Stochastic Mirror Descent on Overparameterized Nonlinear Models

Navid Azizan, *Member, IEEE*, Sahin Lale, *Graduate Student Member, IEEE*, and Babak Hassibi, *Member, IEEE*

***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. In this article, 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*—Deep learning, implicit regularization, mirror descent, overparameterization, stochastic gradient descent (SGD).**

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 are 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 nonconvex—sometimes even nonsmooth—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 also been 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 article, 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

Manuscript received June 8, 2020; revised December 7, 2020 and May 10, 2021; accepted May 24, 2021. This work was supported in part by the National Science Foundation under Grant ECCS-1509977, in part by Qualcomm Inc., in part by the NASA’s Jet Propulsion Laboratory through the President and Director’s Fund, in part by Amazon Web Services Inc., and in part by PIMCO, LLC. This article was presented in part at the 2019 International Conference on Machine Learning (ICML) Generalization Workshop, Long Beach, CA, USA. *(Corresponding author: Navid Azizan.)*

Navid 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).

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

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TNNLS.2021.3087480.

Digital Object Identifier 10.1109/TNNLS.2021.3087480

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 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 solution obtained by ℓ_∞-norm SMD (which uses all the parameters in the already highly overparameterized network) consistently generalizes better than the one obtained by SGD, which in turn outperforms the sparser one obtained by the 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 the 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 SMD algorithm is the closest, compared with 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 the closest to the initialization in the 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 initialized 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 $\{(x_i, y_i) : i = 1, \ldots, n\}$, where $x_i \in \mathbb{R}^d$ are the inputs and $y_i \in \mathbb{R}$ are the labels. The model (which can be, e.g., linear, a deep neural network and so on) 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

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

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

The total loss on the training set (empirical risk) can be expressed as $L(w) = \sum_{i=1}^n L_i(w)$, where $L_i(\cdot) = \ell(y_i, f(x_i, w))$ is the loss on the individual data point i and $\ell(\cdot, \cdot)$ is a differentiable nonnegative 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 and Huber loss). In this case, $L(w) = \sum_{i=1}^n \ell(y_i - f(x_i, w))$. The conventional gradient descent (GD) algorithm, for example, 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., $w_i = \nabla\psi^{-1}(\nabla\psi(w_{i-1}) - \eta\nabla L(w_{i-1}))$. 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.

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

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 nonnegative, 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) $D_\psi(w, w') = \sum_j w_j \log(w_j/w'_j) - \sum_j w_j + \sum_j w'_j$, which corresponds to the exponentiated GD (also known as the exponential weights) algorithm [17].

***p*-Norm 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-norm 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 SMD algorithms, to see whether 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

$$\begin{aligned} w_i[j] = {} & ||w_{i-1}[j]|^{q-1}\operatorname{sign}(w_{i-1}[j]) - \eta\nabla L_i(w_{i-1})[j]|^{\frac{1}{q-1}} \\ & \times \operatorname{sign}(|w_{i-1}[j]|^{q-1}\operatorname{sign}(w_{i-1}[j]) - \eta\nabla L_i(w_{i-1})[j]) \end{aligned} \tag{6}$$

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

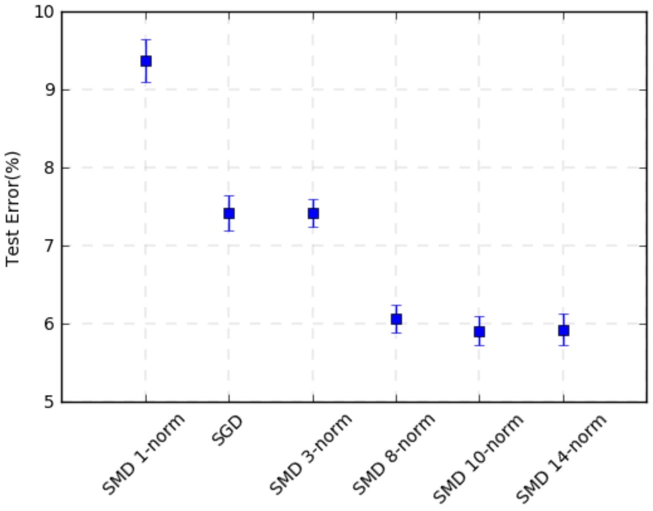


Fig. 1. Generalization performance of different SMD algorithms on the CIFAR-10 dataset using the ResNet-18 neural network. SMDs with higher norms (which are surrogates for ℓ_∞ norm) tend to achieve better generalization performance (lower test error) than the ones with lower norms. In particular, ℓ_{14}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

Note that, for this particular choice of potential function, the update rule is separable, i.e., the jth element of the new weight vector can be computed using only the jth 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 are stored are the weights. Code for SMD, which can be applied to arbitrary models, is available at https://github.com/SahinLale/StochasticMirrorDescent.

A. 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

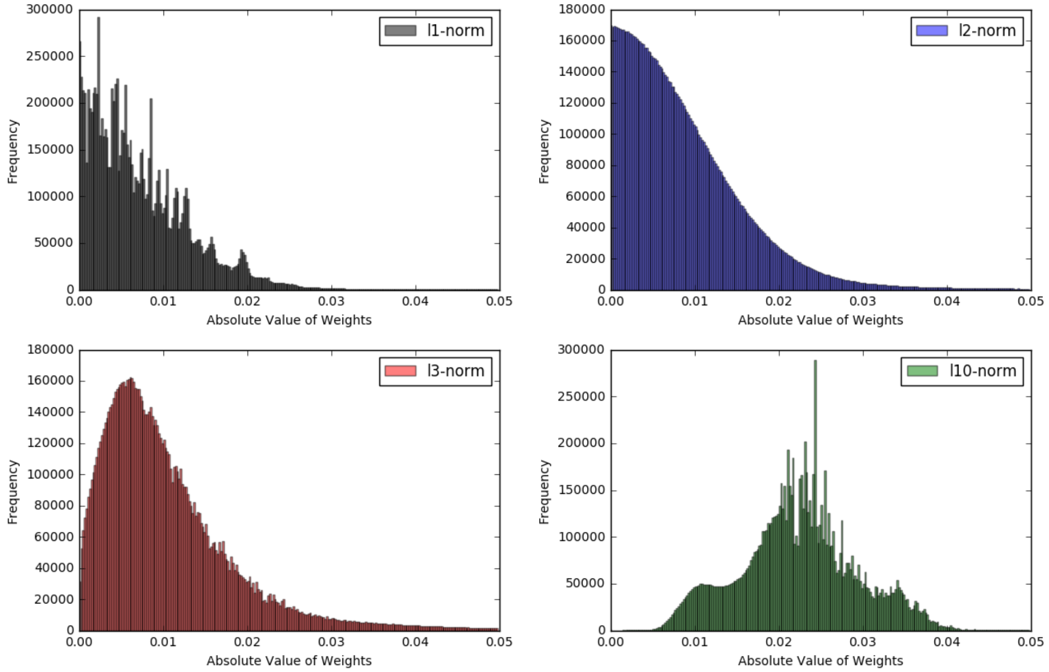


Fig. 2. Histogram of the absolute value of the final weights in the network for different SMD algorithms with different potentials. Note that each of the four histograms corresponds to an 11×10^6 dimensional weight vector that perfectly interpolates the data. Even though the weights remain quite small, 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 nonzero.

CIFAR-10. This is particularly remarkable, given that this very architecture had been designed with training with SGD in mind.

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 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 nonzero) 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 the 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 $\mathcal{W} = \{w \mid y_i = x_i^T w,\ i = 1, \ldots, n\}$, and the loss is $L_i(w) = \ell(y_i - x_i^T w)$. The following result characterizes the solution that SMD converges to ([20] and [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 $\eta \leq 1/\|x_i\|^2$. 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

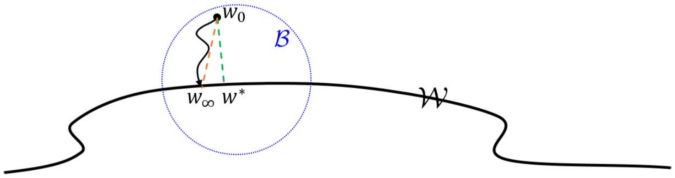


Fig. 3. Illustration of the parameter space. $\mathcal{W}$ represents the set of global minima, w_0 is the initialization, $\mathcal{B}$ is the local neighborhood, w^* is the closest global minimum to w_0 (in Bregman divergence), and w_∞ is the minimum that SMD converges to.

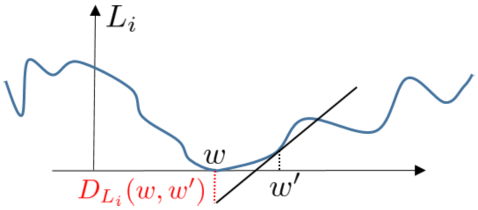


Fig. 4. Illustration of $D_{L_i}(w, w') \geq 0$ in a local region in Assumption 1.

$\psi(\cdot)$. In particular, this means that SGD initialized around zero acts as an ℓ_2-norm regularizer. In what follows, we show that these results continue to hold for highly overparameterized nonlinear models in an approximate sense.

B. Main Results

Let us define

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

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 also the discussion in Appendix A in the Supplementary Material). 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 is inside the intersection of all such neighborhoods for $i = 1, \ldots, n$. In other words, we require w_0 close enough to $\mathcal{W}$. ϵ in Assumption 1 characterizes the closeness.

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

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

This is 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 (see Fig. 3).

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 the Bregman divergence by w^*, i.e.,

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

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_∞ that is "very close" to w^* (see Fig. 3).

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 (Assumption 1), we converge to a point $w_\infty \in \mathcal{W}$ that is $o(\epsilon)$ away from w^*, in Bregman divergence.

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

TABLE I

FIXED INITIALIZATION (THE SETTING SHOWN IN FIG. 5). WE HAVE TRAINED THE NETWORK FROM A COMMON FIXED INITIALIZATION WITH FOUR DIFFERENT SMDs (ℓ_1, ℓ_2, ℓ_3, AND ℓ_{10}) TO OBTAIN FOUR DIFFERENT INTERPOLATING SOLUTIONS. FOR EACH INTERPOLATING SOLUTION, WE CAN COMPUTE ITS DISTANCE FROM THE INITIAL WEIGHT VECTOR. SINCE WE HAVE FOUR DIFFERENT POTENTIALS, WE HAVE FOUR DIFFERENT BREGMAN DIVERGENCES TO ASSESS THE DISTANCE BY. THIS GIVES US A 4 × 4 TABLE. THE COLUMNS CORRESPOND TO THE FOUR 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}

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 (see Appendix A in the Supplementary Material for a proof).

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 the 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 four-layer convolutional neural network (CNN) on the MNIST dataset [30] and the ResNet-18 [31] on the CIFAR-10 dataset [32]. We use the 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 zero 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 six independent initializations for the CNN and eight 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: 1) $q = 1 + 0.01$, as a surrogate for ℓ_1 norm; 2) $q = 2$, which is SGD; 3) $q = 3$; and 4) $q = 10$, as a surrogate for ℓ_∞ norm. We use a fixed step size η, chosen small enough to avoid diverging. See Appendix B in the Supplementary Material for more details on the experiments.

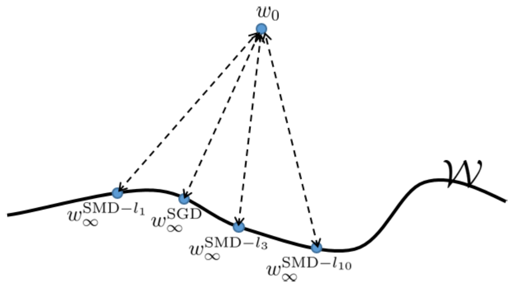

w_0
$\mathcal{W}$
$w_\infty^{\mathrm{SMD}-l_1}$
w_∞^{SGD}
$w_\infty^{\mathrm{SMD}-l_3}$
$w_\infty^{\mathrm{SMD}-l_{10}}$

Fig. 5. Illustration of the experiments in Table I.

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. Tables I and II (as illustrated in Figs. 5 and 6) 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 a 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 (see the Supplementary Material).

It is worth emphasizing that there is virtually no additional overhead in training the networks with ℓ_q-norm SMD,

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE II

FIXED POTENTIAL (THE SETTING SHOWN IN FIG. 6). WE HAVE TRAINED THE NETWORK FROM EIGHT DIFFERENT INITIAL POINTS WITH THE SAME SMD (IN THIS CASE, SGD) TO OBTAIN EIGHT 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

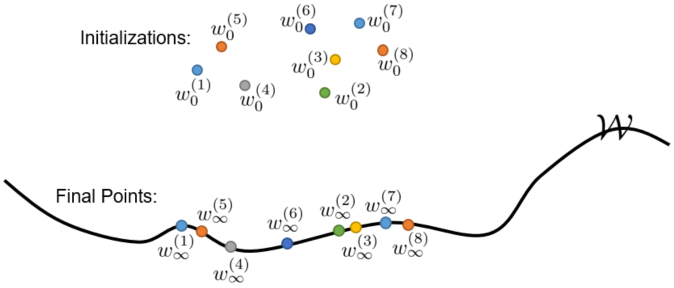


Fig. 6. Illustration of the experiments in Table II.

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.

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 $D_{L_i}(w, w') \geq 0, i = 1, \ldots, n$, for all $w' \in \mathcal{B}$.

Theorem 3: Consider the set of interpolating parameters $\mathcal{W} = \{w \in \mathbb{R}^p \mid f(x_i, w) = y_i, i = 1, \ldots, n\}$, 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 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 nonnegative. 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 nonnegative. Since the loss is nonnegative, $\eta L_i(w_i)$ is always nonnegative. As a result, we have

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

This implies that $D_\psi(w, w_i) \leq \epsilon$, which means that 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 $\mathcal{B}$; 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 $i = 1, \ldots, T$, the first terms on the right- and left-hand side cancel each other telescopically, and we have

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

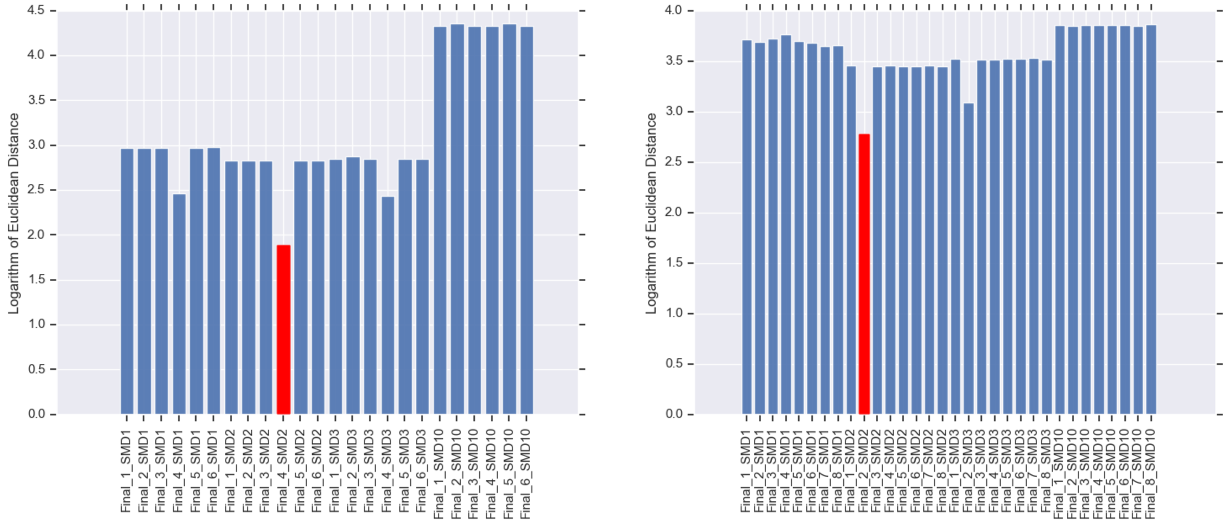
Since $D_\psi(w, w_T) \geq 0$, we have $\sum_{i=1}^{T} [D_{\psi - \eta L_i}(w_i, w_{i-1}) + \eta L_i(w_i) + \eta D_{L_i}(w, w_{i-1})] \leq D_\psi(w, w_0)$. If we take $T \to \infty$, the sum still has to remain bounded, i.e.,

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

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 nonnegative. The second term $\eta L_i(w_i)$ is nonnegative because of the nonnegativity of the loss. Finally, the last term $D_{L_i}(w, w_{i-1})$ is nonnegative because $w_{i-1} \in \mathcal{B}$ for all i. Hence, all the three terms in the summand are nonnegative, and because the sum is bounded, they must go to zero as $i \to \infty$. In particular

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

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, $w_\infty \in \mathcal{W}$. □



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 (six for MNIST and eight for CIFAR-10) initial points with four 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 Tables VIII and IX in Appendix B in the Supplementary Material.

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 the Bregman divergence). The proof is again based on the fundamental identity of SMD.

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

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

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 $D_{L_i}(w, w_{i-1})$:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$\begin{aligned} \alpha\|w_\infty - w_{i-1}\|^2 &\leq (w_\infty - w_{i-1})^T H_{f_i}\big(w_i'\big)(w_\infty - w_{i-1}) \\ &\leq \beta\|w_\infty - w_{i-1}\|^2 \end{aligned} \quad (26)$$

and

$$\begin{aligned} \alpha\|w^* - w_{i-1}\|^2 &\leq (w^* - w_{i-1})^T H_{f_i}\big(w_i''\big)(w^* - w_{i-1}) \\ &\leq \beta\|w^* - w_{i-1}\|^2. \end{aligned} \quad (27)$$

Furthermore, 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., $\big[(w^* - w_{i-1})^T H_{f_i}(w_i'')(w^* - w_{i-1}) - (w_\infty - w_{i-1})^T H_{f_i}(w_i')(w_\infty - w_{i-1})\big]$, is also $O(\epsilon)$, and we have

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

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 $\tilde{w}_i \in$ 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

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

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

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

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

$$D_\psi(w, w_0) = \psi(w) - \psi(w_0) - \nabla\psi(w_0)^T(w - w_0) \quad (31)$$

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

$$D_\psi(w, w_0) = \psi(w) - \psi(w_0). \quad (32)$$

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

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

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 (see [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, rely on the initialization being 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 $\|w^* - w_0\|$. 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] and [21]. Gunasekar *et al.* [21] considered *linear* overparameterized models and showed that if SMD happens to converge to a global minimum, then the global minimum will be the one that is closest in Bregman divergence to the initialization, a result they obtain by examining the KKT conditions. However, they do not provide any conditions for convergence and whether SMD converges with a fixed step size or not. Azizan and Hassibi [20] also studied linear models but derived 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 that 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 article, 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 *et al.*, "Deep speech 2: End-to-end speech recognition in English and mandarin," in *Int. Conf. Mach. Learn.*, vol. 2016, pp. 173–182.

[2] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 6645–6649.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[4] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, p. 529, 2015.

[5] D. Silver *et al.*, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, Jan. 2016.

[6] Y. Wu *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," 2016, *arXiv:1609.08144*. [Online]. Available: http://arxiv.org/abs/1609.08144

[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," 2016, *arXiv:1611.03530*. [Online]. Available: http://arxiv.org/abs/1611.03530

[9] S. Ma, R. Bassily, and M. Belkin, "The power of interpolation: Understanding the effectiveness of SGD in modern over-parametrized learning," in *Proc. 35th Int. Conf. Mach. Learn.*, vol. 80, 2018, pp. 3325–3334.

[10] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht, "Gradient descent only converges to minimizers," in *Proc. Conf. Learn. 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 *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4151–4161.

[12] A. Nemirovski and D. B. Yudin, *Problem Complexity and Method Efficiency in Optimization*. New York, NY, USA: Wiley, 1983.

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

[14] N. Cesa-Bianchi, P. Gaillard, G. Lugosi, and G. Stoltz, "Mirror descent meets fixed share (and feels no regret)," in *Proc. Adv. Neural Inf. Process. Syst.*, 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 *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 7043–7052.

[16] Y. Lei and D.-X. Zhou, "Convergence of online mirror descent," *Appl. Comput. Harmon. Anal.*, vol. 48, no. 1, pp. 343–373, Jan. 2020.

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

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

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

[20] N. Azizan and B. Hassibi, "Stochastic gradient/mirror descent: Minimax optimality and implicit regularization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019, pp. 1–18.

[21] S. Gunasekar, J. Lee, D. Soudry, and N. Srebro, "Characterizing implicit bias in terms of optimization geometry," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1827–1836.

[22] Y. Li and Y. Liang, "Learning overparameterized neural networks via stochastic gradient descent on structured data," in *Proc. Adv. Neural Inf. Process. Syst.*, 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," 2018, *arXiv:1811.03804*. [Online]. Available: http://arxiv.org/abs/1811.03804

[24] Z. Allen-Zhu, Y. Li, and Z. Song, "A convergence theory for deep learning via over-parameterization," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 242–252.

[25] Y. Cao and Q. Gu, "Generalization error bounds of gradient descent for learning over-parameterized deep ReLU networks," 2019, *arXiv:1902.01384*. [Online]. Available: http://arxiv.org/abs/1902.01384

[26] S. Oymak and M. Soltanolkotabi, "Overparameterized nonlinear learning: Gradient descent takes the shortest path," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 4951–4960.

[27] B. Hassibi, A. H. Sayed, and T. Kailath, *Indefinite-Quadratic Estimation Control: A Unified Approach to H2 H-Infinity Theories*, vol. 16. Philadelphia, PA, USA: SIAM, 1999.

[28] D. Simon, *Optimal State Estimation: Kalman, H Infinity, and Nonlinear Approaches*. Hoboken, NJ, USA: Wiley, 2006.

[29] B. Hassibi, A. H. Sayed, and T. Kailath, "Hoo optimality criteria for LMS and backpropagation," in *Proc. Adv. Neural Inf. Process. Syst.*, 1994, pp. 351–358.

[30] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[32] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Citeseer, Univ. Toronto, Toronto, ON, Canada, Tech. Rep., 2009.

[33] A. Achille and S. Soatto, "Emergence of invariance and disentanglement in deep representations," 2017, *arXiv:1706.01350*. [Online]. Available: http://arxiv.org/abs/1706.01350

[34] P. Chaudhari and S. Soatto, "Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks," in *Proc. Inf. Theory Appl. Workshop (ITA)*, Feb. 2018, pp. 1–10.

[35] R. Shwartz-Ziv and N. Tishby, "Opening the black box of deep neural networks via information," 2017, *arXiv:1703.00810*. [Online]. Available: http://arxiv.org/abs/1703.00810

[36] M. Soltanolkotabi, A. Javanmard, and J. D. Lee, "Theoretical insights into the optimization landscape of over-parameterized shallow neural networks," 2017, *arXiv:1707.04926*. [Online]. Available: http://arxiv.org/abs/1707.04926

[37] B. Neyshabur, R. Tomioka, R. Salakhutdinov, and N. Srebro, "Geometry of optimization and implicit regularization in deep learning," 2017, *arXiv:1705.03071*. [Online]. Available: http://arxiv.org/abs/1705.03071

[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 *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 3351–3360.

[39] S. Gunasekar, B. Woodworth, S. Bhojanapalli, B. Neyshabur, and N. Srebro, "Implicit regularization in matrix factorization," in *Proc. Inf. Theory Appl. Workshop (ITA)*, Feb. 2018, pp. 6152–6160.

[40] D. Soudry, E. Hoffer, M. Shpigel Nacson, S. Gunasekar, and N. Srebro, "The implicit bias of gradient descent on separable data," 2017, *arXiv:1710.10345*. [Online]. Available: http://arxiv.org/abs/1710.10345

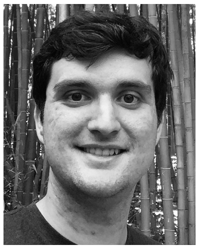
[41] S. Gunasekar, J. Lee, D. Soudry, and N. Srebro, "Implicit bias of gradient descent on linear convolutional networks," 2018, *arXiv:1806.00468*. [Online]. Available: http://arxiv.org/abs/1806.00468

[42] P. Mianjy, R. Arora, and R. Vidal, "On the implicit bias of dropout," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 3537–3545.

Navid Azizan (Member, IEEE) received the B.Sc. degree from the Sharif University of Technology, Tehran, Iran, in 2013, the M.S. degree from the University of Southern California, Los Angeles, CA, USA, in 2015, and the Ph.D. degree from the California Institute of Technology, Pasadena, CA, in 2020.

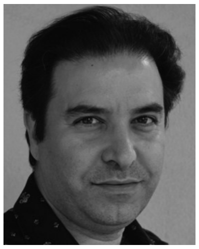
He is currently an incoming Assistant Professor at the Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, and a Post-Doctoral Scholar at Stanford University, Stanford, CA, USA. His research interests include machine learning, mathematical optimization, control theory, and networks.

Dr. Azizan's 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 the gold medalist at the 2008 National Physics Olympiad in Iran.



Sahin Lale (Graduate Student Member, IEEE) was born in İzmir 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, where he is currently pursuing the Ph.D. degree in electrical engineering.

His research interests include reinforcement learning, control theory, machine learning, and information theory.



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

He has been with the California Institute of Technology, Pasadena, CA, since January 2001, where he is currently the Mose and Lilian S. Bohn Professor of Electrical Engineering. From 2013 to 2016, he was the Gordon M. Binder/Amgen Professor of Electrical Engineering. From 2008 to 2015, he was Executive Officer of Electrical Engineering and the 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. From November 1998 to December 2000, he was a Member of the Technical Staff at the Mathematical Sciences Research Center, Bell Laboratories, Murray Hill, NJ, USA. He has held short-term appointments at the Ricoh California Research Center, Menlo Park, CA; the Indian Institute of Science, and Linköping University, Linköping, 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, NY, USA: SIAM, 1999) and *Linear Estimation* (Englewood Cliffs, NJ, USA: Prentice Hall, 2000).

Dr. Hassibi was a recipient of the 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. He was a participant in the 2004 National Academy of Engineering "Frontiers in Engineering" program. He has been a Guest Editor of the IEEE Transactions on Information Theory—Special Issue on Space-Time Transmission, Reception, Coding and Signal Processing. He was an Associate Editor of the IEEE Transactions on Information Theory from 2004 to 2006. He is also an Editor of *Foundations and Trends in Communications and Information Theory* and the IEEE Transactions on Network Science and Engineering. He was a Distinguished Lecturer of the IEEE Information Theory Society from 2016 to 2017 and the Co-Chair of the 2020 IEEE International Symposium on Information Theory (ISIT 2020), Los Angeles, CA.