Convergence and Implicit Regularization of Stochastic Mirror Descent in Overparameterized Models*

Sourav Sahoo[1]

[1]Department of Electrical Engineering, Indian Institute of Technology Madras
ee17b040@smail.iitm.ac.in

Abstract

Stochastic Mirror Descent (SMD) algorithms comprise a family of algorithms that are increasingly being used in multiple domains such as machine learning, optimization, signal processing, and control. A popular member of this family is the Stochastic Gradient Descent (SGD) algorithm, which has gained enormous popularity in recent times due to the unprecedented growth in deep learning. Like SGD, SMD algorithms perform updates along the negative gradient of a stochastically chosen loss function. However, instead of performing updates directly on the objective parameter, updates are performed in a "mirror domain" whose transformation is given by the gradient of a strictly convex function. In this work, we explicitly shed some light on *convergence* and *implicit regularization* due to stochastic mirror descent in over-parameterized linear and non-linear models. All the codes for this work is made available at https://github.com/sourav22899/ee5121-convex/tree/master/term-paper.

1 Introduction

Deep learning (LeCun et al., 2015) has been tremendously successful in a wide variety of tasks ranging from image classification (Krizhevsky et al., 2012), machine translation (Bahdanau et al., 2014), and speech recognition (Graves et al., 2013) to playing games at superhuman level (Silver et al., 2016). Although deep learning techniques have revolutionized multiple fields, the reasons behind the success are arguably unclear. Most of the current state-of-the-art deep networks contain much more parameters than the number of data points, i.e., they are *highly over-parameterized*. Hence, neural networks can fit any random set of data points and targets completely independent of each other (Zhang et al., 2016). Typically, there are infinitely many global minima in a deep network's optimization landscape, but optimization algorithms often converge to a global minimum that also minimizes the generalization error. In recent times, multiple efficient variants of the SGD, such as AdaGrad (Duchi et al., 2011) and Adam (Kingma and Ba, 2014), have been proposed as improved optimization methods. However, all these algorithms have differences in the optimal solution obtained and generalization error (Wilson et al., 2017). These algorithms converge to a solution that approximately minimizes generalization error even in the absence of explicit regularization methods such as dropout, early stopping, and batch normalization (Zhang et al., 2016). So, it can be assumed that these algorithms perform some form of *implicit regularization* during the training procedure.

In this work, we explore the family of stochastic mirror descent (SMD) algorithms and discuss the implicit regularization caused by this family of algorithms both in the linear and non-linear domain. We mainly focus on the over-parameterized models as they are easier to understand and more prevalent in the current times. We reiterate several results proposed in the primary reference in this context and conduct extensive experiments to prove the same.

The rest of this work is organized as follows: we describe the contribution of this work in Section 2. A brief background of SMD and Bregman divergence is present in Section 3. We present the theoretical results on convergence and implicit regularization in Section 4. We perform the experiments in Section 5 and discuss the results in Section 6. Proofs and details of the experiments provided in the appendix A and B respectively.

*Term paper for EE5121 Convex Optimization, Indian Institute of Technology Madras, Spring 2020. This work is primarily based on three papers (Azizan and Hassibi, 2018, 2019; Azizan et al., 2019)

2 Contributions

In this work, we discuss some of the underlying theory behind stochastic mirror descent present in the existing literature (Nemirovsky and Yudin, 1983; Beck and Teboulle, 2003; Cesa-Bianchi et al., 2012). We mainly focus on convergence and implicit regularization of stochastic mirror descent for over-parameterized linear (Azizan and Hassibi, 2018) and non-linear (Azizan and Hassibi, 2019) models. We conduct experiments to prove a few selected results proposed in the papers mentioned above for the linear regime and reproduce the results for the non-linear regime (Azizan et al., 2020). We also design and conduct some novel experiments for the linear models, as described in Section 5.1.2.

3 Background

3.1 Preliminaries

Let the training set be denoted as $\mathcal{D} = \{(x_i, y_i) : i \in [n]\}$ where $x_i \in \mathbb{R}^m$ and $y_i \in \mathbb{R}$. Let the data $y_i = f(x_i, w) + v_i$ where $f(x_i, w)$ denote some model (either linear or non-linear) and v_i represent the noise present in the observations. No assumption is made on the type of noise. As we are interested in the over-parameterized case, we assume $m > n$. So, there exists a subspace $\mathcal{W}$ such that $\mathcal{W} = \{w \in \mathbb{R}^m | y_i = f(x_i, w),\ \forall i \in [n]\}$. The total loss on training set (empirical risk) is given as $L(w) = \sum_{i=1}^{n} L_i(w)$ where $L_i(\cdot)$ denotes the loss due to observation (x_i, y_i). Typically, $L_i(w) = l(y_i - f(x_i, w))$ for some non-negative differentiable function $l(\cdot)$ with $l(0) = 0$. Common examples for l are convex functions with global minima at 0 such as the squared loss (ℓ_2) loss, Huber loss etc. Furthermore, by definition, $L_i(w) = 0$ for $\forall w \in \mathcal{W}$.

3.2 Bregman Divergence

Let $\psi : \mathbb{R}^m \to \mathbb{R}$ be a continuously differentiable, strictly convex function and $p, q \in \mathbb{R}^m$, then Bregman divergence is the difference between the functional value at p and the first-order Taylor expansion around q evaluated at p (Bregman, 1967). Mathematically, the expression is given in (1).

$$D_\psi(p, q) = \psi(p) - \psi(q) - \nabla\psi(q)^\top (p - q) \tag{1}$$

For example, if $\psi(w) = \frac{1}{2}\|w\|_2^2$, $D_\psi(p, q) = \frac{1}{2}\left(\|p\|_2^2 - \|q\|_2^2 - 2q^\top(p - q)\right) = \frac{1}{2}\left(\|p\|_2^2 + \|q\|_2^2 - 2q^\top p\right)$ $= \frac{1}{2}\|p - q\|_2^2$. It is to be noted that as $\psi(\cdot)$ is convex, $D_\psi(\cdot, \cdot) \geqslant 0$ and equality holds *iff* both the arguments are same.

3.3 Stochastic Mirror Descent

Consider a strictly convex differentiable function $\psi(\cdot)$ which is called the *potential function* hereafter. Then SMD updates are given as defined in (2).

$$w_i = \operatorname*{argmin}_w \eta w^\top \nabla L_i(w_{i-1}) + D_\psi(w, w_{i-1}) \tag{2}$$

where $\eta > 0$ is the step size. As the expression on the right in (2) is convex in w and differentiable, if we differentiate w.r.t w and set the derivative to 0, we get:

$$0 = \eta\nabla L_i(w_{i-1}) + \nabla\psi(w) - \nabla\psi(w_{i-1})$$

By definition in (2), this minima is attained at $w = w_i$, so by rearranging the terms, we get that, SMD updates can be re-written as (3).

$$\nabla\psi(w_i) = \nabla\psi(w_{i-1}) - \eta\nabla L_i(w_{i-1}) \tag{3}$$

It is to be noted that as $\psi(\cdot)$ is strictly convex, $\nabla\psi(\cdot)$ generates a one-to-one mapping, so a unique w_i is generated at each recursion step of (3). Another way to view this is by describing $\nabla\psi(\cdot)$ as a transformation applied on w_i. Hence, sometimes $\nabla\psi(w)$ is called the *dual* variable and w is called the *primal* variable. Further, it is trivial to notice that when $\psi(w) = \frac{1}{2}\|w\|_2^2$, then the update rule becomes the familiar SGD algorithm.

Let us now define a term similar to the Bregman divergence w.r.t to the loss function $L_i(w) = l(y_i - f(x_i, w))$ as mentioned in (4).

$$D_{L_i}(w, w') = L_i(w) - L_i(w') - \nabla L_i(w')^\top (w - w') \tag{4}$$

Although the expression looks very similar to actual Bregman Divergence, a significant difference is that here, the function $L_i(\cdot)$ *need not be convex.* So, $D_{L_i}(\cdot, \cdot)$ need not be necessarily non-negative.

3.4 Fundamental Identity of Stochastic Mirror Descent

In this section we describe the fundamental identity of SMD which will be used in the subsequent section to prove some results.

Lemma 1. *For any model $f(\cdot, \cdot)$, any differentiable loss $l(\cdot)$, any parameter w and noise $\{v_i\}_{i=1}^n$ that satisfy $y_i = f(x_i, w) + v_i$, $\forall i \in [n]$ and $\eta > 0$, the SMD iterates satisfy the following:*

$$D_\psi(w, w_{i-1}) + \eta l(v_i) = D_\psi(w, w_i) + E_i(w_i, w_{i-1}) + \eta D_{L_i}(w, w_{i-1}) \tag{5}$$

for $i \geqslant 1$, where

$$E_i(w_i, w_{i-1}) = D_{\psi - nL_i}(w_i, w_{i-1}) + \eta L_i(w_i) \tag{6}$$

This result follows directly from the definition of Bregman Divergence for the functions $\psi(\cdot), L_i(\cdot)$ and $\psi(\cdot) - \eta L_i(\cdot)$ and using the SMD update rule mentioned in (3). The exact proof is mentioned in the appendix A. Adding the expression in (5) for $i = 1, \ldots, T$, we get the result in (7).

Corollary 1.1.

$$D_\psi(w, w_0) + \sum_{i=1}^{T} \eta l(v_i) = D_\psi(w, w_T) + \sum_{i=1}^{T} \left(E_i(w_i, w_{i-1}) + \eta D_{L_i}(w, w_{i-1})\right) \tag{7}$$

4 Convergence and Implicit Regularization

Now that we have described the preliminaries and the fundamental identity of SMD, we will now discuss convergence and implicit regularization in over-parameterized models. We will prove the results for the linear models and qualitatively describe the extension of the results for the non-linear regime (typical deep neural networks). The formal proofs for the non-linear domain are beyond the scope of this work.

4.1 Over-parameterized Linear Models

As the name suggests, over-parameterized models linear models are a system of underdetermined linear equations. In this scenario, $y_i = f(x_i, w) = x_i^\top w$ and $v_i = 0$, $\forall i \in [n]$ as the model perfectly fits the data. By definition, $w \in \mathcal{W}$ as described in Section 3.1. So, Corollary 1.1 reduces to the expression in (8).

$$D_\psi(w, w_0) = D_\psi(w, w_T) + \sum_{i=1}^{T} \left(E_i(w_i, w_{i-1}) + \eta D_{L_i}(w, w_{i-1})\right) \tag{8}$$

It is to be noted that $E_i(w_i, w_{i-1})$ is *independent* of w as clear from (6). Now, we propose that for the linear case the second term of the summand in (8) is also *independent* of w.

Proof. [1]

$$\begin{aligned}
D_{L_i}(w, w_{i-1}) &= L_i(w) - L_i(w_{i-1}) - \nabla L_i(w_{i-1})^\top (w - w_{i-1}) \\
&= 0 - L_i(w_{i-1}) - \nabla L_i(w_{i-1})^\top (w - w_{i-1}) && \textit{as } L_i(w) = 0 \textit{ for } w \in \mathcal{W} \\
&= -l(y_i - x_i^\top w_{i-1}) - l'(y_i - x_i^\top w_{i-1})(-x_i)(w - w_{i-1}) && \textit{as } L_i(w_{i-1}) = l(y_i - x_i^\top w_{i-1}) \\
&= -l(y_i - x_i^\top w_{i-1}) + l'(y_i - x_i^\top w_{i-1})(x_i^\top w - x_i^\top w_{i-1}) \\
&= -l(y_i - x_i^\top w_{i-1}) + l'(y_i - x_i^\top w_{i-1})(y_i - x_i^\top w_{i-1}) && \text{(independent of } w\text{)}
\end{aligned}$$

□

[1] This proof is described as it is present in (Azizan and Hassibi, 2019) without any alteration.

So, to minimize (8) on both sides w.r.t $w \in \mathcal{W}$, we only need to the consider the terms $D_\psi(w, w_0)$ and $D_\psi(w, w_T)$. This leads to the following result.

Theorem 2. *For any differentiable loss $l(\cdot)$, any initialization w_0, any step size $\eta > 0$, if the SMD iterates converges to $w_\infty \in \mathcal{W}$, then*

$$w_\infty = \operatorname*{argmin}_{w \in \mathcal{W}} D_\psi(w, w_0) \tag{9}$$

Proof. Consider the case when $T \to \infty$,

$$\begin{aligned} \operatorname*{argmin}_{w \in \mathcal{W}} D_\psi(w, w_0) &= \operatorname*{argmin}_{w \in \mathcal{W}} D_\psi(w, w_T) + \sum_{i=1}^{T} \left(E_i(w_i, w_{i-1}) + \eta D_{L_i}(w, w_i) \right) && \text{From (8)} \\ &= \operatorname*{argmin}_{w \in \mathcal{W}} D_\psi(w, w_T) = \operatorname*{argmin}_{w \in \mathcal{W}} D_\psi(w, w_\infty) = w_\infty \end{aligned}$$

□

The last equality holds because $D_\psi(w, w_\infty) \geqslant 0$ and $D_\psi(w, w_\infty) = 0$ *iff* $w = w_\infty$. In the constrained subspace $\mathcal{W}$, this minima is actually attained because, by assumption of Thereom 2, $w_\infty \in \mathcal{W}$. So, $\operatorname{argmin}_{w \in \mathcal{W}} D_\psi(w, w_\infty) = w_\infty$. Note that we have not yet proved if $w_\infty \in \mathcal{W}$, i.e., SMD updates even converge to a point that interpolates the data.

Corollary 2.1. *If the initialization of the SMD updates, $w_0 = \operatorname{argmin}_{w \in \mathbb{R}^m} \psi(w)$, then the expression in (9) reduces to:*

$$w_\infty = \operatorname*{argmin}_{w \in \mathcal{W}} \psi(w) \tag{10}$$

Proof.

$$\begin{aligned} w_\infty &= \operatorname*{argmin}_{w \in \mathcal{W}} D_\psi(w, w_0) && \text{From Theorem 2} \\ &= \operatorname*{argmin}_{w \in \mathcal{W}} \psi(w) - \psi(w_0) - \nabla \psi(w_0)^\top (w - w_0) && \text{From definition} \\ &= \operatorname*{argmin}_{w \in \mathcal{W}} \psi(w) - \psi(w_0) && \text{As } w_0 = \operatorname*{argmin}_{w \in \mathbb{R}^m} \psi(w) \implies \nabla \psi(w_0) = 0 \\ &= \operatorname*{argmin}_{w \in \mathcal{W}} \psi(w) \end{aligned}$$

□

From the corollary, we see that the converged point minimizes the potential function in the constrained domain $\mathcal{W}$ i.e., the solution space, which implies that the solution is *implicitly regularized.* For example, if $\psi(w) = \|w\|_1^1$ and initialization $w_0 \sim 0$ (global minima of $\psi(w)$), then the converged solution is sparse. If $\psi(w) = \|w\|_2^2$ and initialization $w_0 \sim 0$ (global minima of $\psi(w)$), then the final result is similar to the one obtained with an explicit ℓ_2-norm regularizer.

Theorem 3. *If $l(\cdot)$ is differentiable and convex[2] and has a unique root at 0, $\psi(\cdot)$ is strictly convex and $\eta > 0$ is such that $\psi - \eta L_i$ is convex $\forall i$, then SMD iterates converge to*

$$w_\infty = \operatorname*{argmin}_{w \in \mathcal{W}} D_\psi(w, w_0) \tag{11}$$

The proof for this result is mentioned in the appendix A.

4.2 Over-parameterized Non-Linear Models

This section extends the results of the previous section to highly over-parameterized non-linear models, which is typically the case for deep neural networks. It is to be noted that we do not provide any proof for the results of this section. An interested reader may refer (Azizan et al., 2019) for the proofs of theorems.

[2] A more generalized version can be proved for a quasi-convex case but it is out of scope of this work.

We consider the cases when $m \gg n$. Since the model is highly over-parameterized, we can say that model perfectly interpolates the data. The two conditions that helped us prove the results for the linear case i) $D_{L_i}(w, w_{i-1})$ being independent of w (necessary for proving implicit regularization) and ii) $D_{L_i}(w, w_{i-1}) \geqslant 0$ (necessary for proving convergence) do not hold true in the non-linear domain. However, these results hold in a *local* sense. So, the necessary conditions for proving the non-linear regime are:

- $D_{L_i}(w, w_{i-1})$ is *weakly dependent* on w for w_{i-1} "close" to w
- $D_{L_i}(w, w_{i-1}) \geqslant 0$ for w_{i-1} "close" to w

It is to be noted that the term "close" can be made more precise which is beyond the scope of the current work. Now, let us define:

$$w_* = \operatorname*{argmin}_{w \in \mathcal{W}} D_\psi(w, w_0) \tag{12}$$

Then we have the following result:

Theorem 4. *There exists $\epsilon > 0$, such that if $\|w_* - w_0\| < \epsilon$, then for sufficiently small step size $\eta > 0$, we have:*

- *SMD iterates converge to $w_\infty \in \mathcal{W}$*
- *$\|w_* - w_\infty\| = o(\epsilon)$*

The most complicated looking part in the theorem is the assumption of w_0 being close to w_*, but in a highly over-parameterized non-linear model, this is actually trivial. This can be observed from the fact that deep networks can perfectly interpolate the training data from almost any arbitrary initialization. It indicates that any arbitrary initial point is "close" enough to the global minima with very high probability. The second point of the proof which states $\|w_* - w_\infty\| = o(\epsilon)$, tells that the convergence point is "very close" to $\operatorname{argmin}_{w \in \mathcal{W}} D_\psi(w, w_0)$ and with an initialization as described in Corollary 2.1, SMD iterates converge to a solution that is implicitly regularized.

5 Experiments

We conducted extensive experiments to validate the theorems on convergence and implicit regularization of SMD. For the linear case, all the experiments have been designed and executed from scratch. For the non-linear domain, we try to reproduce the results obtained in (Azizan et al., 2019).

5.1 Linear Models

We generate a synthetic dataset containing n data points, $y_i = x_i^\top w$ where $x_i \in \mathbb{R}^m$. The standard mean squared error is chosen as the loss function. We consider the ℓ_q-norm potential functions for $q = 1, 2, 3$ and ∞. It is to be noted that if $q = 1$, the potential function is *not* strictly convex, so, $q = 1+\epsilon, \epsilon > 0$ is chosen. In all our experiments, $\epsilon = 0.1$. The potential function obtained for $q = 10$ serves as a surrogate for the $\ell_\infty-$norm potential function. When $q = 2$, SMD reduces to standard SGD. Although, we have stated all the results for a fixed learning rate, the results also hold true for a decaying learning rate. The proofs for the same is beyond the scope of this work. We use a decaying learning rate in certain cases to stabilize the optimization process.

Let us define parameterization ratio $\nu = \frac{m}{n}$. Clearly, if $\nu > 1$, the model is over-parameterized. For all the experiments, $m = 1000$. We initialize the $w_0 \sim \mathcal{N}(0, 10^{-4})$ according the Corollary 2.1 to attain implicit regularization. The algorithm is implemented in Python with JAX (Bradbury et al., 2018) support.

5.1.1 Effect of Potential Function on Convergence and Implicit Regularization

In this experiment, ν is fixed to be 10, i.e. $n = 100$. We finetune the learning rate and decay rate so that the optimization procedure converges within 10^4 iterations. The exact values of the hyperparameters for different values of n and potential functions are provided in appendix B.

The learning curves for the potential functions is plotted in Figure 1. We also plot a histogram of the absolute value of the 1000-dimensional w_∞ vector obtained for the four different potential functions in Figure 2.

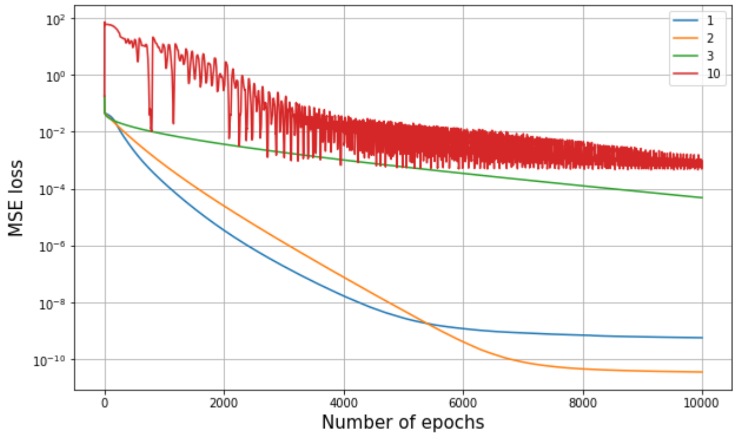


Figure 1: Learning curves for different potential functions for a given parameterization ratio

Table 1: Effect of potential functions on accuracy of the test dataset

Potential Function	Value of q	Test Accuracy
ℓ_1-norm	1.1	89.67%
ℓ_2-norm	2	91.79%
ℓ_3-norm	3	91.75%
$\ell_\infty-$norm	10	**91.96%**

5.1.2 Effect of Parameterization Ratio on Convergence and Implicit Regularization

In this experiment, we vary the parameterization ratio, ν, in a logarithmic scale from 1 ($m = n$) to 1000 (only one observation is given) while keeping the potential function and m fixed. The learning curves for different values of ν for $q = 3$ (ℓ_3-norm) is shown in Figure 3. The corresponding histograms of the absolute value of the 1000-dimensional w_∞ vector is presented in Figure 4. The learning curves and histograms for all the potential functions is provided in appendix B.

5.2 Non-Linear Models

We experimentally show that when the parameters of a deep neural network perfectly interpolate the data points, the final weight vector obtained from SMD updates is implicitly regularized[3]. To this end, a ResNet-18 model (He et al., 2016) having $\sim 11 \times 10^6$ parameters is taken. The training dataset CIFAR-10 (Krizhevsky, 2009) contains 50,000 images and test dataset contains 10,000 images. Hence, $\nu \approx 220$.

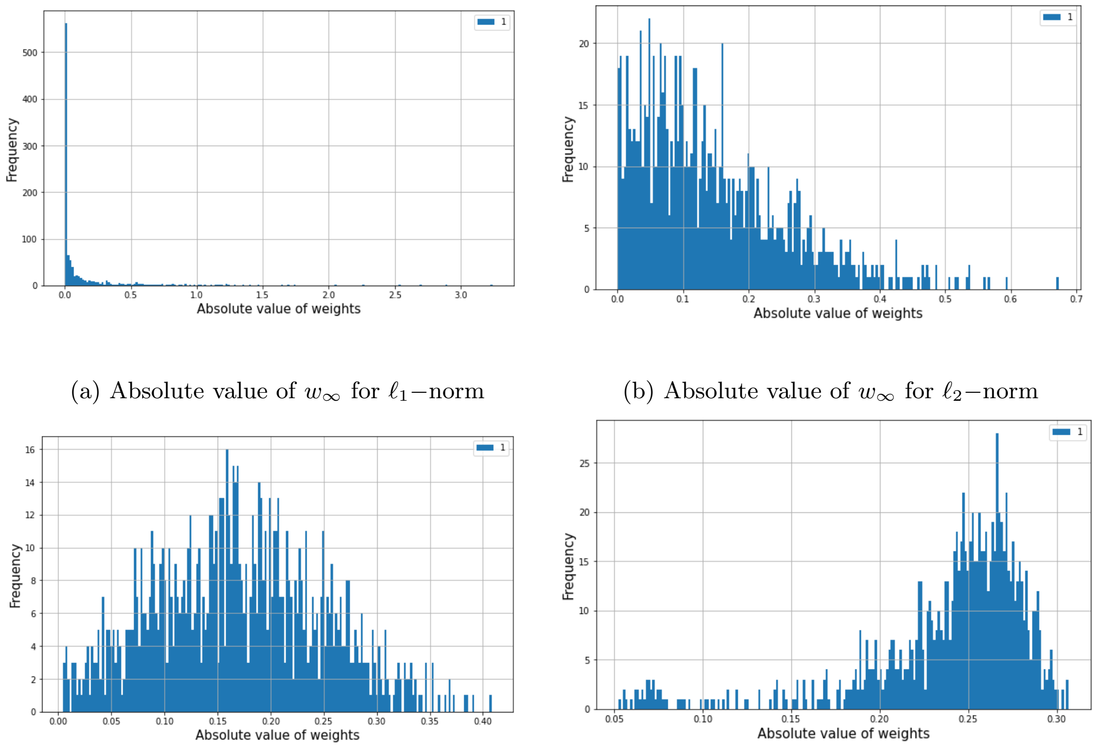
The ResNet-18 model is trained till it achieves $> 99.9\%$ training accuracy[4]. The final test accuracies are mentioned in Table 1 and the histograms of the (almost) interpolating weight vectors are plotted in Figure 5.

6 Discussion and Conclusion

In this section, we draw inferences from the experiments conducted in the previous section. For the chosen loss function (MSE Loss), the rate of convergences followed by the $q-$norm SMD in Figure 1 suggests that for ℓ_1-norm and ℓ_2-norm SMD (SGD) converge to almost zero error i.e. $w \in \mathcal{W}$. For $q = 10$, the surrogate for $\ell_\infty-$ norm, the learning curve is quite noisy even after a minimal learning rate and decay rate applied (Please refer to Table 2 and 3 for the exact values). As seen in Figure 1, it also converges to a sub-optimal solution.

[3] All the experiments in this subsection are meant to reproduce the necessary results of (Azizan et al., 2019). They are obtained using the codes from https://github.com/SahinLale/StochasticMirrorDescent.

[4] Due to resource constraints, perfect interpolation, i.e. 100% training accuracy, could not be attained. Hence, a model achieving > 99.9% train accuracy is assumed to be perfectly interpolating in nature.



(a) Absolute value of w_∞ for ℓ_1-norm

(b) Absolute value of w_∞ for ℓ_2-norm

(c) Absolute value of w_∞ for ℓ_3-norm

(d) Absolute value of w_∞ for $\ell_\infty-$norm

Figure 2: Histograms of absolute value of w_∞ for different potential function

As we have proved theoretically in the previous sections, the converged value is implicitly regularized. It is evident from the fact the absolute value of final weights in case of ℓ_1-norm SMD is concentrated near zero (sparsity is induced) whereas, for $\ell_\infty-$norm SMD, almost no value is close to zero. As the value of q increases, we observe that the histogram slowly shifts towards a higher absolute value of weights.

In the second experiment, when the parameterization ratio is varied, keeping the potential function fixed (ℓ_3-norm), we can observe that all the results we proved, hold for higher ν. As $\nu \to 1$, SMD iterates converge to relatively sub-optimal values. It is also observed that for lower values of ν, learning rates have to be reduced for convergence, and sometimes even a decay factor needs to be included. Similar results are also obtained for the rest of the potential functions.

An interesting phenomenon is observed when we compare the histograms of the absolute value of weights for different ν. As ν increases, most of the weights are concentrated near 0, but the weights start slowly "spreading out" as the parameterization ratio decreases. This trend is clearly observable for ℓ_2-norm and ℓ_3-norm in Figure 7. A plausible explanation for the same could be: in an ideal case, the number of parameters required to satisfy k linear equations is k. So, for higher ν, the number of parameters needed to satisfy the given set of equations is much less than m, the dimensionality of w. Hence, the rest of the entries could be zero or extremely close to it. On the other hand, when the number of equations approaches m, almost all the entries of w need to take non-zero values to satisfy the system of equations. It is to be noted that even when the models are mildly over-parameterized, the shape of the histograms is still retained, as evident in Figure 7.

For non-linear models, we plot the histograms of the absolute value of w_∞ for the different potential functions. Here, the implicit regularization is even more prominent than the linear models. Another impressive result, also pointed out in (Azizan et al., 2019), is that the test accuracy in case of $\ell_\infty-$norm outperforms all the potential functions as seen in Table 1. This result is quite contrary to the existing conventions of using ℓ_1-norm and ℓ_2-norm for regularization in machine learning. A detailed study on the choice of regularizers in deep neural networks and their effect on generalization could be a direction for future research.

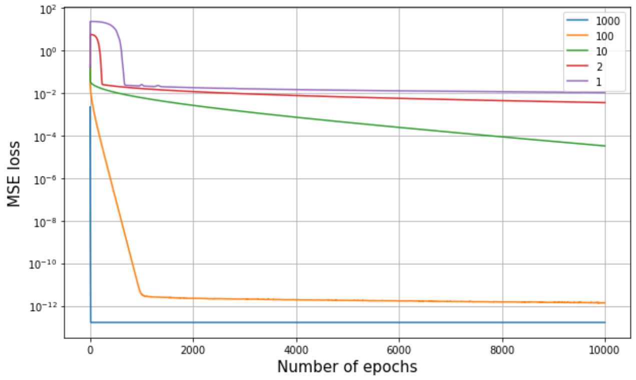


Figure 3: Learning curve of SMD for ℓ_3-norm potential function for different ν

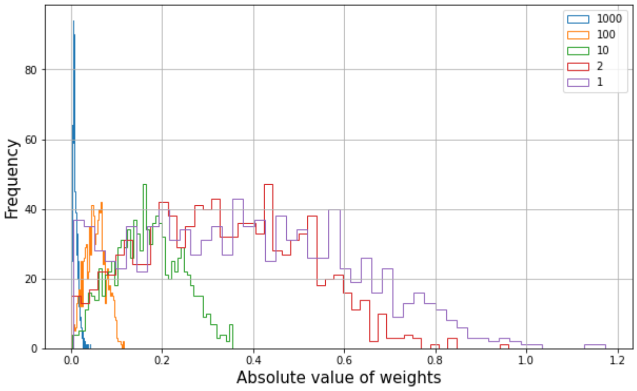


Figure 4: Histograms of absolute value of w_∞ for ℓ_3-norm potential function for different ν

References

Navid Azizan and Babak Hassibi. Stochastic gradient/mirror descent: Minimax optimality and implicit regularization. *arXiv preprint arXiv:1806.00952*, 2018.

Navid Azizan and Babak Hassibi. A characterization of stochastic mirror descent algorithms and their convergence properties. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5167–5171. IEEE, 2019.

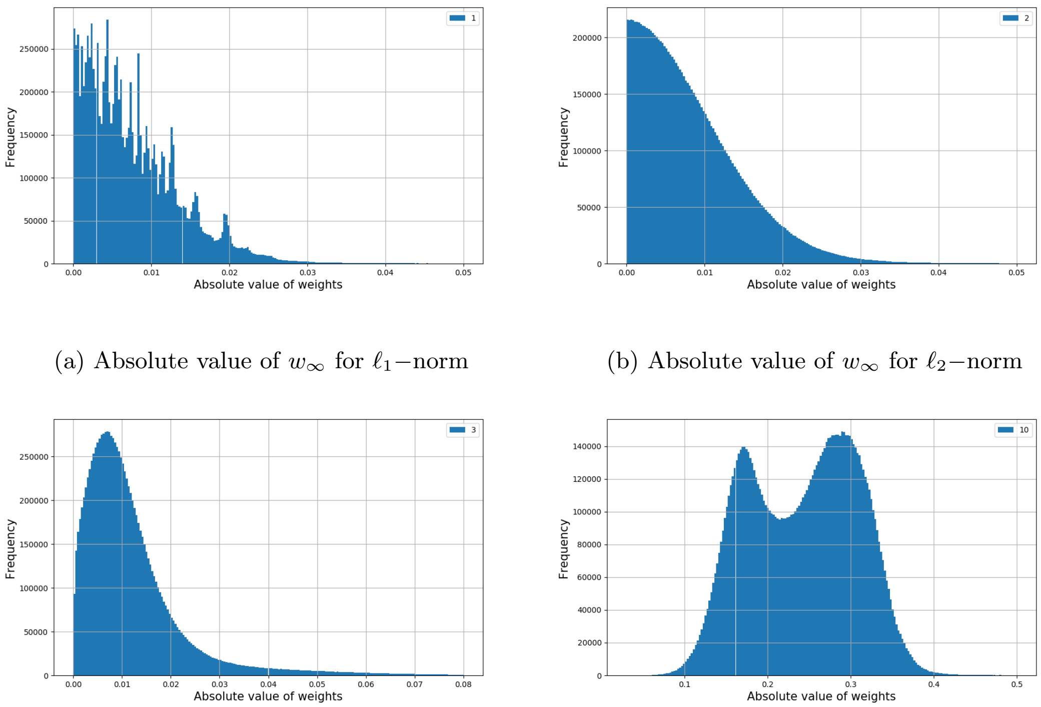
Navid Azizan, Sahin Lale, and Babak Hassibi. Stochastic mirror descent on overparameterized nonlinear models: Convergence, implicit regularization, and generalization. *arXiv preprint arXiv:1906.03830*, 2019.

Navid Azizan, Sahin Lale, and Babak Hassibi. A study of generalization of stochastic mirror descent algorithms on overparameterized nonlinear models. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3132–3136. IEEE, 2020.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, and Skye Wanderman-Milne. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/google/jax.



(a) Absolute value of w_∞ for ℓ_1-norm

(b) Absolute value of w_∞ for ℓ_2-norm

(c) Absolute value of w_∞ for ℓ_3-norm

(d) Absolute value of w_∞ for $\ell_\infty-$norm

Figure 5: Histograms of absolute value of w_∞ for different potential function for non-linear models

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

Nicolo Cesa-Bianchi, Pierre Gaillard, Gábor Lugosi, and Gilles Stoltz. Mirror descent meets fixed share (and feels no regret). In *Advances in Neural Information Processing Systems*, pages 980–988, 2012.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. IEEE, 2013.

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

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Alex Krizhevsky. *Learning Multiple Layers Of Features From Tiny Images.* 2009. URL http://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Arkadiĭ Semenovich Nemirovsky and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Ashia C Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro, and Benjamin Recht. The marginal value of adaptive gradient methods in machine learning. In *Advances in neural information processing systems*, pages 4148–4158, 2017.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

A Proofs

A.1 Proof for Lemma 1

Proof.

$$\begin{aligned} RHS \; = \; & \psi(w) - \cancel{\psi(w_i)} - \nabla\psi(w_i)^\top(w - w_i) + \cancel{\psi(w_i)} - \psi(w_{i-1}) - \nabla\psi(w_{i-1})^\top(w_i - w_{i-1}) \\ & - \cancel{\eta L_i(w_i)} + \cancel{\eta L_i(w_{i-1})} + \eta\nabla L_i(w_{i-1})^\top(w_i - \cancel{w_{i-1}}) + \eta L_i(w) \\ & - \cancel{\eta L_i(w_{i-1})} - \eta\nabla L_i(w_{i-1})^\top(w - \cancel{w_{i-1}}) + \cancel{\eta L_i(w_i)} \end{aligned} \tag{13}$$

$$\begin{aligned} = \; & \psi(w) - \nabla\psi(w_i)^\top(w - w_i) - \psi(w_{i-1}) - \nabla\psi(w_{i-1})^\top(w_i - w_{i-1}) \\ & + \eta L_i(w) - \eta\nabla L_i(w_{i-1})^\top(w - w_i) \end{aligned} \tag{14}$$

Using the SMD update rule $\nabla\psi(w_i) = \nabla\psi(w_{i-1}) - \eta\nabla L_i(w_{i-1})$,

$$\begin{aligned} RHS \; = \; & \psi(w) - \psi(w_{i-1}) - (\nabla\psi(w_{i-1}) - \eta\nabla L_i(w_{i-1}))^\top(w - w_i) - \nabla\psi(w_{i-1})^\top(w_i - w_{i-1}) \\ & + \eta L_i(w) - \eta\nabla L_i(w_{i-1})^\top(w - w_i) \end{aligned} \tag{15}$$

$$\begin{aligned} = \; & \psi(w) - \psi(w_{i-1}) - \nabla\psi(w_{i-1})^\top(w - \cancel{w_i}) + \cancel{\eta\nabla L_i(w_{i-1})^\top(w - w_i)} \\ & - \nabla\psi(w_{i-1})^\top(-\cancel{w_i} - w_{i-1}) + \eta L_i(w) - \cancel{\eta\nabla L_i(w_{i-1})^\top(w - w_i)} \end{aligned} \tag{16}$$

$$= \psi(w) - \psi(w_{i-1}) - \nabla\psi(w_{i-1})^\top(w - w_{i-1}) + \eta L_i(w) \tag{17}$$

$$= D_\psi(w, w_{i-1}) + \eta L_i(w) \tag{18}$$

$$= D_\psi(w, w_{i-1}) + \eta l(y_i - f(x_i, w)) = D_\psi(w, w_{i-1}) + \eta l(\cancel{f(x_i, w)} + v_i - \cancel{f(x_i, w)}) \tag{19}$$

$$= D_\psi(w, w_{i-1}) + \eta l(v_i) \tag{20}$$

□

A.2 Proof for Theorem 3

Proof. Consider the expression in (8):

$$D_\psi(w, w_0) = D_\psi(w, w_T) + \sum_{i=1}^{T}\left(E_i(w_i, w_{i-1}) + \eta D_{L_i}(w, w_{i-1})\right) \tag{21}$$

As $l(\cdot)$ is differentiable and convex, $L_i(\cdot)$ is also convex which implies $D_{L_i}(w, w_{i-1}) \geqslant 0$. Similarly by assumption, $\psi - \eta L_i$ is convex $\forall i$ which implies $E_i(w_i, w_{i-1}) = D_{\psi - nL_i}(w_i, w_{i-1}) + \eta L_i(w_i) \geqslant 0$. As both LHS and RHS are finite in (8), as $T \to \infty$, both the expressions in the summand go to zero. As $D_{L_i}(w, w_{i-1}) \to 0$, $L_i(w_{i-1}) \to 0$. As $L_i(w_{i-1})$ is convex and differentiable, the SMD updates vanish and by definition, $w_{i-1} \to w_\infty$. Furthermore, as $l(\cdot)$ has an unique root at 0, it implies all the data points are being fit with no error, i.e., $w_\infty \in \mathcal{W}$. □

B Details of the Experiments

B.1 Learning Rates and Decay Rates for Linear Models

For stabilizing the optimization procedure, we use an exponential decay rate as given in (22).

$$\eta_i = \eta_0 \cdot (decay)^i \tag{22}$$

where η_i is the learning rate for i^{th} iteration and η_0 is the initial learning rate. The learning rates and decay rates are provided for different configurations of ν and potential functions are provided in Table 2 and Table 3.

B.2 Effect of Parameterization Ratio on Convergence and Implicit Regularization

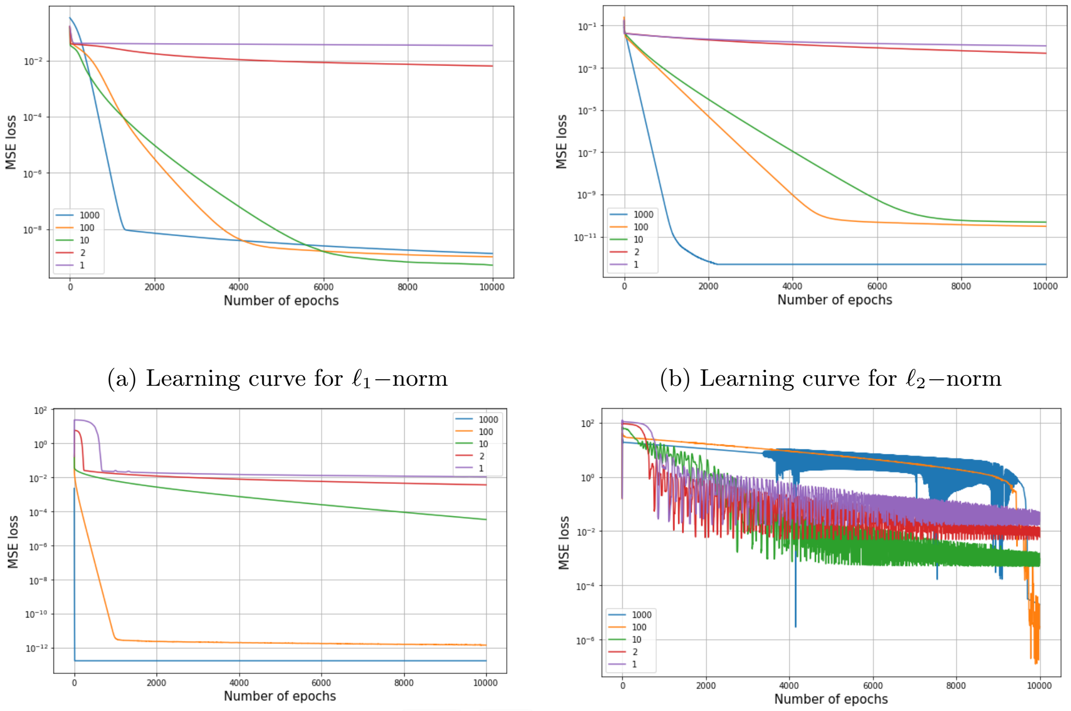
The plots of learning curves and absolute value of w_∞ for different ν for all potential functions in linear models are given in Figure 6 and 7 respectively.

Table 2: Learning Rates for different configurations of ν and potential functions

$q-$norm potential function	Parameterization Ratio				
	1000	500	100	10	1
1	0.01	0.01	0.01	0.001	0.0001
2	0.01	0.01	0.01	0.001	0.001
3	0.01	0.01	0.001	0.001	0.001
10	0.0001	0.0001	0.0001	0.0001	0.0001

Table 3: Decay Rates for different configurations of ν and potential functions

$q-$norm potential function	Parameterization Ratio				
	1000	500	100	10	1
1	1	1	1	0.9999	0.9999
2	1	1	1	1	1
3	1	1	1	1	1
10	0.999	0.999	0.999	0.999	0.999



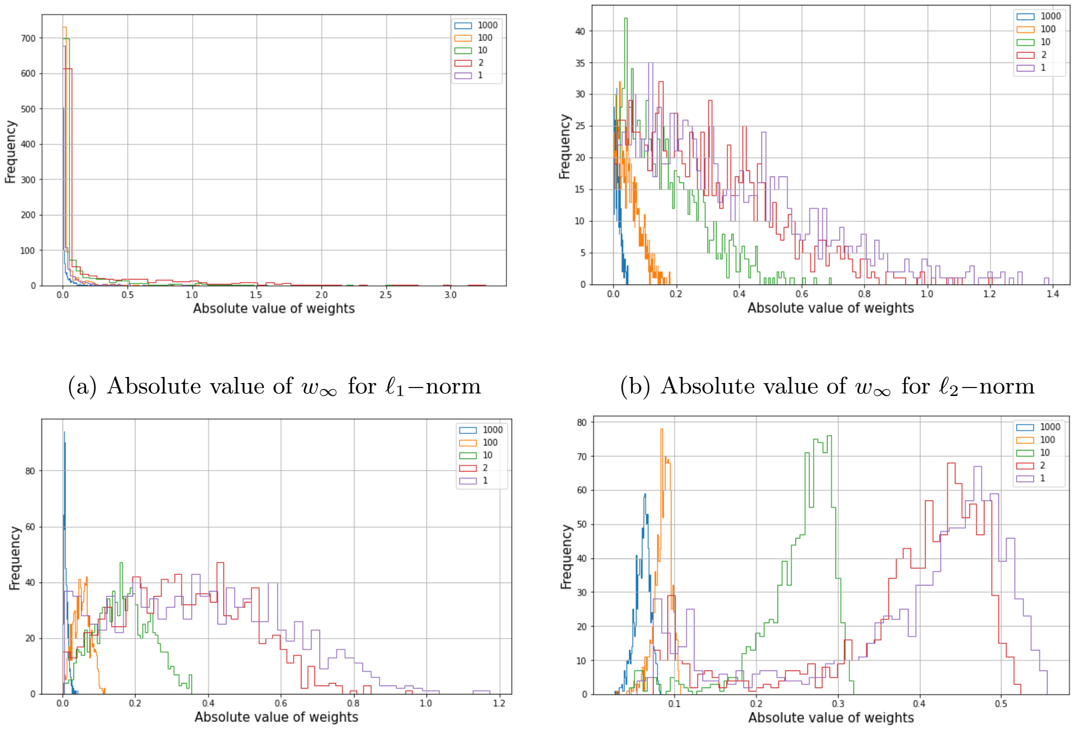
(a) Learning curve for ℓ_1-norm

(b) Learning curve for ℓ_2-norm

(c) Learning curve for ℓ_3-norm

(d) Learning curve for $\ell_\infty-$norm

Figure 6: Learning curves for different potential functions for different ν



(a) Absolute value of w_∞ for ℓ_1-norm

(b) Absolute value of w_∞ for ℓ_2-norm

(c) Absolute value of w_∞ for ℓ_3-norm

(d) Absolute value of w_∞ for $\ell_\infty-$norm

Figure 7: Histograms of absolute value of w_∞ for different potential function for different ν