A COUNTER EXAMPLES

A.1 SIMPLE LINEAR REGRESSION

We first show that the bounded noise model (e.g., Mertikopoulos et al., 2020, Assumption 4) is violated for simple linear regression. Let $Z \in \mathbb{R}$ be a random variable such that $\mathbb{E}\left[Z^2\right] = 1$ and $\mathbb{E}\left[Z^4\right] = 2$. Moreover, let ϵ be an independent random variable with mean zero and variance 1. Finally, let $\theta^* \in \mathbb{R}$ and define $Y = Z\theta^* + \epsilon$. Consider the estimation problem of minimizing $F(\theta)$ where

$$F(\theta) = \frac{1}{2}\mathbb{E}\left[(Z\theta - Y)^2\right] = \frac{1}{2}(\theta - \theta^*)^2 + \frac{1}{2}. \tag{16}$$

Letting $X = (Y, Z)$, let $f(\theta, X) = 0.5(Z\theta - Y)^2$. Therefore, by a straightforward calculation, the variance of $\dot{f}(\theta, X)$ is $(\theta - \theta^*)^2 + 1$. Clearly, the variance scales with the error in the parameter, which violates the common bounded noise model assumption. As a result, any work that makes a globally bounded noise model assumption fails to apply to the simple linear regression problem.

A.2 LINEAR RNN FOR BINARY CLASSIFICATION

Consider observing one of two sequences $(1, 0, 0, 0)$ or $(0, 0, 0, 0)$ with equal probabilities, and suppose that each sequence corresponds to the label 1 or 0, respectively. For convenience, denote the first entry in the sequence by a Bernoulli random variable Z and let Y denote the label. Now consider an 1-dimensional linear recurrent neural network which reads each element of the sequence and uses a logistic output layer to predict either a label of one or zero. If we fix the initial memory state to zero, the input weight to 1, and the bias to zero, then the model predicts the probability of a 1 label as

$$\hat{y}(Z) = \frac{\exp(\theta^3 Z)}{1 + \exp(\theta^3 Z)}, \tag{17}$$

where θ denotes the recurrent weight.

If we use the binary cross entropy loss with ℓ^2 regularization, and let $X = (Y, Z)$, then

$$f(\theta, X) = -Y \log \hat{y}(Z) - (1 - Y)\log[1 - \hat{y}(Z)] + \frac{1}{2}\theta^2 \tag{18}$$

$$= -Y\left[\theta^3 Z - \log(1 + \exp(\theta^3 Z))\right] + (1 - Y)\log(1 + \exp(\theta^3 Z)) + \frac{1}{2}\theta^2 \tag{19}$$

$$= -\theta^3 ZY + \log(1 + \exp(\theta^3 Z)) + \frac{1}{2}\theta^2, \tag{20}$$

and

$$\dot{f}(\theta, X) = -3\theta^2 ZY + \frac{3\theta^2 Z \exp(\theta^3 Z)}{1 + \exp(\theta^3 Z)} + \theta. \tag{21}$$

Taking the expectations, we compute

$$F(\theta) = \frac{1}{2}\left[\log(2) + \log(1 + \exp(\theta^3)) - \theta^3 + \theta^2\right], \tag{22}$$

and

$$\dot{F}(\theta) = \frac{-3\theta^2}{2}\frac{1}{1 + \exp(\theta^3)} + \theta. \tag{23}$$

From this calculation alone, it is easy to see that as $\theta \to -\infty$, $\dot{F}(\theta) \propto -\theta^2$, which is not a Lipschitz continuous function. Moreover, computing the variance of $\dot{f}(\theta, X)$, we recover

$$\mathbb{E}\left[(\dot{f}(\theta, X) - \dot{F}(\theta))^2\right] = \frac{9\theta^4}{4(1 + \exp(\theta^3))^2}, \tag{24}$$

which does not satisfy a bounded variance assumption. Thus, any work that makes either a global Lipschitz bound on the gradient or a global noise model bound fails to apply to this simple recurrent neural network training problem. However, our results *do apply in this context*.

B TECHNICAL LEMMAS

Lemma 5. *Suppose* $\{M_k : k+1 \in \mathbb{N}\}$ *satisfy Properties 1 and 4. Then* $\forall C > 0$, $\exists K \in \mathbb{N}$ *such that* $\forall k \geq K$,

$$\lambda_{\min}(M_k) - \frac{C}{2}\lambda_{\max}(M_k)^{1+\alpha} \geq \frac{1}{2}\lambda_{\min}(M_k). \tag{25}$$

Proof. Fix $C > 0$. Rearranging the conclusion, we see that it is equivalent to prove that $\exists K \in \mathbb{N}$ such that $\forall k \geq K$, $1/C \geq \lambda_{\max}(M_k)^{\alpha}\kappa(M_k)$. This follows from Property 4. □

Lemma 6. *For any* $\theta \in \mathbb{R}^p$, $v \in \mathbb{R}$, $L > 0$ *and* $\alpha \in (0, 1]$,

$$\frac{L}{1+\alpha}v^{1+\alpha} - \left\|\dot{F}(\theta)\right\|_2 v \geq -\frac{\alpha}{1+\alpha}\left[\frac{\left\|\dot{F}(\theta)\right\|_2^{1+\alpha}}{L}\right]^{1/\alpha}. \tag{26}$$

Proof. If we minimize the left hand side of the inequality, we see that a minimum value occurs when $v^{\alpha} = \left\|\dot{F}(\theta)\right\|_2 / L \geq 0$. Solving for v and plugging this back into the left hand side, we conclude that the inequality holds. □

C GLOBAL CONVERGENCE ANALYSIS

We begin by first deriving a recursive relationship between the optimality gap at iteration $k+1$ and the optimality gap at iteration k on the events $\{\mathcal{B}_j(R)\}$ for arbitrary $R \geq 0$. Using this result, we then provide an analysis of the convergence of the objective function. Then, we turn our attention to the gradient function.

C.1 A RECURSIVE RELATIONSHIP

Lemma 7 (Lemma 2)**.** *Let* $\{M_k\}$ *satisfy Property 1. Suppose Assumptions 1 to 4 hold. Let* $\{\theta_k\}$ *satisfy (4). Then,* $\forall R \geq 0$,

$$\begin{aligned}\mathbb{E}\left[\left[F(\theta_{k+1}) - F_{l.b.}\right]\boldsymbol{1}\left[\mathcal{B}_{k+1}(R)\right]\middle|\mathcal{F}_k\right] &\leq \left[F(\theta_k) - F_{l.b.}\right]\boldsymbol{1}\left[\mathcal{B}_k(R)\right] \\ &- \lambda_{\min}(M_k)\left\|\dot{F}(\theta_k)\right\|_2^2 \boldsymbol{1}\left[\mathcal{B}_k(R)\right] + \frac{L_{R+1} + \partial F_R}{1+\alpha}\lambda_{\max}(M_k)^{1+\alpha}G_R,\end{aligned} \tag{27}$$

where $G_R = \sup_{\theta \in \overline{B(R)}} G(\theta) < \infty$ *with* $G(\theta)$*; and* $\partial F_R = \sup_{\theta \in \overline{B(R)}} \|\dot{F}(\theta)\|_2(1+\alpha) < \infty$.

Proof. Fix $R \geq 0$. For any $k+1 \in \mathbb{N}$, the definition of local Hölder continuity implies that L_{R+1} is well defined (see Definition 1). Therefore, Lemma 1 implies

$$\begin{aligned}&\left[F(\theta_{k+1}) - F_{l.b.}\right]\mathbf{1}\left[\mathcal{B}_{k+1}(R+1)\right] \\ &\quad\leq \left(\left[F(\theta_k) - F_{l.b.}\right] + \dot{F}(\theta_k)'(\theta_{k+1} - \theta_k) + \frac{L_{R+1}}{1+\alpha}\left\|\theta_{k+1} - \theta_k\right\|_2^{1+\alpha}\right)\mathbf{1}\left[\mathcal{B}_{k+1}(R+1)\right].\end{aligned} \tag{28}$$

Now, since $\overline{B(R)} \subset \overline{B(R+1)}$, it also holds true that

$$\begin{aligned}&\left[F(\theta_{k+1}) - F_{l.b.}\right]\mathbf{1}\left[\mathcal{B}_{k+1}(R)\right] \\ &\quad\leq \left(\left[F(\theta_k) - F_{l.b.}\right] + \dot{F}(\theta_k)'(\theta_{k+1} - \theta_k) + \frac{L_{R+1}}{1+\alpha}\left\|\theta_{k+1} - \theta_k\right\|_2^{1+\alpha}\right)\mathbf{1}\left[\mathcal{B}_{k+1}(R)\right].\end{aligned} \tag{29}$$

Our goal now is to replace $\mathcal{B}_{k+1}(R)$ on the right hand side by $\mathcal{B}_k(R)$. However, there is a technical difficulty which we must address. First, it follows from the preceding inequality that

$$\begin{aligned}
&[F(\theta_{k+1}) - F_{l.b.}]\mathbf{1}\left[\mathcal{B}_{k+1}(R)\right] \\
&\leq \left([F(\theta_k) - F_{l.b.}] + \dot{F}(\theta_k)'(\theta_{k+1} - \theta_k) + \frac{L_{R+1}}{1+\alpha}\left\|\theta_{k+1} - \theta_k\right\|_2^{1+\alpha}\right) \\
&\quad \times \left(\mathbf{1}\left[\mathcal{B}_{k+1}(R)\right] - \mathbf{1}\left[\mathcal{B}_k(R)\right]\right) \\
&\quad + \left([F(\theta_k) - F_{l.b.}] + \dot{F}(\theta_k)'(\theta_{k+1} - \theta_k) + \frac{L_{R+1}}{1+\alpha}\left\|\theta_{k+1} - \theta_k\right\|_2^{1+\alpha}\right)\mathbf{1}\left[\mathcal{B}_k(R)\right].
\end{aligned} \tag{30}$$

The first term on the right hand side of the inequality only contributes meaningfully if it is positive. Since $\mathbf{1}\left[\mathcal{B}_k(R)\right] \geq \mathbf{1}\left[\mathcal{B}_{k+1}(R)\right]$, then two statements hold: (i) $\mathbf{1}\left[\mathcal{B}_k(R)\right]\mathbf{1}\left[\mathcal{B}_{k+1}(R)\right] = \mathbf{1}\left[\mathcal{B}_{k+1}(R)\right]$; and (ii) the first term of the right hand side of (30) is positive if and only if

$$\left([F(\theta_k) - F_{l.b.}] + \dot{F}(\theta_k)'(\theta_{k+1} - \theta_k) + \frac{L_{R+1}}{1+\alpha}\left\|\theta_{k+1} - \theta_k\right\|_2^{1+\alpha}\right)\mathbf{1}\left[\mathcal{B}_k(R)\right] < 0. \tag{31}$$

By the choice of L_{R+1}, Assumption 1 and Lemma 1 imply that if (31) occurs, then $\|\theta_{k+1}\|_2 > R+1 \geq \|\theta_k\|_2 + 1$. By the reverse triangle inequality and (4), if (31) occurs, then $\|M_k\dot{f}(\theta_k, X_{k+1})\|_2 \geq 1$. Hence,

$$\begin{aligned}
&\left([F(\theta_k) - F_{l.b.}] + \dot{F}(\theta_k)'(\theta_{k+1} - \theta_k) + \frac{L_{R+1}}{1+\alpha}\left\|\theta_{k+1} - \theta_k\right\|_2^{1+\alpha}\right) \\
&\quad \times \left(\mathbf{1}\left[\mathcal{B}_{k+1}(R)\right] - \mathbf{1}\left[\mathcal{B}_k(R)\right]\right) \\
&\leq \left(-[F(\theta_k) - F_{l.b.}] - \dot{F}(\theta_k)'(\theta_{k+1} - \theta_k) - \frac{L_{R+1}}{1+\alpha}\left\|\theta_{k+1} - \theta_k\right\|_2^{1+\alpha}\right) \\
&\quad \times \left(\mathbf{1}\left[\mathcal{B}_k(R)\right] - \mathbf{1}\left[\mathcal{B}_{k+1}(R)\right]\right)\mathbf{1}\left[\mathcal{B}_k(R)\right]\mathbf{1}\left[\left\|M_k\dot{f}(\theta_k, X_{k+1})\right\|_2 \geq 1\right].
\end{aligned} \tag{32}$$

We now compute another coarse upper bound for this inequality. Note, by Assumption 1 and Cauchy-Schwarz,

$$\begin{aligned}
&\left(-[F(\theta_k) - F_{l.b.}] - \dot{F}(\theta_k)'(\theta_{k+1} - \theta_k) - \frac{L_{R+1}}{1+\alpha}\left\|\theta_{k+1} - \theta_k\right\|_2^{1+\alpha}\right) \\
&\quad \times \left(\mathbf{1}\left[\mathcal{B}_k(R)\right] - \mathbf{1}\left[\mathcal{B}_{k+1}(R)\right]\right)\mathbf{1}\left[\mathcal{B}_k(R)\right]\mathbf{1}\left[\left\|M_k\dot{f}(\theta_k, X_{k+1})\right\|_2 \geq 1\right] && (33) \\
&\leq \left\|\dot{F}(\theta_k)\right\|_2\left\|M_k\dot{f}(\theta_k, X_{k+1})\right\|_2\mathbf{1}\left[\mathcal{B}_k(R)\right]\mathbf{1}\left[\left\|M_k\dot{f}(\theta_k, X_{k+1})\right\|_2 \geq 1\right] && (34) \\
&\leq \left\|\dot{F}(\theta_k)\right\|_2\left\|M_k\dot{f}(\theta_k, X_{k+1})\right\|_2^{1+\alpha}\mathbf{1}\left[\mathcal{B}_k(R)\right] && (35) \\
&\leq \frac{\partial F_R}{1+\alpha}\left\|M_k\dot{f}(\theta_k, X_{k+1})\right\|_2^{1+\alpha}\mathbf{1}\left[\mathcal{B}_k(R)\right], && (36)
\end{aligned}$$

where $\partial F_R = \sup_{\theta\in\overline{B(R)}}\|\dot{F}(\theta)\|_2(1+\alpha) < \infty$ given that $\|\dot{F}(\theta)\|_2$ is a continuous function of θ.

Applying this inequality to (30), we conclude

$$\begin{aligned}
&[F(\theta_{k+1}) - F_{l.b.}]\mathbf{1}\left[\mathcal{B}_{k+1}(R)\right] \\
&\leq \left([F(\theta_k) - F_{l.b.}] - \dot{F}(\theta_k)'M_k\dot{f}(\theta_k, X_{k+1}) + \frac{L_{R+1} + \partial F_R}{1+\alpha}\left\|M_k\dot{f}(\theta_k, X_{k+1})\right\|_2^{1+\alpha}\right) \\
&\quad \times \mathbf{1}\left[\mathcal{B}_k(R)\right].
\end{aligned} \tag{37}$$

By Assumption 3,

$$\begin{aligned}&\mathbb{E}\left[\left[F(\theta_{k+1})-F_{l.b.}\right]\mathbf{1}\left[\mathcal{B}_{k+1}(R)\right]\middle|\mathcal{F}_k\right]\\&\leq\left(\left[F(\theta_k)-F_{l.b.}\right]-\dot{F}(\theta_k)'M_k\dot{F}(\theta_k)+\frac{L_{R+1}+\partial F_R}{1+\alpha}\mathbb{E}\left[\left\|M_k\dot{f}(\theta_k,X_{k+1})\right\|_2^{1+\alpha}\middle|\mathcal{F}_k\right]\right)\\&\quad\times\mathbf{1}\left[\mathcal{B}_k(R)\right].\end{aligned}\tag{38}$$

Using Property 1 and Assumption 4,

$$\begin{aligned}&\mathbb{E}\left[\left[F(\theta_{k+1})-F_{l.b.}\right]\mathbf{1}\left[\mathcal{B}_{k+1}(R)\right]\middle|\mathcal{F}_k\right]\\&\leq\left(\left[F(\theta_k)-F_{l.b.}\right]-\lambda_{\min}(M_k)\left\|\dot{F}(\theta_k)\right\|_2^2+\frac{L_{R+1}+\partial F_R}{1+\alpha}\lambda_{\max}(M_k)^{1+\alpha}G(\theta_k)\right)\mathbf{1}\left[\mathcal{B}_k(R)\right].\end{aligned}\tag{39}$$

By Assumption 4, G is upper semicontinuous and $\overline{B(R)}$ is compact, which implies that G_R is well defined and finite. The result follows. □

C.2 Objective Function Analysis

Corollary 1. *Let $\{\theta_k\}$ be defined as in (4) satisfying Properties 1 and 2. Suppose Assumptions 1 to 4 hold. Then, there exists a finite random variable $F_{\lim}$ such that on the event $\{\sup_k\|\theta_k\|_2<\infty\}$, $\lim_{k\to\infty}F(\theta_k)=F_{\lim}$ with probability one.*

Proof. By Lemma 2, for every $R\geq 0$,

$$\begin{aligned}&\mathbb{E}\left[\left[F(\theta_{k+1})-F_{l.b.}\right]\mathbf{1}\left[\mathcal{B}_{k+1}(R)\right]\middle|\mathcal{F}_k\right]\\&\leq\left[F(\theta_k)-F_{l.b.}\right]\mathbf{1}\left[\mathcal{B}_k(R)\right]+\frac{(L_{R+1}+\partial F_R)G_R}{1+\alpha}\lambda_{\max}(M_k)^{1+\alpha}.\end{aligned}\tag{40}$$

By Neveu & Speed (1975, Exercise II.4) (cf. Robbins and Siegmund (1971)) and Property 2, $\lim_{k\to\infty}[F(\theta_k)-F_{l.b.}]\mathbf{1}\left[\mathcal{B}_k(R)\right]$ converges to a finite random variable with probability one. Since $R\geq 0$ is arbitrary, we conclude that there exists a finite random variable $F_{\lim}$ such that $\{\sup_k\|\theta_k\|_2\leq R\}\subset\{\lim_k F(\theta_k)=F_{\lim}\}$ up to a measure zero set. Since the countable union of measure zero sets has measure zero,

$$\left\{\sup_k\|\theta_k\|_2<\infty\right\}=\bigcup_{R\in\mathbb{N}}\left\{\sup_k\|\theta_k\|_2\leq R\right\}\subset\left\{\lim_{k\to\infty}F(\theta_k)=F_{\lim}\right\},\tag{41}$$

up to a measure zero set. The result follows. □

C.3 Gradient Function Analysis

We now prove that the gradient norm evaluated at SGD's iterates must, repeatedly, get arbitrarily close to zero.

Lemma 8. *Let $\{\theta_k\}$ be defined as in (4) satisfying Properties 1 to 3. Suppose Assumptions 1 to 4 hold. Then, $\forall R\geq 0$ and for all $\delta>0$,*

$$\mathbb{P}\left[\left\|\dot{F}(\theta_k)\right\|_2^2\boldsymbol{1}\left[\mathcal{B}_k(R)\right]\leq\delta,\ i.o.\right]=1.\tag{42}$$

Proof. By Lemma 2,

$$\begin{aligned}\lambda_{\min}(M_k)\mathbb{E}\left[\left\|\dot{F}(\theta_k)\right\|_2^2\mathbf{1}\left[\mathcal{B}_k(R)\right]\right]&\leq\mathbb{E}\left[\left[F(\theta_k)-F_{l.b.}\right]\mathbf{1}\left[\mathcal{B}_k(R)\right]\right]\\&\quad-\mathbb{E}\left[\left[F(\theta_{k+1})-F_{l.b.}\right]\mathbf{1}\left[\mathcal{B}_{k+1}(R)\right]\right]+\frac{(L_{R+1}+\partial F_R)G_R}{1+\alpha}\lambda_{\max}(M_k)^{1+\alpha}.\end{aligned}\tag{43}$$

Taking the sum of this equation for all k from 0 to $j \in \mathbb{N}$, we have

$$\sum_{k=0}^{j} \lambda_{\min}(M_k)\mathbb{E}\left[\left\|\dot{F}(\theta_k)\right\|_2^2 \mathbf{1}\left[\mathcal{B}_k(R)\right]\right] \leq \left[F(\theta_0) - F_{l.b.}\right]\mathbf{1}\left[\mathcal{B}_0(R)\right] \\ - \mathbb{E}\left[\left[F(\theta_{j+1}) - F_{l.b.}\right]\mathbf{1}\left[\mathcal{B}_{j+1}(R)\right]\right] + \frac{(L_{R+1} + \partial F_R)G_R}{1+\alpha}\sum_{k=0}^{j}\lambda_{\max}(M_k)^{1+\alpha}. \tag{44}$$

By Assumption 1 and Property 2, the right hand side is bounded by

$$\left[F(\theta_0) - F_{l.b.}\right]\mathbf{1}\left[\mathcal{B}_0(R)\right] + \frac{(L_{R+1} + \partial F_R)G_R}{1+\alpha}S, \tag{45}$$

which is finite. Therefore, $\sum_{k=0}^{\infty}\lambda_{\min}(M_k)\mathbb{E}[\|\dot{F}(\theta_k)\|_2^2\mathbf{1}\left[\mathcal{B}_k(R)\right]]$ is finite. Furthermore, by Property 3, $\liminf_k \mathbb{E}[\|\dot{F}(\theta_k)\|_2^2\mathbf{1}\left[\mathcal{B}_k(R)\right]] = 0$.

Now, for any $\delta > 0$, Markov's inequality implies that for all $j+1 \in \mathbb{N}$,

$$\mathbb{P}\left[\bigcap_{k=j}^{\infty}\left\{\left\|\dot{F}(\theta_k)\right\|_2^2 \mathbf{1}\left[\mathcal{B}_k(R)\right] > \delta\right\}\right] \leq \frac{1}{\delta}\min_{j\leq k}\mathbb{E}\left[\left\|\dot{F}(\theta_k)\right\|_2^2 \mathbf{1}\left[\mathcal{B}_k(R)\right]\right], \tag{46}$$

where the right hand side is zero because $\liminf_k \mathbb{E}[\|\dot{F}(\theta_k)\|_2^2\mathbf{1}\left[\mathcal{B}_k(R)\right]] = 0$.

As the countable union of measure zero sets has measure zero, we conclude that for all $\delta > 0$,

$$\mathbb{P}\left[\left\|\dot{F}(\theta_k)\right\|_2^2 \mathbf{1}\left[\mathcal{B}_k(R)\right] \leq \delta, \ i.o.\right] = 1. \tag{47}$$

□

Unfortunately, Lemma 8 does not guarantee that the gradient norm will be captured within a region of zero. In order to prove this, we first show that it is not possible (i.e., a zero probability event) for the limit supremum and limit infimum of the gradients to be distinct.

Lemma 9. *Let $\{\theta_k\}$ be defined as in (4) satisfying Properties 1 and 2. Suppose Assumptions 1 to 4 hold. Then, $\forall R \geq 0$ and for all $\delta > 0$,*

$$\mathbb{P}\left[\left\|\dot{F}(\theta_{k+1})\right\|_2 \mathbf{1}\left[\mathcal{B}_{k+1}(R)\right] > \delta, \left\|\dot{F}(\theta_k)\right\|_2 \mathbf{1}\left[\mathcal{B}_k(R)\right] \leq \delta, \ i.o.\right] = 0. \tag{48}$$

Proof. Let $\gamma > 0$. Let L_R be as in Definition 1, and G_R be as in Lemma 2. Then, for $\delta > 0$,

$$\mathbb{P}\left[\left\|\dot{F}(\theta_{k+1})\right\|_2 \mathbf{1}\left[\mathcal{B}_{k+1}(R)\right]\mathbf{1}\left[\left\|\dot{F}(\theta_k)\right\|_2 \mathbf{1}\left[\mathcal{B}_k(R)\right] \leq \delta\right] > \delta + L_R\gamma^{\alpha}\right] \tag{49}$$

$$= \mathbb{P}\Bigg[\left(\left\|\dot{F}(\theta_{k+1})\right\|_2 - \left\|\dot{F}(\theta_k)\right\|_2 + \left\|\dot{F}(\theta_k)\right\|_2\right)\mathbf{1}\left[\mathcal{B}_{k+1}(R)\right] \tag{50}$$

$$\times \mathbf{1}\left[\left\|\dot{F}(\theta_k)\right\|_2 \mathbf{1}\left[\mathcal{B}_k(R)\right] \leq \delta\right] > \delta + L_R\gamma^{\alpha}\Bigg] \tag{51}$$

$$\leq \mathbb{P}\left[L_R\left\|\theta_{k+1} - \theta_k\right\|_2^{\alpha}\mathbf{1}\left[\mathcal{B}_{k+1}(R)\right]\mathbf{1}\left[\left\|\dot{F}(\theta_k)\right\|_2 \mathbf{1}\left[\mathcal{B}_k(R)\right] \leq \delta\right] > L_R\gamma^{\alpha}\right] \tag{52}$$

$$= \mathbb{P}\left[\left\|M_k\dot{f}(\theta_k, X_{k+1})\right\|_2 \mathbf{1}\left[\mathcal{B}_{k+1}(R)\right]\mathbf{1}\left[\left\|\dot{F}(\theta_k)\right\|_2 \mathbf{1}\left[\mathcal{B}_k(R)\right] \leq \delta\right] > \gamma\right] \tag{53}$$

$$\leq \mathbb{P}\left[\left\|M_k\dot{f}(\theta_k, X_{k+1})\right\|_2 \mathbf{1}\left[\mathcal{B}_k(R)\right] > \gamma\right] \tag{54}$$

$$\leq \frac{1}{\gamma^{1+\alpha}}\left\|M_k\right\|_2^{1+\alpha}\mathbb{E}\left[\mathbb{E}\left[\left\|\dot{f}(\theta_k, X_{k+1})\right\|_2^{1+\alpha}\middle|\mathcal{F}_k\right]\mathbf{1}\left[\mathcal{B}_k(R)\right]\right] \tag{55}$$

$$\leq \frac{1}{\gamma^{1+\alpha}}\left\|M_k\right\|_2^{1+\alpha}G_R. \tag{56}$$

By Property 2, the sum of the last expression over all $k+1 \in \mathbb{N}$ is finite. By the Borel-Cantelli lemma, for all $R \geq 0$, $\delta > 0$ and $\gamma > 0$,

$$\mathbb{P}\left[\left\|\dot{F}(\theta_{k+1})\right\|_2 \mathbf{1}\left[\mathcal{B}_{k+1}(R)\right] > \delta + L_R\gamma^\alpha, \left\|\dot{F}(\theta_k)\right\|_2 \mathbf{1}\left[\mathcal{B}_k(R)\right] \leq \delta, \ i.o.\right] = 0. \tag{57}$$

Since this holds for any $\gamma > 0$, it will hold for every value in a sequence $\gamma_n \downarrow 0$. Since the countable union of measure zero events has measure zero,

$$\mathbb{P}\left[\left\|\dot{F}(\theta_{k+1})\right\|_2 \mathbf{1}\left[\mathcal{B}_{k+1}(R)\right] > \delta, \left\|\dot{F}(\theta_k)\right\|_2 \mathbf{1}\left[\mathcal{B}_k(R)\right] \leq \delta, \ i.o.\right] = 0. \tag{58}$$

□

We now put together Lemmas 8 and 9 to show that, on the event $\{\sup_k \|\theta_k\|_2 < \infty\}$, $\|\dot{F}(\theta_k)\|_2 \to 0$ with probability one.

Corollary 2. *Let $\{\theta_k\}$ be defined as in (4) satisfying Properties 1 to 3. Suppose Assumptions 1 to 4 hold. Then, on the event $\{\sup_k \|\theta_k\|_2 < \infty\}$, $\lim_{k\to\infty} \|\dot{F}(\theta_k)\|_2 = 0$.*

Proof. For any $R \geq 0$ and $\delta > 0$, Lemma 8 implies

$$\begin{aligned}
&\mathbb{P}\left[\left\|\dot{F}(\theta_{k+1})\right\|_2 \mathbf{1}\left[\mathcal{B}_{k+1}(R)\right] > \delta, \ i.o.\right] \\
&\quad = \mathbb{P}\left[\left\{\left\|\dot{F}(\theta_{k+1})\right\|_2 \mathbf{1}\left[\mathcal{B}_{k+1}(R)\right] > \delta\right\} \cap \left\{\left\|\dot{F}(\theta_k)\right\|_2 \mathbf{1}\left[\mathcal{B}_k(R)\right] \leq \delta, \ i.o.\right\}\right].
\end{aligned} \tag{59}$$

We see that this latter event is exactly,

$$\mathbb{P}\left[\left\|\dot{F}(\theta_{k+1})\right\|_2 \mathbf{1}\left[\mathcal{B}_{k+1}(R)\right] > \delta, \left\|\dot{F}(\theta_k)\right\|_2 \mathbf{1}\left[\mathcal{B}_k(R)\right] \leq \delta, \ i.o.\right], \tag{60}$$

which, by Lemma 9, is zero with probability one. Therefore, $\mathbb{P}[\|\dot{F}(\theta_{k+1})\|_2 \mathbf{1}\left[\mathcal{B}_{k+1}(R)\right] > \delta, \ i.o.]$ is zero. Letting $\delta_n \downarrow 0$ and noting that the countable union of measure zero sets has measure zero, we conclude $\mathbb{P}[\|\dot{F}(\theta_{k+1})\|_2 \mathbf{1}\left[\mathcal{B}_{k+1}(R)\right] > 0, \ i.o.] = 0$.

Therefore, for all $R \geq 0$, $\{\sup_k \|\theta_k\|_2 \leq R\} \subset \{\lim_{k\to\infty} \|\dot{F}(\theta_k)\|_2 = 0\}$ up to a measure zero set. Since $\{\sup_k \|\theta_k\|_2 < \infty\} = \cup_{R\in\mathbb{N}}\{\sup_k \|\theta_k\|_2 \leq R\}$, the result follows. □

C.4 Capture Theorem

The final step in our proof is to study the event $\{\sup_k \|\theta_k\| < \infty\}$.

Theorem 4 (Theorem 1)**.** *Let $\{\theta_k\}$ be defined as in (4), and let $\{M_k\}$ satisfy Properties 1 and 2. If Assumption 4 holds, then either $\{\lim_{k\to\infty} \theta_k \text{ exists}\}$ or $\{\liminf_{k\to\infty} \|\theta_k\|_2 = \infty\}$ must occur.*

Proof. Let $\bar{\theta} \in \mathbb{R}^p$. Fix $R \geq 0$ and let $\gamma > 0$. Then,

$$\begin{aligned}
&\mathbb{P}\left[\left\|\theta_{k+1} - \bar{\theta}\right\|_2 \geq R + \gamma, \left\|\theta_k - \bar{\theta}\right\|_2 \leq R\right] \\
&\quad = \mathbb{P}\left[\left\|\theta_{k+1} - \bar{\theta}\right\|_2 \mathbf{1}\left[\left\|\theta_k - \bar{\theta}\right\|_2 \leq R\right] \geq R + \gamma\right] && (61)\\
&\quad = \mathbb{P}\left[\left(\left\|\theta_{k+1} - \bar{\theta}\right\|_2 - \left\|\theta_k - \bar{\theta}\right\|_2 + \left\|\theta_k - \bar{\theta}\right\|_2\right) \mathbf{1}\left[\left\|\theta_k - \bar{\theta}\right\|_2 \leq R\right] \geq R + \gamma\right] && (62)\\
&\quad \leq \mathbb{P}\left[\left(\left\|\theta_{k+1} - \bar{\theta}\right\|_2 - \left\|\theta_k - \bar{\theta}\right\|_2\right) \mathbf{1}\left[\left\|\theta_k - \bar{\theta}\right\|_2 \leq R\right] + R \geq R + \gamma\right] && (63)\\
&\quad \leq \mathbb{P}\left[\left\|\theta_{k+1} - \theta_k\right\|_2 \mathbf{1}\left[\left\|\theta_k - \bar{\theta}\right\|_2 \leq R\right] \geq \gamma\right] && (64)\\
&\quad \leq \mathbb{P}\left[\left\|M_k \dot{f}(\theta_k, X_{k+1})\right\|_2 \mathbf{1}\left[\left\|\theta_k - \bar{\theta}\right\|_2 \leq R\right] \geq \gamma\right] && (65)\\
&\quad \leq \frac{1}{\gamma^{1+\alpha}} \left\|M_k\right\|_2^{1+\alpha} \mathbb{E}\left[\mathbb{E}\left[\left\|\dot{f}(\theta_k, X_{k+1})\right\|_2^{1+\alpha} \middle| \mathcal{F}_k\right] \mathbf{1}\left[\left\|\theta_k - \bar{\theta}\right\|_2 \leq R\right]\right] && (66)\\
&\quad \leq \frac{1}{\gamma^2} \left\|M_k\right\|_2^2 \mathbb{E}\left[G(\theta_k) \mathbf{1}\left[\left\|\theta_k - \bar{\theta}\right\|_2 \leq R\right]\right] && (67)\\
&\quad \leq \frac{1}{\gamma^2} \left\|M_k\right\|_2^2 G_{R+\|\bar{\theta}\|_2}, && (68)
\end{aligned}$$

where $G_{R+\|\bar{\theta}\|_2} < \infty$ (as defined in Lemma 2), since G is upper semi-continuous. By Property 2, we see that the sum of the probabilities is finite. Together with the Borel-Cantelli lemma, $\mathbb{P}[\|\theta_{k+1} - \bar{\theta}\|_2 \geq R + \gamma,\ \|\theta_k - \bar{\theta}\|_2 \leq R\ i.o.] = 0$. Since $\gamma > 0$ is arbitrary, we can show that this statement holds for a countable sequence of $\gamma_n \downarrow 0$. Since R is arbitrary as well, we conclude that either $\|\theta_k - \bar{\theta}\|_2$ diverges or $\|\theta_k - \bar{\theta}\|_2$ must remain finite. Since this holds for any $\bar{\theta}$, we can choose three distinct values of $\bar{\theta}$ which are not colinear, and, by triangulation, the result holds when the iterates' norms remain finite. □

D Stability Analysis

We begin with a recursive relationship on the events $\{\tau_j > k\}$. We use this result to prove that the objective function converges to a finite limit on these events. Then, we use this result to conclude that the gradient function converges to zero on $\{\nu_j > k\}$. Finally, we study the combination of these events to establish that the two statements above hold with probability one.

D.1 A Recursive Relationship

Lemma 10 (Lemma 3)**.** *Let $\{M_k\}$ satisfy Property 1. Suppose Assumptions 1 to 4 hold. Let $\{\theta_k\}$ satisfy (4). Then, for any $j + 1 \in \mathbb{N}$ and $k > j$,*

$$\begin{aligned}\mathbb{E}\left[\left(F(\theta_{k+1}) - F_{l.b.}\right)\boldsymbol{1}\left[\tau_j > k\right]\middle|\mathcal{F}_k\right] &\leq \left(F(\theta_k) - F_{l.b.} - \dot{F}(\theta_k)' M_k \dot{F}(\theta_k)\right)\boldsymbol{1}\left[\tau_j > k-1\right] \\ &+ \frac{\lambda_{\max}(M_k)^{1+\alpha}}{1+\alpha}\left[\mathcal{L}_\epsilon(\theta_k)G(\theta_k) + \alpha\left[\frac{\left\|\dot{F}(\theta_k)\right\|_2^{1+\alpha}}{\mathcal{L}_\epsilon(\theta_k)}\right]^{1/\alpha}\right]\boldsymbol{1}\left[\tau_j > k-1\right].\end{aligned} \tag{69}$$

Proof. By the construction of τ_j, when $\tau_j > k$, then

$$F(\theta_{k+1}) - F_{l.b.} \leq F(\theta_k) - F_{l.b.} + \dot{F}(\theta_k)'(\theta_{k+1} - \theta_k) + \frac{\mathcal{L}_\epsilon(\theta_k)}{1+\alpha}\left\|\theta_{k+1} - \theta_k\right\|_2^{1+\alpha}. \tag{70}$$

Using this relationship and using $\mathbf{1}\left[\tau_j > k\right] = \mathbf{1}\left[\tau_j > k-1\right] - \mathbf{1}\left[\tau_j = k\right]$,

$$\begin{aligned}&\mathbb{E}\left[\left\{F(\theta_{k+1}) - F_{l.b.}\right\}\mathbf{1}\left[\tau_j > k\right]\middle|\mathcal{F}_k\right] \\ &\leq \mathbb{E}\left[\left\{F(\theta_k) - F_{l.b.} + \dot{F}(\theta_k)'(\theta_{k+1} - \theta_k) + \frac{\mathcal{L}_\epsilon(\theta_k)}{1+\alpha}\left\|\theta_{k+1} - \theta_k\right\|_2^{1+\alpha}\right\}\mathbf{1}\left[\tau_j > k-1\right]\middle|\mathcal{F}_k\right] \\ &- \mathbb{E}\left[\left\{F(\theta_k) - F_{l.b.} + \dot{F}(\theta_k)'(\theta_{k+1} - \theta_k) + \frac{\mathcal{L}_\epsilon(\theta_k)}{1+\alpha}\left\|\theta_{k+1} - \theta_k\right\|_2^{1+\alpha}\right\}\mathbf{1}\left[\tau_j = k\right]\middle|\mathcal{F}_k\right]\end{aligned} \tag{71}$$

For the first time on the right hand side, we can apply Assumptions 3 and 4, Property 1, and (4) to calculate

$$\begin{aligned}&\mathbb{E}\left[\left\{F(\theta_k) - F_{l.b.} + \dot{F}(\theta_k)'(\theta_{k+1} - \theta_k) + \frac{\mathcal{L}_\epsilon(\theta_k)}{1+\alpha}\left\|\theta_{k+1} - \theta_k\right\|_2^{1+\alpha}\right\}\mathbf{1}\left[\tau_j > k-1\right]\middle|\mathcal{F}_k\right] \\ &\leq \left\{F(\theta_k) - F_{l.b.} - \dot{F}(\theta_k)' M_k \dot{F}(\theta_k) + \frac{\lambda_{\max}(M_k)^{1+\alpha}}{1+\alpha}\mathcal{L}_\epsilon(\theta_k)G(\theta_k)\right\}\mathbf{1}\left[\tau_j > k-1\right].\end{aligned} \tag{72}$$

For the second term on the right hand side of (71), we first note $\mathbf{1}\left[\tau_j = k\right] \leq \mathbf{1}\left[\tau_j > k-1\right]$ which implies $\mathbf{1}\left[\tau_j = k\right] = \mathbf{1}\left[\tau_j = k\right]\mathbf{1}\left[\tau_j > k-1\right]$. Second, the Cauchy-Schwarz inequality and

Lemma 6 imply

$$
\begin{aligned}
&- F(\theta_k)' M_k \dot{f}(\theta_k, X_{k+1}) + \frac{\mathcal{L}_\epsilon(\theta_k)}{1+\alpha} \left\| M_k f(\theta_k, X_{k+1}) \right\|_2^{1+\alpha} \\
&\geq - \left\| F(\theta_k) \right\|_2 \left\| M_k \dot{f}(\theta_k, X_{k+1}) \right\|_2 + \frac{\mathcal{L}_\epsilon(\theta_k)}{1+\alpha} \left\| M_k f(\theta_k, X_{k+1}) \right\|_2^{1+\alpha} \qquad (73) \\
&\geq - \frac{\alpha}{1+\alpha} \left[\frac{\left\| \dot{F}(\theta_k) \right\|_2^{1+\alpha}}{\mathcal{L}_\epsilon(\theta_k)} \right]^{1/\alpha} \qquad (74)
\end{aligned}
$$

Hence, using (4),

$$
\begin{aligned}
&- \left[F(\theta_k) - F_{l.b.} \right] - \dot{F}(\theta_k)' (\theta_{k+1} - \theta_k) - \frac{\mathcal{L}_\epsilon(\theta_k)}{1+\alpha} \left\| \theta_{k+1} - \theta_k \right\|_2^{1+\alpha} \\
&\leq - \left[F(\theta_k) - F_{l.b.} \right] + \frac{\alpha}{1+\alpha} \left[\frac{\left\| \dot{F}(\theta_k) \right\|_2^{1+\alpha}}{\mathcal{L}_\epsilon(\theta_k)} \right]^{1/\alpha} \qquad (75)
\end{aligned}
$$

With the help of Theorem 5, we conclude

$$
\begin{aligned}
&- \mathbb{E} \left[\left\{ F(\theta_k) - F_{l.b.} + \dot{F}(\theta_k)' (\theta_{k+1} - \theta_k) + \frac{\mathcal{L}_\epsilon(\theta_k)}{1+\alpha} \left\| \theta_{k+1} - \theta_k \right\|_2^{1+\alpha} \right\} \mathbf{1} \left[\tau_j = k \right] \middle| \mathcal{F}_k \right] \\
&\leq \left\{ - \left[F(\theta_k) - F_{l.b.} \right] + \frac{\alpha}{1+\alpha} \left[\frac{\left\| \dot{F}(\theta_k) \right\|_2^{1+\alpha}}{\mathcal{L}_\epsilon(\theta_k)} \right]^{1/\alpha} \right\} \mathbb{P} \left[\tau_j = k \middle| \mathcal{F}_k \right] \mathbf{1} \left[\tau_j > k-1 \right] \qquad (76) \\
&\leq \frac{\alpha \lambda_{\max}(M_k)^{1+\alpha}}{1+\alpha} \left[\frac{\left\| \dot{F}(\theta_k) \right\|_2^{1+\alpha}}{\mathcal{L}_\epsilon(\theta_k)} \right]^{1/\alpha} \mathbf{1} \left[\tau_j > k-1 \right]. \qquad (77)
\end{aligned}
$$

By putting the bounds on the first and second term together, the result follows. □

By applying Assumption 5 to Lemma 3, we have the following simplified form.

Lemma 11 (Lemma 4)**.** *If Assumptions 1 to 5, and Properties 1 and 4 hold, and $\{\theta_k\}$ satisfy (4), then there exists a $K \in \mathbb{N}$ such that for any $j+1 \in \mathbb{N}$ and any $k \geq \min\{K, j+1\}$,*

$$
\begin{aligned}
&\mathbb{E} \left[(F(\theta_{k+1}) - F_{l.b.}) \mathbf{1} \left[\tau_j > k \right] \middle| \mathcal{F}_k \right] \\
&\leq \left(1 + \lambda_{\max}(M_k)^{1+\alpha} \frac{C_2}{1+\alpha} \right) (F(\theta_k) - F_{l.b.}) \mathbf{1} \left[\tau_j > k-1 \right] \qquad (78) \\
&- \frac{1}{2} \lambda_{\min}(M_k) \left\| \dot{F}(\theta_k) \right\|_2^2 \mathbf{1} \left[\tau_j > k-1 \right] + \lambda_{\max}(M_k)^{1+\alpha} \frac{C_1}{1+\alpha}.
\end{aligned}
$$

Proof. The result follows by first using Assumption 5 in Lemma 3. Then, collecting similar terms, we apply Lemma 5 to find K. □

D.2 Objective Function Analysis

With this recursive formula, we now have the first result.

Corollary 3. *If Assumptions 1 to 5 and Properties 1, 2 and 4 hold, and $\{\theta_k\}$ satisfy (4), then $\lim_{k\to\infty} F(\theta_k)$ exists and is finite on $\cup_{j=0}^{\infty} \{\tau_j = \infty\}$.*

Proof. By Lemma 4 and Robbins & Siegmund (1971); Neveu & Speed (1975, Exercise II.4), the limit as k goes to infinity of $(F(\theta_k) - F_{l.b.})\mathbf{1}[\tau_j > k]$ exists with probability one and is integrable. Therefore, on the event $\{\tau_j = \infty\}$, the limit of $F(\theta_k) - F_{l.b.}$ exists and is integrable. As a result, the limit of $F(\theta_k) - F_{l.b.}$ exists and is finite on $\cup_{j=0}^{\infty}\{\tau_j = \infty\}$. □

Additionally, we can state the following useful result.

Lemma 12. *If Assumptions 1 to 5, and Properties 1, 2 and 4 hold, and $\{\theta_k\}$ satisfy (4), then $\exists K \in \mathbb{N}$ such that for any $j > K$, $\exists N_j > 0$ for which*

$$\sup_{k>j} \mathbb{E}\left[(F(\theta_k) - F_{l.b.})\boldsymbol{1}[\tau_j > k]\right] \leq N_j. \tag{79}$$

Proof. By Lemma 4, Property 2, and any $k \geq j$,

$$\begin{aligned}
&\mathbb{E}\left[(F(\theta_{k+1}) - F_{l.b.})\mathbf{1}[\tau_j > k]\right] + \frac{C_1}{1+\alpha}\sum_{\ell=k+1}^{\infty}\lambda_{\max}(M_\ell)^{1+\alpha} \\
&\leq \exp\left(\frac{C_2}{1+\alpha}\lambda_{\max}(M_k)^{1+\alpha}\right)\left[\mathbb{E}\left[(F(\theta_k) - F_{l.b.})\mathbf{1}[\tau_j > k-1]\right] + \frac{C_1}{1+\alpha}\sum_{\ell=k}^{\infty}\lambda_{\max}(M_\ell)^{1+\alpha}\right].
\end{aligned} \tag{80}$$

Hence,

$$\begin{aligned}
&\mathbb{E}\left[(F(\theta_{k+1}) - F_{l.b.})\mathbf{1}[\tau_j > k]\right] + \frac{C_1}{1+\alpha}\sum_{\ell=k+1}^{\infty}\lambda_{\max}(M_\ell)^{1+\alpha} \\
&\leq \exp\left(\frac{C_2}{1+\alpha}\sum_{\ell=j}^{k}\lambda_{\max}(M_\ell)^{1+\alpha}\right)\left[\mathbb{E}\left[(F(\theta_j) - F_{l.b.})\right] + \frac{C_1}{1+\alpha}\sum_{\ell=j}^{\infty}\lambda_{\max}(M_\ell)^{1+\alpha}\right],
\end{aligned} \tag{81}$$

where we have used $\mathbf{1}[\tau_j > j-1] = 1$. By Property 2, the summation in the exponent is finite, which implies the result. □

D.3 Gradient Function Analysis

Lemma 13. *If Assumptions 1 to 5, and Properties 1 to 4 hold, and $\{\theta_k\}$ satisfy (4), then, for any $\delta > 0$,*

$$\mathbb{P}\left[\left\|\dot{F}(\theta_k)\right\|_2 \boldsymbol{1}[\tau_j > k-1] \leq \delta \ i.o.\right] = 1. \tag{82}$$

Proof. By Lemma 4,

$$\begin{aligned}
\frac{1}{2}\lambda_{\min}(M_k)\mathbb{E}\left[\left\|\dot{F}(\theta_k)\right\|_2^2 \mathbf{1}[\tau_j > k-1]\right] &\leq \mathbb{E}\left[(F(\theta_k) - F_{l.b.})\mathbf{1}[\tau_j > k-1]\right] \\
&- \mathbb{E}\left[(F(\theta_{k+1}) - F_{l.b.})\mathbf{1}[\tau_j > k]\right] + \frac{C_2}{1+\alpha}\lambda_{\max}(M_k)^{1+\alpha}\mathbb{E}\left[(F(\theta_k) - F_{l.b.})\mathbf{1}[\tau_j > k-1]\right] \\
&+ \frac{C_1}{1+\alpha}\lambda_{\max}(M_k)^{1+\alpha}.
\end{aligned} \tag{83}$$

By applying Lemma 12,

$$\begin{aligned}
\frac{1}{2}\lambda_{\min}(M_k)\mathbb{E}\left[\left\|\dot{F}(\theta_k)\right\|_2^2 \mathbf{1}[\tau_j > k-1]\right] &\leq \mathbb{E}\left[(F(\theta_k) - F_{l.b.})\mathbf{1}[\tau_j > k-1]\right] \\
&- \mathbb{E}\left[(F(\theta_{k+1}) - F_{l.b.})\mathbf{1}[\tau_j > k]\right] + \lambda_{\max}(M_k)^{1+\alpha}\left(\frac{C_2 N_j + C_1}{1+\alpha}\right).
\end{aligned} \tag{84}$$

By summing and using Assumption 1,

$$\begin{aligned}\frac{1}{2}\sum_{k=j}^{\infty}\lambda_{\min}(M_k)\mathbb{E}\left[\left\|\dot{F}(\theta_k)\right\|_2^2\mathbf{1}\left[\tau_j>k\right]\right]\\ \leq \mathbb{E}\left[F(\theta_j)-F_{l.b.}\right]+\frac{C_2N_j+C_1}{1+\alpha}\sum_{k=j}^{\infty}\lambda_{\max}(M_k)^{1+\alpha}.\end{aligned}\tag{85}$$

By Property 2, the right hand side is bounded. Now, by Property 3,

$$\liminf_{k\to\infty}\mathbb{E}\left[\left\|\dot{F}(\theta_k)\right\|_2^2\mathbf{1}\left[\tau_j>k\right]\right]=0.\tag{86}$$

Using Markov's inequality, for any $\ell\in\mathbb{N}$ and any $\delta>0$,

$$\mathbb{P}\left[\bigcap_{k=\ell}^{\infty}\left\{\left\|\dot{F}(\theta_k)\right\|_2\mathbf{1}\left[\tau_j>k\right]>\delta\right\}\right]\leq\frac{1}{\delta^2}\min_{k\geq\ell}\mathbb{E}\left[\left\|\dot{F}(\theta_k)\right\|_2^2\mathbf{1}\left[\tau_j>k\right]\right]=0.\tag{87}$$

As the countable union of sets of measure zero have measure zero, the result follows. □

Corollary 4. *If Assumptions 1 to 5, and Properties 1 to 4 hold, and $\{\theta_k\}$ satisfy (4), then, on $\cup_{j=0}^{\infty}\{\nu_j=\infty\}$, $\lim_{k\to\infty}\|\dot{F}(\theta_k)\|=0$.*

Proof. Let $\delta>0$ and $\gamma>0$. Note, $\mathbf{1}\left[\nu_j>k\right]\leq\mathbf{1}\left[\nu_j>k-1\right]$. Then,

$$\mathbb{P}\left[\left\|\dot{F}(\theta_{k+1})\right\|_2\mathbf{1}\left[\nu_j>k\right]>\delta+\mathcal{L}_\epsilon(\theta_k)^{\frac{1}{1+\alpha}}\gamma^{\frac{\alpha}{1+\alpha}},\left\|\dot{F}(\theta_k)\right\|_2\mathbf{1}\left[\nu_j>k-1\right]\leq\delta\middle|\mathcal{F}_k\right]\tag{88}$$

$$\leq\mathbb{P}\Bigg[\left(\left\|\dot{F}(\theta_{k+1})\right\|_2-\left\|\dot{F}(\theta_k)\right\|_2+\left\|\dot{F}(\theta_k)\right\|_2\right)\mathbf{1}\left[\nu_j>k\right]>\delta+\mathcal{L}_\epsilon(\theta_k)^{\frac{1}{1+\alpha}}\gamma^{\frac{\alpha}{1+\alpha}},$$
$$\left\|\dot{F}(\theta_k)\right\|_2\mathbf{1}\left[\nu_j>k-1\right]\leq\delta\middle|\mathcal{F}_k\Bigg]\tag{89}$$

$$\leq\mathbb{P}\Bigg[\mathcal{L}_\epsilon(\theta_k)\left\|\theta_{k+1}-\theta_k\right\|_2^{\alpha}\mathbf{1}\left[\nu_j>k\right]>\mathcal{L}_\epsilon(\theta_k)^{\frac{1}{1+\alpha}}\gamma^{\frac{\alpha}{1+\alpha}},\tag{90}$$

$$\left\|\dot{F}(\theta_k)\right\|_2\mathbf{1}\left[\nu_j>k-1\right]\leq\delta\middle|\mathcal{F}_k\Bigg]\tag{91}$$

$$\leq\mathbb{P}\left[\mathcal{L}_\epsilon(\theta_k)^{\frac{\alpha}{1+\alpha}}\left\|\theta_{k+1}-\theta_k\right\|_2^{\alpha}\mathbf{1}\left[\nu_j>k\right]>\gamma^{\frac{\alpha}{1+\alpha}},\left\|\dot{F}(\theta_k)\right\|_2\mathbf{1}\left[\nu_j>k-1\right]\leq\delta\middle|\mathcal{F}_k\right]\tag{92}$$

$$\leq\frac{1}{\gamma}\lambda_{\max}(M_k)^{1+\alpha}\mathcal{L}_\epsilon(\theta_k)G(\theta_k)\mathbf{1}\left[\nu_j>k-1\right]\mathbf{1}\left[\left\|\dot{F}(\theta_k)\right\|_2\mathbf{1}\left[\nu_j>k-1\right]\leq\delta\right]\tag{93}$$

$$\leq\frac{1}{\gamma}\lambda_{\max}(M_k)^{1+\alpha}\left(C_1+C_2(F(\theta_k)-F_{l.b.})\mathbf{1}\left[\nu_j>k-1\right]+C_3\delta^2\right),\tag{94}$$

where we have made use of Assumption 4, Markov's inequality, and Assumption 5 in the last two lines. Taking expectations, using $\sup_{k>j}\mathbb{E}\left[(F(\theta_k)-F_{l.b.})\mathbf{1}\left[\nu_j>k\right]\right]\leq N_j$ (Lemma 12), and applying Borel-Cantelli with Property 2, we conclude

$$\mathbb{P}\left[\left\|\dot{F}(\theta_{k+1})\right\|_2\mathbf{1}\left[\nu_j>k\right]>\delta+\mathcal{L}_\epsilon(\theta_k)^{\frac{1}{1+\alpha}}\gamma^{\frac{\alpha}{1+\alpha}},\left\|\dot{F}(\theta_k)\right\|_2\mathbf{1}\left[\nu_j>k-1\right]\leq\delta\ i.o.\right]=0.\tag{95}$$

Since γ is arbitrary, the result holds for a collection of $\gamma_n\downarrow 0$. As the countable union of measure zero events has measure zero,

$$\mathbb{P}\left[\left\|\dot{F}(\theta_{k+1})\right\|_2\mathbf{1}\left[\nu_j>k\right]>\delta,\left\|\dot{F}(\theta_k)\right\|_2\mathbf{1}\left[\nu_j>k-1\right]\leq\delta\ i.o.\right]=0.\tag{96}$$

Using this result in conjunction with Lemma 13, $\mathbb{P}\left[\lim_{k\to\infty}\left\|\dot{F}(\theta_{k+1})\right\|_2\mathbf{1}\left[\nu_j>k\right]=0\right]=1$. Hence, on the event $\{\nu_j=\infty\}$, $\lim_{k\to\infty}\|\dot{F}(\theta_k)\|=0$. Thus, on $\cup_{j=0}^{\infty}\{\nu_j=\infty\}$, $\lim_{k\to\infty}\|\dot{F}(\theta_k)\|=0$. □

D.4 STOPPING TIME ANALYSIS

By Lemma 1, $\{\tau_j = k\} \subset \{\nu_j = k\}$. Therefore, $\{\tau_j \leq k\} \subset \{\nu_j \leq k\}$, and $\{\tau_j > k\} \supset \{\nu_j > k\}$. Using these relationships, we can compute the probabilities of $\{\tau_j = k\}$ and $\{\nu_j = k\}$.

Theorem 5. *Let $\{\tau_j : j+1 \in \mathbb{N}\}$ be defined as in (12), and $\{\nu_j : j+1 \in \mathbb{N}\}$ be defined as in (13). If Assumptions 1, 2 and 4 and Property 1 hold, and $\{\theta_k\}$ satisfy (4), then, for any $j+1 \in \mathbb{N}$ and any $k+1 \in \mathbb{N}$,*

$$\mathbb{P}\left[\tau_j = k \middle| \mathcal{F}_k\right] \leq \mathbb{P}\left[\nu_j = k \middle| \mathcal{F}_k\right] \leq \begin{cases} 0 & k \leq j, \\ \lambda_{\max}(M_k)^{1+\alpha} & k > j. \end{cases} \tag{97}$$

Moreover, if Property 2 also holds, then $\mathbb{P}\left[\cup_{j=0}^{\infty}\{\tau_j = \infty\}\right] = \mathbb{P}\left[\cup_{j=0}^{\infty}\{\nu_j = \infty\}\right] = 1$.

Proof. The case of $k \leq j$ is trivial. So consider only $k > j$. By the construction of $L(\cdot,\cdot)$ and $\mathcal{L}_\epsilon(\cdot)$, $\omega \in \{L(\theta_k, \theta_{k+1}) > \mathcal{L}_\epsilon(\theta_k)\}$ implies $\omega \in \{\|\theta_{k+1} - \theta_k\|_2 > (G(\theta_k) \vee \epsilon)^{\frac{1}{1+\alpha}}\}$. Using (4), Markov's inequality, Property 1, we conclude

$$\mathbb{P}\left[\tau_j = k \middle| \mathcal{F}_k\right] \leq \mathbb{P}\left[\nu_j = k \middle| \mathcal{F}_k\right] \leq \mathbb{P}\left[\left\|M_k \dot{f}(\theta_k, X_{k+1})\right\|_2^{1+\alpha} > G(\theta_k) \vee \epsilon \middle| \mathcal{F}_k\right] \tag{98}$$

$$\leq \frac{\lambda_{\max}(M_k)^{1+\alpha}\mathbb{E}\left[\left\|\dot{f}(\theta_k, X_{k+1}\right\|_2^{1+\alpha} \middle| \mathcal{F}_k\right]}{G(\theta_k) \vee \epsilon}. \tag{99}$$

Applying Assumption 4 supplies the bound on $\mathbb{P}\left[\tau_j = k \middle| \mathcal{F}_k\right]$ and $\mathbb{P}\left[\nu_j = k \middle| \mathcal{F}_k\right]$. For the second part, note

$$\mathbb{P}\left[\tau_j = \infty\right] \geq \mathbb{P}\left[\nu_j = \infty\right] \geq 1 - \mathbb{P}\left[\nu_j < \infty\right] \geq 1 - \sum_{k=j+1}^{\infty} \lambda_{\max}(M_k)^{1+\alpha}. \tag{100}$$

Moreover, $\{\nu_j = \infty\} \subset \{\nu_{j+1} = \infty\}$. Therefore,

$$\mathbb{P}\left[\bigcup_{j=0}^{\infty}\{\tau_j = \infty\}\right] \geq \mathbb{P}\left[\bigcup_{j=0}^{\infty}\{\nu_j = \infty\}\right] = \lim_{j\to\infty} \mathbb{P}\left[\nu_j = \infty\right]. \tag{101}$$

Since $\lim_j \mathbb{P}[\nu_j = \infty] \geq 1 - \lim_j \sum_{k=j+1}^{\infty} \lambda_{\max}(M_k)^{1+\alpha}$, applying Property 2 supplies the final result. □