A Details for Counter Examples

A.1 Simple Linear Regression

We begin our exploration of assumptions with a rather simple problem. 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

$$F(\theta) = \frac{1}{2}\mathbb{E}\left[(Z\theta - Y)^2\right] = \frac{1}{2}(\theta - \theta^*)^2 + \frac{1}{2}. \tag{14}$$

Letting $X = (Y, Z)$, let $f(\theta, X) = 0.5(Z\theta - Y)^2$. Now, the variance of $\dot{f}(\theta, X)$ is

$$\begin{aligned}
&\mathbb{E}\left[((Z^2-1)(\theta-\theta^*) - Z\epsilon)^2\right] \\
&= \mathbb{E}\left[(Z^2-1)^2\right](\theta-\theta^*)^2 - 2(\theta-\theta^*)\mathbb{E}\left[(Z^2-1)Z\epsilon\right] + \mathbb{E}\left[Z^2\epsilon^2\right] &(15)\\
&= (\theta-\theta^*)^2 + 1. &(16)
\end{aligned}$$

Clearly, the variance scales with the error in the parameter, which violates the common bounded noise model assumption. In particular, as $|\theta| \to \infty$, the variance diverges.

On the other hand, the simple linear regression problem does satisfy our assumptions. In particular,

1. Assumptions 1 and 3 are easily verified.
2. Given that $\dot{F}$ is globally Lipschitz continuous, it is locally Lipschitz continuous. Therefore, Assumption 2 is satisfied.
3. From the variance calculation of $\dot{f}(\theta, X)$, we conclude

$$\mathbb{E}\left[\dot{f}(\theta, X)^2\right] = 2(\theta - \theta^*)^2 + 1, \tag{17}$$

which is a continuous function. Hence, Assumption 4 is satisfied.

A.2 Feed Forward Network for Binary Classification

We now prove Proposition 1. Consider the binary classification problem with label Y and feature Z where $(Y, Z) = (0, 0)$ with probability $1/2$ and $(Y, Z) = (1, 1)$ with probability $1/2$. We solve this classification problem using the network shown in Fig. 1 with σ linear and φ sigmoid. We will train this model using the binary cross entropy loss function. Letting $X = (Y, Z)$ and $\theta = (W_1, W_2, W_3, W_4)$,

$$f(\theta, X) = -Y\log(\hat{y}) - (1 - Y)\log(1 - \hat{y}) + \frac{1}{2}\sum_{i=1}^{4} W_i^2, \tag{18}$$

and

$$\hat{y} = \begin{cases} \frac{1}{2} & Z = 0 \\ \frac{1}{1+\exp(-W_4W_3W_2W_1)} & Z = 1. \end{cases} \tag{19}$$

From this, we compute,

$$F(\theta) = \frac{1}{2}\log(2) + \frac{1}{2}\log[1 + \exp(-W_4W_3W_2W_1)] + \frac{1}{2}\sum_{i=1}^{4} W_i^2. \tag{20}$$

Moreover,

$$\dot{f}(\theta, X) = \begin{cases} \begin{bmatrix} W_1 \\ W_2 \\ W_3 \\ W_4 \end{bmatrix} & (Y, Z) = (0, 0) \\ \frac{-1}{1+\exp(W_4W_3W_2W_1)} \begin{bmatrix} W_4W_3W_2 \\ W_4W_3W_1 \\ W_4W_2W_1 \\ W_3W_2W_1 \end{bmatrix} + \begin{bmatrix} W_1 \\ W_2 \\ W_3 \\ W_4 \end{bmatrix} & (Y, Z) = (1, 1), \end{cases} \tag{21}$$

and, consequently,

$$\dot{F}(\theta) = \frac{-1/2}{1+\exp(W_4W_3W_2W_1)} \begin{bmatrix} W_4W_3W_2 \\ W_4W_3W_1 \\ W_4W_2W_1 \\ W_3W_2W_1 \end{bmatrix} + \begin{bmatrix} W_1 \\ W_2 \\ W_3 \\ W_4 \end{bmatrix}. \tag{22}$$

Finally, letting $\ddot{F}(\theta) = \nabla^2 F(\psi)|_{\psi=\theta}$,

$$\begin{aligned} \ddot{F}(\theta) = &\frac{-0.5}{1+\exp(W_4W_3W_2W_1)} \begin{bmatrix} 0 & W_4W_3 & W_4W_2 & W_3W_2 \\ W_4W_3 & 0 & W_4W_1 & W_3W_1 \\ W_4W_2 & W_4W_1 & 0 & W_2W_1 \\ W_3W_2 & W_3W_1 & W_2W_1 & 0 \end{bmatrix} \\ &+ \frac{0.5\exp(W_4W_3W_2W_1)}{[1+\exp(W_4W_3W_2W_1)]^2} \begin{bmatrix} W_4W_3W_2 \\ W_4W_3W_1 \\ W_4W_2W_1 \\ W_3W_2W_1 \end{bmatrix} \begin{bmatrix} W_4W_3W_2 \\ W_4W_3W_1 \\ W_4W_2W_1 \\ W_3W_2W_1 \end{bmatrix}' + I_4, \end{aligned} \tag{23}$$

where I_4 is the 4×4 identity matrix.

We first establish that $\dot{F}(\theta)$ is not globally Lipschitz continuous. With $\theta = (1, -1, W_3, W_3)$ and $\phi = (1, -1, W_3, 0)$, it is enough to find a lower bound for the first component of $\dot{F}(\theta) - \dot{F}(\phi)$, denoted by $\dot{F}_1(\theta) - \dot{F}_1(\phi)$. To this end,

$$|\dot{F}_1(\theta) - \dot{F}_1(\phi)| = \frac{0.5W_3^2}{1+\exp(-W_3^2)} \geq \frac{1}{4}|W_3 - 0|^2. \tag{24}$$

Thus, $\dot{F}$ is not globally Lipschitz.

We now establish that F does not satisfy (L_0, L_1)-smoothness. That is, we show that there is no $L_0, L_1 \geq 0$ such that $\|\ddot{F}(\theta)\| \leq L_0\|\dot{F}(\theta)\| + L_1$, where the norms can be chosen arbitrarily owing to the equivalence of norms in finite-dimensional vector spaces. To see this, note that the Frobenius norm of $\ddot{F}(\theta)$ is lower bounded by the absolute value of the $[1, 1]$ entry. Using notation,

$$\frac{0.5\exp(W_4W_3W_2W_1)}{[1+\exp(W_4W_3W_2W_1)]^2}(W_4W_3W_2)^2 + 1 \leq \left\|\ddot{F}(\theta)\right\|_F. \tag{25}$$

Let $\theta = (0, W_4, W_4, W_4)$, then the lower bound is

$$\frac{1}{8}W_4^6 \leq \left\|\ddot{F}(\theta)\right\|_F. \tag{26}$$

Notice, for this same choice of θ, the l^1 norm of the gradient is bounded above by

$$\left\|\dot{F}(\theta)\right\|_1 \leq \frac{1}{4}|W_4|^3 + 3|W_4|. \tag{27}$$

For any choice of $L_0, L_1 > 0$, we conclude that there is a W_4 sufficiently large such that, for this parametrization of θ,

$$L_0\left\|\dot{F}(\theta)\right\| + L_1 \leq L_0[\frac{1}{4}|W_4|^3 + 3|W_4|] + L_1 < \frac{1}{8}W_4^6 \leq \left\|\ddot{F}(\theta)\right\|_F. \tag{28}$$

Thus, we see that no L_0 nor L_1 can exist that will satisfy the (L_0, L_1)-smooth assumption for all choices of θ.

To show that the variance is not bounded, we study the variance of the first component of $\dot{f}(\theta, X)$ which we denote by $\dot{f}_1(\theta, X)$. By direct calculation,

$$\mathbb{E}\left[(\dot{f}_1(\theta, X) - \dot{F}_1(\theta))^2\right] = \frac{1}{4}\frac{W_4^2W_3^2W_2^2}{[1+\exp(W_4W_3W_2W_1)]^2}. \tag{29}$$

We again consider $\theta = (1, -1, W_3, W_3)$, then the variance at this value of θ is

$$\frac{1}{4} \frac{W_3^4}{[1 + \exp(-W_3^2)]^2} \geq \frac{1}{16} W_3^4. \tag{30}$$

Therefore, as $W_3 \to \infty$, the variance goes to infinity. That is, the variance of the stochastic gradients is unbounded.

On the other hand, the problem does satisfy our assumptions. In particular,

1. Assumptions 1 and 3 are easily verified.
2. Given that $\dot{F}$ is continuously differentiable, then compactness and continuity of the derivative of $\dot{F}$ imply that it is locally Lipschitz continuous. Therefore, Assumption 2 is satisfied.
3. Given the computation of the variance for the first component, we have $\mathbb{E}[\dot{f}_1(\theta, X)^2]$ is

$$\frac{1}{4} \frac{W_4^2 W_3^2 W_2^2}{[1 + \exp(W_4 W_3 W_2 W_1)]^2} + \dot{F}_1(\theta)^2, \tag{31}$$

which is a continuous function. By repeating this argument for each component, we conclude that Assumption 4 is satisfied.

A.3 Recurrent Neural Network 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. Now consider Fig. 2 to be a 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 $H_0 = 0$ and $W_3 = 1$, then the model predicts the probability of a 1 label as

$$\hat{y}(Z_0, Z_1, Z_2, Z_3) = \frac{\exp(W_1^3 W_2 Z_0)}{1 + \exp(W_1^3 W_2 Z_0)}. \tag{32}$$

If we use the binary cross entropy loss with ℓ^2 regularization, and let $X = (Y, Z_0, Z_1, Z_2, Z_3)$ and $\theta = (W_1, W_2)$ then

$$f(\theta, X) = -Y \log \hat{y}(Z_0, Z_1, Z_2, Z_3) - (1 - Y) \log[1 - \hat{y}(Z_0, Z_1, Z_2, Z_3)] + \frac{1}{2}(W_1^2 + W_2^2) \tag{33}$$

$$= -Y \left[W_1^3 W_2 Z_0 - \log(1 + \exp(W_1^3 W_2 Z_0))\right] + (1 - Y) \log(1 + \exp(W_1^3 W_2 Z_0)) + \frac{1}{2}(W_1^2 + W_2^2) \tag{34}$$

$$= -W_1^3 W_2 Z_0 Y + \log(1 + \exp(W_1^3 W_2 Z_0)) + \frac{1}{2}(W_1^2 + W_2^2), \tag{35}$$

and

$$\dot{f}(\theta, X) = \begin{bmatrix} -3W_1^2 W_2 Z_0 Y + \frac{3W_1^2 W_2 Z_0 \exp(W_1^3 W_2 Z_0)}{1+\exp(W_1^3 W_2 Z_0)} + W_1 \\ -W_1^3 Z_0 Y + \frac{W_1^3 Z_0 \exp(W_1^3 W_2 Z_0)}{1+\exp(W_1^3 W_2 Z_0)} + W_2 \end{bmatrix} \tag{36}$$

Taking the expectations, we compute

$$F(\theta) = \frac{1}{2}\left[\log(2) + \log(1 + \exp(W_1^3 W_2)) - W_1^3 W_2 + W_1^2 + W_2^2\right], \tag{37}$$

and

$$\dot{F}(\theta) = \begin{bmatrix} \frac{-3W_1^2 W_2}{2} \frac{1}{1+\exp(W_1^3 W_2)} + W_1 \\ \frac{-W_1^3}{2} \frac{1}{1+\exp(W_1^3 W_2)} + W_2 \end{bmatrix}. \tag{38}$$

Taking another derivative and letting $\ddot{F}(\theta) = \nabla^2 F(\psi)|_{\psi=\theta}$,

$$\ddot{F}(\theta) = \begin{bmatrix} \frac{9W_1^4 W_2^2 \exp(W_1^3 W_2)}{2(1+\exp(W_1^3 W_2))^2} - \frac{3W_1 W_2}{1+\exp(W_1^3 W_2)} + 1 & \frac{3W_1^5 W_2 \exp(W_1^3 W_2)}{2(1+\exp(W_1^3 W_2))^2} - \frac{3W_1^2}{2}\frac{1}{1+\exp(W_1^3 W_2)} \\ \frac{3W_1^5 W_2 \exp(W_1^3 W_2)}{2(1+\exp(W_1^3 W_2))^2} - \frac{3W_1^2}{2}\frac{1}{1+\exp(W_1^3 W_2)} & \frac{W_1^6 \exp(W_1^3 W_2)}{2(1+\exp(W_1^3 W_2))^2} + 1 \end{bmatrix}. \tag{39}$$

We first establish that $\dot{F}$ is not globally Lipschitz continuous. Notice, if we set $W_2 = 1$, then the first and second component of $\dot{F}(\theta)$ are proportional to $-W_1^2$ and $-W_1^3$ respectively, which are not globally Lipschitz continuous functions.

We now show that F also does not satisfy (L_0, L_1)-smoothness. Notice that, using the bottom right entry of $\ddot{F}(\theta)$,

$$\frac{W_1^6 \exp(W_1^3 W_2)}{2(1+\exp(W_1^3 W_2))^2} < \left\| \ddot{F}(\theta) \right\|_F, \tag{40}$$

and

$$\left\| \dot{F}(\theta) \right\|_1 \leq \frac{3W_1^2 |W_2| + |W_1|^3}{2[1+\exp(W_1^3 W_2)]} + |W_1| + |W_2|. \tag{41}$$

If we choose $W_2 = 0$, then, for any $L_0, L_1 \geq 0$ there exists a $|W_1|$ sufficiently large such that

$$\frac{W_1^6}{8} < \left\| \ddot{F}(\theta) \right\|_F \not\leq L_0 \left\| \dot{F}(\theta) \right\|_1 + L_1 \leq L_0 \left(\frac{|W_1|^3}{4} + |W_1| \right) + L_1. \tag{42}$$

Hence, $F(\theta)$ is not (L_0, L_1)-smooth.

Moreover, computing the trace of the variance of $\dot{f}(\theta, X)$, we recover

$$\mathbb{E}\left[\left\| \dot{f}(\theta, X) - \dot{F}(\theta) \right\|_2^2 \right] = \left(\frac{3W_1^2 W_2}{2[1+\exp(W_1^3 W_2)]} \right)^2 + \left(\frac{W_1^3}{2[1+\exp(W_1^3 W_2)]} \right)^2, \tag{43}$$

which does not satisfy a bounded variance assumption (choose $W_2 = 0$ and let $W_1 \to \infty$). 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.

On the other hand, the problem does satisfy our assumptions. In particular,

1. Assumptions 1 and 3 are easily verified.
2. Given that $\dot{F}$ is continuously differentiable, then compactness and continuity of the derivative of $\dot{F}$ imply that it is locally Lipschitz continuous. Therefore, Assumption 2 is satisfied.
3. Moreover,

$$\mathbb{E}\left[\left\| \dot{f}(\theta, X) \right\|_2^2 \right] = \left(\frac{3W_1^2 W_2}{2[1+\exp(W_1^3 W_2)]} \right)^2 + \left(\frac{W_1^3}{2[1+\exp(W_1^3 W_2)]} \right)^2 + \left\| \dot{F}(\theta) \right\|_2^2, \tag{44}$$

which is a continuous function. Hence, Assumption 4 is satisfied.

A.4 Poisson Regression

Here, we consider the task of estimating a Poisson regression model for data $X = (Y, Z)$ where Y is a count response variable and Z is the covariate. To make this problem simpler, we will assume that both Y and Z are independent Poisson random variables with mean 1, which implies that the parameter in the model, $\theta^* = 0$. If we use a likelihood framework, then, up to a constant depending on Y,

$$f(\theta, X) = -YZ\theta + \exp(\theta Z), \tag{45}$$

and

$$\dot{f}(\theta, X) = -YZ + Z\exp(\theta Z). \tag{46}$$

From this, we compute

$$F(\theta) = -\theta + \exp(\exp(\theta) - 1), \tag{47}$$

$$\dot{F}(\theta) = -1 + \exp(\exp(\theta) + \theta - 1), \tag{48}$$

and, letting $\nabla^2 F(\psi)|_{\psi=\theta} = \ddot{F}(\theta)$,

$$\ddot{F}(\theta) = (\exp(\theta) + 1)\exp(\exp(\theta) + \theta - 1). \tag{49}$$

We begin by showing that $\dot{F}(\theta)$ is not globally Lipschitz continuous. To do so, for any $\theta > 0$, note

$$|\dot{F}(\theta) - \dot{F}(0)| = \exp(\exp(\theta) + \theta - 1) - 1 > \exp(\theta) - 1 \geq \theta + \theta^2/2. \tag{50}$$

Thus, for any $L > 0$ there exists a $\theta > 0$ such that $|\dot{F}(\theta) - \dot{F}(0)| > L|\theta|$.

We now show that $F(\theta)$ does not satisfy the (L_0, L_1)-smooth assumption. Note, for $\theta \geq 0$,

$$\exp(\exp(\theta) + 2\theta - 1) < \ddot{F}(\theta), \tag{51}$$

and

$$\dot{F}(\theta) < \exp(\exp(\theta) + \theta - 1). \tag{52}$$

It follows that for any $L_0, L_1 > 0$, there exists a $\theta > 0$ such that $L_0|\dot{F}(\theta)| + L_1 < \ddot{F}(\theta)$.

For the noise, we compute the second moment of $\dot{f}(\theta, X)$. That is,

$$\begin{aligned}
\mathbb{E}\left[\dot{f}(\theta, X)^2\right] &= \mathbb{E}\left[Y^2Z^2 - 2YZ^2\exp(\theta Z) + Z^2\exp(2\theta Z)\right] && (53)\\
&= 4 - 2\mathbb{E}\left[Z^2\exp(\theta Z)\right] + \mathbb{E}\left[Z^2\exp(2\theta Z)\right] && (54)\\
&= 4 - 2(\exp(\theta) + 1)\exp(\exp(\theta) + \theta - 1)\\
&\quad + (\exp(2\theta) + 1)\exp(\exp(2\theta) + 2\theta - 1). && (55)
\end{aligned}$$

It is clear from this calculation that the variance (computed by subtracting off $\dot{F}(\theta)^2$) will diverge as θ tends to infinity. To show that [Bottou et al., 2018, Assumption 4.3c] does not apply, it is enough to show that its generalization, [Khaled and Richtárik, 2020, Assumption 2] does not apply. To this end, we must show that there does not exists a $C_0, C_1, C_2 \geq 0$ such that, $\forall\theta$,

$$\mathbb{E}\left[\dot{f}(\theta, X)^2\right] \leq C_0 + C_1 F(\theta) + C_2|\dot{F}(\theta)|^2. \tag{56}$$

From our calculations, it is easy to verify that $F(\theta)$ and $\dot{F}(\theta)$ are dominated by $\exp(2\exp(\theta))$, and that the second moment of the stochastic gradient is bounded from below by $\exp(\exp(2\theta))$ for $\theta \geq \log(4)$. Hence, for any $C_0, C_1, C_2 \geq 0$, there exists θ sufficiently large such that

$$C_0 + C_1 F(\theta) + C_2|\dot{F}(\theta)|^2 \leq C_0 + (C_1 + C_2)\exp(2\exp(\theta)) < \exp(\exp(2\theta)) \leq \mathbb{E}\left[\dot{f}(\theta, X)^2\right]. \tag{57}$$

Thus, [Bottou et al., 2018, Assumption 4.3c] and [Khaled and Richtárik, 2020, Assumption 2] do not hold.

On the other hand, the problem does satisfy our assumptions. In particular,

1. Assumptions 1 and 3 are easily verified.
2. Given that $\ddot{F}$ is continuous, Assumption 2 is satisfied.
3. Moreover, we can use the calculated value $\mathbb{E}[\dot{f}(\theta, X)^2]$, which is a continuous function, as $G(\theta)$ to satisfy Assumption 4.

A.5 Noiseless Feed Forward Network for Binary Classification

Out of interest, we reconsider the second example but construct a different data distribution that produces noiseless stochastic gradient. Consider the binary classification problem with label Y and feature Z where $(Y, Z) = (0, -1)$ with probability $1/2$ and $(Y, Z) = (1, 1)$ with probability $1/2$. We solve this classification problem using the network shown in Fig. 1 with σ linear and φ sigmoid.

We will train this model using the binary cross entropy loss function. Letting $X = (Y, Z)$ and $\theta = (W_1, W_2, W_3, W_4)$,

$$f(\theta, X) = -Y \log(\hat{y}) - (1 - Y)\log(1 - \hat{y}) + \frac{1}{2}\sum_{i=1}^{4} W_i^2, \tag{58}$$

and

$$\hat{y} = \frac{1}{1 + \exp(-W_4 W_3 W_2 W_1 Z)}. \tag{59}$$

Moreover,

$$\dot{f}(\theta, X) = Z(\hat{y} - Y)\begin{bmatrix} W_4 W_3 W_2 \\ W_4 W_3 W_1 \\ W_4 W_2 W_1 \\ W_3 W_2 W_1 \end{bmatrix} + \begin{bmatrix} W_1 \\ W_2 \\ W_3 \\ W_4 \end{bmatrix}, \tag{60}$$

and, consequently,

$$\dot{F}(\theta) = \frac{-1}{1 + \exp(W_4 W_3 W_2 W_1)}\begin{bmatrix} W_4 W_3 W_2 \\ W_4 W_3 W_1 \\ W_4 W_2 W_1 \\ W_3 W_2 W_1 \end{bmatrix} + \begin{bmatrix} W_1 \\ W_2 \\ W_3 \\ W_4 \end{bmatrix}. \tag{61}$$

We first establish that $\dot{F}(\theta)$ is not globally Lipschitz continuous. With $\theta = (1, -1, W_3, W_3)$ and $\phi = (1, 0, 0, 0)$, it is enough to find a lower bound for the first component of $\dot{F}(\theta) - \dot{F}(\phi)$, denoted by $\dot{F}_1(\theta) - \dot{F}_1(\phi)$. To this end,

$$|\dot{F}_1(\theta) - \dot{F}_1(\phi)| = \frac{W_3^2}{1 + \exp(-W_3^2)} \geq \frac{1}{2}|W_3 - 0|^2. \tag{62}$$

Thus, $\dot{F}$ is not globally Lipschitz.

On the other hand, the problem does satisfy our assumptions. In particular,

1. Assumptions 1 and 3 are easily verified.
2. Given that $\dot{F}$ is continuously differentiable, then compactness and continuity of the derivative of $\dot{F}$ imply that it is locally Lipschitz continuous. Therefore, Assumption 2 is satisfied.
3. Moreover, $\dot{f}(\theta, Z) = \dot{F}(\theta)$—that is, there $\dot{f}(\theta, Z)$ has zero variance for the distribution that we have constructed. Therefore,

$$\mathbb{E}\left[\left\|\dot{f}(\theta, X)\right\|_2^2\right] = \left\|\dot{F}(\theta)\right\|_2^2, \tag{63}$$

which is a continuous function. Hence, Assumption 4 is satisfied.

B Technical Lemmas

Lemma 5 (Lemma 1)**.** *Suppose Assumptions 1 and 2 hold. Then, for any $\theta, \varphi \in \mathbb{R}^p$,*

$$F(\varphi) - F_{l.b.} \leq F(\theta) - F_{l.b.} + \dot{F}(\theta)'(\varphi - \theta) + \frac{L(\theta, \varphi)}{1 + \alpha}\left\|\varphi - \theta\right\|_2^{1+\alpha}. \tag{64}$$

Proof. By Taylor's theorem,

$$F(\varphi) - F_{l.b.} = F(\theta) - F_{l.b.} + \int_0^1 \dot{F}(\theta + t(\varphi - \theta))'(\varphi - \theta)dt. \tag{65}$$

Now, add and subtract $\dot{F}(\theta)$ to $\dot{F}(\theta + t(\varphi - \theta))$ in the integral, then apply Assumption 2. By Definition 1,

$$\left\|\dot{F}(\theta + t(\varphi - \theta)) - \dot{F}(\theta)\right\|_2 \leq L(\theta, \varphi)t^{\alpha}\left\|\theta - \varphi\right\|_2^{\alpha}. \tag{66}$$

We conclude,

$$F(\varphi) - F_{l.b.} \leq F(\theta) - F_{l.b.} + \dot{F}(\theta)'(\varphi - \theta) + L(\theta,\varphi)\left\|\varphi - \theta\right\|_2^{1+\alpha} \int_0^1 t^{\alpha} dt. \tag{67}$$

By computing the integral, the result follows. □

Lemma 6. *Suppose $\{M_k : k+1 \in \mathbb{N}\}$ satisfy Properties 1 and 4. Then $\forall C > 0$, $\exists K \in \mathbb{N}$ such that $\forall k \geq K$,*

$$\lambda_{\min}(M_k) - \frac{C}{2}\lambda_{\max}(M_k)^{1+\alpha} \geq \frac{1}{2}\lambda_{\min}(M_k). \tag{68}$$

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

$$\frac{L}{1+\alpha}v^{1+\alpha} - \left\|\dot{F}(\theta)\right\|_2 v \geq -\frac{\alpha}{1+\alpha}\left[\frac{\left\|\dot{F}(\theta)\right\|_2^{1+\alpha}}{L}\right]^{1/\alpha}. \tag{69}$$

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 $\{\mathcal{B}_j(R)\}$ as defined in (8) 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. Note, $B(\theta, r)$ is the open ball around θ of radius r.

C.1 A Recursive Relationship

Lemma 8 (Lemma 2)**.** *Let $\{M_k\}$ satisfy Property 1. Suppose Assumptions 1 to 4 hold. Let $\{\theta_k\}$ satisfy (5). Then, $\forall R \geq 0$,*

$$\begin{aligned}\mathbb{E}\left[\left[F(\theta_{k+1}) - F_{l.b.}\right]\mathbf{1}\left[\mathcal{B}_{k+1}(R)\right]\middle|\, \mathcal{F}_k\right] &\leq \left[F(\theta_k) - F_{l.b.}\right]\mathbf{1}\left[\mathcal{B}_k(R)\right] \\ &- \lambda_{\min}(M_k)\left\|\dot{F}(\theta_k)\right\|_2^2 \mathbf{1}\left[\mathcal{B}_k(R)\right] + \frac{L_{R+1} + \partial F_R}{1+\alpha}\lambda_{\max}(M_k)^{1+\alpha} G_R,\end{aligned} \tag{70}$$

where $G_R = \sup_{\theta \in \overline{B(0,R)}} G(\theta) < \infty$ with $G(\theta)$; and $\partial F_R = \sup_{\theta \in \overline{B(0,R)}} \|\dot{F}(\theta)\|_2 (1+\alpha) < \infty$.

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

$$\begin{aligned}&\left[F(\theta_{k+1}) - F_{l.b.}\right]\mathbf{1}\left[\mathcal{B}_{k+1}(R+1)\right] \\ &\leq \left(\left[F(\theta_k) - F_{l.b.}\right] + \dot{F}(\theta_k)'(\theta_{k+1} - \theta_k) + \frac{L_{R+1}}{1+\alpha}\left\|\theta_{k+1} - \theta_k\right\|_2^{1+\alpha}\right)\mathbf{1}\left[\mathcal{B}_{k+1}(R+1)\right].\end{aligned} \tag{71}$$

Now, since $\overline{B(0,R)} \subset \overline{B(0,R+1)}$, it also holds true that

$$\begin{aligned}&\left[F(\theta_{k+1}) - F_{l.b.}\right]\mathbf{1}\left[\mathcal{B}_{k+1}(R)\right] \\ &\leq \left(\left[F(\theta_k) - F_{l.b.}\right] + \dot{F}(\theta_k)'(\theta_{k+1} - \theta_k) + \frac{L_{R+1}}{1+\alpha}\left\|\theta_{k+1} - \theta_k\right\|_2^{1+\alpha}\right)\mathbf{1}\left[\mathcal{B}_{k+1}(R)\right].\end{aligned} \tag{72}$$

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

$$\begin{aligned}
&[F(\theta_{k+1}) - F_{l.b.}]\mathbf{1}\left[\mathcal{B}_{k+1}(R)\right] \\
&\leq \left([F(\theta_k) - F_{l.b.}] + \dot{F}(\theta_k)'(\theta_{k+1} - \theta_k) + \frac{L_{R+1}}{1+\alpha} \left\|\theta_{k+1} - \theta_k\right\|_2^{1+\alpha} \right) \\
&\quad \times \left(\mathbf{1}\left[\mathcal{B}_{k+1}(R)\right] - \mathbf{1}\left[\mathcal{B}_k(R)\right] \right) \\
&\quad + \left([F(\theta_k) - F_{l.b.}] + \dot{F}(\theta_k)'(\theta_{k+1} - \theta_k) + \frac{L_{R+1}}{1+\alpha} \left\|\theta_{k+1} - \theta_k\right\|_2^{1+\alpha} \right) \mathbf{1}\left[\mathcal{B}_k(R)\right].
\end{aligned} \tag{73}$$

The first term on the right hand side of the inequality only contributes meaningfully if it is positive. Since $\mathbf{1}[\mathcal{B}_k(R)] \geq \mathbf{1}[\mathcal{B}_{k+1}(R)]$, then two statements hold: (i) $\mathbf{1}[\mathcal{B}_k(R)]\,\mathbf{1}[\mathcal{B}_{k+1}(R)] = \mathbf{1}[\mathcal{B}_{k+1}(R)]$; and (ii) the first term of the right hand side of (73) is positive if and only if

$$\left([F(\theta_k) - F_{l.b.}] + \dot{F}(\theta_k)'(\theta_{k+1} - \theta_k) + \frac{L_{R+1}}{1+\alpha} \left\|\theta_{k+1} - \theta_k\right\|_2^{1+\alpha} \right) \mathbf{1}\left[\mathcal{B}_k(R)\right] < 0. \tag{74}$$

By the choice of L_{R+1}, Assumption 1 and Lemma 1 imply that if (74) occurs, then $\|\theta_{k+1}\|_2 > R + 1 \geq \|\theta_k\|_2 + 1$. By the reverse triangle inequality and (5), if (74) occurs, then $\|M_k \dot{f}(\theta_k, X_{k+1})\|_2 \geq 1$. Hence,

$$\begin{aligned}
&\left([F(\theta_k) - F_{l.b.}] + \dot{F}(\theta_k)'(\theta_{k+1} - \theta_k) + \frac{L_{R+1}}{1+\alpha} \left\|\theta_{k+1} - \theta_k\right\|_2^{1+\alpha} \right) \\
&\quad \times \left(\mathbf{1}\left[\mathcal{B}_{k+1}(R)\right] - \mathbf{1}\left[\mathcal{B}_k(R)\right] \right) \\
&\leq \left(-[F(\theta_k) - F_{l.b.}] - \dot{F}(\theta_k)'(\theta_{k+1} - \theta_k) - \frac{L_{R+1}}{1+\alpha} \left\|\theta_{k+1} - \theta_k\right\|_2^{1+\alpha} \right) \\
&\quad \times \left(\mathbf{1}\left[\mathcal{B}_k(R)\right] - \mathbf{1}\left[\mathcal{B}_{k+1}(R)\right] \right) \mathbf{1}\left[\mathcal{B}_k(R)\right] \mathbf{1}\left[\left\| M_k \dot{f}(\theta_k, X_{k+1}) \right\|_2 \geq 1 \right].
\end{aligned} \tag{75}$$

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

$$\begin{aligned}
&\left(-[F(\theta_k) - F_{l.b.}] - \dot{F}(\theta_k)'(\theta_{k+1} - \theta_k) - \frac{L_{R+1}}{1+\alpha} \left\|\theta_{k+1} - \theta_k\right\|_2^{1+\alpha} \right) \\
&\quad \times \left(\mathbf{1}\left[\mathcal{B}_k(R)\right] - \mathbf{1}\left[\mathcal{B}_{k+1}(R)\right] \right) \mathbf{1}\left[\mathcal{B}_k(R)\right] \mathbf{1}\left[\left\| M_k \dot{f}(\theta_k, X_{k+1}) \right\|_2 \geq 1 \right]
\end{aligned} \tag{76}$$

$$\leq \left\| \dot{F}(\theta_k) \right\|_2 \left\| M_k \dot{f}(\theta_k, X_{k+1}) \right\|_2 \mathbf{1}\left[\mathcal{B}_k(R)\right] \mathbf{1}\left[\left\| M_k \dot{f}(\theta_k, X_{k+1}) \right\|_2 \geq 1 \right] \tag{77}$$

$$\leq \left\| \dot{F}(\theta_k) \right\|_2 \left\| M_k \dot{f}(\theta_k, X_{k+1}) \right\|_2^{1+\alpha} \mathbf{1}\left[\mathcal{B}_k(R)\right] \tag{78}$$

$$\leq \frac{\partial F_R}{1+\alpha} \left\| M_k \dot{f}(\theta_k, X_{k+1}) \right\|_2^{1+\alpha} \mathbf{1}\left[\mathcal{B}_k(R)\right], \tag{79}$$

where $\partial F_R = \sup_{\theta \in \overline{B(0,R)}} \|\dot{F}(\theta)\|_2 (1+\alpha) < \infty$ given that $\|\dot{F}(\theta)\|_2$ is a continuous function of θ.

Applying this inequality to (73), we conclude

$$\begin{aligned}
&[F(\theta_{k+1}) - F_{l.b.}]\mathbf{1}\left[\mathcal{B}_{k+1}(R)\right] \\
&\leq \left([F(\theta_k) - F_{l.b.}] - \dot{F}(\theta_k)' M_k \dot{f}(\theta_k, X_{k+1}) + \frac{L_{R+1} + \partial F_R}{1+\alpha} \left\| M_k \dot{f}(\theta_k, X_{k+1}) \right\|_2^{1+\alpha} \right) \\
&\quad \times \mathbf{1}\left[\mathcal{B}_k(R)\right].
\end{aligned} \tag{80}$$

By Assumption 3,

$$\begin{aligned}
&\mathbb{E}\left[[F(\theta_{k+1}) - F_{l.b.}]\mathbf{1}\left[\mathcal{B}_{k+1}(R)\right] \middle| \mathcal{F}_k \right] \\
&\leq \left([F(\theta_k) - F_{l.b.}] - \dot{F}(\theta_k)' M_k \dot{F}(\theta_k) + \frac{L_{R+1} + \partial F_R}{1+\alpha} \mathbb{E}\left[\left\| M_k \dot{f}(\theta_k, X_{k+1}) \right\|_2^{1+\alpha} \middle| \mathcal{F}_k \right] \right) \\
&\quad \times \mathbf{1}\left[\mathcal{B}_k(R)\right].
\end{aligned} \tag{81}$$

Using Property 1 and Assumption 4,

$$\mathbb{E}\left[[F(\theta_{k+1}) - F_{l.b.}]\mathbf{1}\left[\mathcal{B}_{k+1}(R)\right] \middle| \mathcal{F}_k\right] \\ \leq \left([F(\theta_k) - F_{l.b.}] - \lambda_{\min}(M_k)\left\|\dot{F}(\theta_k)\right\|_2^2 + \frac{L_{R+1}+\partial F_R}{1+\alpha}\lambda_{\max}(M_k)^{1+\alpha}G(\theta_k)\right)\mathbf{1}\left[\mathcal{B}_k(R)\right]. \tag{82}$$

By Assumption 4, G is upper semicontinuous and $\overline{B(0,R)}$ 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 (5) 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 $R \geq 0$,

$$\mathbb{E}\left[[F(\theta_{k+1}) - F_{l.b.}]\mathbf{1}\left[\mathcal{B}_{k+1}(R)\right] \middle| \mathcal{F}_k\right] \\ \leq [F(\theta_k) - F_{l.b.}]\mathbf{1}\left[\mathcal{B}_k(R)\right] + \frac{(L_{R+1}+\partial F_R)G_R}{1+\alpha}\lambda_{\max}(M_k)^{1+\alpha}. \tag{83}$$

By Neveu and 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 the set $\{\sup_k \|\theta_k\|_2 \leq R\}$ is a subset of $\{\lim_k F(\theta_k) = F_{\lim}\}$ up to a measure zero set. Since the countable union of measure zero sets has measure zero,

$$\left\{\sup_k \|\theta_k\|_2 < \infty\right\} = \bigcup_{R\in\mathbb{N}} \left\{\sup_k \|\theta_k\|_2 \leq R\right\} \subset \left\{\lim_{k\to\infty} F(\theta_k) = F_{\lim}\right\}, \tag{84}$$

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 9. *Let $\{\theta_k\}$ be defined as in (5) satisfying Properties 1 to 3. Suppose Assumptions 1 to 4 hold. Then, $\forall R \geq 0$ and for all $\delta > 0$,*

$$\mathbb{P}\left[\left\|\dot{F}(\theta_k)\right\|_2^2 \mathbf{1}\left[\mathcal{B}_k(R)\right] \leq \delta,\ i.o.\right] = 1. \tag{85}$$

Proof. By Lemma 2,

$$\lambda_{\min}(M_k)\mathbb{E}\left[\left\|\dot{F}(\theta_k)\right\|_2^2 \mathbf{1}\left[\mathcal{B}_k(R)\right]\right] \leq \mathbb{E}\left[[F(\theta_k) - F_{l.b.}]\mathbf{1}\left[\mathcal{B}_k(R)\right]\right] \\ - \mathbb{E}\left[[F(\theta_{k+1}) - F_{l.b.}]\mathbf{1}\left[\mathcal{B}_{k+1}(R)\right]\right] + \frac{(L_{R+1}+\partial F_R)G_R}{1+\alpha}\lambda_{\max}(M_k)^{1+\alpha}. \tag{86}$$

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

$$\sum_{k=0}^{j}\lambda_{\min}(M_k)\mathbb{E}\left[\left\|\dot{F}(\theta_k)\right\|_2^2 \mathbf{1}\left[\mathcal{B}_k(R)\right]\right] \leq [F(\theta_0) - F_{l.b.}]\mathbf{1}\left[\mathcal{B}_0(R)\right] \\ - \mathbb{E}\left[[F(\theta_{j+1}) - F_{l.b.}]\mathbf{1}\left[\mathcal{B}_{j+1}(R)\right]\right] + \frac{(L_{R+1}+\partial F_R)G_R}{1+\alpha}\sum_{k=0}^{j}\lambda_{\max}(M_k)^{1+\alpha}. \tag{87}$$

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

$$[F(\theta_0) - F_{l.b.}]\mathbf{1}\left[\mathcal{B}_0(R)\right] + \frac{(L_{R+1}+\partial F_R)G_R}{1+\alpha}S, \tag{88}$$

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

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

$$\mathbb{P}\left[\bigcap_{k=j}^{\infty}\left\{\left\|\dot{F}(\theta_k)\right\|_2^2 \mathbf{1}[\mathcal{B}_k(R)] > \delta\right\}\right] \leq \mathbb{P}\left[\left\|\dot{F}(\theta_k)\right\|_2^2 \mathbf{1}[\mathcal{B}_k(R)] > \delta\right] \tag{89}$$

$$\leq \frac{1}{\delta}\mathbb{E}\left[\left\|\dot{F}(\theta_k)\right\|_2^2 \mathbf{1}[\mathcal{B}_k(R)]\right]. \tag{90}$$

Since the last inequality holds for every $k \geq j$, then, in particular, for all $j+1 \in \mathbb{N}$,

$$\mathbb{P}\left[\bigcap_{k=j}^{\infty}\left\{\left\|\dot{F}(\theta_k)\right\|_2^2 \mathbf{1}[\mathcal{B}_k(R)] > \delta\right\}\right] \leq \frac{1}{\delta}\min_{j \leq k}\mathbb{E}\left[\left\|\dot{F}(\theta_k)\right\|_2^2 \mathbf{1}[\mathcal{B}_k(R)]\right], \tag{91}$$

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

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

$$\mathbb{P}\left[\left\|\dot{F}(\theta_k)\right\|_2^2 \mathbf{1}[\mathcal{B}_k(R)] \leq \delta,\ i.o.\right] = 1. \tag{92}$$

□

Unfortunately, Lemma 9 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 10. *Let $\{\theta_k\}$ be defined as in (5) satisfying Properties 1 and 2. Suppose Assumptions 1 to 4 hold. Then, $\forall R \geq 0$ and for all $\delta > 0$,*

$$\mathbb{P}\left[\left\|\dot{F}(\theta_{k+1})\right\|_2 \mathbf{1}[\mathcal{B}_{k+1}(R)] > \delta, \left\|\dot{F}(\theta_k)\right\|_2 \mathbf{1}[\mathcal{B}_k(R)] \leq \delta,\ i.o.\right] = 0. \tag{93}$$

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

$$\mathbb{P}\left[\left\|\dot{F}(\theta_{k+1})\right\|_2 \mathbf{1}[\mathcal{B}_{k+1}(R)]\,\mathbf{1}\left[\left\|\dot{F}(\theta_k)\right\|_2 \mathbf{1}[\mathcal{B}_k(R)] \leq \delta\right] > \delta + L_R\gamma^{\alpha}\right] \tag{94}$$

$$= \mathbb{P}\Bigg[\left(\left\|\dot{F}(\theta_{k+1})\right\|_2 - \left\|\dot{F}(\theta_k)\right\|_2 + \left\|\dot{F}(\theta_k)\right\|_2\right)\mathbf{1}[\mathcal{B}_{k+1}(R)] \tag{95}$$

$$\times\,\mathbf{1}\left[\left\|\dot{F}(\theta_k)\right\|_2 \mathbf{1}[\mathcal{B}_k(R)] \leq \delta\right] > \delta + L_R\gamma^{\alpha}\Bigg]. \tag{96}$$

Using the reverse triangle inequality, $\|\dot{F}(\theta_{k+1})\|_2 - \|\dot{F}(\theta_k)\|_2 \leq \|\dot{F}(\theta_{k+1}) - \dot{F}(\theta_k)\|_2$. Now, making use of the restriction to $\mathcal{B}_{k+1}(R)$, $\|\dot{F}(\theta_{k+1}) - \dot{F}(\theta_k)\|_2 \leq L_R\|\theta_{k+1} - \theta_k\|_2^{\alpha}$. Moreover, on $\|\dot{F}(\theta_k)\|_2 \leq \delta$, $\|\dot{F}(\theta_k)\|_2 \mathbf{1}[\mathcal{B}_{k+1}(R)] \leq \delta$. Putting these two observations together,

$$\mathbb{P}\Bigg[\left(\left\|\dot{F}(\theta_{k+1})\right\|_2 - \left\|\dot{F}(\theta_k)\right\|_2 + \left\|\dot{F}(\theta_k)\right\|_2\right)\mathbf{1}[\mathcal{B}_{k+1}(R)] \tag{97}$$

$$\times\,\mathbf{1}\left[\left\|\dot{F}(\theta_k)\right\|_2 \mathbf{1}[\mathcal{B}_k(R)] \leq \delta\right] > \delta + L_R\gamma^{\alpha}\Bigg] \tag{98}$$

$$\leq \mathbb{P}\left[L_R\|\theta_{k+1} - \theta_k\|_2^{\alpha}\,\mathbf{1}[\mathcal{B}_{k+1}(R)]\,\mathbf{1}\left[\left\|\dot{F}(\theta_k)\right\|_2 \mathbf{1}[\mathcal{B}_k(R)] \leq \delta\right] > L_R\gamma^{\alpha}\right] \tag{99}$$

$$= \mathbb{P}\left[\left\|M_k\dot{f}(\theta_k, X_{k+1})\right\|_2 \mathbf{1}[\mathcal{B}_{k+1}(R)]\,\mathbf{1}\left[\left\|\dot{F}(\theta_k)\right\|_2 \mathbf{1}[\mathcal{B}_k(R)] \leq \delta\right] > \gamma\right]. \tag{100}$$

Now, using $\mathbf{1}\left[\mathcal{B}_{k+1}(R)\right]\mathbf{1}\left[\left\|\dot{F}(\theta_k)\right\|_2\mathbf{1}\left[\mathcal{B}_k(R)\right]\right] \leq \mathbf{1}\left[\mathcal{B}_k(R)\right]$,

$$\mathbb{P}\left[\left\|M_k\dot{f}(\theta_k, X_{k+1})\right\|_2\mathbf{1}\left[\mathcal{B}_{k+1}(R)\right]\mathbf{1}\left[\left\|\dot{F}(\theta_k)\right\|_2\mathbf{1}\left[\mathcal{B}_k(R)\right] \leq \delta\right] > \gamma\right] \tag{101}$$

$$\leq \mathbb{P}\left[\left\|M_k\dot{f}(\theta_k, X_{k+1})\right\|_2\mathbf{1}\left[\mathcal{B}_k(R)\right] > \gamma\right] \tag{102}$$

$$\leq \mathbb{P}\left[\left\|M_k\dot{f}(\theta_k, X_{k+1})\right\|_2^{1+\alpha}\mathbf{1}\left[\mathcal{B}_k(R)\right] > \gamma^{1+\alpha}\right] \tag{103}$$

$$\leq \frac{1}{\gamma^{1+\alpha}}\left\|M_k\right\|_2^{1+\alpha}\mathbb{E}\left[\mathbb{E}\left[\left\|\dot{f}(\theta_k, X_{k+1})\right\|_2^{1+\alpha}\middle|\mathcal{F}_k\right]\mathbf{1}\left[\mathcal{B}_k(R)\right]\right], \tag{104}$$

where the last inequality is a consequence of Markov's inequality, $\|M_k\dot{f}(\theta_k, X_{k+1})\|_2 \leq \|M_k\|_2\|\dot{f}(\theta_k, X_{k+1})\|_2$, and $\mathbf{1}\left[\mathcal{B}_k(R)\right]$ being measurable with respect to $\mathcal{F}_k$.

By Assumption 4, $\mathbb{E}\left[\|\dot{f}(\theta_k, X_{k+1})\|_2^{1+\alpha}\middle|\mathcal{F}_k\right] \leq G(\theta_k)$. Moreover, on $\mathcal{B}_k(R)$, $G(\theta_k) \leq G_R$. Using this in the expectation, we conclude

$$\mathbb{P}\left[\left\|\dot{F}(\theta_{k+1})\right\|_2\mathbf{1}\left[\mathcal{B}_{k+1}(R)\right]\mathbf{1}\left[\left\|\dot{F}(\theta_k)\right\|_2\mathbf{1}\left[\mathcal{B}_k(R)\right] \leq \delta\right] > \delta + L_R\gamma^{\alpha}\right] \leq \frac{1}{\gamma^{1+\alpha}}\left\|M_k\right\|_2^{1+\alpha}G_R. \tag{105}$$

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 $\gamma > 0$,

$$\mathbb{P}\left[\left\|\dot{F}(\theta_{k+1})\right\|_2\mathbf{1}\left[\mathcal{B}_{k+1}(R)\right] > \delta + L_R\gamma^{\alpha}, \left\|\dot{F}(\theta_k)\right\|_2\mathbf{1}\left[\mathcal{B}_k(R)\right] \leq \delta,\ i.o.\right] = 0. \tag{106}$$

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, for any $R \geq 0$ and $\delta > 0$,

$$\mathbb{P}\left[\left\{\left\|\dot{F}(\theta_{k+1})\right\|_2\mathbf{1}\left[\mathcal{B}_{k+1}(R)\right] > \delta, \left\|\dot{F}(\theta_k)\right\|_2\mathbf{1}\left[\mathcal{B}_k(R)\right] \leq \delta,\ i.o.\right\} \cap \Omega_\delta^c\right] = 0, \tag{107}$$

where $\Omega_\delta = \{\limsup_k \|\dot{F}(\theta_{k+1})\|_2\mathbf{1}\left[\mathcal{B}_{k+1}(R)\right] = \delta\}$.

We now show that Ω_δ is a probability zero event. Notice, by Lemma 9 and the definition of Ω_δ,

$$\Omega_\delta \subset \left\{\left\|\dot{F}(\theta_{k+1})\right\|_2\mathbf{1}\left[\mathcal{B}_{k+1}(R)\right] > \delta/2, \left\|\dot{F}(\theta_k)\right\|_2\mathbf{1}\left[\mathcal{B}_k(R)\right] \leq \delta/2,\ i.o.\right\} \cap \Omega_{\delta/2}^c, \tag{108}$$

up to a set of measure zero. By applying (107) with $\delta/2$, $\mathbb{P}[\Omega_\delta] = 0$. The conclusion of the result follows. ☐

We now put together Lemmas 9 and 10 to show that, on the event $\{\sup_k \|\theta_k\|_2 < \infty\}$, $\|\dot{F}(\theta_k)\|_2$ converges to 0 with probability one.

Corollary 2. *Let $\{\theta_k\}$ be defined as in (5) 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 9 implies

$$\begin{aligned}&\mathbb{P}\left[\left\|\dot{F}(\theta_{k+1})\right\|_2\mathbf{1}\left[\mathcal{B}_{k+1}(R)\right] > \delta,\ i.o.\right]\\ &\quad= \mathbb{P}\left[\left\{\left\|\dot{F}(\theta_{k+1})\right\|_2\mathbf{1}\left[\mathcal{B}_{k+1}(R)\right] > \delta,\ i.o.\right\} \cap \left\{\left\|\dot{F}(\theta_k)\right\|_2\mathbf{1}\left[\mathcal{B}_k(R)\right] \leq \delta,\ i.o.\right\}\right].\end{aligned} \tag{109}$$

We see that this latter event is exactly,

$$\mathbb{P}\left[\left\|\dot{F}(\theta_{k+1})\right\|_2\mathbf{1}\left[\mathcal{B}_{k+1}(R)\right] > \delta, \left\|\dot{F}(\theta_k)\right\|_2\mathbf{1}\left[\mathcal{B}_k(R)\right] \leq \delta,\ i.o.\right], \tag{110}$$

which, by Lemma 10, 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 (5), 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,

$$\begin{aligned}
&\mathbb{P}\left[\left\|\theta_{k+1}-\bar\theta\right\|_2 \geq R+\gamma, \left\|\theta_k-\bar\theta\right\|_2 \leq R\right] \\
&\quad= \mathbb{P}\left[\left\|\theta_{k+1}-\bar\theta\right\|_2 \mathbf{1}\left[\left\|\theta_k-\bar\theta\right\|_2 \leq R\right] \geq R+\gamma\right] && (111)\\
&\quad= \mathbb{P}\left[\left(\left\|\theta_{k+1}-\bar\theta\right\|_2 - \left\|\theta_k-\bar\theta\right\|_2 + \left\|\theta_k-\bar\theta\right\|_2\right) \mathbf{1}\left[\left\|\theta_k-\bar\theta\right\|_2 \leq R\right] \geq R+\gamma\right]. && (112)
\end{aligned}$$

Now, $\|\theta_k-\bar\theta\|_2 \mathbf{1}\left[\|\theta_k-\bar\theta\|_2 \leq R\right] \leq R$. Therefore,

$$\begin{aligned}
&\mathbb{P}\left[\left(\left\|\theta_{k+1}-\bar\theta\right\|_2 - \left\|\theta_k-\bar\theta\right\|_2 + \left\|\theta_k-\bar\theta\right\|_2\right) \mathbf{1}\left[\left\|\theta_k-\bar\theta\right\|_2 \leq R\right] \geq R+\gamma\right] && (113)\\
&\quad\leq \mathbb{P}\left[\left(\left\|\theta_{k+1}-\bar\theta\right\|_2 - \left\|\theta_k-\bar\theta\right\|_2\right) \mathbf{1}\left[\left\|\theta_k-\bar\theta\right\|_2 \leq R\right] + R \geq R+\gamma\right] && (114)\\
&\quad\leq \mathbb{P}\left[\left\|\theta_{k+1}-\theta_k\right\|_2 \mathbf{1}\left[\left\|\theta_k-\bar\theta\right\|_2 \leq R\right] \geq \gamma\right], && (115)
\end{aligned}$$

where the last line follows by applying the reverse triangle inequality. By using (5) and Markov's inequality,

$$\begin{aligned}
&\mathbb{P}\left[\left\|\theta_{k+1}-\theta_k\right\|_2 \mathbf{1}\left[\left\|\theta_k-\bar\theta\right\|_2 \leq R\right] \geq \gamma\right] && (116)\\
&\quad\leq \mathbb{P}\left[\left\|M_k \dot f(\theta_k, X_{k+1})\right\|_2 \mathbf{1}\left[\left\|\theta_k-\bar\theta\right\|_2 \leq R\right] \geq \gamma\right] && (117)\\
&\quad\leq \frac{1}{\gamma^{1+\alpha}} \left\|M_k\right\|_2^{1+\alpha} \mathbb{E}\left[\mathbb{E}\left[\left\|\dot f(\theta_k, X_{k+1})\right\|_2^{1+\alpha} \middle| \mathcal{F}_k\right] \mathbf{1}\left[\left\|\theta_k-\bar\theta\right\|_2 \leq R\right]\right]. && (118)
\end{aligned}$$

By applying Assumption 4, $\mathbb{E}\left[\|\dot f(\theta_k, X_{k+1})\|_2^{1+\alpha} \middle| \mathcal{F}_k\right] \leq G(\theta_k)$. Moreover, on $\|\theta_k-\bar\theta\|_2 \leq R$, $G(\theta_k) \leq \sup_{\theta:\|\theta\|_2 \leq R+\|\bar\theta\|_2} G(\theta) =: G_{R+\|\bar\theta\|_2} < \infty$ since G is upper semi-continuous. Combining these steps,

$$\mathbb{P}\left[\left\|\theta_{k+1}-\bar\theta\right\|_2 \geq R+\gamma, \left\|\theta_k-\bar\theta\right\|_2 \leq R\right] \leq \frac{1}{\gamma^{1+\alpha}} \left\|M_k\right\|_2^{1+\alpha} G_{R+\|\bar\theta\|_2}, \tag{119}$$

By Property 2, we see that the sum of the probabilities is finite. Together with the Borel-Cantelli lemma, $\forall R \geq 0$, $\forall \gamma > 0$, and for all $\bar\theta \in \mathbb{R}^p$,

$$\mathbb{P}\left[\|\theta_{k+1}-\bar\theta\|_2 \geq R+\gamma,\ \|\theta_k-\bar\theta\|_2 \leq R,\ i.o.\right] = 0. \tag{120}$$

Since $\gamma > 0$ is arbitrary, we can show that this statement holds for a countable sequence of $\gamma_n \downarrow 0$. Therefore, $\forall R \geq 0$ and all $\bar\theta \in \mathbb{R}^p$,

$$\mathbb{P}\left[\limsup_k \left\|\theta_k-\bar\theta\right\|_2 > R, \liminf_k \left\|\theta_k-\bar\theta_2\right\|_2 \leq R\right] = 0. \tag{121}$$

Since R is arbitrary, we conclude that for any ordering of positive rational numbers, $\{R_n\}$, $\mathbb{P}[\limsup_k \|\theta_{k+1}-\bar\theta\|_2 > R_n, \liminf_k \|\theta_k-\bar\theta\|_2 \leq R_n] = 0$ for every n. Again, the countable union of measure zero sets is measure zero. Hence, we conclude that $\mathbb{P}[\limsup_k \|\theta_k-\bar\theta\|_2 > \liminf_k \|\theta_k-\bar\theta\|_2] = 0$. Thus, either $\lim_k \|\theta_k-\bar\theta\|_2$ exists and is either infinite or finite.

Moreover, on the event that the limit is finite, since $\bar\theta$ is arbitrary, we can choose $p+1$ distinct values of $\bar\theta$ which do not belong to a hyperplane of dimension smaller than p, and, by triangulation, the $\lim_k \theta_k$ converges to a fixed point (up to a set of measure zero). □

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 visits to a region of zero on the same event. Finally, we study this event to establish that the two statements above hold with probability one.

D.1 A Recursive Relationship

Lemma 11 (Lemma 3)**.** *Let $\{M_k\}$ satisfy Property 1. Suppose Assumptions 1 to 4 hold. Let $\{\theta_k\}$ satisfy (5). Then, for any $j+1 \in \mathbb{N}$ and $k > j$,*

$$
\begin{aligned}
\mathbb{E}\left[(F(\theta_{k+1}) - F_{l.b.}) \mathbf{1}[\tau_j > k] \middle| \mathcal{F}_k \right] &\leq \left(F(\theta_k) - F_{l.b.} - \dot{F}(\theta_k)' M_k \dot{F}(\theta_k) \right) \mathbf{1}[\tau_j > k-1] \\
&+ \frac{\lambda_{\max}(M_k)^{1+\alpha}}{1+\alpha} \left[\mathcal{L}_\epsilon(\theta_k) G(\theta_k) + \alpha \left[\frac{\left\| \dot{F}(\theta_k) \right\|_2^{1+\alpha}}{\mathcal{L}_\epsilon(\theta_k)} \right]^{1/\alpha} \right] \mathbf{1}[\tau_j > k-1].
\end{aligned} \tag{122}
$$

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

$$
F(\theta_{k+1}) - F_{l.b.} \leq F(\theta_k) - F_{l.b.} + \dot{F}(\theta_k)'(\theta_{k+1} - \theta_k) + \frac{\mathcal{L}_\epsilon(\theta_k)}{1+\alpha} \|\theta_{k+1} - \theta_k\|_2^{1+\alpha}. \tag{123}
$$

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

$$
\begin{aligned}
&\mathbb{E}\left[\{F(\theta_{k+1}) - F_{l.b.}\} \mathbf{1}[\tau_j > k] \middle| \mathcal{F}_k \right] \\
&\leq \mathbb{E}\left[\left\{ F(\theta_k) - F_{l.b.} + \dot{F}(\theta_k)'(\theta_{k+1} - \theta_k) + \frac{\mathcal{L}_\epsilon(\theta_k)}{1+\alpha} \|\theta_{k+1} - \theta_k\|_2^{1+\alpha} \right\} \mathbf{1}[\tau_j > k-1] \middle| \mathcal{F}_k \right] \\
&- \mathbb{E}\left[\left\{ F(\theta_k) - F_{l.b.} + \dot{F}(\theta_k)'(\theta_{k+1} - \theta_k) + \frac{\mathcal{L}_\epsilon(\theta_k)}{1+\alpha} \|\theta_{k+1} - \theta_k\|_2^{1+\alpha} \right\} \mathbf{1}[\tau_j = k] \middle| \mathcal{F}_k \right]
\end{aligned} \tag{124}
$$

For the first term on the right hand side, we can apply Assumptions 3 and 4, Property 1, and (5) to calculate

$$
\begin{aligned}
&\mathbb{E}\left[\left\{ F(\theta_k) - F_{l.b.} + \dot{F}(\theta_k)'(\theta_{k+1} - \theta_k) + \frac{\mathcal{L}_\epsilon(\theta_k)}{1+\alpha} \|\theta_{k+1} - \theta_k\|_2^{1+\alpha} \right\} \mathbf{1}[\tau_j > k-1] \middle| \mathcal{F}_k \right] \\
&\leq \left\{ F(\theta_k) - F_{l.b.} - \dot{F}(\theta_k)' M_k \dot{F}(\theta_k) + \frac{\lambda_{\max}(M_k)^{1+\alpha}}{1+\alpha} \mathcal{L}_\epsilon(\theta_k) G(\theta_k) \right\} \mathbf{1}[\tau_j > k-1].
\end{aligned} \tag{125}
$$

For the second term on the right hand side of (124), we require two facts. The first fact is $\mathbf{1}[\tau_j = k] \leq \mathbf{1}[\tau_j > k-1]$ which implies $\mathbf{1}[\tau_j = k] = \mathbf{1}[\tau_j = k]\mathbf{1}[\tau_j > k-1]$. For the second fact, the Cauchy-Schwarz inequality and Lemma 7 imply

$$
\begin{aligned}
&- F(\theta_k)' M_k \dot{f}(\theta_k, X_{k+1}) + \frac{\mathcal{L}_\epsilon(\theta_k)}{1+\alpha} \|M_k f(\theta_k, X_{k+1})\|_2^{1+\alpha} \\
&\geq - \|F(\theta_k)\|_2 \left\| M_k \dot{f}(\theta_k, X_{k+1}) \right\|_2 + \frac{\mathcal{L}_\epsilon(\theta_k)}{1+\alpha} \|M_k f(\theta_k, X_{k+1})\|_2^{1+\alpha} \qquad (126) \\
&\geq - \frac{\alpha}{1+\alpha} \left[\frac{\left\| \dot{F}(\theta_k) \right\|_2^{1+\alpha}}{\mathcal{L}_\epsilon(\theta_k)} \right]^{1/\alpha}. \qquad (127)
\end{aligned}
$$

Hence, using (5),

$$
\begin{aligned}
&- [F(\theta_k) - F_{l.b.}] - \dot{F}(\theta_k)'(\theta_{k+1} - \theta_k) - \frac{\mathcal{L}_\epsilon(\theta_k)}{1+\alpha} \|\theta_{k+1} - \theta_k\|_2^{1+\alpha} \\
&\leq - [F(\theta_k) - F_{l.b.}] + \frac{\alpha}{1+\alpha} \left[\frac{\left\| \dot{F}(\theta_k) \right\|_2^{1+\alpha}}{\mathcal{L}_\epsilon(\theta_k)} \right]^{1/\alpha}.
\end{aligned} \tag{128}
$$

Putting together these two preceding facts together,

$$
\begin{aligned}
&-\mathbb{E}\left[\left\{F(\theta_k) - F_{l.b.} + \dot{F}(\theta_k)'(\theta_{k+1}-\theta_k) + \frac{\mathcal{L}_\epsilon(\theta_k)}{1+\alpha}\left\|\theta_{k+1}-\theta_k\right\|_2^{1+\alpha}\right\}\mathbf{1}\left[\tau_j = k\right]\middle|\mathcal{F}_k\right] \\
&\leq \left\{-\left[F(\theta_k) - F_{l.b.}\right] + \frac{\alpha}{1+\alpha}\left[\frac{\left\|\dot{F}(\theta_k)\right\|_2^{1+\alpha}}{\mathcal{L}_\epsilon(\theta_k)}\right]^{1/\alpha}\right\}\mathbb{P}\left[\tau_j = k|\,\mathcal{F}_k\right]\mathbf{1}\left[\tau_j > k-1\right] \qquad (129) \\
&\leq \frac{\alpha\lambda_{\max}(M_k)^{1+\alpha}}{1+\alpha}\left[\frac{\left\|\dot{F}(\theta_k)\right\|_2^{1+\alpha}}{\mathcal{L}_\epsilon(\theta_k)}\right]^{1/\alpha}\mathbf{1}\left[\tau_j > k-1\right], \qquad (130)
\end{aligned}
$$

where we bound $\mathbb{P}[\tau_j = k|\mathcal{F}_k]$ using Theorem 5. By applying the bounds on the first term, (125), and second term, (130), to (124), the result follows. □

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

Lemma 12 (Lemma 4)**.** *If Assumptions 1 to 5, and Properties 1 and 4 hold, and $\{\theta_k\}$ satisfy (5), then there exists a $K \in \mathbb{N}$ such that for any $j+1 \in \mathbb{N}$ and any $k \geq \min\{K, j+1\}$,*

$$
\begin{aligned}
&\mathbb{E}\left[(F(\theta_{k+1}) - F_{l.b.})\mathbf{1}\left[\tau_j > k\right]|\,\mathcal{F}_k\right] \\
&\leq \left(1 + \lambda_{\max}(M_k)^{1+\alpha}\frac{C_2}{1+\alpha}\right)(F(\theta_k) - F_{l.b.})\mathbf{1}\left[\tau_j > k-1\right] \\
&-\frac{1}{2}\lambda_{\min}(M_k)\left\|\dot{F}(\theta_k)\right\|_2^2\mathbf{1}\left[\tau_j > k-1\right] + \lambda_{\max}(M_k)^{1+\alpha}\frac{C_1}{1+\alpha}.
\end{aligned} \qquad (131)
$$

Proof. The result follows by first using Assumption 5 in Lemma 3. Then, collecting similar terms, we apply Lemma 6 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 (5), 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 and Siegmund [1971], Neveu and Speed [1975, Exercise II.4], the limit as k goes to infinity of $(F(\theta_k) - F_{l.b.})\mathbf{1}\left[\tau_j > k-1\right]$ 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 13. *If Assumptions 1 to 5, and Properties 1, 2 and 4 hold, and $\{\theta_k\}$ satisfy (5), then $\exists K \in \mathbb{N}$ such that for any $j > K$, $\exists N_j > 0$ for which*

$$
\sup_{k>j}\mathbb{E}\left[(F(\theta_k) - F_{l.b.})\mathbf{1}\left[\tau_j > k-1\right]\right] \leq N_j. \qquad (132)
$$

Proof. In Lemma 4, we upper bound the right hand side by removing the negative term, and, by Property 2, we add $C_1(1+\alpha)^{-1}\sum_{\ell=k+1}^\infty \lambda_{\max}(M_\ell)^{1+\alpha}$ to both side. Then, for all $k \geq j$,

$$
\begin{aligned}
&\mathbb{E}\left[(F(\theta_{k+1}) - F_{l.b.})\mathbf{1}\left[\tau_j > k\right]\right] + \frac{C_1}{1+\alpha}\sum_{\ell=k+1}^\infty \lambda_{\max}(M_\ell)^{1+\alpha} \\
&\leq \left(1 + \lambda_{\max}(M_k)^{1+\alpha}\frac{C_2}{1+\alpha}\right)\mathbb{E}\left[(F(\theta_k) - F_{l.b.})\mathbf{1}\left[\tau_j > k-1\right]\right] + \frac{C_1}{1+\alpha}\sum_{\ell=k}^\infty \lambda_{\max}(M_\ell)^{1+\alpha}.
\end{aligned} \qquad (133)
$$

Using $1 + C_2(1+\alpha)^{-1}\lambda_{\max}(M_k)^{1+\alpha} \leq \exp(C_2(1+\alpha)^{-1}\lambda_{\max}(M_k)^{1+\alpha})$, it follows

$$\begin{aligned} &\mathbb{E}\left[(F(\theta_{k+1}) - F_{l.b.})\mathbf{1}\left[\tau_j > k\right]\right] + \frac{C_1}{1+\alpha}\sum_{\ell=k+1}^{\infty}\lambda_{\max}(M_\ell)^{1+\alpha} \\ &\leq \exp\left(\frac{C_2}{1+\alpha}\lambda_{\max}(M_k)^{1+\alpha}\right)\left[\mathbb{E}\left[(F(\theta_k) - F_{l.b.})\mathbf{1}\left[\tau_j > k-1\right]\right] + \frac{C_1}{1+\alpha}\sum_{\ell=k}^{\infty}\lambda_{\max}(M_\ell)^{1+\alpha}\right]. \end{aligned} \tag{134}$$

Hence,

$$\begin{aligned} &\mathbb{E}\left[(F(\theta_{k+1}) - F_{l.b.})\mathbf{1}\left[\tau_j > k\right]\right] + \frac{C_1}{1+\alpha}\sum_{\ell=k+1}^{\infty}\lambda_{\max}(M_\ell)^{1+\alpha} \\ &\leq \exp\left(\frac{C_2}{1+\alpha}\sum_{\ell=j}^{k}\lambda_{\max}(M_\ell)^{1+\alpha}\right)\left[\mathbb{E}\left[(F(\theta_j) - F_{l.b.})\right] + \frac{C_1}{1+\alpha}\sum_{\ell=j}^{\infty}\lambda_{\max}(M_\ell)^{1+\alpha}\right], \end{aligned} \tag{135}$$

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

D.3 Gradient Function Analysis

Lemma 14. *If Assumptions 1 to 5, and Properties 1 to 4 hold, and $\{\theta_k\}$ satisfy (5), then, for any $\delta > 0$,*

$$\mathbb{P}\left[\left\|\dot{F}(\theta_k)\right\|_2 \mathbf{1}\left[\tau_j > k-1\right] \leq \delta \text{ } i.o.\right] = 1. \tag{136}$$

Proof. By Lemma 4,

$$\begin{aligned} &\frac{1}{2}\lambda_{\min}(M_k)\mathbb{E}\left[\left\|\dot{F}(\theta_k)\right\|_2^2 \mathbf{1}\left[\tau_j > k-1\right]\right] \leq \mathbb{E}\left[(F(\theta_k) - F_{l.b.})\mathbf{1}\left[\tau_j > k-1\right]\right] \\ &- \mathbb{E}\left[(F(\theta_{k+1}) - F_{l.b.})\mathbf{1}\left[\tau_j > k\right]\right] + \frac{C_2}{1+\alpha}\lambda_{\max}(M_k)^{1+\alpha}\mathbb{E}\left[(F(\theta_k) - F_{l.b.})\mathbf{1}\left[\tau_j > k-1\right]\right] \\ &+ \frac{C_1}{1+\alpha}\lambda_{\max}(M_k)^{1+\alpha}. \end{aligned} \tag{137}$$

By applying Lemma 13,

$$\begin{aligned} \frac{1}{2}\lambda_{\min}(M_k)\mathbb{E}\left[\left\|\dot{F}(\theta_k)\right\|_2^2 \mathbf{1}\left[\tau_j > k-1\right]\right] &\leq \mathbb{E}\left[(F(\theta_k) - F_{l.b.})\mathbf{1}\left[\tau_j > k-1\right]\right] \\ - \mathbb{E}\left[(F(\theta_{k+1}) - F_{l.b.})\mathbf{1}\left[\tau_j > k\right]\right] &+ \lambda_{\max}(M_k)^{1+\alpha}\left(\frac{C_2 N_j + C_1}{1+\alpha}\right). \end{aligned} \tag{138}$$

By summing and using Assumption 1,

$$\begin{aligned} &\frac{1}{2}\sum_{k=j}^{\infty}\lambda_{\min}(M_k)\mathbb{E}\left[\left\|\dot{F}(\theta_k)\right\|_2^2 \mathbf{1}\left[\tau_j > k-1\right]\right] \\ &\leq \mathbb{E}\left[F(\theta_j) - F_{l.b.}\right] + \frac{C_2 N_j + C_1}{1+\alpha}\sum_{k=j}^{\infty}\lambda_{\max}(M_k)^{1+\alpha}. \end{aligned} \tag{139}$$

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

$$\liminf_{k\to\infty}\mathbb{E}\left[\left\|\dot{F}(\theta_k)\right\|_2^2 \mathbf{1}\left[\tau_j > k-1\right]\right] = 0. \tag{140}$$

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

$$\mathbb{P}\left[\bigcap_{k=\ell}^{\infty}\left\{\left\|\dot{F}(\theta_k)\right\|_2 \mathbf{1}\left[\tau_j > k-1\right] > \delta\right\}\right] \leq \frac{1}{\delta^2}\min_{k\geq\ell}\mathbb{E}\left[\left\|\dot{F}(\theta_k)\right\|_2^2 \mathbf{1}\left[\tau_j > k-1\right]\right] = 0. \tag{141}$$

As the countable union of sets of measure zero have measure zero, the result follows. $\square$

D.4 Stopping Time Analysis

WE compute the probability of $\{\tau_j = k\}$.

Theorem 5. *Let $\{\tau_j : j+1 \in \mathbb{N}\}$ be defined as in (11). If Assumptions 1, 2 and 4 and Property 1 hold, and $\{\theta_k\}$ satisfy (5), then, for any $j+1 \in \mathbb{N}$ and any $k+1 \in \mathbb{N}$,*

$$\mathbb{P}\left[\tau_j = k \middle| \mathcal{F}_k\right] \leq \begin{cases} 0 & k \leq j, \\ \lambda_{\max}(M_k)^{1+\alpha} & k > j. \end{cases} \tag{142}$$

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

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 (5), Markov's inequality, Property 1, we conclude

$$\mathbb{P}\left[\tau_j = k \middle| \mathcal{F}_k\right] \leq \mathbb{P}\left[\left\| M_k \dot{f}(\theta_k, X_{k+1}) \right\|_2^{1+\alpha} > G(\theta_k) \vee \epsilon \middle| \mathcal{F}_k \right] \tag{143}$$

$$\leq \frac{\lambda_{\max}(M_k)^{1+\alpha}\mathbb{E}\left[\left\| \dot{f}(\theta_k, X_{k+1} \right\|_2^{1+\alpha} \middle| \mathcal{F}_k \right]}{G(\theta_k) \vee \epsilon}. \tag{144}$$

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

$$\mathbb{P}\left[\tau_j = \infty\right] \geq 1 - \mathbb{P}\left[\tau_j < \infty\right] \geq 1 - \sum_{k=j+1}^{\infty} \lambda_{\max}(M_k)^{1+\alpha}. \tag{145}$$

Therefore,

$$\mathbb{P}\left[\bigcup_{j=0}^{\infty}\{\tau_j = \infty\}\right] = \lim_{j\to\infty} \mathbb{P}\left[\tau_j = \infty\right]. \tag{146}$$

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