LINEAR CONVERGENCE AND IMPLICIT REGULARIZATION OF GENERALIZED MIRROR DESCENT WITH TIME-DEPENDENT MIRRORS

Anonymous authors
Paper under double-blind review

ABSTRACT

The following questions are fundamental to understanding the properties of over-parameterization in modern machine learning: (1) Under what conditions and at what rate does training converge to a global minimum? (2) What form of implicit regularization occurs through training? While significant progress has been made in answering both of these questions for gradient descent, they have yet to be answered more completely for general optimization methods. In this work, we establish sufficient conditions for linear convergence and obtain approximate implicit regularization results for *generalized mirror descent* (GMD), a generalization of mirror descent with a possibly time-dependent mirror. GMD subsumes popular first order optimization methods including gradient descent, mirror descent, and preconditioned gradient descent methods such as Adagrad. By using the Polyak-Lojasiewicz inequality, we first present a simple analysis under which non-stochastic GMD converges linearly to a global minimum. We then present a novel, Taylor-series based analysis to establish sufficient conditions for linear convergence of stochastic GMD. As a corollary, our result establishes sufficient conditions and provides learning rates for linear convergence of stochastic mirror descent and Adagrad. Lastly, we obtain approximate implicit regularization results for GMD by proving that GMD converges to an interpolating solution that is approximately the closest interpolating solution to the initialization in ℓ_2-norm in the dual space, thereby generalizing the result of Azizan, Lale, and Hassibi (2019) in the full batch setting.

1 INTRODUCTION

Recent work has established the optimization and generalization benefits of over-parameterization in machine learning (Belkin et al., 2019; Liu et al., 2020; Zhang et al., 2017). In particular, several works including Vaswani et al. (2019); Du et al. (2018); Liu et al. (2020); Li & Liang (2018) have demonstrated that over-parameterized models converge to a global minimum when trained using stochastic gradient descent and that such convergence can occur at a linear rate. Independently, other work, such as Gunasekar et al. (2018), have characterized *implicit regularization* of over-parameterized models, i.e., the properties of the solution selected by a given optimization method, without proving convergence.

Recently, Azizan & Hassibi (2019); Azizan et al. (2019) simultaneously proved convergence and analyzed approximate implicit regularization for mirror descent (Beck & Teboulle, 2003; Nemirovsky & Yudin, 1983). In particular, by using the fundamental identity of stochastic mirror descent (SMD), they proved that SMD converges to an interpolating solution that is approximately the closest one to the initialization in Bregman divergence. However, these works do not provide a rate of convergence for SMD and assume that there exists an interpolating solution within ϵ in Bregman divergence from the initialization. In this work, we generalize results from Azizan & Hassibi (2019); Azizan et al. (2019) by providing sufficient conditions for *linear* convergence and obtain approximate implicit regularization results for *generalized mirror descent* (GMD), an extension of mirror descent that introduces (1) a potential-free update rule and (2) a time-dependent mirror; namely, GMD with invertible $\phi : \mathbb{R}^d \to \mathbb{R}^d$ and learning rate η is used to minimize a real valued loss function, f, according to the update rule:

$$\phi^{(t)}(w^{(t+1)}) = \phi^{(t)}(w^{(t)}) - \eta \nabla f(w^{(t)}). \tag{1}$$

We discuss the stochastic version of GMD (SGMD) in Section 3. GMD generalizes both mirror descent and preconditioning methods. Namely, if for all t, $\phi^{(t)} = \nabla\psi$ for some strictly convex function ψ, then GMD corresponds to mirror descent with potential ψ; if $\phi^{(t)} = G^{(t)}$ for some invertible matrix $G^{(t)} \in \mathbb{R}^{d\times d}$, then the update rule in equation (1) reduces to

$$w^{(t+1)} = w^{(t)} - \eta {G^{(t)}}^{-1} \nabla f(w^{(t)})$$

and hence represents applying a pre-conditioner to gradient updates. The following is a summary of our results:

1. We provide a simple proof for linear convergence of GMD under the Polyak-Lojasiewicz inequality (Theorem 1).
2. We provide sufficient conditions under which SGMD converges linearly under an adaptive learning rate (Theorems 2 and 3)[1].
3. As corollaries to Theorems 1 and 3, in Section 5 we provide sufficient conditions for linear convergence of stochastic mirror descent as well as stochastic preconditioner methods such as Adagrad (Duchi et al., 2011).
4. We prove the existence of an interpolating solution and linear convergence of GMD to this solution for non-negative loss functions that locally satisfy the PL* inequality (Liu et al., 2020). This result (Theorem 4) provides approximate implicit regularization results for GMD: GMD converges linearly to an interpolating solution that is approximately the closest interpolating solution to the initialization in ℓ_2 norm in the dual space induced by $\phi^{(t)}$. For full batch mirror descent, our result recovers the approximate implicit regularization result from Azizan et al. (2019).

2 Related Work

Recent work (Azizan et al., 2019) established convergence of stochastic mirror descent (SMD) for nonlinear optimization problems. It characterized the implicit bias of mirror descent by demonstrating that SMD converges to a global minimum that is within epsilon of the closest interpolating solution in Bregman divergence. The analysis in Azizan et al. (2019) relies on the fundamental identity of SMD and does not provide explicit learning rates or establish a rate of convergence for SMD in the nonlinear setting. The work in Azizan & Hassibi (2019) provided explicit learning rates for the convergence of SMD in the linear setting under strongly convex potential, again without a rate of convergence. While these works established convergence of SMD, prior work by Gunasekar et al. (2018) analyzed the implicit bias of SMD without proving convergence.

A potential-based version of generalized mirror descent with time-varying regularizes was presented for online problems in Orabona et al. (2015). That work is primarily concerned with establishing regret bounds for the online learning setting, which differs from our setting of minimizing a loss function given a set of known data points. A potential-free formulation of GMD for the flow was presented in Gunasekar et al. (2020).

The Polyak-Lojasiewicz (PL) inequality (Lojasiewicz, 1963; Polyak, 1963) serves as a simple condition for linear convergence in non-convex optimization problems and is satisfied in a number of settings including over-parameterized neural networks (Liu et al., 2020). Work by Karimi et al. (2016) demonstrated linear convergence of a number of descent methods (including gradient descent) under the PL inequality. Similarly, Vaswani et al. (2019) proved linear convergence of stochastic gradient descent (SGD) under the PL inequality and the strong growth condition (SGC), and Bassily et al. (2018) established the same rate for SGD under just the PL inequality. Soltanolkotabi et al. (2019) also used the PL inequality to establish a local linear convergence result for gradient descent on 1 hiddden layer over-parameterized neural networks.

Recently, Xie et al. (2020) established linear convergence for a norm version of Adagrad (Adagrad-Norm) using the PL inequality, while Wu et al. (2019) established linear convergence for Adagrad-Norm in the particular setting of over-parameterized neural networks with one hidden layer. An alternate analysis for Adagrad-Norm for smooth, non-convex functions was presented in Ward et al. (2019), resulting in a sub-linear convergence rate.

[1] We also provide a fixed learning rate for monotonically decreasing gradients $\nabla f(w^{(t)})$.

Instead of focusing on a specific method, the goal of this work is to establish sufficient conditions for linear convergence by applying the PL inequality to a more general setting (SGMD). We arrive at linear convergence for specific methods such as mirror descent and preconditioned gradient descent methods as corollaries. Moreover, our local convergence results provide an intuitive formulation of approximate implicit regularization for GMD and thus mirror descent. Namely, instead of resorting to Bregman divergence, we prove that GMD converges to an interpolating solution that is approximately the closest interpolating solution to the initialization in ℓ_2 norm in the dual space induced by $\phi^{(t)}$.

3 Algorithm Description and Preliminaries

We begin with a formal description of SGMD. Let $f_i : \mathbb{R}^d \to \mathbb{R}$ denote real-valued, differentiable loss functions and let $f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x)$. In addition, let $\phi^{(t)} : \mathbb{R}^d \to \mathbb{R}^d$ be an invertible function for all non-negative integers t. We solve the optimization problem

$$\arg\min_{x\in\mathbb{R}^d} f(x)$$

using **stochastic generalized mirror descent** with learning rate η[2]:

$$\phi^{(t)}(w^{(t+1)}) = \phi^{(t)}(w^{(t)}) - \eta\nabla f_{i_t}(w^{(t)}), \tag{2}$$

where $i_t \in [n]$ is chosen uniformly at random. As described in the introduction, the above algorithm generalizes both gradient descent (where $\phi(x) = x$) and mirror descent (where $\phi^{(t)}(x) = \nabla\psi(x)$ for some strictly convex potential function ψ). In the case where $\phi^{(t)}(x) = G^{(t)}x$ for an invertible matrix $G^{(t)} \in \mathbb{R}^{d\times d}$, the update rule in equation (2) reduces to:

$$w^{(t+1)} = w^{(t)} - \eta {G^{(t)}}^{-1}\nabla f_{i_t}(w^{(t)})$$

Hence, when $\phi^{(t)}$ is an invertible linear transformation, Equation (2) is equivalent to pre-conditioned gradient descent. We now present the Polyak-Lojasiewicz inequality and lemmas from optimization theory that will be used in our proofs[3].

Polyak-Lojasiewicz (PL) Inequality. *A function $f : \mathbb{R}^d \to \mathbb{R}$ is μ-PL if for some $\mu > 0$:*

$$\frac{1}{2}\|\nabla f(x)\|^2 \geq \mu(f(x) - f(x^*)) \;\; \forall x \in \mathbb{R}^d, \tag{3}$$

where $x^* \in \mathbb{R}^d$ is a global minimizer for f.

A useful variation of the PL inequality is the PL* inequality introduced in Liu et al. (2020) which does not require knowledge of $f(x^*)$.

Definition. *A function $f : \mathbb{R}^d \to \mathbb{R}$ is μ-PL* if for some $\mu > 0$:*

$$\frac{1}{2}\|\nabla f(x)\|^2 \geq \mu f(x) \;\; \forall x \in \mathbb{R}^d, \tag{4}$$

A function that is μ-PL* is also μ-PL when f is non-negative. Additionally, we will typically assume that f is L-smooth (with L-Lipschitz continuous derivative).

Definition. *A function $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth for $L > 0$ if for all $x, y \in \mathbb{R}^d$:*

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|.$$

If $\phi^{(t)}(x) = x$ for any t and $x \in \mathbb{R}^d$ then SGMD reduces to SGD. If f is L-smooth and satisfies the PL-Inequality, then SGD converges linearly to a global minimum (Bassily et al., 2018; Karimi et al., 2016; Vaswani et al., 2019). Moreover, the following lemma (proven in Appendix A) shows that the PL* condition implies the existence of a global minimum x^* for non-negative, L-smooth f.

Lemma 1. *If $f : \mathbb{R}^d \to \mathbb{R}$ is μ-PL*, L-smooth and $f(x) \geq 0$ for all $x \in \mathbb{R}^d$, then gradient descent with learning rate $\eta < \frac{2}{L}$ converges linearly to x^* satisfying $f(x^*) = 0$.*

[2]The framework also allows for adaptive learning rates by using $\eta^{(t)}$ to denote a time-dependent step size.

[3]We assume all norms are the 2-norm unless stated otherwise.

Hence, in cases where the loss function is nonnegative (for example the squared loss), we can remove the usual assumption about the existence of a global minimum, x^*, and instead assume that f satisfies the PL* inequality. We now reference standard properties of L-smooth functions (Zhou, 2018), which will be used in our proofs.

Lemma 2. *If $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth, then for all $x, y \in \mathbb{R}^d$:*

$$(a)\ \ f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2,$$
$$(b)\ \ \|\nabla f(x)\|^2 \leq 2L(f(x) - f(x^*)).$$

The following lemma relates μ and L (the proof is in Appendix B).

Lemma 3. *If $f : \mathbb{R}^d \to \mathbb{R}$ is μ-PL and L-smooth, then $\mu \leq L$.*

Using Lemma 2b in place of the strong growth condition (i.e. $\mathbb{E}_i[\|\nabla f_i(x)\|^2] \leq \rho\|\nabla f(x)\|^2$) yields slightly different learning rates when establishing convergence of stochastic descent methods (as is apparent from the different learning rates between Bassily et al. (2018) and Vaswani et al. (2019)). The following simple lemma will be used in the proof of Theorem 3.

Lemma 4. *If $f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x)$ where $f_i : \mathbb{R}^d \to \mathbb{R}$ are L_i-smooth , then f is $\sup_i L_i$-smooth.*

Note that there could exist some other constant $L' < \sup_i L_i$ for which f is L'-smooth, but this upper bound suffices for our proof of Theorem 3. Lastly, we define and reference standard properties of strongly convex functions (Zhou, 2018), which will be useful in demonstrating how our GMD results generalize those for mirror descent.

Definition. *For $\alpha > 0$, a differentiable function, $\psi : \mathbb{R}^d \to \mathbb{R}$, is α-strongly convex if for all x, y,*

$$\psi(y) \geq \psi(x) + \langle \nabla\psi(x), y - x \rangle + \frac{\alpha}{2}\|y - x\|^2.$$

Lemma 5. *If $\psi : \mathbb{R}^d \to \mathbb{R}$ is α-strongly convex, then for all x, y:*

$$\psi(y) \leq \psi(x) + \langle \nabla\psi(x), y - x \rangle + \frac{1}{2\alpha}\|\nabla\psi(y) - \nabla\psi(x)\|^2.$$

With these preliminaries in hand, we now present our proofs for linear convergence of SGMD using the PL-Inequality.

4 Sufficient Conditions for Linear Convergence of SGMD

In this section, we provide sufficient conditions to establish (expected) linear convergence for (stochastic) GMD. We first provide simple conditions under which GMD converges linearly by extending the proof strategy from Karimi et al. (2016). We then present alternate conditions for linear convergence of GMD, which can be naturally extended to the stochastic setting.

4.1 Simple Conditions for Linear Convergence of GMD

We begin with a simple set of conditions under which (non-stochastic) GMD converges linearly (the full proof is presented in Appendix C). The main benefit of this analysis is that it is a straightforward extension of the proof of linear convergence for gradient descent under the PL-Inequality presented in Karimi et al. (2016).

Theorem 1. *Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth and μ-PL and $\phi^{(t)} : \mathbb{R}^d \to \mathbb{R}^d$ is an invertible, $\alpha_u^{(t)}$-Lipschitz function where $\lim_{t\to\infty} \alpha_u^{(t)} < \infty$. If for all $x, y \in \mathbb{R}^d$ and for all timesteps t there exist $\alpha_l^{(t)} > 0$ such that*

$$\langle \phi^{(t)}(x) - \phi^{(t)}(y), x - y \rangle \geq \alpha_l^{(t)}\|x - y\|^2,$$

and $\lim_{t\to\infty} \alpha_l^{(t)} > 0$, then generalized mirror descent converges linearly to a global minimum for any $\eta^{(t)} < \frac{2\alpha_l^{(t)}}{L}$.

Remark. Theorem 1 yields a fixed learning rate provided that $\alpha_l^{(t)}$ is uniformly bounded. In addition, note that Theorem 1 applies also under weaker assumptions, namely when $\phi^{(t)}$ is locally Lipschitz. Finally, the provided learning rate can be computed exactly for settings such as linear regression, since it only requires knowledge of L and $\alpha_l^{(t)}$ (see Section 7). When $\eta = \frac{\alpha_l^{(t)}}{L}$ and given w^* a minimizer of f, the proof of Theorem 1 implies that:

$$f(w^{(t+1)}) - f(w^*) \leq \left(1 - \frac{\mu {\alpha_l^{(t)}}^2}{L {\alpha_u^{(t)}}^2}\right)(f(w^{(t)}) - f(w^*)).$$

Letting $\kappa^{(t)} = \frac{L {\alpha_u^{(t)}}^2}{\mu {\alpha_l^{(t)}}^2}$ thus generalizes the condition number introduced in Definition 4.1 of Liu et al. (2020) for gradient descent. Provided that $\lim_{t\to\infty} \kappa^{(t)} > 0$, then Theorem 1 guarantees linear convergence to a global minimum.

4.2 Taylor Series Analysis for Linear Convergence in GMD

Although the proof of Theorem 1 is succinct, it is nontrivial to extend to the stochastic setting[4]. In order to develop a convergence result for the stochastic setting, we turn to an alternate set of conditions for linear convergence by using the Taylor expansion of ϕ^{-1}. We use $\mathbf{J}_\phi$ to denote the Jacobian of ϕ. For ease of notation, we consider non-time-dependent α_l, α_u, but our results are trivially extendable to the setting when these quantities are time-dependent.

Theorem 2. *Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth and μ-PL and $\phi : \mathbb{R}^d \to \mathbb{R}^d$ is an infinitely differentiable, analytic function with analytic inverse, ϕ^{-1}. If there exist $\alpha_l, \alpha_u > 0$ such that*

$$(a)\ \ \alpha_l \mathbf{I} \preccurlyeq \mathbf{J}_\phi \preccurlyeq \alpha_u \mathbf{I},$$

$$(b)\ \ |\partial_{i_1,\dots i_k} \phi_j^{-1}(x)| \leq \frac{k!}{2\alpha_u d}\ \ \forall x \in \mathbb{R}^d, i_1, \dots i_k \in [d], j \in [d], k \geq 2,$$

then generalized mirror descent converges linearly for any $\eta^{(t)} < \min\left(\frac{4\alpha_l^2}{5L\alpha_u}, \frac{1}{2\sqrt{d}\|f(w^{(t)})\|}\right)$.

The full proof is provided in Appendix D. Importantly, the adaptive component of the learning rate is only used to ensure that the sum of the higher order terms for the Taylor expansion converges. In particular, if $\phi^{(t)}$ is a linear function, then our learning rate no longer needs to be adaptive. Note that alternatively, we can establish linear convergence for a fixed learning rate given that the gradients monotonically decrease. We analyze this case in Appendix E and provide an explicit condition on μ and L under which this holds.

4.3 Taylor Series Analysis for Linear Convergence in Stochastic GMD

The main benefit of the above Taylor series analysis is that it naturally extends to the stochastic setting as demonstrated in the following result (with proof presented in Appendix F).

Theorem 3. *Suppose $f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x)$, where $f_i : \mathbb{R}^d \to \mathbb{R}$ are L_i-smooth functions with $L = \sup_{i\in[n]} L_i$ and f is μ-PL. Let $\phi : \mathbb{R}^d \to \mathbb{R}^d$ be an infinitely differentiable, analytic function with analytic inverse, ϕ^{-1}. SGMD is used to minimize f according to the updates:*

$$\phi(w^{(t+1)}) = \phi(w^{(t)}) - \eta^{(t)} \nabla f_{i_t}(w^{(t)}),$$

where $i_t \in [n]$ is chosen uniformly at random and $\eta^{(t)}$ is an adaptive step size. If there exist $\alpha_l, \alpha_u > 0$ such that:

$$(a)\ \ \alpha_l \mathbf{I} \preccurlyeq \mathbf{J}_\phi \preccurlyeq \alpha_u \mathbf{I},$$

$$(b)\ \ |\partial_{i_1,\dots i_k} \phi_j^{-1}(x)| \leq \frac{k!}{2\alpha_u d}\ \ \forall x \in \mathbb{R}^d, i_1, \dots i_k \in [d], j \in [d], k \geq 2,$$

then SGMD converges linearly to a global minimum for any $\eta^{(t)} < \min\left(\frac{4\mu\alpha_l^2}{5L^2\alpha_u}, \frac{1}{2\sqrt{d}\|f_{i_t}(w^{(t)})\|}\right)$.

[4]The main difficulty is relating $w^{(t+1)} - w^{(t)}$ to the gradient at timestep t.

Remark. Note that there is a slight difference between the learning rate in Theorem 2 and Theorem 3 due to a multiplicative factor of μ. Consistent with the difference in learning rates between Bassily et al. (2018) and Vaswani et al. (2019), we can make the learning rate between the two theorems match if we assume the strong growth condition (i.e. $\mathbb{E}_i[\|\nabla f_i(x)\|^2] \leq \rho\|\nabla f(x)\|^2$) with $\rho = \mu$ instead of using Lemma 2b.

5 Corollaries of Linear Convergence in SGMD

We now present how the linear convergence results established by Theorems 1, 2, and 3 apply to commonly used optimization algorithms including mirror descent and Adagrad. In this section, we primarily extend the analysis from Theorem 1 for the non-stochastic case. However, our results can be extended analogously to give expected linear convergence in the stochastic case by using the extension provided in Theorem 3.

Gradient Descent. For the case of gradient descent, $\phi(x) = x$ and so $\alpha_l = \alpha_u = 1$. Hence, we see that gradient descent converges linearly under the conditions of Theorem 1 with $\eta < \frac{2}{L}$, which is consistent with the analysis in Karimi et al. (2016).

Mirror Descent. Let $\psi : \mathbb{R}^d \to \mathbb{R}$ be a strictly convex potential. Thus, $\phi(x) = \nabla\psi(x)$ is an invertible function. If ψ is α_l-strongly convex and (locally) α_u-Lipschitz and f is L-smooth and μ-PL, then the conditions of Theorem 1 are satisfied. Moreover, since the α_u-Lipschitz condition holds locally for most potentials considered in practice, our result implies linear convergence for mirror descent with α_l-strongly convex potential ψ.

Adagrad. Let $\phi^{(t)} = \mathcal{G}^{(t)\,\frac{1}{2}}$ where $\mathcal{G}^{(t)}$ is a diagonal matrix such that

$$\mathcal{G}_{i,i}^{(t)} = \sum_{k=0}^{t} \nabla f_i(w^{(k)})^2.$$

Then GMD corresponds to Adagrad. In this case, we can apply Theorem 1 to establish linear convergence of Adagrad under the PL-Inequality provided that $\phi^{(t)}$ satisfies the condition of Theorem 1. The following corollary proves that this condition holds and hence that Adagrad converges linearly.

Corollary 1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be an L-smooth function that is μ-PL. Let ${\alpha_l^{(t)}}^2 = \min_{i\in[d]} \mathcal{G}_{i,i}^{(t)}$ and ${\alpha_u^{(t)}}^2 = \max_{i\in[d]} \mathcal{G}_{i,i}^{(t)}$. If $\lim_{t\to\infty} \frac{\alpha_l^{(t)}}{\alpha_u^{(t)}} \neq 0$, then Adagrad converges linearly for adaptive step size $\eta^{(t)} = \frac{\alpha_l^{(t)}}{L}$.*

The proof is presented in Appendix H. While Corollary 1 can be extended to the stochastic setting via Theorem 3, it requires knowledge of μ to setup the learning rate, and the resulting learning rate provided is typically smaller than what we can use in practice. We analyze this case further in Section 7. Additionally, since the condition $\lim_{t\to\infty} \frac{\alpha_l^{(t)}}{\alpha_u^{(t)}} \neq 0$ is difficult to verify in practice, we provide Corollary 2 in Appendix H, which presents a verifiable condition under which Adagrad converges linearly.

6 Local Convergence and Implicit Regularization in GMD

In the previous sections, we established linear convergence for GMD for real-valued loss, $f : \mathbb{R}^d \to \mathbb{R}$, that is μ-PL for all $x \in \mathbb{R}^d$. In this section, we show that f need only satisfy the PL inequality locally (i.e. within a ball of fixed radius around the initialization) in order to establish linear convergence. The following theorem (proof in Appendix G) extends Theorem 4.2 from Liu et al. (2020) to GMD and uses the PL* condition to establish both the existence of a global minimum and linear convergence to this global minimum under GMD[5]. We use $\mathcal{B}(x, R) = \{z \,;\, z \in \mathbb{R}^d,\ \|x - z\|_2 \leq R\}$ to denote the ball of radius R centered at x.

Theorem 4. *Suppose $\phi : \mathbb{R}^d \to \mathbb{R}^d$ is an invertible, α_u-Lipschitz function and that $f : \mathbb{R}^d \to \mathbb{R}$ is non-negative, L-smooth, and μ-PL* on $\tilde{\mathcal{B}} = \{x \,;\, \phi(x) \in \mathcal{B}(\phi(w^{(0)}), R)\}$ with*

[5]We require additional assumptions on $\phi^{(t)}$ for the case of time-dependent mirrors (see Appendix G.)

$R = \frac{2\sqrt{2L}\sqrt{f(w^{(0)})}\alpha_u^2}{\alpha_l \mu}$. *If for all* $x, y \in \mathbb{R}^d$ *there exists* $\alpha_l > 0$ *such that*

$$\langle \phi(x) - \phi(y), x - y \rangle \geq \alpha_l \|x - y\|^2,$$

then,

(1) *There exists a global minimum* $w^{(\infty)} \in \tilde{\mathcal{B}}$.

(2) *GMD converges linearly to* $w^{(\infty)}$ *for* $\eta = \dfrac{\alpha_l}{L}$.

(3) *If* $w^* = \underset{w \in \tilde{\mathcal{B}}\ ;\ f(w)=0}{\arg\min} \|\phi(w) - \phi(w^{(0)})\|$, *then,* $\|\phi(w^*) - \phi(w^{(\infty)})\| \leq 2R$.

Approximate Implicit Regularization in GMD. When R is small, we can view the result of Theorem 4 as a characterization of the solution selected by GMD, thereby obtaining approximate implicit regularization results for GMD. Namely, for $\delta = \frac{R}{2}$, we have $\|\phi(w^*) - \phi(w^\infty)\| \leq \delta$. Hence provided that R is small (which holds for small $f(w^{(0)})$), GMD selects an interpolating solution that is close to w^* in ℓ_2-norm in the dual space induced by ϕ. This view is consistent with the characterization of approximate implicit regularization in Azizan et al. (2019), as is shown by Corollary 3 in Appendix I). In particular, Corollary 3 strengthens Theorem 4 of Azizan et al. (2019) for the full batch case by proving (1) the existence of such a w^*, (2) *linear* convergence of $w^{(0)}$ to w^*, and (3) providing explicit forms for the $o(\epsilon)$ terms (where $\epsilon = R^2$ above). Importantly, the approximate implicit regularization result for mirror descent does not need to be stated in terms of Bregman divergence, but can be viewed more naturally as $\|\nabla\psi(w^{(\infty)}) - \nabla\psi(w^*)\|_2$ being small.

7 Experimental Verification of our Theoretical Results

We now present a simple set of experiments under which we can explicitly compute the learning rates in our theorems. We will show that in accordance with our theory, both fixed and adaptive versions of these learning rates yield linear convergence. We focus on computing learning rates for Adagrad in the noiseless regression setting used in Xie et al. (2020). Namely, we are given $(X, y) \in \mathbb{R}^{n \times d} \times \mathbb{R}^n$ such that there exists a $w^* \in \mathbb{R}^d$ such that $Xw^* = y$. If $n < d$, then the system is over-parameterized, and if $n \geq d$, the system is sufficiently parameterized and has a unique solution.

In this setting, the squared loss (MSE) is L-smooth with $L = \lambda_{\max}(XX^T)$, and it is μ-PL with $\mu = \lambda_{\min}(XX^T)$ where λ_{max} and λ_{min} refer to the largest and smallest non-zero eigenvalues, respectively[6]. Moreover, for Adagrad, we can compute $\alpha_l^{(t)} = \min_{i \in [d]} (\sum_{k=0}^t \nabla f_i(w^{(k)}))^2$ and $\alpha_u^{(t)} = \max_{i \in [d]} (\sum_{k=0}^t \nabla f_i(w^{(k)}))^2$ at each timestep. Hence for Adagrad in the noiseless linear regression setting, we can explicitly compute the learning rate provided in Theorem 3 for the stochastic setting and in Corollary 1 for the full batch setting.

Figure 1 demonstrates that in both, the over-parameterized and sufficiently parameterized settings, our provided learning rates yield linear convergence. In the stochastic setting, the theory for fixed learning rates suggests a very small rate ($\approx 10^{-9}$ for Figure 1d) and hence we chose to only present the more reasonable adaptive step size as a comparison. In the full batch setting, the learning rate obtained from our theorems out-performs using the standard fixed learning rate of 0.1, while performance is comparable for the stochastic setting. Interestingly, our theory suggests an adaptive learning rate that is increasing (in contrast to the usual decreasing learning rate schedules). In particular, while the suggested learning rate for Figure 1a starts at 0.99, it increases to 1.56 at the end of training.

In Appendix J, we present experiments on over-parameterized neural networks. While the PL-condition holds in this setting (Liu et al., 2020), it can be difficult to compute the smoothness parameter L (which was the motivation for developing Adagrad-Norm). Interestingly, our experiments demonstrate that our increasing adaptive learning rate from Theorem 1, using an approximation for L, provides convergence for Adagrad in over-parameterized networks. The link to the code is provided in Appendix J.

[6]We take μ as the smallest non-zero eigenvalue since Adagrad updates keep parameters in the span of the data.

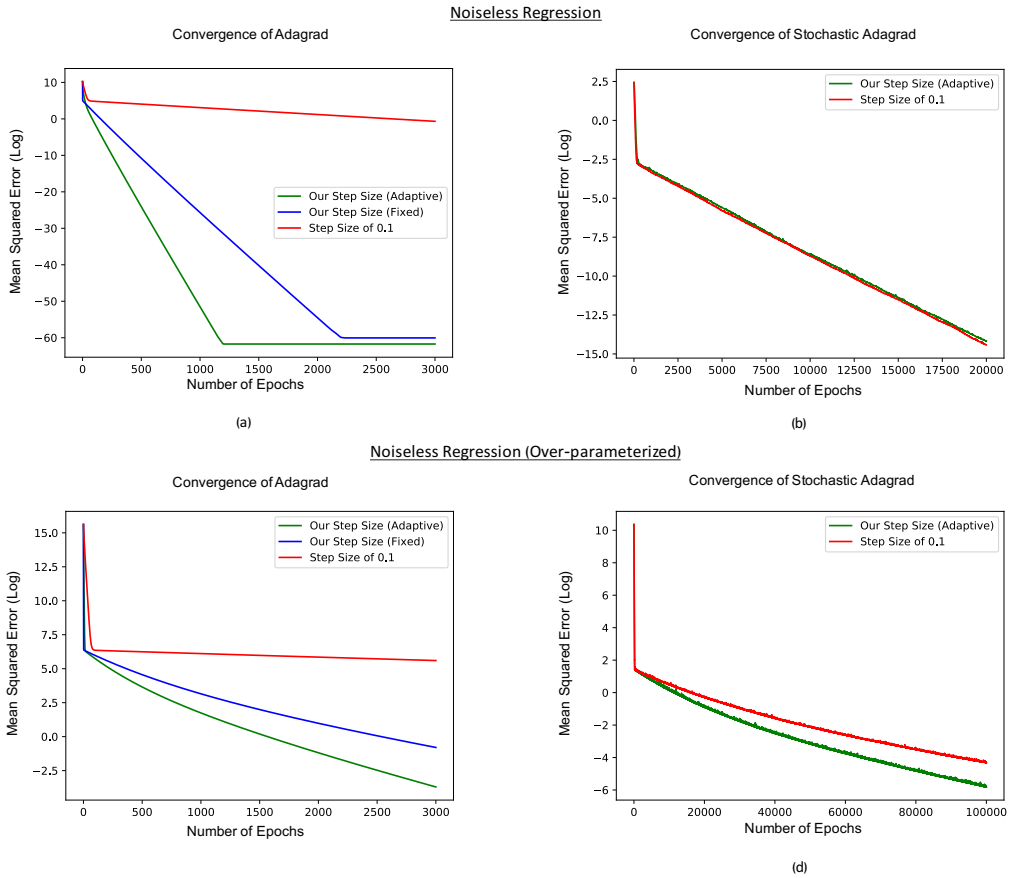

Figure 1: Using the rates provided by Corollary 1 leads to linear convergence for (Stochastic) Adagrad in the noiseless linear regression setting also considered in Xie et al. (2020). (a, b) Noiseless linear regression on 2000 examples in 20 dimensions. (c, d) Noiseless linear regression on 200 examples in 1000 dimensions.

8 Conclusion

In this work, we presented stochastic generalized mirror descent, which generalizes both mirror descent and pre-conditioner methods. By using the PL-condition and a Taylor-series based analysis, we provided sufficient conditions for linear convergence of SGMD in the non-convex setting. As a corollary, we obtained sufficient conditions for linear convergence of both mirror descent and pre-conditioner methods such as Adagrad. Lastly, we prove the existence of an interpolating solution and linear convergence of GMD to this solution for non-negative loss functions that are locally PL*. Importantly, our local convergence results allow us to obtain approximate implicit regularization results for GMD. Namely, we prove that GMD linearly converges to an interpolating solution that is approximately the closest interpolating solution to the initialization in ℓ_2 norm in the dual space. For the full batch setting, this result generalizes the approximate implicit regularization results from Azizan et al. (2019) for mirror descent to generalized mirror descent, and provides a more natural characterization of implicit regularization in terms of ℓ_2 norm in the dual space, as opposed to Bregman divergence.

Looking ahead, we envision that the generality of our analysis (and the PL-condition) could provide useful in the analysis of other commonly used adaptive methods such as Adam (Kingma & Ba, 2015). Moreover, since the PL-condition holds in varied settings including over-parameterized neural networks (Liu et al., 2020), it would be interesting to analyze whether the learning rates obtained here provide an improvement for convergence in these modern settings.

REFERENCES

Navid R. Azizan and Babak Hassibi. Stochastic Gradient/Mirror Descent: Minimax Optimality and Implicit Regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Navid R. Azizan, Sahin Lale, and Babak Hassibi. Stochastic Mirror Descent on Overparameterized Nonlinear Models. In *International Conference on Machine Learning (ICML) Generalization Workshop*, 2019.

Raef Bassily, Mikhail Belkin, and Siyuan Ma. On exponential convergence of sgd in non-convex over-parametrized learning. *arXiv preprint arXiv:1811.02564*, 2018.

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

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks, 2018. arXiv:1810.02054.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive Subgradient Methods forOnline Learning and Stochastic Optimization. *Journal of Machine Learning Research (JMLR)*, 12:2121–2159, 2011.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning (ICML)*, 2018.

Suriya Gunasekar, Blake Woodworth, and Nathan Srebro. Mirrorless Mirror Descent: A More Natural Discretization of Riemannian Gradient Flow. *arXiv preprint arXiv:2004.01025*, 2020.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear Convergence of Gradient and Proximal-Gradient Methods Under the Polyak-Lojasiewicz Condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 795–811. Springer, 2016.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 8157–8166, 2018.

Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Toward a theory of optimization for over-parameterized systems of non-linear equations: the lessons of deep learning. *arXiv preprint arXiv:2003.00307*, 2020.

Stanislaw Lojasiewicz. A topological property of real analytic subsets (in French). *Les équations aux dérivées partielles.*, 117:87–89, 1963.

Arkadi S. Nemirovsky and David B. Yudin. Problem Complexity and Method Efficiency in Optimization. *SIAM*, 27(2):264–265, 1983.

Francesco Orabona, Koby Crammer, and Nicolò Cesa-Bianchi. Mirror descent and nonlinear projected subgradient methodsfor convex optimization. *Machine Learning*, 99:411–435, 2015.

Boris Polyak. Gradient methods for minimizing functionals (in Russian). *Zh. Vychisl. Mat. Mat. Fiz.*, 3(4):643–653, 1963.

Mahdi Soltanolkotabi, Adel Javanmard, and Jason D. Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE transaction on Information Theory (2018)*, 65(2):742 – 769, 2019.

Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and Faster Convergence of SGD for Over-Parameterized Models and an Accelerated Perceptron. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

Rachel Ward, Xiaoxia Wu, and Léon Bottou. AdaGrad Stepsizes: Sharp Convergence Over Non-convex Landscapes. In *International Conference on Machine Learning (ICML)*, 2019.

Xiaoxia Wu, Simon S Du, and Rachel Ward. Global convergence of adaptive gradient methods for an over-parameterized neural network. *arXiv preprint arXiv:1902.07111*, 2019.

Yuege Xie, Xiaoxia Wu, and Rachel Ward. Linear Convergence of Adaptive Stochastic Gradient Descent. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolution network, 2015. arXiv:1505.00853.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations (ICLR)*, 2017.

Xingyu Zhou. On the Fenchel Duality between Strong Convexity and Lipschitz Continuous Gradient. *arXiv preprint arXiv:1803.06573*, 2018.