w^*.

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

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

In other words, if we start with an initialization that is $O(\epsilon)$ away from $\mathcal{W}$ (in Bregman divergence), we converge to a point $w_\infty \in \mathcal{W}$ 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, $w^* = \arg\min_{w \in \mathcal{W}} \psi(w)$ and the following holds:*

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

C. Fundamental Identity of SMD

An important tool used in our proofs is a "fundamental identity" that governs the behavior of the iterates of SMD, which holds under very general conditions.

TABLE I

FIXED INITIALIZATION (THE SETTING DEPICTED IN FIG. 5). WE HAVE TRAINED THE NETWORK FROM A COMMON FIXED INITIALIZATION WITH 4 DIFFERENT SMDS (ℓ_1, ℓ_2, ℓ_3, AND ℓ_{10}) TO OBTAIN 4 DIFFERENT INTERPOLATING SOLUTIONS. FOR EACH INTERPOLATING SOLUTION, WE CAN COMPUTE ITS DISTANCE FROM THE INITIAL WEIGHT VECTOR. SINCE WE HAVE 4 DIFFERENT POTENTIALS, WE HAVE 4 DIFFERENT BREGMAN DIVERGENCES TO ASSESS THE DISTANCE BY. THIS GIVES US A 4-BY-4 TABLE. THE COLUMNS CORRESPOND TO THE 4 DIFFERENT INTERPOLATING SOLUTIONS (ONE FOR EACH SMD) AND THE ROWS CORRESPOND TO THE DIFFERENT BREGMAN DIVERGENCES. AS CAN BE SEEN, THE SMALLEST ENTRY IN EACH ROW IS THE ONE WHERE THE POTENTIALS CORRESPONDING TO THE ALGORITHM AND THE BREGMAN DIVERGENCE MATCH. IN OTHER WORDS, FOR EACH BREGMAN DIVERGENCE, THE CLOSEST INTERPOLATING SOLUTION TO THE INITIALIZATION IS THE ONE THAT IS OBTAINED FROM THE SMD CORRESPONDING TO THAT PARTICULAR BREGMAN DIVERGENCE.

	SMD 1-norm	SMD 2-norm (SGD)	SMD 3-norm	SMD 10-norm
1-norm BD	141	9.19×10^3	4.1×10^4	2.34×10^5
2-norm BD	3.15×10^3	562	1.24×10^3	6.89×10^3
3-norm BD	4.31×10^4	107	53.5	1.85×10^2
10-norm BD	6.83×10^{13}	972	7.91×10^{-5}	2.72×10^{-8}

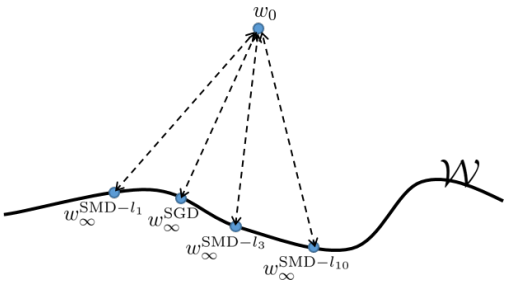


Fig. 5. An illustration of the experiments in Table I.

Lemma 6. *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 $\{w_i\}$*

$$D_\psi(w, w_{i-1}) = D_\psi(w, w_i) + D_{\psi-\eta L_i}(w_i, w_{i-1}) + \eta L_i(w_i) + \eta D_{L_i}(w, w_{i-1}), \quad (10)$$

for all $i \geq 1$.

This identity allows one to prove the results in a remarkably simple and direct way. The ideas behind it are related to H_∞ estimation theory [27], [28], which was originally developed in the 1990s in the context of robust control theory. In fact, it has connections to the minimax optimality of SGD, which was shown in [29] for linear models, and recently extended to nonlinear models and general mirrors in [20].

V. EXPERIMENTAL VALIDATION

In this section, we evaluate the theoretical claims of Section IV, by running extensive experiments for different initializations and different mirrors and computing the distances between each global minimum achieved and each initialization, in different Bregman divergences.

The theoretical results suggest that SMD converges to (almost) the closest point in the corresponding 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 4-layer convolutional neural network (CNN) on the MNIST dataset [30], and the ResNet-18 [31] on the CIFAR-10 dataset [32]. We use cross-entropy loss as the loss function in our training. We train the models from different initializations, and with different SMDs from each particular initialization, until we reach 0 training error, i.e., a point on $\mathcal{W}$. We randomly initialize the parameters of the networks around zero with $\mathcal{N}(0, 0.0001)$ for the weights in the convolutional and batch-norm layers, and $\mathcal{U}(-0.01, 0.01)$ for the weights in the linear layers. We choose 6 independent initializations for the CNN, and 8 for ResNet-18, and for each initialization, we run different SMD algorithms defined by the norm potential function $\psi(w) = \frac{1}{q}\|w\|_q^q$ for the following values of q: (a) $q = 1+0.01$, as a surrogate for ℓ_1 norm, (b) $q = 2$, which is SGD, (c) $q = 3$, and (d) $q = 10$, as a surrogate for ℓ_∞ norm. We use a fixed step-size η, chosen small enough to avoid diverging. See Appendix B for more details on the experiments.

In all the cases, provided the learning rate was small enough, the algorithm converged to an interpolating solution. We measure the distances between the initializations and the global minima obtained from different mirrors and different initializations, in different Bregman divergences. Table I, and Table II show some examples among different mirrors and different initializations, respectively. Fig. 7 shows the 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, which follows what Theorems 3 and 4 predict. This trend is very consistent among all our experiments, which can be found in Appendix B.

It is worth emphasizing that there is virtually no additional overhead in training the networks with ℓ_q-norm SMD, compared to SGD. The computational and memory complexity of every iteration is the same. We empirically observed that larger values of q require smaller step sizes, and in fact, this is also what the theoretical condition on the step size suggests. For instance, we have the step sizes for SGD and ℓ_{10}-SMD as 10^{-2} and 10^{-9}, respectively. However, the number of iterations required for ℓ_{10} SMD is *not* significantly higher (1000 iterations, compared to 500 for SGD).

VI. PROOFS

In this section, we prove the main theoretical results discussed in Section IV.

TABLE II

FIXED POTENTIAL (THE SETTING DEPICTED IN FIG. 6). WE HAVE TRAINED THE NETWORK FROM 8 DIFFERENT INITIAL POINTS WITH THE SAME SMD (IN THIS CASE, SGD) TO OBTAIN 8 DIFFERENT INTERPOLATING SOLUTIONS. THE ROWS CORRESPOND TO THE INITIAL POINTS, THE COLUMNS CORRESPOND TO THE INTERPOLATING SOLUTIONS, AND EACH ENTRY IS THE DISTANCE BETWEEN THE TWO, ALL MEASURED IN THE SAME BREGMAN DIVERGENCE (IN THIS CASE, EUCLIDEAN). AS CAN BE SEEN, THE SMALLEST ENTRY IN EACH ROW IS THE ONE WHERE THE INITIAL POINT AND THE FINAL POINT MATCH. IN OTHER WORDS, THE CLOSEST FINAL POINT TO EACH INITIAL POINT i, AMONG ALL THE EIGHT FINAL POINTS, IS THE ONE OBTAINED BY THE ALGORITHM FROM THE INITIAL POINT i.

	Final 1	Final 2	Final 3	Final 4	Final 5	Final 6	Final 7	Final 8
Initial 1	6×10^2	2.9×10^3	2.8×10^3	2.8×10^3	2.8×10^3	2.8×10^3	2.8×10^3	2.8×10^3
Initial 2	2.8×10^3	6.1×10^2	2.8×10^3	2.8×10^3	2.8×10^3	2.8×10^3	2.8×10^3	2.8×10^3
Initial 3	2.8×10^3	2.9×10^3	5.6×10^2	2.8×10^3	2.8×10^3	2.8×10^3	2.8×10^3	2.8×10^3
Initial 4	2.8×10^3	2.9×10^3	2.8×10^3	5.9×10^2	2.8×10^3	2.8×10^3	2.8×10^3	2.8×10^3
Initial 5	2.8×10^3	2.9×10^3	2.8×10^3	2.8×10^3	5.7×10^2	2.8×10^3	2.8×10^3	2.8×10^3
Initial 6	2.8×10^3	2.9×10^3	2.8×10^3	2.8×10^3	2.8×10^3	5.6×10^2	2.8×10^3	2.8×10^3
Initial 7	2.8×10^3	2.9×10^3	2.8×10^3	2.8×10^3	2.8×10^3	2.8×10^3	6×10^2	2.8×10^3
Initial 8	2.8×10^3	2.9×10^3	2.8×10^3	2.8×10^3	2.8×10^3	2.8×10^3	2.8×10^3	5.8×10^2

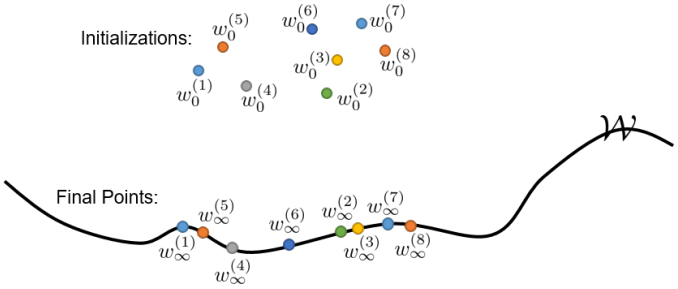


Fig. 6. An illustration of the experiments in Table II.

A. Convergence of SMD to the Interpolating Set

Let us first prove the convergence of SMD to the set of solutions.

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

Theorem 3. *Consider the set of interpolating parameters $\mathcal{W} = \{w \in \mathbb{R}^p \mid f(x_i, w) = y_i, i = 1, \dots, n\}$, 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 for all i, the following holds.*

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

Proof of Theorem 3. First we show that all the iterates will remain in $\mathcal{B}$. Recall the identity (10) from Lemma 6, which holds for all $w \in \mathcal{W}$. If w_{i-1} is in the region $\mathcal{B}$, 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 $D_{\psi - \eta L_i}(w_i, w_{i-1})$ 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

$$D_\psi(w, w_{i-1}) \geq D_\psi(w, w_i), \tag{11}$$

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

Next, we prove that the iterates converge and $w_\infty \in \mathcal{W}$. If we sum up the identity (10) for all $i = 1, \dots, T$, the first terms on the right- and left-hand side cancel each other telescopically, and we have

$$D_\psi(w, w_0) = D_\psi(w, w_T) + \sum_{i=1}^{T} [D_{\psi - \eta L_i}(w_i, w_{i-1}) + \eta L_i(w_i) + \eta D_{L_i}(w, w_{i-1})]. \tag{12}$$

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

$$\sum_{i=1}^{\infty} [D_{\psi - \eta L_i}(w_i, w_{i-1}) + \eta L_i(w_i) + \eta D_{L_i}(w, w_{i-1})] \leq D_\psi(w, w_0). \tag{13}$$

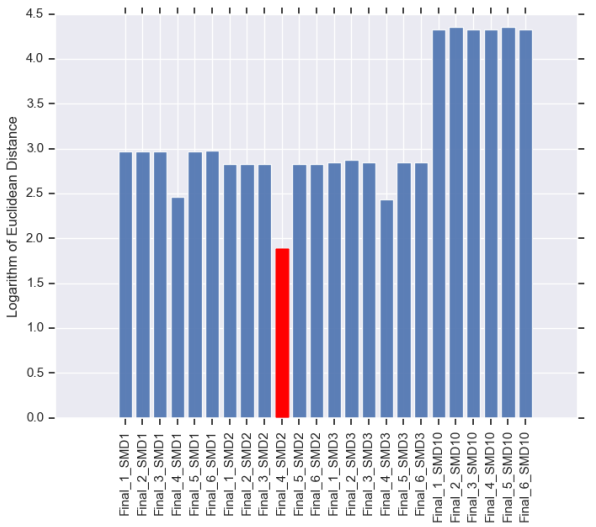
Since the step size is small enough that $\psi(\cdot) - \eta L_i(\cdot)$ is strictly convex for all i, the first term $D_{\psi - \eta L_i}(w_i, w_{i-1})$ is non-negative. The second term $\eta L_i(w_i)$ is non-negative because of the non-negativity of the loss. Finally, the last term $D_{L_i}(w, w_{i-1})$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

$$D_{\psi - \eta L_i}(w_i, w_{i-1}) \to 0, \text{ and } \eta L_i(w_i) \to 0. \tag{14}$$

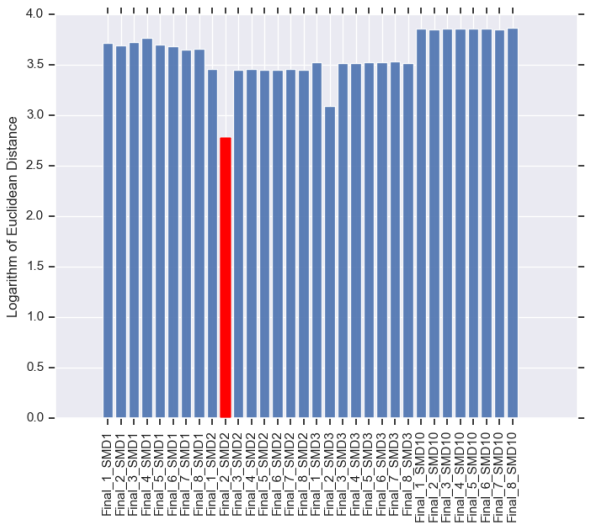
This implies convergence ($w_i \to w_\infty$), and that all the individual losses are going to zero. Since every data point is being revisited after some steps, all the data points are being fit. Therefore, $w_\infty \in \mathcal{W}$. □

B. Closeness of the Final Point to the Regularized Solution

Next, we show that with the additional Assumption 2 (which is roughly equivalent to $f_i(\cdot)$ having bounded Hessian in $\mathcal{B}$), not only do the iterates remain in $\mathcal{B}$ and converge to the set $\mathcal{W}$, 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 the fundamental identity of SMD.



MNIST. The (Euclidean) distance of different interpolating solutions from the initial point 4.



CIFAR-10. The (Euclidean) distance of different interpolating solutions from the initial point 4.

Fig. 7. We have trained the network from a few (6 for MNIST, and 8 for CIFAR-10) initial points with 4 different SMDs, to obtain a number of interpolating solutions (24 for MNIST, and 32 for CIFAR-10). The plot shows the distance between a particular initial point (initial point 2 for MNIST, and initial point 4 for CIFAR-10) and each of the interpolating solutions. The smallest distance, among all the interpolating solutions, corresponds exactly to the final point obtained from the particular 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 Appendix B).

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 $w' \in \operatorname{conv} \mathcal{B}$.*

Theorem 4. *Define $w^* = \arg\min_{w \in \mathcal{W}} D_\psi(w, w_0)$. 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. Recall the identity (10) from Lemma 6. Summing the identity for all $i \geq 1$, we have

$$D_\psi(w, w_0) = D_\psi(w, w_\infty) + \sum_{i=1}^{\infty} [D_{\psi - \eta L_i}(w_i, w_{i-1}) + \eta L_i(w_i) + \eta D_{L_i}(w, w_{i-1})]. \tag{15}$$

for all $w \in \mathcal{W}$. 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 $\mathcal{B}$, the dependence on w in the last term is "weak."

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

$$\begin{aligned} D_{L_i}(w, w_{i-1}) &= 0 - L_i(w_{i-1}) - \nabla L_i(w_{i-1})^T (w - w_{i-1}) \\ &= -L_i(w_{i-1}) \\ &\quad + \ell'(y_i - f_i(w_{i-1})) \nabla f_i(w_{i-1})^T (w - w_{i-1}) \end{aligned} \tag{16}$$

for all $w \in \mathcal{W}$, where the first equality comes from the definition of $D_{L_i}(\cdot, \cdot)$ and the fact that $L_i(w) = 0$ for $w \in \mathcal{W}$. 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

$$\begin{aligned} f_i(w) &= f_i(w_{i-1}) + \nabla f_i(w_{i-1})^T (w - w_{i-1}) \\ &\quad + \frac{1}{2}(w - w_{i-1})^T H_{f_i}(\hat{w}_i)(w - w_{i-1}), \end{aligned} \tag{17}$$

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

$$\begin{aligned} \nabla f_i(w_{i-1})^T (w - w_{i-1}) = y_i - f_i(w_{i-1}) - \frac{1}{2}(w \\ - w_{i-1})^T H_{f_i}(\hat{w}_i)(w - w_{i-1}), \end{aligned} \tag{18}$$

for all $w \in \mathcal{W}$. Plugging this into (16), we have

$$\begin{aligned} D_{L_i}(w, w_{i-1}) &= -L_i(w_{i-1}) \\ &\quad + \ell'(y_i - f_i(w_{i-1})) \Big(y_i - f_i(w_{i-1}) \\ &\quad - \frac{1}{2}(w - w_{i-1})^T H_{f_i}(\hat{w}_i)(w - w_{i-1}) \Big) \end{aligned} \tag{19}$$

for all $w \in \mathcal{W}$. Finally, by plugging this back into the identity (15), we have

$$\begin{aligned} D_\psi(w, w_0) = D_\psi(w, w_\infty) + \sum_{i=1}^{\infty} \Big[& D_{\psi - \eta L_i}(w_i, w_{i-1}) \\ & + \eta L_i(w_i) - \eta L_i(w_{i-1}) \\ & + \eta \ell'(y_i - f_i(w_{i-1}))(y_i - f_i(w_{i-1}) \\ & - \frac{1}{2}(w - w_{i-1})^T H_{f_i}(\hat{w}_i)(w - w_{i-1})) \Big]. \end{aligned} \tag{20}$$

for all $w \in \mathcal{W}$. Note that this can be expressed as

$$D_\psi(w, w_0) = D_\psi(w, w_\infty) + C - \sum_{i=1}^{\infty} \frac{1}{2}\eta \ell'(y_i - f_i(w_{i-1}))(w - w_{i-1})^T H_{f_i}(\hat{w}_i)(w - w_{i-1}), \tag{21}$$

for all $w \in \mathcal{W}$, where C does not depend on w:

$$C = \sum_{i=1}^{\infty} \left[D_{\psi - \eta L_i}(w_i, w_{i-1}) + \eta L_i(w_i) - \eta L_i(w_{i-1}) + \eta \ell'(y_i - f_i(w_{i-1}))(y_i - f_i(w_{i-1})) \right]. \tag{22}$$

From Theorem 3, we know that $w_\infty \in \mathcal{W}$. Therefore, by plugging it into equation (21), and using the fact that $D_\psi(w_\infty, w_\infty) = 0$, we have

$$D_\psi(w_\infty, w_0) = C - \sum_{i=1}^{\infty} \frac{1}{2}\eta \ell'(y_i - f_i(w_{i-1}))(w_\infty - w_{i-1})^T H_{f_i}(w_i')(w_\infty - w_{i-1}), \tag{23}$$

where w_i' is a point in the convex hull of w_∞ and w_{i-1} (and therefore also in $\operatorname{conv} \mathcal{B}$), for all i. Similarly, by plugging w^*, which is also in $\mathcal{W}$, into (21), we have

$$D_\psi(w^*, w_0) = D_\psi(w^*, w_\infty) + C - \sum_{i=1}^{\infty} \frac{1}{2}\eta \ell'(y_i - f_i(w_{i-1}))(w^* - w_{i-1})^T H_{f_i}(w_i'')(w^* - w_{i-1}), \tag{24}$$

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

$$\begin{aligned} &D_\psi(w_\infty, w_0) - D_\psi(w^*, w_0) \\ &= -D_\psi(w^*, w_\infty) + \sum_{i=1}^{\infty} \frac{1}{2}\eta \ell'(y_i - f_i(w_{i-1})) \left[(w^* - w_{i-1})^T H_{f_i}(w_i'')(w^* - w_{i-1}) - (w_\infty - w_{i-1})^T H_{f_i}(w_i')(w_\infty - w_{i-1}) \right]. \end{aligned} \tag{25}$$

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

$$\alpha \|w_\infty - w_{i-1}\|^2 \leq (w_\infty - w_{i-1})^T H_{f_i}(w_i')(w_\infty - w_{i-1}) \leq \beta \|w_\infty - w_{i-1}\|^2, \tag{26}$$

and

$$\alpha \|w^* - w_{i-1}\|^2 \leq (w^* - w_{i-1})^T H_{f_i}(w_i'')(w^* - w_{i-1}) \leq \beta \|w^* - w_{i-1}\|^2. \tag{27}$$

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

$$D_\psi(w_\infty, w_0) - D_\psi(w^*, w_0) = -D_\psi(w^*, w_\infty) + \sum_{i=1}^{\infty} \eta \ell'(y_i - f_i(w_{i-1}))O(\epsilon). \tag{28}$$

Now note that $\ell'(y_i - f_i(w_{i-1})) = \ell'(f_i(w) - f_i(w_{i-1})) = \ell'(\nabla f_i(\tilde{w}_i)^T(w - w_{i-1}))$ for some $\tilde{w}_i \in \operatorname{conv} \mathcal{B}$. 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

$$(D_\psi(w_\infty, w_0) - D_\psi(w^*, w_0)) + D_\psi(w^*, w_\infty) = o(\epsilon). \tag{29}$$

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

$$\begin{cases} D_\psi(w_\infty, w_0) - D_\psi(w^*, w_0) = o(\epsilon) \\ D_\psi(w^*, w_\infty) = o(\epsilon) \end{cases} \tag{30}$$

which concludes the proof. □

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

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

Proof of Corollary 5. The proof is a straightforward application of Theorem 4. Note that we have

$$D_\psi(w, w_0) = \psi(w) - \psi(w_0) - \nabla\psi(w_0)^T(w - w_0) \tag{31}$$

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

$$D_\psi(w, w_0) = \psi(w) - \psi(w_0). \tag{32}$$

In particular, by plugging in w_∞ 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

$$D_\psi(w_\infty, w_0) - D_\psi(w^*, w_0) = \psi(w_\infty) - \psi(w^*), \tag{33}$$

which, along with the application of Theorem 4, concludes the proof. □

VII. Related Work

There have been many efforts in the past few years to study deep learning from an optimization perspective, e.g., [9], [20], [22]–[26], [33]–[35]. 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 [9], [24], [26], [36]. Almost all these works, similar to ours, assume that the initialization is close to the solution space (of global minima), which is reasonable in highly overparameterized models. However, our results are more general because they extend to SMD.

On the other hand, even for the case of SGD, our results are stronger than those in this 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_∞, say, is close to the closest interpolating weight vector, w^*, say. Denoting the initialization by w_0, Oymak and Soltanolkotabi [26] 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_∞ 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 results.

There exist a number of results that characterize the implicit regularization properties of different algorithms in different contexts [20], [21], [37]–[42]. The closest ones to our results, since they concern mirror descent, are the works of [20], [21]. The authors in [21] 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. [20] also study linear models, but derive conditions on the step size for which SMD converges to the aforementioned global minimum. Our current results extend the aforementioned to *nonlinear* overparametrized 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 known for nonlinear models.

VIII. Conclusion

In this paper, we studied the convergence and implicit regularization properties of the family of stochastic mirror descent (SMD) for highly overparameterized nonlinear models. From a theoretical perspective, we showed that, under reasonable assumptions, SMD with sufficiently small step size (1) converges to a global minimum and (2) the global minimum converged to is approximately the closest to the initialization in Bregman divergence sense. Furthermore, our extensive experimental results, on various initializations, various mirror descents, and various Bregman divergences, revealed that this phenomenon indeed happens in practical scenarios in deep learning. This further implies that different mirror descent algorithms act as different regularizers, a property that is referred to as *implicit regularization*. The fact that the ℓ_∞-regularized solution showed a better generalization performance than the other ones, while ℓ_1 was the opposite, 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.

References

[1] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International Conference on Machine Learning*, 2016, pp. 173–182.

[2] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 6645–6649.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[4] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, p. 529, 2015.

[5] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot *et al.*, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[6] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.

[7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

[8] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," *arXiv preprint arXiv:1611.03530*, 2016.

[9] S. Ma, R. Bassily, and M. Belkin, "The power of interpolation: Understanding the effectiveness of SGD in modern over-parametrized learning," in *Proceedings of the 35th International Conference on Machine Learning*, vol. 80. PMLR, 2018, pp. 3325–3334.

[10] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht, "Gradient descent only converges to minimizers," in *Conference on Learning Theory*, 2016, pp. 1246–1257.

[11] A. C. Wilson, R. Roelofs, M. Stern, N. Srebro, and B. Recht, "The marginal value of adaptive gradient methods in machine learning," in *Advances in Neural Information Processing Systems*, 2017, pp. 4151–4161.

[12] A. Nemirovski and D. B. Yudin, "Problem complexity and method efficiency in optimization." 1983.

[13] A. Beck and M. Teboulle, "Mirror descent and nonlinear projected subgradient methods for convex optimization," *Operations Research Letters*, vol. 31, no. 3, pp. 167–175, 2003.

[14] N. Cesa-Bianchi, P. Gaillard, G. Lugosi, and G. Stoltz, "Mirror descent meets fixed share (and feels no regret)," in *Advances in Neural Information Processing Systems*, 2012, pp. 980–988.

[15] Z. Zhou, P. Mertikopoulos, N. Bambos, S. Boyd, and P. W. Glynn, "Stochastic mirror descent in variationally coherent optimization problems," in *Advances in Neural Information Processing Systems*, 2017, pp. 7043–7052.

[16] Y. Lei and D.-X. Zhou, "Convergence of online mirror descent," *Applied and Computational Harmonic Analysis*, vol. 48, no. 1, pp. 343–373, 2020.

[17] J. Kivinen and M. K. Warmuth, "Exponentiated gradient versus gradient descent for linear predictors," *Information and Computation*, vol. 132, no. 1, pp. 1–63, 1997.

[18] A. J. Grove, N. Littlestone, and D. Schuurmans, "General convergence results for linear discriminant updates," *Machine Learning*, vol. 43, no. 3, pp. 173–210, 2001.

[19] C. Gentile, "The robustness of the p-norm algorithms," *Machine Learning*, vol. 53, no. 3, pp. 265–299, 2003.

[20] N. Azizan and B. Hassibi, "Stochastic gradient/mirror descent: Minimax optimality and implicit regularization," in *International Conference on Learning Representations (ICLR)*, 2019.

[21] S. Gunasekar, J. Lee, D. Soudry, and N. Srebro, "Characterizing implicit bias in terms of optimization geometry," in *International Conference on Machine Learning*, 2018, pp. 1827–1836.
[22] Y. Li and Y. Liang, "Learning overparameterized neural networks via stochastic gradient descent on structured data," in *Advances in Neural Information Processing Systems*, 2018, pp. 8157–8166.
[23] S. S. Du, J. D. Lee, H. Li, L. Wang, and X. Zhai, "Gradient descent finds global minima of deep neural networks," *arXiv preprint arXiv:1811.03804*, 2018.
[24] Z. Allen-Zhu, Y. Li, and Z. Song, "A convergence theory for deep learning via over-parameterization," in *Proceedings of the 36th International Conference on Machine Learning*. PMLR, 2019.
[25] Y. Cao and Q. Gu, "A generalization theory of gradient descent for learning over-parameterized deep relu networks," *arXiv preprint arXiv:1902.01384*, 2019.
[26] S. Oymak and M. Soltanolkotabi, "Overparameterized nonlinear learning: Gradient descent takes the shortest path?" in *Proceedings of the 36th International Conference on Machine Learning*. PMLR, 2019.
[27] B. Hassibi, A. H. Sayed, and T. Kailath, *Indefinite-Quadratic Estimation and Control: A Unified Approach to H2 and H-infinity Theories*. SIAM, 1999, vol. 16.
[28] D. Simon, *Optimal state estimation: Kalman, H infinity, and nonlinear approaches*. John Wiley & Sons, 2006.
[29] B. Hassibi, A. H. Sayed, and T. Kailath, "Hoo optimality criteria for LMS and backpropagation," in *Advances in Neural Information Processing Systems 6*, 1994, pp. 351–358.
[30] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[32] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Citeseer, Tech. Rep., 2009.
[33] A. Achille and S. Soatto, "On the emergence of invariance and disentangling in deep representations," *arXiv preprint arXiv:1706.01350*, 2017.
[34] P. Chaudhari and S. Soatto, "Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks," in *International Conference on Learning Representations*, 2018.
[35] R. Shwartz-Ziv and N. Tishby, "Opening the black box of deep neural networks via information," *arXiv preprint arXiv:1703.00810*, 2017.
[36] M. Soltanolkotabi, A. Javanmard, and J. D. Lee, "Theoretical insights into the optimization landscape of over-parameterized shallow neural networks," *arXiv preprint arXiv:1707.04926*, 2017.
[37] B. Neyshabur, R. Tomioka, R. Salakhutdinov, and N. Srebro, "Geometry of optimization and implicit regularization in deep learning," *arXiv preprint arXiv:1705.03071*, 2017.
[38] C. Ma, K. Wang, Y. Chi, and Y. Chen, "Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix completion," in *International Conference on Machine Learning*, 2018, pp. 3351–3360.
[39] S. Gunasekar, B. E. Woodworth, S. Bhojanapalli, B. Neyshabur, and N. Srebro, "Implicit regularization in matrix factorization," in *Advances in Neural Information Processing Systems*, 2017, pp. 6152–6160.
[40] D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro, "The implicit bias of gradient descent on separable data," *arXiv preprint arXiv:1710.10345*, 2017.
[41] S. Gunasekar, J. Lee, D. Soudry, and N. Srebro, "Implicit bias of gradient descent on linear convolutional networks," *arXiv preprint arXiv:1806.00468*, 2018.
[42] P. Mianjy, R. Arora, and R. Vidal, "On the implicit bias of dropout," in *International Conference on Machine Learning*, 2018, pp. 3537–3545.

Navid Azizan is an incoming Assistant Professor at the Massachusetts Institute of Technology (MIT), and a postdoctoral scholar at Stanford University. He received the B.Sc. degree form Sharif University of Technology, Tehran, Iran, the M.S. degree from the University of Southern California, Los Angeles, CA, and the Ph.D. degree from the California Institute of Technology, Pasadena, CA, in 2013, 2015, and 2020, respectively. His research interests lie in machine learning, mathematical optimization, control theory, and networks.

His work has been recognized with several awards, including the Amazon Fellowship in Artificial Intelligence, the PIMCO Fellowship in Data Science, the 2016 ACM GREENMETRICS Best Student Paper Award, and the 2020 Information Theory and Applications (ITA) Gold Graduation Award. He was also the first-place winner and a gold medalist at the 2008 National Physics Olympiad in Iran.

Sahin Lale was born in Izmir, Turkey, in 1993. He received the B.S. degree in electrical and electronics engineering from Bogazici University, Istanbul, Turkey, in 2015 and the M.S. degree in electrical engineering from the California Institute of Technology (Caltech), Pasadena, CA, USA, in 2016. He is currently a Ph.D. Candidate in Electrical Engineering at Caltech. His research interests include reinforcement learning, control theory, machine learning and information theory.

Babak Hassibi was born in Tehran, Iran, in 1967. He received the B.S. degree from the University of Tehran in 1989, and the M.S. and Ph.D. degrees from Stanford University in 1993 and 1996, respectively, all in electrical engineering.

He has been with the California Institute of Technology since January 2001, where he is currently the Mose and Lilian S. Bohn Professor of Electrical Engineering. From 2013-2016 he was the Gordon M. Binder/Amgen Professor of Electrical Engineering and from 2008-2015 he was Executive Officer of Electrical Engineering, as well as Associate Director of Information Science and Technology. From October 1996 to October 1998 he was a research associate at the Information Systems Laboratory, Stanford University, and from November 1998 to December 2000 he was a Member of the Technical Staff in the Mathematical Sciences Research Center at Bell Laboratories, Murray Hill, NJ. He has also held short-term appointments at Ricoh California Research Center, the Indian Institute of Science, and Linkoping University, Sweden. His research interests include communications and information theory, control and network science, and signal processing and machine learning. He is the coauthor of the books (both with A.H. Sayed and T. Kailath) *Indefinite Quadratic Estimation and Control: A Unified Approach to H^2 and H^∞ Theories* (New York: SIAM, 1999) and *Linear Estimation* (Englewood Cliffs, NJ: Prentice Hall, 2000). He is a recipient of an Alborz Foundation Fellowship, the 1999 O. Hugo Schuck best paper award of the American Automatic Control Council (with H. Hindi and S.P. Boyd), the 2002 National Science Foundation Career Award, the 2002 Okawa Foundation Research Grant for Information and Telecommunications, the 2003 David and Lucille Packard Fellowship for Science and Engineering, the 2003 Presidential Early Career Award for Scientists and Engineers (PECASE), and the 2009 Al-Marai Award for Innovative Research in Communications, and was a participant in the 2004 National Academy of Engineering "Frontiers in Engineering" program.

He has been a Guest Editor for the IEEE Transactions on Information Theory special issue on "space-time transmission, reception, coding and signal processing", was an Associate Editor for Communications of the IEEE Transactions on Information Theory during 2004-2006, and is currently an Editor for the Journal "Foundations and Trends in Information and Communication" and for the IEEE Transactions on Network Science and Engineering. He was an IEEE Information Theory Society Distinguished Lecturer for 2016-2017 and the co-chair of the 2020 IEEE International Symposium on Information Theory (ISIT 2020) in Los Angeles, CA.

Supplementary Material

APPENDIX A
ADDITIONAL PROOFS

A. Fundamental Identity of SMD

Lemma 6. *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 $\{w_i\}$*

$$D_\psi(w, w_{i-1}) = D_\psi(w, w_i) + D_{\psi-\eta L_i}(w_i, w_{i-1}) + \eta L_i(w_i) + \eta D_{L_i}(w, w_{i-1}), \tag{10}$$

for all $i \geq 1$.

Proof of Lemma 6. 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

$$D_\psi(w, w_i) = \psi(w) - \psi(w_i) - \nabla\psi(w_{i-1})^T(w - w_i) + \eta\nabla L_i(w_{i-1})^T(w - w_i). \tag{34}$$

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

$$\begin{aligned} D_\psi(w, w_i) &= D_\psi(w, w_{i-1}) + \psi(w_{i-1}) + \nabla\psi(w_{i-1})^T(w - w_{i-1}) - \psi(w_i) \\ &\qquad - \nabla\psi(w_{i-1})^T(w - w_i) + \eta\nabla L_i(w_{i-1})^T(w - w_i) \qquad (35)\\ &= D_\psi(w, w_{i-1}) + \psi(w_{i-1}) - \psi(w_i) + \nabla\psi(w_{i-1})^T(w_i - w_{i-1}) \\ &\qquad + \eta\nabla L_i(w_{i-1})^T(w - w_i) \qquad (36)\\ &= D_\psi(w, w_{i-1}) - D_\psi(w_i, w_{i-1}) + \eta\nabla L_i(w_{i-1})^T(w - w_i). \qquad (37) \end{aligned}$$

Expanding the last term using $w - w_i = (w - w_{i-1}) - (w_i - w_{i-1})$, and following the definition of $D_{L_i}(.,.)$ from (8) for (w, w_{i-1}) and (w_i, w_{i-1}), we have

$$\begin{aligned} D_\psi(w, w_i) &= D_\psi(w, w_{i-1}) - D_\psi(w_i, w_{i-1}) + \eta\nabla L_i(w_{i-1})^T(w - w_{i-1}) \\ &\qquad - \eta\nabla L_i(w_{i-1})^T(w_i - w_{i-1}) \qquad (38)\\ &= D_\psi(w, w_{i-1}) - D_\psi(w_i, w_{i-1}) + \eta\left(L_i(w) - L_i(w_{i-1}) - D_{L_i}(w, w_{i-1})\right) \\ &\qquad - \eta\left(L_i(w_i) - L_i(w_{i-1}) - D_{L_i}(w_i, w_{i-1})\right) \qquad (39)\\ &= D_\psi(w, w_{i-1}) - D_\psi(w_i, w_{i-1}) + \eta\left(L_i(w) - D_{L_i}(w, w_{i-1})\right) \\ &\qquad - \eta\left(L_i(w_i) - D_{L_i}(w_i, w_{i-1})\right) \qquad (40) \end{aligned}$$

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

$$D_\psi(w, w_i) = D_\psi(w, w_{i-1}) - D_\psi(w_i, w_{i-1}) - \eta D_{L_i}(w, w_{i-1}) - \eta L_i(w_i) + \eta D_{L_i}(w_i, w_{i-1}). \tag{41}$$

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

$$D_\psi(w, w_i) = D_\psi(w, w_{i-1}) - D_{\psi-\eta L_i}(w_i, w_{i-1}) - \eta D_{L_i}(w, w_{i-1}) - \eta L_i(w_i), \tag{42}$$

for all $w \in \mathcal{W}$, which concludes the proof. □

B. 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, with high probability, close to the solution set $\mathcal{W}$ [20], [22]–[25]. In the highly overparameterized regime, we have $p \gg n$, 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

$$\begin{aligned} \min_w \quad & \|w - w_0\|^2 \\ \text{s.t.} \quad & y = f(x, w) \end{aligned}$$

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

$$\begin{aligned} \min_w \quad & \|w - w_0\|^2 \\ \text{s.t.} \quad & y \approx f(x, w_0) + \nabla f(x, w_0)^T(w - w_0) \end{aligned}$$

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

$$\|w^* - w_0\|^2 \approx (y - f(x,w_0))^T \left(\nabla f(x,w_0)^T \nabla f(x,w_0)\right)^{-1} (y - f(x,w_0)) \tag{43}$$

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, with high probability,

$$\lambda_{\min}\left(\nabla f(x,w_0)^T \nabla f(x,w_0)\right) = \Omega(p),$$

from which we conclude

$$\|w^* - w_0\|^2 \approx \|y - f(x,w_0)\|^2 \cdot O(\frac{1}{p}) = O(\frac{n}{p}), \tag{44}$$

since $y - f(x,w_0)$ is n-dimensional. The above implies that w_0 is close to w^* and hence to $\mathcal{W}$.

APPENDIX B
ADDITIONAL DETAILS ON THE EXPERIMENTAL RESULTS

In order to evaluate the theoretical claims, we ran systematic experiments on standard deep learning problems.

Datasets. We use the standard MNIST [30] and CIFAR-10 [32] 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 [31] 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 0 training error, i.e., until we hit $\mathcal{W}$.

Initialization. We randomly initialize the parameters of the networks around zero ($\mathcal{N}(0, 0.0001)$). 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 $\psi(w) = \frac{1}{q}\|w\|_q^q$ for the following four different norms: (a) ℓ_1 norm, i.e., $q = 1+\epsilon$, (b) ℓ_2 norm, i.e., $q = 2$ (which is SGD), (c) ℓ_3 norm, i.e., $q = 3$, (d) ℓ_{10} norm, i.e., $q = 10$ (as a surrogate for ℓ_∞ norm). The update rule can be expressed as follows.

$$w_i[j] = \left||w_{i-1}[j]|^{q-1}\text{sign}(w_{i-1}[j]) - \eta \nabla L_i(w_{i-1})[j]\right|^{\frac{1}{q-1}} \text{sign}\left(|w_{i-1}[j]|^{q-1}\text{sign}(w_{i-1}[j]) - \eta \nabla L_i(w_{i-1})[j]\right), \tag{45}$$

where $w_i[j]$ denotes the j-th element of the w_i vector.

We use a fixed step size η. The step size is chosen to obtain convergence to global minima.

A. MNIST Experiments

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 III, IV, V, VI, VII, and VIII 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 III
MNIST INITIAL POINT 1.

	SMD 1-norm	SMD 2-norm (SGD)	SMD 3-norm	SMD 10-norm
1-norm BD	2.767	937.8	1.05×10^4	1.882×10^5
2-norm BD	301.6	58.61	261.3	2.118×10^4
3-norm BD	1720	37.45	7.143	2518
10-norm BD	7.453×10^8	773.4	0.2939	0.003545

TABLE IV
MNIST INITIAL POINT 2.

	SMD 1-norm	SMD 2-norm (SGD)	SMD 3-norm	SMD 10-norm
1-norm BD	2.78	945	1.37×10^4	2.01×10^5
2-norm BD	292	59.3	374	2.29×10^4
3-norm BD	1.51×10^3	38.6	11.6	2.71×10^3
10-norm BD	1.06×10^8	831	0.86	0.00321

TABLE V
MNIST INITIAL POINT 3.

	SMD 1-norm	SMD 2-norm (SGD)	SMD 3-norm	SMD 10-norm
1-norm BD	3.02	968	1.06×10^4	1.9×10^5
2-norm BD	291	60.9	272	2.12×10^4
3-norm BD	1.49×10^3	39.1	7.82	2.49×10^3
10-norm BD	1.1×10^8	900	0.411	0.00318

TABLE VI
MNIST INITIAL POINT 4.

	SMD 1-norm	SMD 2-norm (SGD)	SMD 3-norm	SMD 10-norm
1-norm BD	2.78	1.21×10^3	1.08×10^4	1.92×10^5
2-norm BD	291	77.3	271	2.15×10^4
3-norm BD	1.48×10^3	49.7	7.56	2.52×10^3
10-norm BD	9.9×10^7	1.72×10^3	0.352	0.00296

TABLE VII
MNIST INITIAL POINT 5.

	SMD 1-norm	SMD 2-norm (SGD)	SMD 3-norm	SMD 10-norm
1-norm BD	2.79	958	1.08×10^4	2×10^5
2-norm BD	292	60.4	271	2.28×10^4
3-norm BD	1.49×10^3	39	7.52	2.69×10^3
10-norm BD	9.09×10^7	846	0.342	0.00309

TABLE VIII
MNIST INITIAL POINT 6.

	SMD 1-norm	SMD 2-norm (SGD)	SMD 3-norm	SMD 10-norm
1-norm BD	2.96	930	1.08×10^4	1.9×10^5
2-norm BD	308	59	271	2.12×10^4
3-norm BD	1.63×10^3	38.6	7.46	2.47×10^3
10-norm BD	1.65×10^8	864	0.334	0.00295

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 IX, X, XI, and XII 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 IX
MNIST 1-NORM BREGMAN DIVERGENCE BETWEEN THE INITIAL POINTS AND THE FINAL POINTS OBTAINED BY SMD 1-NORM.

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

TABLE X
MNIST 2-NORM BREGMAN DIVERGENCE BETWEEN THE INITIAL POINTS AND THE FINAL POINTS OBTAINED BY SMD 2-NORM (SGD).

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

TABLE XI
MNIST 3-NORM BREGMAN DIVERGENCE BETWEEN THE INITIAL POINTS AND THE FINAL POINTS OBTAINED BY SMD 3-NORM.

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

TABLE XII
MNIST 10-NORM BREGMAN DIVERGENCE BETWEEN THE INITIAL POINTS AND THE FINAL POINTS OBTAINED BY SMD 10-NORM.

	Final 1	Final 2	Final 3	Final 4	Final 5	Final 6
Initial Point 1	0.00354	0.37	0.403	0.286	0.421	0.408
Initial Point 2	0.33	0.00321	0.369	0.383	0.415	0.422
Initial Point 3	0.347	0.318	0.00318	0.401	0.312	0.406
Initial Point 4	0.282	0.38	0.458	0.00296	0.491	0.376
Initial Point 5	0.405	0.418	0.354	0.484	0.00309	0.48
Initial Point 6	0.403	0.353	0.422	0.331	0.503	0.00295

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.

	F1 SMD 1	F2 SMD 1	F3 SMD 1	F4 SMD 1	F5 SMD 1	F6 SMD 1	F1 SMD 2	F2 SMD 2	F3 SMD 2	F4 SMD 2	F5 SMD 2	F6 SMD 2	F1 SMD 3	F2 SMD 3	F3 SMD 3	F4 SMD 3	F5 SMD 3	F6 SMD 3	F1 SMD 10	F2 SMD 10	F3 SMD 10	F4 SMD 10	F5 SMD 10	F6 SMD 10
I1 1-norm BD	2.767105	20310.58	20266.27	20330.6	20340.2	20281.51	937.7902	20501.09	20453.6	20615.37	20505.63	20451.42	10500.44	24298.6	22690.41	22883.13	22928.17	22930.01	188233.4	200749.8	189599.3	192017.6	200332.6	189842.1
I2 1-norm BD	20332.47	2.777443	20280.59	20298.8	20312	20322.66	20477.15	944.8926	20467.58	20572.54	20486.79	20481.46	22902.71	13736.89	22683.03	22823.09	22927.75	22951.2	188019	200838.7	189406.7	191694.7	200319.4	189452.9
I3 1-norm BD	20319.38	20312.19	3.018036	20343.8	20308.9	20322.02	20443.74	20487.21	967.6324	20612.98	20486.93	20485.8	22897.06	24300.62	10609.4	22876.31	22901.84	22949.55	187883.2	201071.8	189571	192131.8	199958.1	189571.5
I4 1-norm BD	20338.77	20279.16	20309.78	2.78104	20321.1	20297.36	20476.14	20461.38	20499.16	1214.917	20499.88	20469.45	22910.51	24283.22	22733.45	10756.58	22928.43	22938.72	187740.6	200692.5	189522.4	192082.9	200434.4	189653.4
I5 1-norm BD	20347.03	20317.23	20273.07	20316.4	2.79019	20310.78	20498.73	20496.97	20464.54	20600.07	957.8013	20484.67	22921.69	24335.41	22722.83	22877.07	10812.1	22955.94	188056.5	200743.9	189707.6	192056.4	200478.6	189883
I6 1-norm BD	20343.59	20322.62	20339.82	20318.4	20320.9	2.964027	20493.68	20504.14	20535.06	20590.71	20491.61	930.2714	22926.8	24311.73	22713.74	22837.8	22900.31	10848.27	187959.4	200482.2	189602.3	192052.5	200309.6	189738.7
I1 2-norm BD	301.6218	928.1953	922.246	925.889	929.909	940.8018	58.60796	670.7482	667.0325	684.1751	671.3561	667.8379	261.2823	760.2361	701.1998	706.1766	704.5516	704.0641	21179.18	22902.48	21188.34	21536.64	22803.44	21162.22
I2 2-norm BD	938.5225	291.6223	924.8324	925.561	926.34	944.1569	669.8414	59.31496	669.1617	682.0358	669.4517	669.9774	703.9956	373.9718	702.4789	703.8165	703.9578	705.4268	21164.67	22901.68	21187.33	21523.37	22797.37	21152.21
I3 2-norm BD	936.4615	928.4752	290.902	926.259	924.438	943.7131	666.3494	670.2202	60.85767	683.4393	667.5668	669.9933	700.8777	758.6538	272.0649	705.6848	701.1583	705.155	21164.46	22904.93	21186.56	21536.03	22787.11	21151.72
I4 2-norm BD	938.7566	926.655	926.2601	290.552	928.945	944.0035	669.7086	668.8569	671.186	77.27538	670.3311	669.7023	703.3976	757.9133	704.6345	270.9099	703.842	704.6346	21161.99	22898.71	21186.54	21541.32	22799.92	21152.3
I5 2-norm BD	940.8469	928.4445	923.667	927.336	291.765	939.7045	671.1005	669.1169	668.446	683.6102	60.39	666.0443	705.3977	758.6884	702.3112	705.3715	270.8719	701.3619	21166.8	22898.1	21191.65	21533.87	22805.13	21162.55
I6 2-norm BD	937.93	929.7885	929.404	927.348	925.181	307.5172	669.4556	670.9225	671.5908	684.3248	667.3748	59.04266	702.8038	759.7584	703.6673	705.2996	700.8271	271.1133	21166.69	22894.56	21188.54	21530.77	22796.98	21153.93
I1 3-norm BD	1719.866	1543.515	1516.246	1512.4	1521.08	1656.464	37.45108	67.57934	66.73737	78.02365	67.95686	66.51245	7.14298	35.30229	32.07697	32.65884	32.64842	32.30852	2517.617	2706.617	2491.476	2519.086	2688.245	2470.969
I2 3-norm BD	1751.333	1510.961	1516.163	1514.81	1518.79	1658.074	66.28766	38.64332	66.75334	78.01804	66.94847	67.09068	32.50659	11.57823	32.25632	32.32539	32.2253	32.45956	2516.606	2705.533	2491.199	2518.034	2687.31	2470.926
I3 3-norm BD	1751.98	1544.446	1486.664	1513.27	1517.48	1658.303	65.47958	67.39749	39.096	78.03239	66.49712	67.24052	31.59447	34.64265	7.823877	32.52136	31.58038	32.51863	2517.107	2706.491	2489.415	2519.598	2687.107	2470.339
I4 3-norm BD	1751.523	1543.899	1517.328	1483.49	1522.07	1659.334	66.36948	67.31509	67.59354	49.69977	67.96119	67.20248	32.30269	34.81075	32.93691	7.558935	32.61658	32.28448	2517.248	2706.852	2491.392	2518.947	2687.751	2470.657
I5 3-norm BD	1753.311	1545.901	1516.143	1515.92	1488.06	1657.359	66.56918	67.42434	67.07494	78.55313	39.04714	66.25287	32.67308	34.67835	32.07084	32.73818	7.518829	31.55844	2517.357	2706.916	2491.048	2519.073	2687.064	2471.216
I6 3-norm BD	1751.224	1544.936	1520.698	1514.66	1519.78	1626.957	66.33501	67.47943	67.81073	78.43179	67.07613	38.58941	32.11641	34.73071	32.37629	32.43067	31.69857	7.459286	2517.511	2706.82	2490.098	2518.297	2687.431	2469.509
I1 10-norm BD	7.45E+08	1.06E+08	1.1E+08	9.9E+07	9.1E+07	1.65E+08	773.3514	831.1445	900.464	1718.299	846.4625	864.5718	0.293932	1.233024	0.782131	0.615488	0.748684	0.707943	0.003545	0.370181	0.403135	0.28582	0.421482	0.408148
I2 10-norm BD	7.45E+08	1.06E+08	1.1E+08	9.9E+07	9.1E+07	1.65E+08	773.7523	830.5577	900.2781	1718.625	846.2303	864.6849	0.61333	0.860265	0.732687	0.725046	0.727329	0.727967	0.330493	0.003207	0.368537	0.382603	0.415105	0.422372
I3 10-norm BD	7.45E+08	1.06E+08	1.1E+08	9.9E+07	9.1E+07	1.65E+08	773.8534	831.2133	900.141	1718.575	846.1995	864.7488	0.63865	1.196859	0.410611	0.735479	0.634941	0.718673	0.347069	0.317821	0.00318	0.400619	0.311682	0.405827
I4 10-norm BD	7.45E+08	1.06E+08	1.1E+08	9.9E+07	9.1E+07	1.65E+08	773.8442	831.1647	900.4524	1718.06	846.5443	864.7191	0.585811	1.241436	0.824513	0.351863	0.819103	0.694199	0.281852	0.379772	0.457535	0.002963	0.49113	0.376261
I5 10-norm BD	7.45E+08	1.06E+08	1.1E+08	9.9E+07	9.1E+07	1.65E+08	773.8727	831.1471	900.3779	1718.562	845.8668	864.717	0.691508	1.273044	0.735977	0.814864	0.342117	0.783273	0.40501	0.417533	0.353985	0.483609	0.003094	0.479598
I6 10-norm BD	7.45E+08	1.06E+08	1.1E+08	9.9E+07	9.1E+07	1.65E+08	773.9967	831.1642	900.7065	1718.531	846.6509	864.2966	0.703456	1.22497	0.82352	0.679747	0.840384	0.33448	0.403348	0.352543	0.421798	0.330755	0.503257	0.002948

Fig. 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. CIFAR-10 Experiments

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 XIII, XIV, XV, XVI, XVII, XVIII, XIX, and XX 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 XIII
CIFAR-10 INITIAL POINT 1.

	SMD 1-norm	SMD 2-norm (SGD)	SMD 3-norm	SMD 10-norm
1-norm BD	189	9.58×10^3	4.19×10^4	2.34×10^5
2-norm BD	3.12×10^3	597	1.28×10^3	6.92×10^3
3-norm BD	4.31×10^4	119	55.8	1.87×10^2
10-norm BD	1.35×10^{14}	869	6.34×10^{-5}	2.64×10^{-8}

TABLE XIV
CIFAR-10 INITIAL POINT 2.

	SMD 1-norm	SMD 2-norm (SGD)	SMD 3-norm	SMD 10-norm
1-norm BD	275	9.86×10^3	4.09×10^4	2.38×10^5
2-norm BD	4.89×10^3	607	1.23×10^3	7.03×10^3
3-norm BD	9.21×10^4	104	53.5	1.88×10^2
10-norm BD	1.17×10^{15}	225	0.000102	2.65×10^{-8}

TABLE XV
CIFAR-10 INITIAL POINT 3.

	SMD 1-norm	SMD 2-norm (SGD)	SMD 3-norm	SMD 10-norm
1-norm BD	141	9.19×10^3	4.1×10^4	2.34×10^5
2-norm BD	3.15×10^3	562	1.24×10^3	6.89×10^3
3-norm BD	4.31×10^4	107	53.5	1.85×10^2
10-norm BD	6.83×10^{13}	972	7.91×10^{-5}	2.72×10^{-8}

TABLE XVI
CIFAR-10 INITIAL POINT 4.

	SMD 1-norm	SMD 2-norm (SGD)	SMD 3-norm	SMD 10-norm
1-norm BD	255	9.77×10^3	4.18×10^4	2.36×10^5
2-norm BD	3.64×10^3	594	1.26×10^3	6.96×10^3
3-norm BD	5.5×10^4	116	54	1.87×10^2
10-norm BD	3.74×10^{14}	640	5.33×10^{-5}	2.67×10^{-8}

TABLE XVII
CIFAR-10 INITIAL POINT 5.

	SMD 1-norm	SMD 2-norm (SGD)	SMD 3-norm	SMD 10-norm
1-norm BD	113	9.48×10^3	4.15×10^4	2.32×10^5
2-norm BD	2.95×10^3	572	1.27×10^3	6.85×10^3
3-norm BD	3.68×10^4	109	56.2	1.84×10^2
10-norm BD	2.97×10^{13}	151	5.74×10^{-5}	2.61×10^{-8}

TABLE XVIII
CIFAR-10 INITIAL POINT 6.

	SMD 1-norm	SMD 2-norm (SGD)	SMD 3-norm	SMD 10-norm
1-norm BD	128	9.25×10^3	4.25×10^4	2.34×10^5
2-norm BD	2.71×10^3	558	1.29×10^3	6.89×10^3
3-norm BD	3.34×10^4	104	55.3	1.85×10^2
10-norm BD	2.61×10^{13}	612	4.74×10^{-5}	2.62×10^{-8}

TABLE XIX
CIFAR-10 INITIAL POINT 7.

	SMD 1-norm	SMD 2-norm (SGD)	SMD 3-norm	SMD 10-norm
1-norm BD	223	9.76×10^3	4.38×10^4	2.27×10^5
2-norm BD	2.41×10^3	599	1.37×10^3	6.65×10^3
3-norm BD	2.3×10^4	116	61	1.78×10^2
10-norm BD	4.22×10^{12}	679	6.42×10^{-5}	2.55×10^{-8}

TABLE XX
CIFAR-10 INITIAL POINT 8.

	SMD 1-norm	SMD 2-norm (SGD)	SMD 3-norm	SMD 10-norm
1-norm BD	145	9.37×10^3	4.17×10^4	2.36×10^5
2-norm BD	2.48×10^3	576	1.26×10^3	6.99×10^3
3-norm BD	2.85×10^4	108	54.5	1.89×10^2
10-norm BD	1.81×10^{13}	1.22×10^3	5.2×10^{-5}	2.64×10^{-8}

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 XXI, XXII, XXIII, XXIV 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 XXI
CIFAR-10 1-NORM BREGMAN DIVERGENCE BETWEEN THE INITIAL POINTS AND THE FINAL POINTS OBTAINED BY SMD 1-NORM.

	Final 1	Final 2	Final 3	Final 4	Final 5	Final 6	Final 7	Final 8
Initial 1	1.9×10^2	8.1×10^4	8.1×10^4	8.4×10^4	8×10^4	8.2×10^4	7.8×10^4	7.8×10^4
Initial 2	8.1×10^4	2.7×10^2	8.1×10^4	8.3×10^4	8×10^4	8.2×10^4	7.8×10^4	7.9×10^4
Initial 3	8.1×10^4	8.1×10^4	1.4×10^2	8.4×10^4	8×10^4	8.1×10^4	7.8×10^4	7.8×10^4
Initial 4	8.1×10^4	8.1×10^4	8.1×10^4	2.5×10^2	8×10^4	8.2×10^4	7.8×10^4	7.9×10^4
Initial 5	8.1×10^4	8.1×10^4	8.1×10^4	8.3×10^4	1.1×10^2	8.1×10^4	7.8×10^4	7.8×10^4
Initial 6	8.1×10^4	8.1×10^4	8.1×10^4	8.4×10^4	8×10^4	1.3×10^2	7.8×10^4	7.8×10^4
Initial 7	8.1×10^4	8.1×10^4	8.1×10^4	8.4×10^4	8×10^4	8.1×10^4	2.2×10^2	7.8×10^4
Initial 8	8.1×10^4	8.1×10^4	8.1×10^4	8.4×10^4	7.9×10^4	8.1×10^4	7.8×10^4	1.5×10^2

TABLE XXII
CIFAR-10 2-NORM BREGMAN DIVERGENCE BETWEEN THE INITIAL POINTS AND THE FINAL POINTS OBTAINED BY SMD 2-NORM (SGD).

	Final 1	Final 2	Final 3	Final 4	Final 5	Final 6	Final 7	Final 8
Initial 1	6×10^2	2.9×10^3	2.8×10^3	2.8×10^3	2.8×10^3	2.8×10^3	2.8×10^3	2.8×10^3
Initial 2	2.8×10^3	6.1×10^2	2.8×10^3	2.8×10^3	2.8×10^3	2.8×10^3	2.8×10^3	2.8×10^3
Initial 3	2.8×10^3	2.9×10^3	5.6×10^2	2.8×10^3	2.8×10^3	2.8×10^3	2.8×10^3	2.8×10^3
Initial 4	2.8×10^3	2.9×10^3	2.8×10^3	5.9×10^2	2.8×10^3	2.8×10^3	2.8×10^3	2.8×10^3
Initial 5	2.8×10^3	2.9×10^3	2.8×10^3	2.8×10^3	5.7×10^2	2.8×10^3	2.8×10^3	2.8×10^3
Initial 6	2.8×10^3	2.9×10^3	2.8×10^3	2.8×10^3	2.8×10^3	5.6×10^2	2.8×10^3	2.8×10^3
Initial 7	2.8×10^3	2.9×10^3	2.8×10^3	2.8×10^3	2.8×10^3	2.8×10^3	6×10^2	2.8×10^3
Initial 8	2.8×10^3	2.9×10^3	2.8×10^3	2.8×10^3	2.8×10^3	2.8×10^3	2.8×10^3	5.8×10^2

TABLE XXIII
CIFAR-10 3-NORM BREGMAN DIVERGENCE BETWEEN THE INITIAL POINTS AND THE FINAL POINTS OBTAINED BY SMD 3-NORM.

	Final 1	Final 2	Final 3	Final 4	Final 5	Final 6	Final 7	Final 8
Initial 1	55.844	103.47	103.61	104.05	106.2	105.32	110.88	104.56
Initial 2	105.87	53.455	103.68	104.04	106.31	105.34	110.93	104.58
Initial 3	105.89	103.59	53.527	104.09	106.29	105.35	110.99	104.55
Initial 4	105.83	103.54	103.64	53.978	106.23	105.3	110.87	104.54
Initial 5	105.82	103.55	103.64	104	56.161	105.34	110.88	104.55
Initial 6	105.91	103.6	103.66	104.1	106.28	55.316	110.94	104.55
Initial 7	105.87	103.51	103.67	103.98	106.26	105.25	61.045	104.5
Initial 8	105.77	103.54	103.59	104.04	106.25	105.28	110.88	54.509

TABLE XXIV
CIFAR-10 10-NORM BREGMAN DIVERGENCE BETWEEN THE INITIAL POINTS AND THE FINAL POINTS OBTAINED BY SMD 10-NORM.

	Final 1	Final 2	Final 3	Final 4	Final 5	Final 6	Final 7	Final 8
Initial 1	2.64×10^{-8}	2.89×10^{-8}	2.99×10^{-8}	2.81×10^{-8}	2.85×10^{-8}	2.82×10^{-8}	2.66×10^{-8}	2.82×10^{-8}
Initial 2	2.79×10^{-8}	2.65×10^{-8}	2.83×10^{-8}	2.83×10^{-8}	2.71×10^{-8}	2.74×10^{-8}	2.69×10^{-8}	2.88×10^{-8}
Initial 3	2.89×10^{-8}	2.87×10^{-8}	2.72×10^{-8}	2.94×10^{-8}	2.84×10^{-8}	2.89×10^{-8}	2.78×10^{-8}	2.94×10^{-8}
Initial 4	2.79×10^{-8}	2.86×10^{-8}	2.92×10^{-8}	2.67×10^{-8}	2.84×10^{-8}	2.81×10^{-8}	2.69×10^{-8}	2.85×10^{-8}
Initial 5	2.76×10^{-8}	2.88×10^{-8}	2.95×10^{-8}	2.93×10^{-8}	2.61×10^{-8}	2.73×10^{-8}	2.66×10^{-8}	2.83×10^{-8}
Initial 6	2.80×10^{-8}	2.76×10^{-8}	2.93×10^{-8}	2.79×10^{-8}	2.76×10^{-8}	2.62×10^{-8}	2.71×10^{-8}	2.85×10^{-8}
Initial 7	2.73×10^{-8}	2.76×10^{-8}	2.82×10^{-8}	2.79×10^{-8}	2.71×10^{-8}	2.77×10^{-8}	2.55×10^{-8}	2.83×10^{-8}
Initial 8	2.73×10^{-8}	2.79×10^{-8}	2.85×10^{-8}	2.78×10^{-8}	2.75×10^{-8}	2.74×10^{-8}	2.73×10^{-8}	2.64×10^{-8}

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.

	F1 SMD 1	F2 SMD 1	F3 SMD 1	F4 SMD 1	F5 SMD 1	F6 SMD 1	F7 SMD 1	F8 SMD 1	F1 SMD 2	F2 SMD 2	F3 SMD 2	F4 SMD 2	F5 SMD 2	F6 SMD 2	F7 SMD 2	F8 SMD 2	F1 SMD 3	F2 SMD 3	F3 SMD 3	F4 SMD 3	F5 SMD 3	F6 SMD 3	F7 SMD 3	F8 SMD 3	F1 SMD 10	F2 SMD 10	F3 SMD 10	F4 SMD 10	F5 SMD 10	F6 SMD 10	F7 SMD 10	F8 SMD 10
I1 1-norm BD	189.328	81283.3	81471.4	83604.7	79579.4	81574.5	78218.4	78389.4	9580.35	96492.8	95885.6	96458.8	96180.1	95912	96403.9	95970	41925.9	114691	114844	115624	115215	116010	116934	115410	233735	243939	240590	241635	240066	240783	236041	242224.4
I2 1-norm BD	80718.8	274.637	81361.3	83423	79566.4	81521.8	78162.2	78526.3	96199.3	9858.87	95881.2	96432.7	96177.9	96011	96360.3	96073	115489	40871	114857	115444	115270	116042	117056	115479	240760	238156	240430	241386	239851	240885	236146	242360.5
I3 1-norm BD	80753.1	81232	140.651	83587.2	79590.2	81486.7	78193	78409	96271.6	96499.5	9189.09	96431.2	96189.8	95954.8	96437.4	96029.7	115549	114774	40954.3	115560	115315	116047	117093	115429	240778	243890	233638	241525	239917	240999	236223	242303.6
I4 1-norm BD	80851.2	81366.7	81446.9	254.781	79561.5	81544.5	78242.4	78550.7	96292.8	96533.1	95833.1	9770.32	96166.7	95905	96399.4	96071	115538	114746	114886	41757.1	115308	116051	117009	115446	240852	243762	240360	235800	239705	240813	236021	242218.3
I5 1-norm BD	80735.9	81095.3	81265.8	83396	112.929	81449.6	78178.8	78416.2	96233.6	96518.6	95873.1	96389.1	9476.29	95930.1	96421.5	96077.4	115579	114753	114937	115436	41482.2	116134	116895	115460	240672	243753	240605	241462	232340	240884	236114	242169.7
I6 1-norm BD	80610.2	81289.8	81424	83502.5	79528.5	127.523	78185.1	78419	96169.5	96635.1	95860.7	96465.6	96195.5	9252.54	96447	95938.9	115554	114848	114825	115542	115272	42498.8	116984	115293	240769	243866	240463	241597	239922	233795	236173	242261.8
I7 1-norm BD	80712.9	81231.3	81427.1	83570.8	79616.1	81490.4	223.16	78465.8	96193.4	96429.7	95832.9	96458.2	96115.9	95851.2	9758.27	95989.1	115562	114705	114924	115462	115271	115967	43834.6	115374	240605	243785	240599	241628	239886	240793	227206	242312.7
I8 1-norm BD	80733	81342.6	81374.9	83662	79457.7	81452.1	78241.5	145.142	96221.4	96491.7	95916.8	96440.2	96122.5	95852.2	96417.2	9365.13	115506	114698	114829	115547	115293	115973	116956	41691.3	240719	243866	240445	241632	239894	240812	236172	236014.5
I1 2-norm BD	3120.14	6974.61	5249.23	5782.27	5001.56	4812.6	4416.03	4494.71	597.342	2854.45	2810.65	2842.48	2820.28	2807.59	2847.02	2822.15	1280.42	3267.8	3274.47	3295.87	3304.95	3316.2	3388.3	3296.81	6916.86	7306.33	7200.89	7231.33	7181.46	7201.92	7032.73	7276.412
I2 2-norm BD	5192.44	4890.69	5245.16	5780.78	4997.38	4808.08	4409.99	4497.59	2845.99	607.493	2811.68	2843.32	2821.78	2809.34	2846	2824.65	3310.89	1234.3	3276.02	3294.08	3308.61	3317	3390.21	3297.57	7229.1	7029.11	7197.54	7229.07	7178.02	7205.71	7036.86	7281.481
I3 2-norm BD	5188.32	6962.86	3146.41	5779.56	4997.19	4801.63	4411	4490.97	2846.96	2856.09	562.461	2842.86	2822.19	2808.17	2848.66	2824.24	3312.22	3271.6	1238.4	3296.24	3308.08	3317.55	3392.26	3296.88	7229.01	7307.63	6889.44	7230.78	7179.44	7210.22	7037.09	7279.456
I4 2-norm BD	5193.64	6980.54	5251.91	3637.04	5004.97	4815.21	4416.43	4501.13	2846.63	2855.25	2810.84	594.372	2820.12	2806.85	2846.75	2824.84	3311.1	3270.4	3276.34	1260.59	3306.82	3316.69	3388.77	3297.26	7227.94	7304.05	7196.27	6961.64	7174.85	7204.47	7030.67	7276.045
I5 2-norm BD	5184.99	6957.65	5235.89	5770.93	2948.17	4797.47	4405.55	4490.22	2844.86	2856.32	2811.19	2840.89	571.718	2807.17	2846.79	2824.71	3311.24	3271.1	3276.49	3293.23	1271.14	3318.64	3388	3297.79	7226.45	7304.07	7201.92	7230.03	6845.38	7205.5	7034.49	7277.01
I6 2-norm BD	5190.72	6974.64	5251.31	5782.08	4998.22	2710.92	4413.17	4496.44	2845.9	2857.39	2811.76	2843.42	2821.57	557.986	2847.93	2822.11	3312.84	3272.1	3275.54	3296.07	3307.67	1286.14	3389.88	3295.92	7228.18	7307.39	7199.14	7230.83	7177.58	6888.1	7035.99	7278.548
I7 2-norm BD	5191.64	6968.14	5244.83	5780.7	5002.59	4807.16	2407.1	4496.42	2844.65	2853.81	2811.15	2841.67	2819.3	2805.37	599.485	2823.2	3312.49	3269.3	3277.22	3293.05	3307.42	3314.74	1366.29	3295.35	7221.98	7303.57	7200.85	7232.19	7176.33	7205.15	6654.25	7281.245
I8 2-norm BD	5191.97	6975.22	5247.43	5785.94	4995.53	4806.71	4415.62	2475.94	2843.69	2855.67	2811.02	2841.96	2820.31	2805.18	2847.1	575.791	3308.79	3270.4	3274.15	3295.42	3307.34	3315.04	3388.73	1263.73	7224.78	7305.98	7197.26	7233.95	7175.87	7204.39	7035.03	6987.731
I1 3-norm BD	43111	92154.2	43157.5	55093.1	36896.9	33445.9	23038.8	28562.7	119.1	157.976	160.31	169.291	163.148	157.375	169.689	162.119	55.8444	103.47	103.613	104.053	106.201	105.324	110.875	104.558	186.84	199.04	196.541	197.293	195.919	196.22	191.327	199.1721
I2 3-norm BD	43160.3	92104.2	43157	55093	36896.5	33445.6	23038.2	28563	172.928	104.241	160.359	169.353	163.209	157.416	169.69	162.184	105.868	53.455	103.676	104.044	106.313	105.337	110.933	104.575	197.924	188.573	196.48	197.283	195.88	196.337	191.481	199.3378
I3 3-norm BD	43159.3	92152.5	43106.2	55092.4	36896.2	33444.7	23038	28562.2	172.919	158.038	106.548	169.325	163.224	157.389	169.749	162.168	105.891	103.59	53.5269	104.086	106.287	105.354	110.985	104.549	197.914	199.105	185.391	197.332	195.919	196.475	191.44	199.2385
I4 3-norm BD	43159.8	92154.2	43157.3	55041.3	36897.1	33446.1	23038.5	28563	172.89	157.991	160.313	115.535	163.144	157.333	169.674	162.168	105.832	103.54	103.639	53.9784	106.233	105.3	110.87	104.544	197.856	198.994	196.416	187.302	195.801	196.303	191.255	199.1668
I5 3-norm BD	43159	92151.9	43155.9	55091.7	36847	33444.2	23037.4	28562	172.832	158.005	160.305	169.254	109.372	157.31	169.661	162.143	105.818	103.55	103.638	104.002	56.1612	105.343	110.875	104.548	197.836	199.009	196.562	197.275	184.152	196.306	191.368	199.1883
I6 3-norm BD	43160.1	92154.1	43157.6	55092.9	36896.3	33395.6	23038.3	28562.9	172.923	158.052	160.364	169.35	163.205	103.541	169.742	162.125	105.909	103.6	103.664	104.096	106.281	55.3157	110.939	104.554	197.901	199.117	196.51	197.293	195.832	184.944	191.437	199.2369
I7 3-norm BD	43159.8	92153.1	43156.5	55092.5	36896.7	33445.3	22990.5	28562.7	172.853	157.961	160.332	169.264	163.126	157.3	115.924	162.126	105.875	103.51	103.667	103.98	106.255	105.25	61.0451	104.499	197.719	198.984	196.51	197.306	195.81	196.319	178.337	199.3073
I8 3-norm BD	43160	92154	43157.3	55093.1	36896.3	33445.5	23038.5	28514.4	172.807	158.005	160.31	169.275	163.147	157.297	169.677	108.372	105.771	103.54	103.594	104.043	106.255	105.275	110.882	54.5092	197.798	199.066	196.446	197.388	195.808	196.308	191.362	188.6573
I1 10-norm BD	1.4E+14	1.2E+15	6.8E+13	3.7E+14	3E+13	2.6E+13	4.2E+12	1.8E+13	868.965	225.129	972.063	640.205	151.146	612.265	678.628	1219.04	6.3E-05	0.0001	7.9E-05	5.3E-05	5.7E-05	4.7E-05	6.4E-05	5.2E-05	2.64E-08	2.89E-08	2.99E-08	2.81E-08	2.85E-08	2.82E-08	2.66E-08	2.82E-08
I2 10-norm BD	1.4E+14	1.2E+15	6.8E+13	3.7E+14	3E+13	2.6E+13	4.2E+12	1.8E+13	868.965	225.129	972.063	640.205	151.146	612.265	678.628	1219.04	6.3E-05	0.0001	7.9E-05	5.3E-05	5.7E-05	4.7E-05	6.4E-05	5.2E-05	2.79E-08	2.65E-08	2.83E-08	2.83E-08	2.71E-08	2.74E-08	2.69E-08	2.88E-08
I3 10-norm BD	1.4E+14	1.2E+15	6.8E+13	3.7E+14	3E+13	2.6E+13	4.2E+12	1.8E+13	868.965	225.129	972.063	640.205	151.146	612.265	678.628	1219.04	6.3E-05	0.0001	7.9E-05	5.3E-05	5.7E-05	4.7E-05	6.4E-05	5.2E-05	2.89E-08	2.87E-08	2.72E-08	2.94E-08	2.84E-08	2.89E-08	2.78E-08	2.94E-08
I4 10-norm BD	1.4E+14	1.2E+15	6.8E+13	3.7E+14	3E+13	2.6E+13	4.2E+12	1.8E+13	868.965	225.129	972.063	640.205	151.146	612.265	678.628	1219.04	6.3E-05	0.0001	7.9E-05	5.3E-05	5.7E-05	4.7E-05	6.4E-05	5.2E-05	2.79E-08	2.86E-08	2.92E-08	2.67E-08	2.84E-08	2.81E-08	2.69E-08	2.85E-08
I5 10-norm BD	1.4E+14	1.2E+15	6.8E+13	3.7E+14	3E+13	2.6E+13	4.2E+12	1.8E+13	868.965	225.129	972.063	640.205	151.146	612.265	678.628	1219.04	6.3E-05	0.0001	7.9E-05	5.3E-05	5.7E-05	4.7E-05	6.4E-05	5.2E-05	2.76E-08	2.88E-08	2.95E-08	2.93E-08	2.61E-08	2.73E-08	2.66E-08	2.83E-08
I6 10-norm BD	1.4E+14	1.2E+15	6.8E+13	3.7E+14	3E+13	2.6E+13	4.2E+12	1.8E+13	868.965	225.129	972.063	640.205	151.146	612.265	678.628	1219.04	6.3E-05	0.0001	7.9E-05	5.3E-05	5.7E-05	4.7E-05	6.4E-05	5.2E-05	2.80E-08	2.76E-08	2.93E-08	2.79E-08	2.76E-08	2.62E-08	2.71E-08	2.85E-08
I7 10-norm BD	1.4E+14	1.2E+15	6.8E+13	3.7E+14	3E+13	2.6E+13	4.2E+12	1.8E+13	868.965	225.129	972.063	640.205	151.146	612.265	678.628	1219.04	6.3E-05	0.0001	7.9E-05	5.3E-05	5.7E-05	4.7E-05	6.4E-05	5.2E-05	2.73E-08	2.76E-08	2.82E-08	2.79E-08	2.71E-08	2.77E-08	2.55E-08	2.83E-08
I8 10-norm BD	1.4E+14	1.2E+15	6.8E+13	3.7E+14	3E+13	2.6E+13	4.2E+12	1.8E+13	868.965	225.129	972.063	640.205	151.146	612.265	678.628	1219.04	6.3E-05	0.0001	7.9E-05	5.3E-05	5.7E-05	4.7E-05	6.4E-05	5.2E-05	2.73E-08	2.79E-08	2.85E-08	2.78E-08	2.75E-08	2.74E-08	2.73E-08	2.64E-08

Fig. 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.

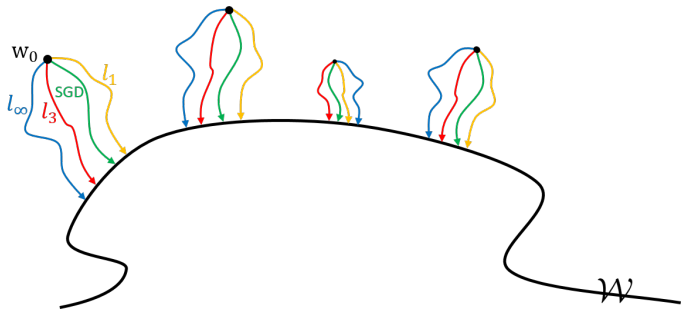


Fig. 10. An illustration of the experimental results. For each initialization w_0, we ran different SMD algorithms until convergence to a point on the set $\mathcal{W}$ (zero training error). We then measured all the pairwise distances from different w_∞ to different w_0, in different Bregman divergences. 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.

C. Distribution of the Final Weights of the Network

One may be curious to see how the final weights obtained by these different mirrors look, and whether, for example, mirror descent corresponding to the ℓ_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), and (d) show the histogram of the weights for ℓ_1-SMD, ℓ_2-SMD (SGD), ℓ_3-SMD, and ℓ_{10}-SMD, respectively. Note that each of the four histograms corresponds to an 11×10^6-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 ℓ_1-SMD has more weights at and close to zero, which again confirms that it induces sparsity. The histogram of the ℓ_2-SMD (SGD) looks almost identical to the histogram of the initialization, whereas the ℓ_3 and ℓ_{10} histograms are shifted to the right, so much so that almost all weights in the ℓ_{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).

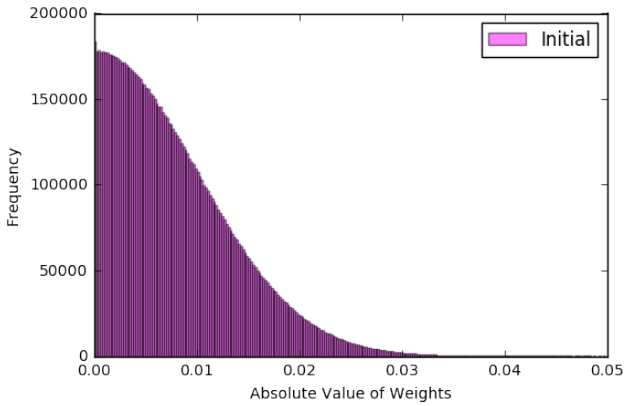


Fig. 11. Histogram of the absolute value of the initial weights in the network (half-normal distribution).

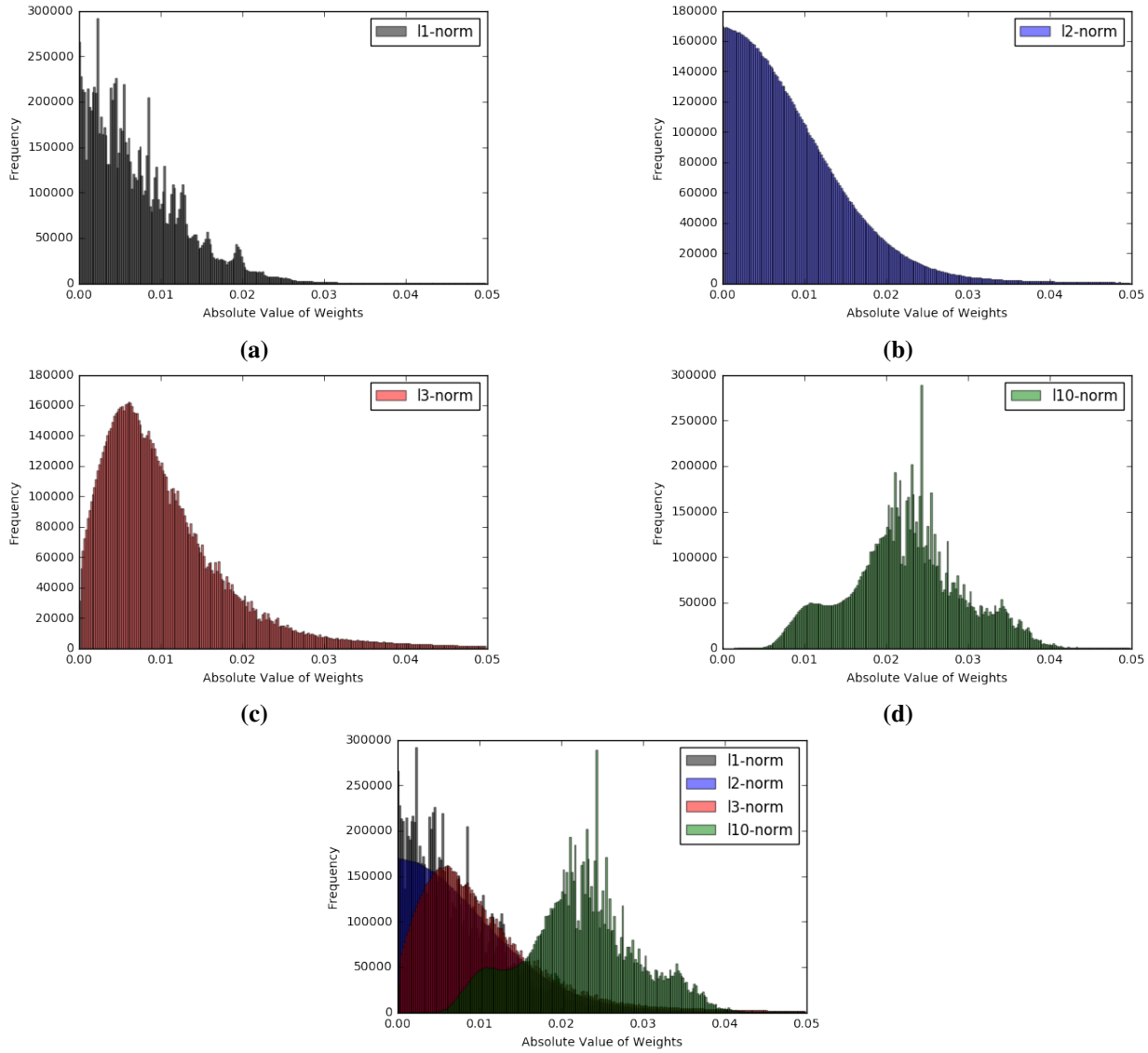


Fig. 12. Histogram of the absolute value of the final weights in the network for different SMD algorithms: (a) ℓ_1-SMD, (b) ℓ_2-SMD (SGD), (c) ℓ_3-SMD, and (d) ℓ_{10}-SMD. Note that each of the four histograms corresponds to an 11×10^6-dimensional weight vector that perfectly interpolates the data. Even though the weights remain quite small, the histograms are drastically different. ℓ_1-SMD induces sparsity on the weights, as expected. SGD does not seem to change the distribution of the weights significantly. ℓ_3-SMD starts to reduce the sparsity, and ℓ_{10} shifts the distribution of the weights significantly, so much so that almost all the weights are non-zero.

D. Generalization Errors of Different Mirrors/Regularizers

Here, we show the performance of the SMD algorithms discussed before for each individual run.

For MNIST, perhaps not surprisingly, all the four SMD algorithms achieve around 99% or higher accuracy for every individual run. For CIFAR-10, however, as noted before, there is a notable 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. The ℓ_{10} performs consistently the best, while the ℓ_1 performs consistently the worst. 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.

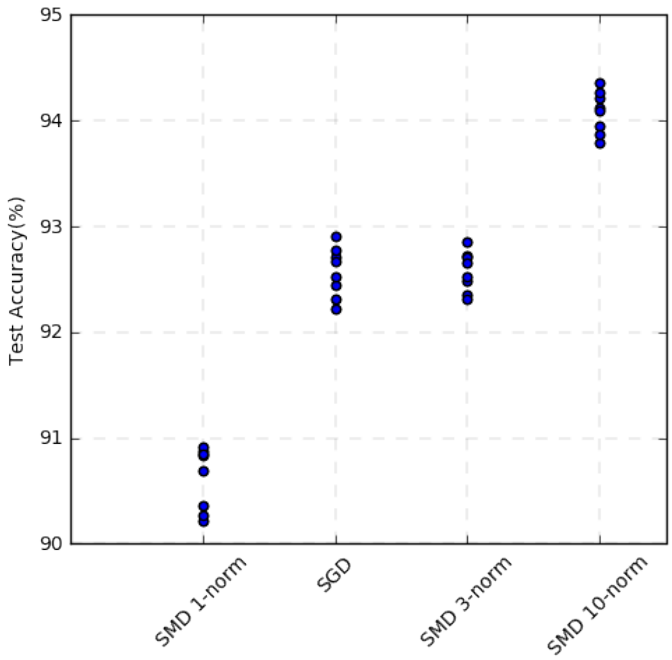


Fig. 13. Generalization performance of different SMD algorithms on the CIFAR-10 dataset using ResNet-18. ℓ_{10} performs consistently better, while ℓ_1 performs consistently worse.