# A CHARACTERIZATION OF STOCHASTIC MIRROR DESCENT ALGORITHMS AND THEIR CONVERGENCE PROPERTIES

*Navid Azizan, Babak Hassibi*

California Institute of Technology, Pasadena, CA 91125

## ABSTRACT

Stochastic mirror descent (SMD) algorithms have recently garnered a great deal of attention in optimization, signal processing, and machine learning. They are similar to stochastic gradient descent (SGD), in that they perform updates along the negative gradient of an instantaneous (or stochastically chosen) loss function. However, rather than update the parameter (or weight) vector directly, they update it in a "mirrored" domain whose transformation is given by the gradient of a strictly convex differentiable potential function. SMD was originally conceived to take advantage of the underlying geometry of the problem as a way to improve the convergence rate over SGD. In this paper, we study SMD, for linear models and convex loss functions, through the lens of $H^\infty$ estimation theory and come up with a minimax interpretation of the SMD algorithm which is the counterpart of the $H^\infty$-optimality of the SGD algorithm for linear models and quadratic loss. In doing so, we identify a fundamental conservation law that SMD satisfies and use it to study the convergence properties of the algorithm. For constant step size SMD, when the linear model is over-parameterized, we give a deterministic proof of convergence for SMD and show that from any initial point, it converges to the closest point in the space of all parameter vectors that interpolate the data, where closest is in the sense of the Bregman divergence of the potential function. This property is referred to as *implicit regularization*: with an appropriate choice of the potential function one can guarantee convergence to the minimizer of any desired convex regularizer. For vanishing step size SMD, and in the standard stochastic optimization setting, we give a direct and elementary proof of convergence for SMD to the "true" parameter vector which avoids ergodic averaging or appealing to stochastic differential equations.

***Index Terms***— Stochastic gradient descent, mirror descent, minimax optimality, convergence, implicit regularization

## 1. PRELIMINARIES

Denote the training dataset by $\{(x_i, y_i) : i = 1, \dots, n\}$, where $x_i \in \mathbb{R}^m$ are the inputs, and $y_i \in \mathbb{R}$ are the labels. We assume that the data is generated through a linear model with a parameter vector $w \in \mathbb{R}^m$, plus some noise $v_i$, i.e., $y_i = x_i^T w + v_i$ for $i = 1, \dots, n$. The noise can be due to actual measurement error, or it can be due to modeling error (if the model is not rich enough to fully represent the data), or it can be a combination of both. As a result, we do not make any assumptions on the noise (such as stationarity, whiteness, Gaussianity, etc.) for now.

We are often interested in the over-parameterized (so-called interpolating) regime, i.e., when $m > n$. In this case, there are many parameter vectors $w$ (in fact, uncountably infinitely many) that are consistent with the observations. We denote the set of these parameter vectors by $\mathcal{W} = \left\{w \in \mathbb{R}^m \mid y_i = x_i^T w,\ i = 1, \dots, n\right\}$ (Note the absence of the noise term, since in this regime we can fully interpolate the data).

The total loss (empirical risk) on the training set can be denoted by $L(w) = \sum_{i=1}^n L_i(w)$, where $L_i(\cdot)$ is the loss on the individual data point $i$. We assume that the loss $L_i(\cdot)$ depends only on the residual, i.e., the difference between the prediction and the true label. In other words, $L_i(w) = l(y_i - x_i^T w)$ , where $l(\cdot)$ can be any nonnegative differentiable function with $l(0) = 0$. Typical examples of $l(\cdot)$ include square ($l_2$) loss, Huber loss, etc. We remark that, in the interpolating regime, every parameter vector in the set $\mathcal{W}$ renders each individual loss zero, i.e., $L_i(w) = 0$, for all $w \in \mathcal{W}$.

We will often consider two uncertainties, or error terms, $e_i$ and $e_{p,i}$, defined as follows.

$$e_i := y_i - x_i^T w_{i-1}, \text{ and } e_{p,i} := x_i^T w - x_i^T w_{i-1}.$$

$e_i$ is often referred to as the *innvovations* and is the error in predicting $y_i$, given the input $x_i$. $e_{p,i}$ is sometimes called the *prediction error*, since it is the error in predicting the noiseless output $x_i^T w$, i.e., in predicting what the best output of the model is. In the absence of noise, $e_i$ and $e_{p,i}$ coincide.

## 2. FUNDAMENTAL IDENTITY OF STOCHASTIC MIRROR DESCENT

Stochastic Mirror Descent (SMD) [1, 2, 3, 4] is one of the most widely used family of algorithms for stochastic optimization, which includes SGD as a special case.

For any strictly convex and differentiable potential $\psi(\cdot)$, the corresponding SMD updates are defined as

$$w_i = \arg\min_w \ \eta_i w^T \nabla L_i(w_{i-1}) + D_\psi(w, w_{i-1}), \qquad (1)$$

for $i \geq 1$ (we cycle through the data, or select them at random, for $i > n$), where

$$D_\psi(w, w_{i-1}) = \psi(w) - \psi(w_{i-1}) - \nabla\psi(w_{i-1})^T(w - w_{i-1}) \qquad (2)$$

Email: azizan@caltech.edu

is the Bregman divergence with respect to the potential function $\psi(\cdot)$, and $\eta_i \geq 0$ is the step size (learning rate). Since the potential function is strictly convex, the updates can be equivalently written as

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

which are uniquely defined because of the invertibility of $\nabla\psi$ (implied by the strict convexity of $\psi(\cdot)$). In other words, stochastic mirror descent can be thought of as transforming the variable $w$, with a *mirror map* $\nabla\psi(\cdot)$, into $\nabla\psi(w)$, and performing SGD on the new variable. For this reason, $\nabla\psi(w)$ is often referred to as the *dual* variable, while $w$ is the *primal* variable.

Different choices of the potential function $\psi(\cdot)$ yield different optimization algorithms, which, as we will see, result in different implicit regularizations. To name a few examples: 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:

$$w_i = w_{i-1} - \eta_i \nabla L_i(w_{i-1}), \tag{4}$$

For $\psi(w) = \sum_j w_j \log w_j$, the Bregman divergence becomes the unnormalized relative entropy (Kullback-Leibler divergence) $D_\psi(w, w') = \sum_j w_j \log \frac{w_j}{w'_j} - \sum_j w_j + \sum_j w'_j$, which corresponds to the exponentiated gradient descent (aka the exponential weights) algorithm. Other examples include $\psi(w) = \frac{1}{2}\|w\|_Q^2 = \frac{1}{2}w^T Q w$ for a positive definite matrix $Q$, which yields $D_\psi(w, w') = \frac{1}{2}(w - w')^T Q (w - w')$, and the $q$-norm squared $\psi(w) = \frac{1}{2}\|w\|_q^2$ with $\frac{1}{p} + \frac{1}{q} = 1$, which yields the $p$-norm algorithms [5, 6].

Let us define the Bregman divergence with respect to the loss function $L_i$ (note that $L_i(w) = l(y_i - x_i^T w)$ is convex when $l(\cdot)$ is convex)

$$D_{L_i}(w, w') := L_i(w) - L_i(w') - \nabla L_i(w')^T (w - w'), \tag{5}$$

The following result is an identity that characterizes SMD updates [7].

**Lemma 1.** *For any differentiable loss $l(\cdot)$, any parameter $w$ and noise values $\{v_i\}$ that satisfy $y_i = x_i^T w + v_i$ for $i = 1, \ldots, n$, and any sequence of step sizes $\{\eta_i\}$, the following relation holds for the stochastic mirror descent updates $\{w_i\}$ given in Eq.* (3)

$$\begin{aligned} D_\psi(w, w_{i-1}) + \eta_i l(v_i) = D_\psi(w, w_i) + &\\ E_i(w_i, w_{i-1}) + \eta_i D_{L_i}(w, w_{i-1}),& \end{aligned} \tag{6}$$

*for all $i \geq 1$, where*

$$\begin{aligned} E_i(w_i, w_{i-1}) := D_\psi(w_i, w_{i-1}) - \eta_i D_{L_i}(w_i, w_{i-1}) + &\\ \eta_i L_i(w_i).& \end{aligned} \tag{7}$$

Note that $E_i(w_i, w_{i-1})$ is not a function of $w$. Furthermore, even though it does not have to be nonnegative in general, for $\eta_i$ sufficiently small, it becomes nonnegative, because the Bregman divergence $D_\psi(.,.)$ is nonnegative. The interpretation of this result, as illustrated in Fig. 1, is that each step of SMD can be thought of as a lossless transformation of the input uncertainties to the output uncertainties, with specific coefficients that depend on the step size.

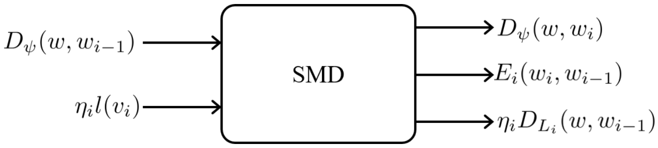

**Fig. 1**. Each step of SMD can be viewed as a transformation of the uncertainties.

Summing Equation (6) over all $i = 1, \ldots, T$ leads to the following identity.

**Lemma 2.** *For any differentiable loss $l(\cdot)$, any parameter $w$ and noise values $\{v_i\}$ that satisfy $y_i = x_i^T, w + v_i$ for $i = 1, \ldots, n$, any initialization $w_0$, any sequence of step sizes $\{\eta_i\}$, and any number of steps $T \geq 1$, the following relation holds for the stochastic mirror descent updates $\{w_i\}$ given in Eq.* (3)

$$\boxed{\begin{aligned} D_\psi(w, w_0) + \sum_{i=1}^T \eta_i l(v_i) = D_\psi(w, w_T) + &\\ \sum_{i=1}^T \left( E_i(w_i, w_{i-1}) + \eta_i D_{L_i}(w, w_{i-1}) \right).& \end{aligned}} \tag{8}$$

This is a fundamental property of SMD, which as will be shown in the subsequent section, can be used to prove many important results, in a very direct way.

## 3. MINIMAX OPTMIALITY OF STOCHASTIC MIRROR DESCENT

In particular, using Lemma 2, one can show that SMD with sufficiently small step size is the optimal solution to a minimax problem [7].

**Theorem 3.** *Consider any differentiable loss $l(\cdot)$ with property $l(0) = l'(0) = 0$, and any initialization $w_0$. For sufficiently small sequence of step sizes $\{\eta_i\}$, i.e., one for which $\psi(w) - \eta_i L_i(w)$ is convex for all i, and for any number of steps $T \geq 1$, the stochastic mirror descent iterates $\{w_i\}$ given by Eq.* (3), *w.r.t. any strictly convex potential $\psi(\cdot)$, are the optimal solution to the following minimization problem*

$$\min_{\{w_i\}} \max_{w, \{v_i\}} \frac{D_\psi(w, w_T) + \sum_{i=1}^T \eta_i D_{L_i}(w, w_{i-1})}{D_\psi(w, w_0) + \sum_{i=1}^T \eta_i l(v_i)}. \tag{9}$$

*Furthermore, the optimal value (achieved by SMD) is* 1.

For SGD and square loss, this result reduces to the following.

**Corollary 4.** *For any initialization $w_0$, any sequence of step size $0 < \eta_i \leq \frac{1}{\|x_i\|^2}$, and any number of steps $T \geq 1$, the* *stochastic gradient descent iterates* $\{w_i\}$ *given in Eq.* (4) *are the optimal solution to the following minimization problem*

$$\min_{\{w_i\}} \max_{w,\{v_i\}} \frac{\|w-w_T\|^2+\sum_{i=1}^{T}\eta_i e_{p,i}^2}{\|w-w_0\|^2+\sum_{i=1}^{T}\eta_i v_i^2}. \tag{10}$$

*Furthermore, the optimal value (achieved by SGD) is* 1*.*

This result in fact states that SGD is choosing the best estimate that safeguards against the worst-case disturbances, which is a conservative choice. However, this choice may actually be the rational thing to do in situations when we do not have much knowledge about the disturbances.

The above result holds for any horizon $T \geq 1$. A variation of this result, i.e., when $T \to \infty$ and without the $\|w-w_T\|^2$ term in the numerator, was first shown in [8, 9]. In that case, the ratio $\frac{\sum_{i=1}^{\infty}\eta_i e_{p,i}^2}{\|w-w_0\|^2+\sum_{i=1}^{\infty}\eta_i v_i^2}$ in the minimax problem is in fact the $H^\infty$ *norm* of the transfer operator that maps the unknown disturbances $(w-w_0, \{\sqrt{\eta_i}v_i\})$ to the prediction errors $\{\sqrt{\eta_i}e_{p,i}\}$ [10, 11].

Theorem 3 also generalizes the result of [12], which is the special case for the $p$-norm algorithms, again, with square loss. Another interesting connection to the literature is that it was shown in [13] that SGD is *locally* minimax optimal, with respect to the $H^\infty$ norm. Strictly speaking, this result is not a generalization of that result; however, Theorem 3 can be interpreted as SGD/SMD being *globally* minimax optimal, but with respect to a different metric in the numerator and denominator.

## 4. DETERMINISTIC CONVERGENCE AND IMPLICIT REGULARIZATION IN OVER-PARAMETERIZED MODELS

In this section, we show some other implications of the fundamental identity of Section 2. In particular, we show convergence and implicit regularization, in the over-parameterized (so-called interpolating) regime, for general SMD algorithms.

The over-parameterized (interpolating) linear regression regime is a simple but instructive setting, recently considered in some papers [14, 15]. In this setting, since the model is over-parameterized, it is assumed that it can perfectly match (interpolate) the training data, and therefore the $v_i$ are zero. The set of parameter vectors that interpolate the data is given by $\mathcal{W} = \left\{w \mid y_i = x_i^T w,\ i=1,\ldots,n\right\}$, and further $L_i(w) = l(y_i - x_i^T w)$, with any differentiable loss $l(\cdot)$. Therefore, Eq. (8) reduces to

$$\begin{aligned} D_\psi(w,w_0) = D_\psi(w,w_T) + \\ \sum_{i=1}^{T}\left(E_i(w_i,w_{i-1}) + \eta_i D_{L_i}(w,w_{i-1})\right), \end{aligned} \tag{11}$$

for all $w \in \mathcal{W}$, where

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

which is notably *independent of* $w$. As a result, we can easily minimize both sides of Eq. (11) with respect to $w \in \mathcal{W}$, which leads to the following result.

**Proposition 5.** *For any differentiable loss* $l(\cdot)$*, any initialization* $w_0$*, and any sequence of step sizes* $\{\eta_i\}$*, consider the stochastic mirror descent iterates given in Eq.* (3) *with respect to any strictly convex potential* $\psi(\cdot)$*. If the iterates converge to a solution* $w_\infty \in \mathcal{W}$*, then*

$$w_\infty = \arg\min_{w\in\mathcal{W}} D_\psi(w,w_0). \tag{12}$$

**Remark.** *In particular, for the initialization* $w_0 = \arg\min_{w\in\mathbb{R}^m}\psi(w)$*, if the iterates converge to a solution* $w_\infty \in \mathcal{W}$*, then*

$$w_\infty = \arg\min_{w\in\mathcal{W}} \psi(w). \tag{13}$$

An equivalent form of Proposition 5 has been shown recently in, e.g., [14][1]. Note that the result of [14] does not say anything about *whether the algorithm converges or not*. However, our fundamental identity of SMD (Lemma 2) allows us to also establish convergence to the regularized point in a deterministic sense, for some common cases, which will be shown next.

What Proposition 5 says is that depending on the choice of the potential function $\psi(\cdot)$, the optimization algorithm can perform an implicit regularization without any explicit regularization term. In other words, for any desired regularizer, if one chooses a potential function that approximates the regularizer, we can run the optimization without explicit regularization, and if it converges to a solution, the solution must be the one with the minimum potential.

Next we establish *convergence to the regularized point* for the convex case.

**Proposition 6.** *Consider the following two cases.*

*(i)* $l(\cdot)$ *is differentiable and convex and has a unique root at 0,* $\psi(\cdot)$ *is strictly convex, and positive sequence* $\{\eta_i\}$ *is such that* $\psi - \eta_i L_i$ *is convex for all i, or*

*(ii)* $l(\cdot)$ *is differentiable and quasi-convex and has zero derivative only at 0,* $\psi(\cdot)$ *is* $\alpha$*-strongly convex, and* $0 < \eta_i \leq \frac{\alpha|y_i - x_i^T w_{i-1}|}{\|x_i\|^2 |l'(y_i - x_i^T w_{i-1})|}, \forall i$.

*If either (i) or (ii) holds, then for any initialization* $w_0$*, the stochastic mirror descent iterates given in Eq.* (3) *converge to*

$$w_\infty = \arg\min_{w\in\mathcal{W}} D_\psi(w,w_0). \tag{14}$$

### 4.1. Example: Compressed Sensing via Stochastic Mirror Descent

The implicit regularization results discussed earlier suggest that with a proper choice of potential function $\psi(\cdot)$, one may

[1] To be precise, the authors in [14] assume convergence to a global minimizer of the loss function $L(w) = \sum_{i=1}^{n} l(y_i - x_i^T w)$, which with their assumption of the loss function $l(\cdot)$ having a unique finite root is equivalent to assuming convergence to a point $w_\infty \in \mathcal{W}$.

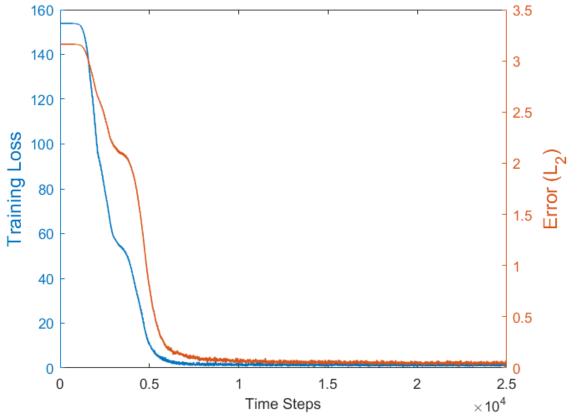

**Fig. 2**. The training loss and actual error of stochastic mirror descent for compressed sensing (Section 4.1). SMD recovers the actual sparse signal.

run stochastic mirror descent for any desired regularization, without any explicit regularization. As an example, we consider the popular problem of compressed sensing and see if it can be solved via SMD, something that, to the best of our knowledge, has not been studied in previous work.

In compressed sensing, one seeks the sparsest solution to an under-determined (over-parameterized) system of linear equations. The surrogate problem is

$$\begin{aligned} \min_{w} \quad & \|w\|_1 \\ \text{s.t.} \quad & x_i^T w = y_i, \; i = 1, \dots, n. \end{aligned} \tag{15}$$

A natural question to ask is can we do compressed sensing using SMD, with $\psi(w) = \|w\|_1$? The answer, unfortunately, is no, because $\|w\|_1$ is neither differentiable nor strictly convex. However, one can choose the potential function to be $\psi(w) = \|w\|_{1+\epsilon}^{1+\epsilon}$ for any $\epsilon > 0$, along with a suitable choice of loss function, e.g. $l(z) = |z|^{1+\epsilon}$.

To evaluate the performance of this approach, we consider a k-sparse signal with k=10 nonzero parameters in m=100 dimensions (10% sparsity), and n=50 Gaussian measurements (data points). As demonstrated in Figure 2, both the training loss and the actual error (the difference from the true signal) decrease as SMD progresses, and they converge to zero. SMD indeed recovers to the true underlying sparse signal.

## 5. STOCHASTIC CONVERGENCE IN UNDER-PARAMETERIZED MODELS

In Section 4, we showed several implications of the fundamental identity of SMD in the over-parameterized (so-called interpolating) regime. In this section, we consider the under-parameterized (online streaming) linear regression setting, and show that the fundamental identity (8) yields a simple proof for mean-square convergence of SMD to the ground truth, for decaying step size sequence that satisfies the Robbins-Monro summability condition ($\sum_{i=1}^{\infty} \eta_i = \infty, \sum_{i=1}^{\infty} \eta_i^2 < \infty$).

In this setting, there is an online stream of data $y_i = x_i^T w + v_i$ for $i = 1, 2, \dots$, where $v_i$ are iid with $\mathbb{E}[v_i] = 0$ and $\mathbb{E}\left[v_i^2\right] = \sigma^2$, and the inputs are "persistently exciting," i.e., for any $\delta > 0$, there exists $T > 0$ s.t. $\sum_{i=1}^{T} x_i x_i^T \succeq \delta I$.

**Proposition 7.** *Consider $y_i = x_i^T w + v_i, i \geq 1$, where $\mathbb{E}[v_i] = 0$, $\mathbb{E}[v_i v_j] = \sigma^2 \delta_{ij}$, and the $x_i$ are persistently exciting. For any step size sequence $\{\eta_i\}$ such that $\sum_{i=1}^{\infty} \eta_i = \infty, \sum_{i=1}^{\infty} \eta_i^2 < \infty$, the stochastic mirror descent iterates given in Eq.* (3) *with respect to any strongly convex potential $\psi(\cdot)$, for a square loss, converge to $w$ in a mean-square sense.*

To see that, note that for the square loss and a linear model, the identity (8), after some simple algebra, reduces to the following form.

$$\begin{aligned} D_\psi(w, w_0) = D_\psi(w, w_T) + \sum_{i=1}^{T} \Big( & D_\psi(w_i, w_{i-1}) + \\ & \eta_i e_{p,i} v_i - \eta_i (e_{p,i} + v_i) x_i^T (w_i - w_{i-1}) + \eta_i e_{p,i}^2 \Big), \end{aligned} \tag{16}$$

where we have used the fact that $e_i = e_{p,i} + v_i$.

On the other hand, the update rule $\nabla\psi(w_i) = \nabla\psi(w_{i-1}) + \eta_i (e_{p,i} + v_i) x_i$ can be expressed, using a Taylor expansion, as

$$\begin{aligned} w_i &= \nabla\psi^{-1}\left(\nabla\psi(w_{i-1}) + \eta_i (e_{p,i} + v_i) x_i\right) \\ &= w_{i-1} + \eta_i M_i (e_{p,i} + v_i) x_i + O(\eta_i^2), \end{aligned}$$

where $M_i := \nabla^2\psi(w_{i-1})^{-1}$. This implies that $D_\psi(w_i, w_{i-1}) = \frac{1}{2}(w_i - w_{i-1})^T \nabla^2\psi(w_{i-1})(w_i - w_{i-1}) + O(\eta_i^3) = \frac{1}{2}\eta_i^2 (e_{p,i} + v_i)^2 x_i^T M_i x_i + O(\eta_i^3)$. Plugging this into (16) leads to

$$\begin{aligned} D_\psi(w, w_0) = D_\psi(w, w_T) + \sum_{i=1}^{T} \Big( & \eta_i e_{p,i} v_i \\ & - \frac{1}{2}\eta_i^2 (e_{p,i} + v_i)^2 x_i^T M_i x_i + \eta_i e_{p,i}^2 + O(\eta_i^3) \Big). \end{aligned} \tag{17}$$

Taking expected values from both sides, noting that $e_{p,i}$ and $w_{i-1}$ are independent of $v_i$, we get

$$\begin{aligned} \mathbb{E}\left[D_\psi(w, w_0)\right] = \mathbb{E}\left[D_\psi(w, w_T)\right] + \sum_{i=1}^{T} \Big( & -\frac{\sigma^2}{2}\eta_i^2 \mathbb{E}\left[x_i^T M_i x_i\right] \\ & - \frac{1}{2}\eta_i^2 \mathbb{E}\left[e_{p,i}^2 x_i^T M_i x_i\right] + \eta_i \mathbb{E}\left[e_{p,i}^2\right] + O(\eta_i^3) \Big). \end{aligned} \tag{18}$$

From strong convexity of $\psi(\cdot)$, we have $\nabla^2\psi(w_{i-1}) \succeq \alpha I$, and therefore $\mathbb{E}\left[x_i^T M_i x_i\right] \leq \frac{1}{\alpha}\|x_i\|^2$ and $\mathbb{E}\left[e_{p,i}^2 x_i^T M_i x_i\right] \leq \frac{1}{\alpha}\|x_i\|^2 \mathbb{E}\left[e_{p,i}^2\right]$. As a result, we have that $\sum_{i=1}^{T} \eta_i (1 - \frac{\|x_i\|^2}{2\alpha}\eta_i) \mathbb{E}\left[e_{p,i}^2\right] < \infty$ because $\sum_{i=1}^{T} \eta_i^2 < \infty$ and $\sum_{i=1}^{T} O(\eta_i^3) < \infty$, which implies that $\mathbb{E}\left[e_{p,i}^2\right]$ goes to zero. If the inputs are persistently exciting, this implies that $\mathbb{E}\left[\|w - w_{i-1}\|^2\right] \to 0$, which means SMD converges to the true parameter, in mean-square sense.

## 6. REFERENCES

[1] Arkadii Nemirovskii, David Borisovich Yudin, and Edgar Ronald Dawson, "Problem complexity and method efficiency in optimization," 1983.

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

[3] 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*, 2012, pp. 980–988.

[4] Zhengyuan Zhou, Panayotis Mertikopoulos, Nicholas Bambos, Stephen Boyd, and Peter W Glynn, "Stochastic mirror descent in variationally coherent optimization problems," in *Advances in Neural Information Processing Systems*, 2017, pp. 7043–7052.

[5] Adam J Grove, Nick Littlestone, and Dale Schuurmans, "General convergence results for linear discriminant updates," *Machine Learning*, vol. 43, no. 3, pp. 173–210, 2001.

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

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

[8] Babak Hassibi, Ali H. Sayed, and Thomas Kailath, "Hoo optimality criteria for LMS and backpropagation," in *Advances in Neural Information Processing Systems 6*, pp. 351–358. 1994.

[9] Babak Hassibi, Ali H Sayed, and Thomas Kailath, "Hoo optimality of the LMS algorithm," *IEEE Transactions on Signal Processing*, vol. 44, no. 2, pp. 267–280, 1996.

[10] Babak Hassibi, Ali H Sayed, and Thomas Kailath, *Indefinite-Quadratic Estimation and Control: A Unified Approach to H2 and H-infinity Theories*, vol. 16, SIAM, 1999.

[11] Dan Simon, *Optimal state estimation: Kalman, H infinity, and nonlinear approaches*, John Wiley & Sons, 2006.

[12] Jyrki Kivinen, Manfred K Warmuth, and Babak Hassibi, "The p-norm generalization of the LMS algorithm for adaptive filtering," *IEEE Transactions on Signal Processing*, vol. 54, no. 5, pp. 1782–1793, 2006.

[13] Babak Hassibi and Thomas Kailath, "Hoo optimal training algorithms and their relation to backpropagation," in *Advances in Neural Information Processing Systems 7*, pp. 191–198. 1995.

[14] Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro, "Characterizing implicit bias in terms of optimization geometry," *arXiv preprint arXiv:1802.08246*, 2018.

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