APPENDIX

A PROOF OF LEMMA 1

We restate the lemma below.

Lemma. *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 $f(x^*) = 0$.*

Proof. The proof follows exactly from Theorem 1 of Karimi et al. (2016). Since f is L-smooth, by Lemma 2a it holds that:

$$\begin{aligned}
&f(w^{(t+1)}) - f(w^{(t)}) \leq \langle \nabla f(w^{(t)}), w^{(t+1)} - w^{(t)} \rangle + \frac{L}{2} \|w^{(t+1)} - w^{(t)}\|^2. \\
\implies &f(w^{(t+1)} - f(w^{(t)}) \leq -\eta \|\nabla f(w^{(t)})\|^2 + \frac{L}{2}\eta^2 \|\nabla f(w^{(t)})\|^2 \\
\implies &f(w^{(t+1)} - f(w^{(t)}) \leq \left(-\eta + \frac{\eta^2 L}{2}\right) 2\mu f(w^{(t)}) \\
\implies &f(w^{(t+1)} \leq \left(1 - 2\mu\eta + \mu\eta^2 L\right) f(w^{(t)})
\end{aligned}$$

Hence, if $\eta < \frac{2}{L}$, then $C = \left(1 - 2\mu\eta + \mu\eta^2 L\right) < 1$. Thus, we have $f(w^{(t+1)}) \leq C f(w^{(t)})$ for $C < 1$. Thus, as f is bounded below by 0 and the sequence $\{f(w^{(t)})\}_{t\in\mathbb{N}}$ monotonically decreases with infimum 0, the monotone convergence theorem implies $\lim_{t\to\infty} f(w^{(t)}) = 0$. □

B PROOF OF LEMMA 3

Proof. From Lemma 2 and from the PL condition, we have:

$$2\mu(f(x) - f(x^*)) \leq \|\nabla f(x)\|^2 \leq 2L(f(x) - f(x^*)) \implies \mu \leq L \quad \square$$

C PROOF OF THEOREM 1

Proof. Since f is L-smooth, by Lemma 2a it holds that:

$$f(w^{(t+1)}) - f(w^{(t)}) \leq \langle \nabla f(w^{(t)}), w^{(t+1)} - w^{(t)} \rangle + \frac{L}{2} \|w^{(t+1)} - w^{(t)}\|^2. \tag{5}$$

Now by the condition on $\phi^{(t)}$ in Theorem 1, we bound the first term on the right as follows:

$$\begin{aligned}
&\langle \phi^{(t)}(w^{(t+1)}) - \phi^{(t)}(w^{(t)}), w^{(t+1)} - w^{(t)} \rangle \geq \alpha_l^{(t)} \|w^{(t+1)} - w^{(t)}\|^2 \\
\implies &\langle -\eta \nabla f(w^{(t)}), w^{(t+1)} - w^{(t)} \rangle \geq \alpha_l^{(t)} \|w^{(t+1)} - w^{(t)}\|^2 \text{ using Equation (2)} \\
\implies &\langle \nabla f(w^{(t)}), w^{(t+1)} - w^{(t)} \rangle \leq -\frac{\alpha_l^{(t)}}{\eta} \|w^{(t+1)} - w^{(t)}\|^2.
\end{aligned}$$

Substituting this bound back into the inequality in (5), we obtain

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

Since the learning rate is selected so that the coefficient of $\|w^{(t+1)} - w^{(t)}\|^2$ on the right is negative, we obtain

$$\begin{aligned}
f(w^{(t+1)}) - f(w^{(t)}) &\leq \left(-\frac{\alpha_l^{(t)}}{\eta} + \frac{L}{2}\right)\|w^{(t+1)} - w^{(t)}\|^2 \\
&\leq \left(-\frac{\alpha_l^{(t)}}{\eta} + \frac{L}{2}\right)\frac{1}{{\alpha_u^{(t)}}^2}\|\phi^{(t)}(w^{(t+1)}) - \phi^{(t)}(w^{(t)})\|^2 \\
&= \left(-\frac{\alpha_l^{(t)}}{\eta} + \frac{L}{2}\right)\frac{1}{{\alpha_u^{(t)}}^2}\|-\eta\nabla f(w^{(t)})\|^2 \text{ using Equation (1)} \\
&\leq \left(-\frac{\alpha_l^{(t)}}{\eta} + \frac{L}{2}\right)2\mu\frac{\eta^2}{{\alpha_u^{(t)}}^2}(f(w^{(t)}) - f(w^*)) \text{ as } f \text{ is } \mu\text{-PL} \\
\implies f(w^{(t+1)}) - f(w^*) &\leq \left(1 - 2\mu\frac{\eta\alpha_l^{(t)}}{{\alpha_u^{(t)}}^2} + \mu\frac{L\eta^2}{{\alpha_u^{(t)}}^2}\right)(f(w^{(t)}) - f(w^*)),
\end{aligned}$$

where the second inequality follows since $\phi^{(t)}$ is $\alpha_u^{(t)}$-Lipschitz. For linear convergence, we need.

$$0 < 1 - 2\mu\frac{\eta\alpha_l^{(t)}}{{\alpha_u^{(t)}}^2} + \mu\frac{L\eta^2}{{\alpha_u^{(t)}}^2} < 1. \tag{6}$$

From Lemma 3, $\mu < \frac{{\alpha_u^{(t)}}^2 L}{\alpha_l^{(t)}}$ always holds and implies that the left inequality in (6) is satisfied for all $\eta^{(t)}$. The right inequality holds by our assumption that $\eta^{(t)} < \frac{2\alpha_l^{(t)}}{L}$, which completes the proof. $\square$

D Proof of Theorem 2

We repeat the theorem below for convenience.

Theorem. *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:*

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

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

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

Proof. Since f is L-smooth, it holds by Lemma that 2:

$$f(w^{(t+1)}) - f(w^{(t)}) \leq \langle\nabla f(w^{(t)}), w^{(t+1)} - w^{(t)}\rangle + \frac{L}{2}\|w^{(t+1)} - w^{(t)}\|^2.$$

Next, we want to bound the two quantities on the right hand side by a multiple of $\|\nabla f(w^{(t)})\|^2$. We do so by expanding $w^{(t+1)} - w^{(t)}$ using the Taylor series for ϕ^{-1} as follows:

$$\begin{aligned}
w^{(t+1)} - w^{(t)} &= \phi^{-1}(\phi(w^{(t)}) - \eta\nabla f(w^{(t)})) - w^{(t)} \\
&= -\eta\mathbf{J}_{\phi^{-1}}(\phi(w^{(t)}))\nabla f(w^{(t)}) \\
&\quad + \sum_{k=2}^{\infty}\frac{1}{k!}\left[\textstyle\sum_{i_1,i_2\ldots i_k=1}^{d}(-\eta)^k\partial_{i_1,\ldots i_k}\phi_j^{-1}(\phi(w^{(t)}))(\nabla f(w^{(t)})_{i_1}\ldots\nabla f(w^{(t)})_{i_k})\right].
\end{aligned}$$

The quantity in brackets is a column vector where we only wrote out the j^{th} coordinate for $j \in [d]$. Now we bound the term $\langle\nabla f(w^{(t)}), w^{(t+1)} - w^{(t)}\rangle$:

$$\begin{aligned}
\langle\nabla f(w^{(t)}), w^{(t+1)} - w^{(t)}\rangle &= -\eta\nabla f(w^{(t)})^T\mathbf{J}_\phi^{-1}(w^{(t)})\nabla f(w^{(t)}) \\
&+\nabla f(w^{(t)})^T\sum_{k=2}^{\infty}\frac{1}{k!}\left[\sum_{i_1,i_2\ldots i_k=1}^{d}(-\eta)^k\partial_{i_1,\ldots i_k}\phi_j^{-1}(\phi(w^{(t)}))(\nabla f(w^{(t)})_{i_1}\ldots\nabla f(w^{(t)})_{i_k})\right].
\end{aligned}$$

We have separated the first order term from the other orders because we will bound them separately using conditions (a) and (b) respectively. Namely, we first have:

$$-\eta \nabla f(w^{(t)})^T \mathbf{J}_\phi^{-1}(w^{(t)}) \nabla f(w^{(t)}) \leq -\frac{\eta}{\alpha_u} \|\nabla f(w^{(t)})\|^2.$$

Next, we use the Cauchy-Schwarz inequality on inner products to bound the inner product of $\nabla f(w^{(t)})$ and the higher order terms. In the following, we use α to denote $\frac{1}{2\alpha_u d}$.

$$\begin{aligned}
&\nabla f(w^{(t)})^T \sum_{k=2}^{\infty} \frac{1}{k!} \left[\textstyle\sum_{i_1,i_2\dots i_k=1}^{d} (-\eta)^k \partial_{i_1,\dots i_k} \phi_j^{-1}(\phi(w^{(t)}))(\nabla f(w^{(t)})_{i_1} \dots \nabla f(w^{(t)})_{i_k})\right] \\
\leq &\|\nabla f(w^{(t)})\| \sum_{k=2}^{\infty} \frac{1}{k!} \left\| \left[\textstyle\sum_{i_1,i_2\dots i_k=1}^{d} (-\eta)^k \partial_{i_1,\dots i_k} \phi_j^{-1}(\phi(w^{(t)}))(\nabla f(w^{(t)})_{i_1} \dots \nabla f(w^{(t)})_{i_k})\right] \right\| \\
\leq &\|\nabla f(w^{(t)})\| \sum_{k=2}^{\infty} \frac{\alpha k!}{k!} (\eta)^k \left\| \left[\textstyle\sum_{i_1,i_2\dots i_k=1}^{d} (|\nabla f(w^{(t)})_{i_1}| \dots |\nabla f(w^{(t)})_{i_k}|)\right] \right\| \\
= &\|\nabla f(w^{(t)})\| \alpha \sum_{k=2}^{\infty} \sqrt{d} (\eta)^k (|\nabla f(w^{(t)})_1| + \dots |\nabla f(w^{(t)}))_d|)^k \\
= &\|\nabla f(w^{(t)})\| \alpha \sum_{k=2}^{\infty} (\eta)^k \sqrt{d} |\langle \begin{bmatrix} |\nabla f(w^{(t)})_1| \\ \vdots \\ |\nabla f(w^{(t)})_d| \end{bmatrix}, \mathbf{1} \rangle|^k \\
\leq &\|\nabla f(w^{(t)})\| \alpha \sum_{k=2}^{\infty} (\eta)^k \sqrt{d} \|\nabla f(w^{(t)})\|^k (\sqrt{d})^k \\
= &\alpha \sum_{k=2}^{\infty} (\sqrt{d})^{k+1} (\eta)^k \|\nabla f(w^{(t)})\|^{k+1} \\
= &\alpha (\sqrt{d})^3 (\eta)^2 \|\nabla f(w^{(t)})\|^3 \sum_{k=0}^{\infty} (\sqrt{d})^k (\eta)^k \|\nabla f(w^{(t)})\|^k = \frac{\alpha(\sqrt{d})^3 (\eta)^2 \|\nabla f(w^{(t)})\|^3}{1 - \sqrt{d}\eta \|\nabla f(w^{(t)})\|}.
\end{aligned}$$

Hence we can select $\eta < \frac{1}{2\sqrt{d}\|\nabla f(w^{(t)})\|}$ such that:

$$\frac{\alpha(\sqrt{d})^3 (\eta)^2 \|\nabla f(w^{(t)})\|^3}{1 - \sqrt{d}\eta \|\nabla f(w^{(t)})\|} \leq \frac{\alpha(\sqrt{d})^3 (\eta)^2 \|\nabla f(w^{(t)})\|^3}{\sqrt{d}\eta \|\nabla f(w^{(t)})\|} = d\alpha\eta \|\nabla f(w^{(t)})\|^2.$$

Thus, we have established the following bound:

$$\langle \nabla f(w^{(t)}), w^{(t+1)} - w^{(t)} \rangle \leq \left(-\frac{\eta}{\alpha_u} + d\alpha\eta \right) \|\nabla f(w^{(t)})\|^2 = \left(-\frac{\eta}{2\alpha_u} \right) \|\nabla f(w^{(t)})\|^2.$$

Proceeding analogously as above, we establish a bound on $\|w^{(t+1)} - w^{(t)}\|^2$:

$$\|w^{(t+1)} - w^{(t)}\|^2 \leq \left(\frac{\eta^2}{\alpha_l^2} + \alpha^2 d^2 \eta^2 \right) \|\nabla f(w^{(t)})\|^2 = \left(\frac{\eta^2}{\alpha_l^2} + \frac{\eta^2}{4\alpha_u^2} \right) \|\nabla f(w^{(t)})\|^2.$$

Putting the bounds together we obtain:

$$f(w^{(t+1)}) - f(w^{(t)}) \leq \left(-\frac{\eta}{2\alpha_u} + \frac{L\eta^2}{2\alpha_l^2} + \frac{L\eta^2}{8\alpha_u^2} \right) \|\nabla f(w^{(t)})\|^2.$$

We select our learning rate to make the coefficient of $\|\nabla f(w^{(t)}\|^2$ negative, and thus by the PL-inequality (4), we have:

$$\begin{aligned}
&f(w^{(t+1)}) - f(w^{(t)}) \leq \left(-\frac{\eta}{2\alpha_u} + \frac{L\eta^2}{2\alpha_l^2} + \frac{L\eta^2}{8\alpha_u^2} \right) 2\mu (f(w^{(t)}) - f(w^*)) \\
\implies &f(w^{(t+1)}) - f(w^*) \leq \left(1 - \frac{\mu\eta}{\alpha_u} + \frac{\mu L\eta^2}{\alpha_l^2} + \frac{\mu L\eta^2}{4\alpha_u^2} \right) (f(w^{(t)}) - f(w^*)).
\end{aligned}$$

Hence, $w^{(t)}$ converges linearly when:

$$0 < 1 - \frac{\mu\eta}{\alpha_u} + \frac{\mu L\eta^2}{\alpha_l^2} + \frac{\mu L\eta^2}{4\alpha_u^2} < 1.$$

To show that the left hand side is true, we analyze when the discriminant is negative. Namely, we have that the left side holds if:

$$\begin{aligned}
&\frac{\mu^2}{\alpha_u^2} - \frac{4\mu L}{\alpha_l^2} - \frac{\mu L}{\alpha_u^2} < 0 \\
\implies &\frac{\mu}{\alpha_u^2} < \frac{4L}{\alpha_l^2} + \frac{L}{\alpha_u^2} \\
\implies &\mu < \frac{4L\alpha_u^2}{\alpha_l^2} + L.
\end{aligned}$$

Since $\mu < L$ by Lemma 3, this is always true. The right hand side holds when $\eta < \frac{4\alpha_l^2}{5L\alpha_u}$, which holds by the assumption of the theorem, thereby completing the proof. □

Note that if f is non-negative and μ-PL*, then we have:

$$\eta^{(t)} \leq \frac{1}{2\sqrt{2Ld}\sqrt{f(w^{(0)})}} \leq \frac{1}{2\sqrt{2Ld}\sqrt{f(w^{(t)})}} \leq \frac{1}{2\sqrt{d}\|\nabla f(w^{(t)})\|}$$

Hence, we can use a fixed learning rate of $\eta = \min\left(\frac{4\alpha_l^2}{5L\alpha_u}, \frac{1}{2\sqrt{2Ld}\sqrt{f(w^{(0)})}}\right)$ in this setting.

E Conditions for Monotonically Decreasing Gradients

As discussed in the remarks after Theorem 2, we can provide a fixed learning rate for linear convergence provided that the gradients are monotonically decreasing. As we show below, this requires special conditions on the PL constant, μ, and the smoothness constant, L, for f.

Proposition 1. *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) $|\partial_{i_1,\ldots i_k}\phi_j^{-1}(x)| \leq \dfrac{k!}{2\alpha_u d}$ $\forall x \in \mathbb{R}^d, i_1, \ldots i_k \in [d], j \in [d], k \geq 2$,

(c) $\dfrac{\mu}{L} > \dfrac{4\alpha_u^2 + \alpha_l^2}{4\alpha_u^2 + 2\alpha_l^2}$,

then generalized mirror descent converges linearly for any $\eta < \min\left(\frac{4\alpha_l^2}{5L\alpha_u}, \frac{1}{2\sqrt{d}\|\nabla f(w^{(0)})\|}\right)$.

Proof. Let $C = 1 - \frac{\mu\eta}{\alpha_u} + \frac{\mu L\eta^2}{\alpha_l^2} + \frac{\mu L\eta^2}{4\alpha_u^2}$. We follow exactly the proof of Theorem 2 except that at each timestep we need $C < \frac{\mu}{L}$ (which is less than 1 by Lemma 3) in order for the gradients to converge monotonically since:

$$\begin{aligned}
\|\nabla f(w^{(t+1)})\|^2 &\leq 2L(f(w^{(t+1)}) - f(w^*)) \text{ See Lemma 2} \\
&\leq 2LC(f(w^{(t)}) - f(w^*)) \\
&\leq \frac{LC}{\mu}\|\nabla f(w^{(t)})\|^2 \text{ As } f \text{ is } \mu\text{-PL.}
\end{aligned}$$

Hence in order for $\|\nabla f(w^{(t+1)})\|^2 < \|\nabla f(w^{(t)})\|^2$, we need $C < \frac{\mu}{L}$. Thus, we select our learning rate such that:

$$0 < 1 - \frac{\mu\eta}{\alpha_u} + \frac{\mu L\eta^2}{\alpha_l^2} + \frac{\mu L\eta^2}{4\alpha_u^2} < \frac{\mu}{L}.$$

Now, in order to have a solution to this system, we must ensure that the discriminant of the quadratic equation in η when considering the right hand side inequality is larger than zero. In particular we require:

$$\frac{\mu^2}{\alpha_u^2} - 4\left(1 - \frac{\mu}{L}\right)\left(\frac{\mu L}{\alpha_l^2} + \frac{\mu L}{4\alpha_u^2}\right) > 0$$
$$\implies \frac{\mu}{L} > \frac{4\alpha_u^2 + \alpha_l^2}{4\alpha_u^2 + 2\alpha_l^2},$$

which completes the proof. □

F Proof of Theorem 3

We repeat the theorem below for convenience.

Theorem. *Suppose $f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x)$ where $f_i : \mathbb{R}^d \to \mathbb{R}$ are non-negative, 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:

$$(a)\ \ \alpha_l \mathbf{I} \preccurlyeq \mathbf{J}_\phi \preccurlyeq \alpha_u \mathbf{I},$$
$$(b)\ \ |\partial_{i_1,\dots i_k}\phi_j^{-1}(x)| \leq \frac{k!\,\mu}{2\alpha_u d L}\ \ \forall x \in \mathbb{R}^d, i_1, \dots i_k \in [d], j \in [d], k \geq 2,$$

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

Proof. We follow the proof of Theorem 2. Namely, Lemma 4 implies that f is L-smooth and hence

$$f(w^{(t+1)}) - f(w^{(t)}) \leq \langle \nabla f(w^{(t)}), w^{(t+1)} - w^{(t)}\rangle + \frac{L}{2}\|w^{(t+1)} - w^{(t)}\|^2.$$

As before, we want to bound the two quantities on the right by $\|\nabla f(w^{(t)})\|^2$. Following the bounds from the proof of Theorem 2, provided $\eta^{(t)} < \frac{1}{2\sqrt{d}\|\nabla f_i(w^{(t)})\|}$, we have

$$\nabla f(w^{(t)})^T \sum_{k=2}^{\infty} \frac{1}{k!}\left[\sum_{i_1,i_2\dots i_k=1}^d (-\eta)^k \partial_{l_1,\dots l_k}\phi_j^{-1}(\phi(w^{(t)}))(\nabla f_{i_t}(w^{(t)})_{l_1}\dots \nabla f_{i_t}(w^{(t)})_{l_k})\right]$$
$$\leq \frac{\eta^{(t)}\mu}{2\alpha_u L}\|\nabla f(w^{(t)})\|\|\nabla f_{i_t}(w^{(t)})\|.$$

To remove the dependence of $\eta^{(t)}$ on i_t, we take $\eta^{(t)} < \frac{1}{2\sqrt{d}\max_i\|\nabla f_i(w^{(t)})\|}$. Since f is $\mu-$PL* and f_i is non-negative for all $i \in [n]$, $\|\nabla f_i(w^{(t)}\| \leq \sqrt{2Lf_i(w^{(t)})}$. Thus, we can take

$$\eta^{(t)} < \frac{1}{2\sqrt{2dLn}\sqrt{f(w^{(t)})}} \leq \frac{1}{2\sqrt{d}\max_i\|\nabla f_i(w^{(t)})\|}$$

This implies the following bounds:

$$\langle \nabla f(w^{(t)}), w^{(t+1)} - w^{(t)}\rangle \leq -\eta^{(t)}\nabla f(w^{(t)})^T \mathbf{J}_\phi^{-1}(w^{(t)})\nabla f_{i_t}(w^{(t)}) + \left(\frac{\eta^{(t)}\mu}{2\alpha_u L}\right)\|\nabla f(w^{(t)})\|\|\nabla f_{i_t}(w^{(t)})\|,$$

$$\|w^{(t+1)} - w^{(t)}\|^2 \leq \left(\frac{{\eta^{(t)}}^2}{\alpha_l^2} + \frac{{\eta^{(t)}}^2}{4\alpha_u^2}\right)\|\nabla f_{i_t}(w^{(t)})\|^2.$$

Putting the bounds together we obtain:

$$
\begin{aligned}
f(w^{(t+1)}) - f(w^{(t)}) &\leq -\eta^{(t)} \nabla f(w^{(t)})^T \mathbf{J}_\phi^{-1}(w^{(t)}) \nabla f_{i_t}(w^{(t)}) + \left(\frac{\eta^{(t)}\mu}{2\alpha_u L}\right) \|\nabla f(w^{(t)})\| \|\nabla f_{i_t}(w^{(t)})\| \\
&\quad + \left(\frac{{\eta^{(t)}}^2}{\alpha_l^2} + \frac{{\eta^{(t)}}^2}{4\alpha_u^2}\right) \|\nabla f_{i_t}(w^{(t)})\|^2 \\
&\leq -\eta^{(t)} \nabla f(w^{(t)})^T \mathbf{J}_\phi^{-1}(w^{(t)}) \nabla f_{i_t}(w^{(t)}) + \left(\frac{\eta^{(t)}\mu}{2\alpha_u L}\right) 2L\sqrt{f(w^{(t)}) f_{i_t}(w^{(t)})} \\
&\quad + \left(\frac{{\eta^{(t)}}^2}{\alpha_l^2} + \frac{{\eta^{(t)}}^2}{4\alpha_u^2}\right) \|\nabla f_{i_t}(w^{(t)})\|^2
\end{aligned}
$$

Now taking expectation over i_t, we obtain

$$
\begin{aligned}
\mathbb{E}[f(w^{(t+1)})] - f(w^{(t)}) &\leq \left(-\frac{\eta^{(t)}}{\alpha_u}\right) \|\nabla f(w^{(t)})\|^2 + \left(\frac{\eta^{(t)}\mu}{\alpha_u}\right) \sqrt{f(w^{(t)})} \mathbb{E}\left[\sqrt{f_{i_t}(w^{(t)})}\right] \\
&\quad + \left(\frac{L{\eta^{(t)}}^2}{2\alpha_l^2} + \frac{L{\eta^{(t)}}^2}{8\alpha_u^2}\right) \mathbb{E}[\|\nabla f_{i_t}(w^{(t)})\|^2] \\
&\leq \left(-\frac{\eta^{(t)}}{\alpha_u}\right) \|\nabla f(w^{(t)})\|^2 + \left(\frac{\eta^{(t)}\mu}{\alpha_u}\right) f(w^{(t)}) \\
&\quad + \left(\frac{L{\eta^{(t)}}^2}{2\alpha_l^2} + \frac{L{\eta^{(t)}}^2}{8\alpha_u^2}\right) \mathbb{E}[\|\nabla f_{i_t}(w^{(t)})\|^2] \\
&\leq \left(-\frac{2\mu\eta^{(t)}}{\alpha_u}\right) f(w^{(t)}) + \left(\frac{\eta^{(t)}\mu}{\alpha_u}\right) f(w^{(t)}) \\
&\quad + \left(\frac{L{\eta^{(t)}}^2}{2\alpha_l^2} + \frac{L{\eta^{(t)}}^2}{8\alpha_u^2}\right) \mathbb{E}[2L(f_{i_t}(w^{(t)}) - f_{i_t}(w^*))] \\
&\leq \left(-\frac{\mu\eta^{(t)}}{\alpha_u} + \frac{L^2{\eta^{(t)}}^2}{\alpha_l^2} + \frac{L^2{\eta^{(t)}}^2}{4\alpha_u^2}\right) (f(w^{(t)})).
\end{aligned}
$$

where the second inequality follows from Jensen's inequality and the third inequality follows from Lemma 2. Hence, we have:

$$
\mathbb{E}[f(w^{(t+1)})] \leq \left(1 - \frac{\mu\eta^{(t)}}{\alpha_u} + \frac{L^2{\eta^{(t)}}^2}{\alpha_l^2} + \frac{L^2{\eta^{(t)}}^2}{4\alpha_u^2}\right) (f(w^{(t)})).
$$

Now let $C = \left(-\frac{\mu\eta^{(t)}}{\alpha_u} + \frac{L^2{\eta^{(t)}}^2}{\alpha_l^2} + \frac{L^2{\eta^{(t)}}^2}{4\alpha_u^2}\right)$. Then taking expectation with respect to $i_t, i_{t-1}, \ldots i_1$, yields

$$
\begin{aligned}
\mathbb{E}_{i_t,\ldots,i_1}[f(w^{(t+1)})] &\leq (1+C)(\mathbb{E}_{i_t,\ldots,i_1}[f(w^{(t)})] \\
&= (1+C)(\mathbb{E}_{i_{t-1},\ldots,i_1}[\mathbb{E}_{i_t|i_{t-1},\ldots i_1}[f(w^{(t)})]]) \\
&= (1+C)(\mathbb{E}_{i_{t-1},\ldots,i_1} f(w^{(t)})]).
\end{aligned}
$$

Hence, we can proceed inductively to conclude that

$$
\mathbb{E}_{i_t,\ldots,i_1}[f(w^{(t+1)})] \leq (1+C)^{t+1}(f(w^{(0)}))).
$$

Thus if $0 < 1 + C < 1$, we establish linear convergence. The left hand side is satisfied since $\mu < L$, and the right hand side is satisfied for $\eta^{(t)} < \frac{4\mu\alpha_l^2}{5L^2\alpha_u}$, which holds by the theorem's assumption, thereby completing the proof. □

G Proof of Theorem 4

We restate the theorem below.

Theorem. *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 $R = \frac{2\sqrt{2L}\sqrt{f(w^{(0)})}\alpha_u^2}{\alpha_l \mu}$. 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, $\|\phi(w^*) - \phi(w^{(\infty)})\| \leq 2R$.*

Proof. The proof follows from the proofs of Lemma 1, Theorem 1, and Theorem 4.2 from Liu et al. (2020). Namely, we will proceed by strong induction. Let $\kappa = \frac{L{\alpha_u}^2}{\mu{\alpha_l}^2}$. At timestep 0, we trivially have that $w^{(0)} \in \tilde{\mathcal{B}}$ and $f(w^{(0)}) \leq f(w^{(0)})$. At timestep t, we assume that $w^{(0)}, w^{(1)}, \ldots w^{(t)} \in \tilde{\mathcal{B}}$ and that $f(w^{(i)}) \leq (1 - \kappa^{-1}) f(w^{(i-1)})$ for $i \in [t]$. Then at timestep $t + 1$, from the proofs of Lemma 1 and Theorem 1, we have:

$$f(w^{(t+1)}) \leq (1 - \kappa^{-1}) f(w^{(t)})$$

Next, we need to show that $w^{(t+1)} \in \tilde{\mathcal{B}}$. We have that:

$$\begin{aligned}
\|\phi(w^{(t+1)}) - \phi(w^{(0)})\| &= \left\| \sum_{i=0}^{t} -\eta \nabla f(w^{(i)}) \right\| \\
&\leq \eta \sum_{i=0}^{t} \|\nabla f(w^{(i)})\| \quad \text{By the Triangle Inequality} \\
&\leq \eta \sqrt{2 \frac{L\alpha_u^2}{\alpha_l^2}} \sum_{i=0}^{t} \sqrt{f(w^{(t)}) - f(w^{(t+1)})} \qquad (7) \\
&\leq \eta \sqrt{2 \frac{L\alpha_u^2}{\alpha_l^2}} \sum_{i=0}^{t} \sqrt{f(w^{(t)})} \\
&\leq \eta \sqrt{2L} \frac{\alpha_u}{\alpha_l} \sum_{i=0}^{t} \sqrt{(1 - \kappa^{-1})^i} \sqrt{f(w^{(0)})} \\
&= \eta \sqrt{2Lf(w^{(0)})} \frac{\alpha_u}{\alpha_l} \sum_{i=0}^{t} (1 - \kappa^{-1})^{\frac{i}{2}} \\
&\leq \eta \sqrt{2Lf(w^{(0)})} \frac{\alpha_u}{\alpha_l} \frac{1}{1 - \sqrt{1 - \kappa^{-1}}} \\
&\leq \eta \sqrt{2Lf(w^{(0)})} \frac{\alpha_u}{\alpha_l} \frac{2}{\kappa^{-1}} \\
&= \frac{\alpha_l}{L} \sqrt{2Lf(w^{(0)})} \frac{\alpha_u}{\alpha_l} 2 \frac{\alpha_u L}{\alpha_l \mu} \\
&= \frac{2\sqrt{2L}\sqrt{f(w^{(0)})}\alpha_u^2}{\alpha_l \mu} = R
\end{aligned}$$

The identity in (7) follows from the proof of $f(w^{(t+1)}) \leq (1-\kappa^{-1})f(w^{(t)})$. Namely,

$$\begin{aligned} f(w^{(t+1)}) - f(w^{(t)}) &\leq -\frac{L}{2\alpha_u^2}\|-\eta\nabla f(w^{(t)})\|^2 \\ \implies \|\nabla f(w^{(t)})\| &\leq \sqrt{\frac{2\alpha_u^2}{L}}\sqrt{f(w^{(t)}) - f(w^{(t+1)})} \\ \implies \|\nabla f(w^{(t)})\| &\leq \eta\sqrt{\frac{2L\alpha_u^2}{\alpha_l^2}}\sqrt{f(w^{(t)}) - f(w^{(t+1)})} \end{aligned}$$

Hence we conclude that $w^{(t+1)} \in \tilde{\mathcal{B}}$ and so induction is complete. $\square$

In the case that $\phi^{(t)}$ is time-dependent, we establish a similar convergence result by assuming that $\left\|\sum_{i=1}^{\infty} \phi^{(i)}(w^{(i)}) - \phi^{(i-1)}(w^{(i)})\right\| = \delta < \infty$. Additionally if $\alpha_u^{(t)}$ has a uniform upper bound and $\alpha_l^{(t)}$ has a uniform lower bound, then:

$$\begin{aligned} \|\phi^{(t)}(w^{(t+1)}) - \phi^{(0)}(w^{(0)})\| &= \|\phi^{(t)}(w^{(t+1)}) - \phi^{(t)}(w^{(t)}) + \phi^{(t)}(w^{(t)}) - \phi^{(t-1)}(w^{(t)}) \\ &\quad + \phi^{(t-1)}(w^{(t)}) - \phi^{(t-1)}(w^{(t-1)}) + \dots \phi^{(0)}(w^{(1)}) - \phi^{(0)}(w^{(0)})\| \\ &\leq \left\|\sum_{i=0}^{t} \phi^{(i)}(w^{(i+1)}) - \phi^{(i)}(w^{(i)})\right\| + \left\|\sum_{i=1}^{t} \phi^{(i)}(w^{(i)}) - \phi^{(i-1)}(w^{(i)})\right\| \\ &\leq R + \delta \end{aligned}$$

Hence we would conclude that $\phi^{(t)}(w^{(t+1)}) \in \mathcal{B}(\phi^{(0)}(w^{(0)}), R+\delta)$.

H PROOF OF COROLLARY 1 AND COROLLARY 2

We repeat Corollary 1 below.

Corollary. *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 $\lim\limits_{t\to\infty} \frac{\alpha_l^{(t)}}{\alpha_u^{(t)}} \neq 0$, then Adagrad converges linearly for adaptive step size $\eta^{(t)} = \frac{\alpha_l^{(t)}}{L}$.*

Proof. By definition of $\mathcal{G}^{(t)}$, we have that:

$$(1)\ \ {\alpha_l^{(t)}}^2 = \min_{i\in[d]} \mathcal{G}_{i,i}^{(t)}$$

$$(2)\ \ {\alpha_u^{(t)}}^2 = \max_{i\in[d]} \mathcal{G}_{i,i}^{(t)}$$

From the proof of Theorem 1, using learning rate $\eta^{(t)} = \frac{\alpha_l^{(t)}}{L}$ at timestep t gives:

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

Let $\kappa^{(t)} = \frac{\mu {\alpha_l^{(t)}}^2}{L{\alpha_u^{(t)}}^2}$. Although we have that $(1-\kappa^{(t)}) < 1$ for all t, we need to ensure that $\prod_{i=0}^{\infty}(1-\kappa^{(i)}) = 0$ (otherwise we would not get convergence to a global minimum). Using the assumption that $\lim\limits_{t\to\infty} \frac{\alpha_l^{(t)}}{\alpha_u^{(t)}} \neq 0$, let $\lim\limits_{t\to\infty}(1-\kappa^{(t)}) = 1-c < 1$. Then using the definition of the limit, for $0 < \epsilon < c$, there exists N such that for $t > N$, $\left|\kappa^{(t)} - c\right| < \epsilon$. Hence, letting

$c^* = \min\left(c - \epsilon, \min_{t \in \{0,1,\dots N\}} \kappa^{(t)}\right)$, implies that $(1 - \kappa^{(t)}) < 1 - c^*$ for all timesteps t. Thus, we have that:

$$\prod_{i=0}^{\infty}(1 - \kappa^{(i)}) < \prod_{i=0}^{\infty}(1 - c^*) = 0$$

Thus, Adagrad converges linearly to a global minimum. □

We present Corollary 2 below.

Corollary 2. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be an L-smooth function that is μ-PL. Let ${\alpha_l^{(t)}}^2 = \min_{i \in [d]} \mathcal{G}_{i,i}^{(t)}$. Then Adagrad converges linearly for adaptive step size $\eta^{(t)} = \frac{\alpha_l^{(t)}}{L}$ or fixed step size $\eta = \frac{\alpha_l^{(0)}}{L}$ if $\frac{{\alpha_l^{(0)}}^2}{2L(f(w^{(0)}) - f(w^*))} > \frac{L}{\mu}$.*

Proof. By definition of $\mathcal{G}^{(t)}$, we have that:

$$(1)\ \ {\alpha_l^{(t)}}^2 = \min_{i \in [d]} \mathcal{G}_{i,i}^{(t)}$$
$$(2)\ \ {\alpha_u^{(t)}}^2 = \max_{i \in [d]} \mathcal{G}_{i,i}^{(t)}$$

In particular, we can choose $\alpha_l = \alpha_l^{(0)}$ uniformly. We need to now ensure that $\alpha_u^{(t)}$ does not diverge. We prove this by using strong induction to show that ${\alpha_u^{(t)}}^2 \leq S$ uniformly for some $S > 0$. The base case holds by Lemma 2 since we have:

$${\alpha_u^{(0)}}^2 \leq \|\nabla f(w^{(0)})\|^2 = S$$

Now assume that ${\alpha_u^{(i)}}^2 < S$ for $i \in \{0, 1, \dots t-1\}$. Then we have:

$$\begin{aligned}
{\alpha_u^{(t)}}^2 &\leq \sum_{i=0}^{t} \|\nabla f(w^{(i)})\|^2 \\
&\leq \sum_{i=0}^{t} 2L(f(w^{(i)}) - f(w^*)) \ \text{ by Lemma 2} \\
&\leq 2L(f(w^{(0)}) - f(w^*)) \sum_{i=0}^{t-1} \prod_{j=0}^{i} \left(1 - \frac{\mu {\alpha_l^{(j)}}^2}{L {\alpha_u^{(j)}}^2}\right) \\
&\leq 2L(f(w^{(0)}) - f(w^*)) \sum_{i=0}^{t-1} \prod_{j=0}^{i} \left(1 - \frac{\mu {\alpha_l^{(0)}}^2}{LS}\right) \\
&\leq 2L(f(w^{(0)}) - f(w^*)) \frac{1}{1 - 1 + \frac{\mu {\alpha_l^{(0)}}^2}{LS}} \\
&= 2L(f(w^{(0)}) - f(w^*)) \frac{LS}{\mu {\alpha_l^{(0)}}^2} < S \ \text{ by assumption}
\end{aligned}$$

Hence, by induction, $\alpha_u^{(t)}$ is bounded uniformly for all timesteps t.

□

I Proof of Corollary 3

We present the corollary below.

Corollary 3. *Suppose ψ is an α_l-strongly convex function and that $\nabla\psi$ is α_u-Lipschitz. Let $D_\psi(x,y) = \psi(x) - \psi(y) - \nabla\psi(y)^T(x-y)$ denote the Bregman divergence for $x, y \in \mathbb{R}^d$. If $f : \mathbb{R}^d \to \mathbb{R}$ is non-negative, L-smooth, and μ-PL* on $\tilde{\mathcal{B}} = \{x\ ;\ \nabla\psi(x) \in \mathcal{B}(\nabla\psi(w^{(0)}), R)\}$ with $R = \frac{2\sqrt{2L}\sqrt{f(w^{(0)})}\alpha_u^2}{\alpha_l \mu}$, then:*

(1) *There exists a global minimum $w^{(\infty)} \in \tilde{\mathcal{B}}$ such that $D_\psi(w^{(\infty)}, w^{(0)}) \leq \dfrac{R^2}{2\alpha_l}$.*

(2) *Mirror descent with potential ψ converges linearly to $w^{(\infty)}$ for $\eta = \dfrac{\alpha_l}{L}$.*

(3) *If $w^* = \underset{\{w\ ;\ f(w)=0\}}{\arg\min}\ D_\psi(w, w^{(0)})$, then $D(w^*, w^{(\infty)}) \leq \dfrac{\alpha_u R^2}{\alpha_l^3} + \dfrac{R^2}{\alpha_l}$.*

Proof. The proof of existence and linear convergence follow immediately from Theorem 4. All that remains is to show that $D_\psi(w^{(\infty)}, w^{(0)}) \leq \frac{R^2}{2\mu}$. As ψ is α_l-strongly convex, we have:

$$\psi(w^{(\infty)}) \leq \psi(w^{(0)}) + \langle \nabla\psi(w^{(0)}), w^{(\infty)} - w^{(0)} \rangle + \frac{1}{2\alpha_l}\|\nabla\psi(w^{(\infty)}) - \nabla\psi(w^{(0)})\|^2 \quad \text{By Lemma 5}$$

$$\implies D_\psi(w^{(\infty)}, w^{(0)}) \leq \frac{1}{2\alpha_l}\|\nabla\psi(w^{(\infty)}) - \nabla\psi(w^{(0)})\|^2 \leq \frac{R^2}{2\alpha_l}$$

Now let $w^* = \arg\min_{\{w\ ;\ f(w)=0\}} D_\psi(w, w^{(0)})$. Hence $D_\psi(w^*, w^{(0)}) < \frac{R^2}{2\alpha_l}$ by definition. Then we have:

$$\begin{aligned}
D_\psi(w^*, w^{(\infty)}) &\leq \frac{1}{2\alpha_l}\|\nabla\psi(w^*) - \nabla\psi(w^{(\infty)})\|^2 \\
&\leq \frac{1}{2\alpha_l}(2\|\nabla\psi(w^*) - \nabla\psi(w^{(0)})\|^2 + 2\|\nabla\psi(w^{(0)}) - \nabla\psi(w^{(\infty)})\|^2) \\
&\leq \frac{\alpha_u}{\alpha_l}\|w^* - w^{(0)}\|^2 + \frac{R^2}{\alpha_l} \\
&\leq \frac{\alpha_u}{\alpha_l}\frac{2}{\alpha_l}D_\psi(w^*, w^{(0)}) + \frac{R^2}{\alpha_l} \quad \text{By Definition 3} \\
&\leq \frac{\alpha_u R^2}{\alpha_l^3} + \frac{R^2}{\alpha_l}
\end{aligned}$$

□

J Experiments on Over-parameterized Neural Networks

Below, we present experiments in which we apply the learning rate given by Corollary 1 to over-parameterized neural networks. Since the main difficulty is estimating the parameter L in neural networks, we instead provide a crude approximation for L by setting $L^{(t)} = .99\frac{\|\nabla f(w^{(t)})\|^2}{2f(w^{(t)})}$. The intuition for this approximation comes from Lemma 2. While there are no guarantees that this approximation yields linear convergence according to our theory, Figure 2 suggests empirically that this approximation provides convergence. Moreover, this approximation allows us to compute our adaptive learning rate in practice.

Code for all experiments is available at:

https://anonymous.4open.science/r/cef30260-473d-4116-bda1-1debdcc4e00a/

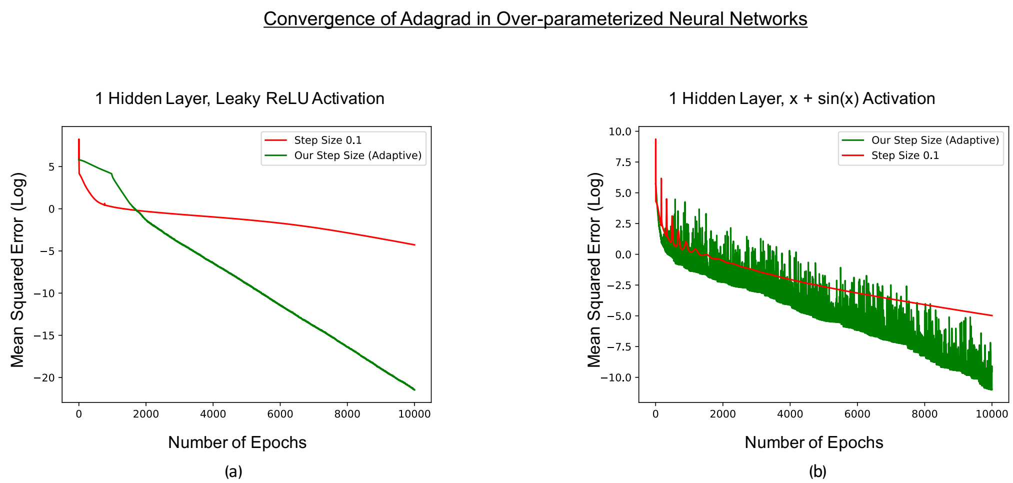

Figure 2: Using the adaptive rate provided by Corollary 1 with L approximated by $L^{(t)} = .99\frac{\|\nabla f(w^{(t)})\|^2}{2f(w^{(t)})}$ leads to convergence for Adagrad in the noisy linear regression setting (60 examples in 50 dimensions with uniform noise applied to the labels). (a) 1 hidden layer network with Leaky ReLU activation Xu et al. (2015) and 100 hidden units. (b) 1 hidden layer network with $x + \sin(x)$ activation with 100 hidden units. All networks were trained using a single Titan Xp, but can be trained on a laptop as well.