

Additional Notation: For a matrix \mathbf{M} , $\lambda_{\min}(\mathbf{M})$ denotes its minimum eigenvalue, and $\lambda_r(\mathbf{M})$ denotes its r -th largest eigenvalue. Thus $\lambda_1(\mathbf{M}) \geq \lambda_2(\mathbf{M}) \geq \dots$. We denote the $k \times k$ identity matrix by \mathbf{I}_k . We use $\mathbb{B}^k(\mathbf{p}, R)$ to denote the full k -dimensional l_2 -ball centered at $\mathbf{p} \in \mathbb{R}^k$ with radius R , including the boundary. When k is not specified explicitly, $\mathbb{B}(\mathbf{p}, R)$ refers to the l_2 -ball in \mathbb{R}^d , following Notation. All logarithms in the following are the natural logarithm. For an event \mathcal{S} , $1_{\mathcal{S}}$ denotes the indicator function. In the following, the norm $\|\cdot\|$ of matrices and higher-order tensors refers to the operator norm unless otherwise stated. The norm $\|\cdot\|$ of vectors refers to l_2 -Euclidean norm.

Contents

1	Introduction	1
1.1	Our Contributions	2
2	Main Idea	3
2.1	High Level Idea	3
2.2	The Formal Framework	4
2.3	Examples Subsumed by Framework	5
3	Convergence Results	7
3.1	Gradient Descent	8
3.2	Adaptive Gradient Descent	8
3.3	Stochastic Gradient Descent	9
3.4	Perturbed Gradient Descent	9
3.5	Restarted Stochastic Gradient Descent	9
3.6	Examples	10
3.7	Practical Implications and Simulations	10
4	Conclusion	10
5	Acknowledgments	10
A	Technical Preliminaries	23
A.1	Helpful Background Lemmas	23
A.2	Comparison of Assumptions with Literature	25
A.3	Proofs of Technical Results	29
B	Proof of Framework	31
C	First Order Convergence Proofs	33
C.1	Proofs for Adaptive GD	33
C.2	Proofs for SGD for FOSPs	34
D	Perturbed GD finding Second Order Stationary Points	42
D.1	Proof using the Framework	42
D.2	Proving the key Lemma	45
D.3	Proof of Escaping Saddles Lemmas	50
E	Restarted SGD finding Second Order Stationary Points	56
E.1	Notation and Parameters	56
E.2	Result	58
E.3	Preliminaries	60
E.4	Escaping Saddles	63
E.5	Faster Descent	71
E.6	Finding Second Order Stationary Points	84
F	Examples	86
F.1	Phase Retrieval	86
F.2	Matrix PCA	88

G Simulations	90
G.1 Synthetic Simulations with GD	91
G.2 Synthetic Simulations with SGD	95

A Technical Preliminaries

A.1 Helpful Background Lemmas

We will use the following classical inequalities from optimization to show we still have some notion of control if we have local bounds on the relevant derivatives.

Lemma A.1. *Suppose F is twice differentiable, and for all $\mathbf{u} \in \overline{\mathbf{x}\mathbf{y}}$ (the line segment) we have $\|\nabla^2 F(\mathbf{u})\|_{\text{op}} \leq L$. Then, we have*

$$F(\mathbf{y}) \leq F(\mathbf{x}) + \langle \nabla F(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|^2.$$

Proof. This follows by the proof of Lemma 3.4 in [Bubeck et al. \(2015\)](#). In particular, one can readily verify that $\mathbf{x} + t(\mathbf{y} - \mathbf{x}) \in \overline{\mathbf{x}\mathbf{y}}$ for all $t \in [0, 1]$. Hence for all $t \in [0, 1]$ and \mathbf{u} in the line segment between \mathbf{x} and $\mathbf{x} + t(\mathbf{y} - \mathbf{x})$, $\|\nabla^2 F(\mathbf{u})\|_{\text{op}} \leq L$. Thus,

$$\begin{aligned} |F(\mathbf{y}) - F(\mathbf{x}) - \langle \nabla F(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle| &= \left| \int_0^1 \langle \nabla F(\mathbf{x} + t(\mathbf{y} - \mathbf{x})), \mathbf{y} - \mathbf{x} \rangle dt - \langle \nabla F(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \right| \\ &= \left| \int_0^1 \langle \nabla F(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla F(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle dt \right| \\ &\leq \left| \int_0^1 Lt \|\mathbf{y} - \mathbf{x}\|^2 dt \right| = \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|^2. \end{aligned}$$

This gives the desired result. \square

Analogously, one can show the following by considering the local second-order approximation around \mathbf{x} .

Lemma A.2. *Suppose F is twice differentiable, and for all $\mathbf{u} \in \overline{\mathbf{x}\mathbf{y}}$ (again the line segment), we have*

$$\|\nabla^2 F(\mathbf{u}) - \nabla^2 F(\mathbf{x})\|_{\text{op}} \leq L \|\mathbf{u} - \mathbf{x}\|.$$

Then,

$$F(\mathbf{y}) \leq F(\mathbf{x}) + \langle \nabla F(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{1}{2} (\mathbf{y} - \mathbf{x})^\top \nabla^2 F(\mathbf{x}) (\mathbf{y} - \mathbf{x}) + \frac{L}{6} \|\mathbf{y} - \mathbf{x}\|^3.$$

Proof. Similarly to the proof of [Lemma A.1](#), we show this via the proof of Lemma 1 in [Nesterov and Polyak \(2006\)](#). Analogously as in the proof of [Lemma A.1](#), one can readily verify that for any $\mathbf{y}' \in \overline{\mathbf{x}\mathbf{y}}$, $\mathbf{x} + t(\mathbf{y}' - \mathbf{x}) \in \overline{\mathbf{x}\mathbf{y}}$ holds for all $t \in [0, 1]$. Hence for all $t \in [0, 1]$, applying the condition of this Lemma,

$$\|\nabla^2 F(\mathbf{x} + t(\mathbf{y}' - \mathbf{x})) - \nabla^2 F(\mathbf{x})\|_{\text{op}} \leq Lt \|\mathbf{y}' - \mathbf{x}\|.$$

Thus for any $\mathbf{y}' \in \overline{\mathbf{x}\mathbf{y}}$, by Cauchy-Schwartz and the above, we obtain

$$\begin{aligned} \|\nabla F(\mathbf{y}') - \nabla F(\mathbf{x}) - \langle \nabla^2 F(\mathbf{x}), \mathbf{y}' - \mathbf{x} \rangle\| &= \left\| \int_0^1 \langle \nabla^2 F(\mathbf{x} + t(\mathbf{y}' - \mathbf{x})), \mathbf{y}' - \mathbf{x} \rangle dt - \langle \nabla^2 F(\mathbf{x}), \mathbf{y}' - \mathbf{x} \rangle \right\| \\ &= \left\| \int_0^1 \langle \nabla^2 F(\mathbf{x} + t(\mathbf{y}' - \mathbf{x})) - \nabla^2 F(\mathbf{x}), \mathbf{y}' - \mathbf{x} \rangle dt \right\| \\ &\leq \left| \int_0^1 Lt \|\mathbf{y}' - \mathbf{x}\|^2 dt \right| = \frac{L}{2} \|\mathbf{y}' - \mathbf{x}\|^2. \end{aligned}$$

Applying the above relation for $\mathbf{y}' = \mathbf{x} + t(\mathbf{y} - \mathbf{x})$ which is in $\overline{\mathbf{x}\mathbf{y}}$ for all $t \in [0, 1]$, we obtain

$$\begin{aligned} &\left| F(\mathbf{y}) - F(\mathbf{x}) - \langle \nabla F(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle - \frac{1}{2} \langle \nabla^2 F(\mathbf{x})(\mathbf{y} - \mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \right| \\ &= \left| \int_0^1 \langle \nabla F(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla F(\mathbf{x}) - t \nabla^2 F(\mathbf{x})(\mathbf{y} - \mathbf{x}), \mathbf{y} - \mathbf{x} \rangle dt \right| \end{aligned}$$

$$\begin{aligned}
&= \left| \int_0^1 \langle \nabla F(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla F(\mathbf{x}) - \nabla^2 F(\mathbf{x}) \cdot t(\mathbf{y} - \mathbf{x}), \mathbf{y} - \mathbf{x} \rangle dt \right| \\
&\leq \int_0^1 \|\mathbf{y} - \mathbf{x}\| \cdot \frac{L}{2} \|t(\mathbf{y} - \mathbf{x})\|^2 dt = \frac{L}{6} \|\mathbf{y} - \mathbf{x}\|^3.
\end{aligned}$$

This gives the desired result. \square

We will also use the following Lemmas.

Lemma A.3. For vectors \mathbf{a}, \mathbf{b} , the matrix operator norm $\|\mathbf{a}\mathbf{b}^\top\|_{\text{op}} \leq \|\mathbf{a}\| \|\mathbf{b}\|$.

Proof. Consider any unit vector \mathbf{x} . By Cauchy-Schwartz and associativity, we have

$$\mathbf{x}^\top (\mathbf{a}\mathbf{b}^\top) \mathbf{x} \leq \langle \mathbf{x}, \mathbf{a} \rangle \langle \mathbf{x}, \mathbf{b} \rangle \leq \|\mathbf{x}\|^2 \|\mathbf{a}\| \|\mathbf{b}\| = \|\mathbf{a}\| \|\mathbf{b}\|.$$

The conclusion follows by definition of operator norm. \square

Lemma A.4. Consider any non-negative, continuous function $g(x)$ such that $\lim_{x \rightarrow \infty} g(x) = \infty$ and such that $g(x) > 0$ on $[1, \infty)$. Then on $[1, \infty)$, $g(x)$ can be lower bounded by a strictly positive, infinitely differentiable, strictly increasing function $\tilde{g}(x)$, where \tilde{g} has domain $[1, \infty)$.

Proof. We will explicitly construct such a \tilde{g} in terms of g . First, since $\lim_{x \rightarrow \infty} g(x) = \infty$, for all $i \geq 1$, there exists $t_i \in [1, \infty)$ such that $g(x) \geq i + 1$ for all $x \geq t_i$. We furthermore can clearly assume $2 \leq t_1 < t_2 < \dots$, by increasing each t_N if necessary. Also let $t_0 = 1$. Thus $\cup_{i \geq 0} [t_i, t_{i+1})$ forms a disjoint union of $[1, \infty)$.

Now, let $c = \min(1, \inf_{x \in [1, t_1]} g(x)) > 0$; the strict inequality here holds as $t_1 < \infty$ and as g is continuous. Define a sequence $\{b_i\}_{i \geq 0}$ by $b_0 = c/2, b_1 = c$, and $b_i = i$ for all $i \geq 2$. Thus $b_0 < b_1 < \dots$. Furthermore, this construction of $\{b_i\}_{i \geq 0}$ implies for all $i \geq 0$, we have $g(x) \geq b_{i+1}$ for all $x \in [t_i, t_{i+1}]$.

Now construct $\tilde{g}(x)$ as follows. For all $i \geq 0$, we let $\tilde{g}(x)$ equal a function $h_i(x)$ defined on $[t_i, t_{i+1}]$ such that $h_i(t_i) = b_i, h_i(t_{i+1}) = b_{i+1}$, where we define h_i as follows. We first define $h : [0, 1] \rightarrow [0, 1]$ such that h is infinitely differentiable, $h(0) = 0, h(1) = 1, h^{(n)}(0) = h^{(n)}(1) = 0$ for all $n \geq 1$ where $h^{(n)}$ denotes the n -th derivative, and $h'(x) > 0$ for all $x \in (0, 1)$. To this end we use a construction from [Chen and Sridharan \(2025\)](#): let

$$h(x) = \frac{e^{-\frac{1}{x^2}}}{e^{-\frac{1}{x^2}} + e^{-\frac{1}{1-x^2}}} \text{ on } (0, 1),$$

and extend h to $[0, 1]$ by $h(0) = 0, h(1) = 1$. We justify these claims about h shortly below. Now we let

$$h_i(x) = (b_{i+1} - b_i) \cdot h\left(\frac{x - t_i}{t_{i+1} - t_i}\right) + b_i \text{ for all } i \geq 0.$$

We now check h satisfies the claimed properties.

- In [Chen and Sridharan \(2025\)](#), it is argued that h maps to $[0, 1]$, $h(0) = 0, h(1) = 1$, and that h is infinitely differentiable. It is also argued in [Chen and Sridharan \(2025\)](#), Lemma 11.5, that $h'(x)$ (which is called $\tilde{p}(x)$ there) is non-negative on $[0, 1]$.
- Next, we check $h^{(n)}(0) = h^{(n)}(1) = 0$ for all $n \geq 1$. Via a straightforward induction outlined in [Chen and Sridharan \(2025\)](#), one can check that $\left(e^{-\frac{1}{x^2}}\right)^{(n)} = 0, \left(e^{-\frac{1}{1-x^2}}\right)^{(n)} = 0$ for all $n \geq 1$ (following the standard convention in analysis that $0 \cdot \infty = 0$, see e.g. [Folland \(1999\)](#)). Now let $f(x) = e^{-\frac{1}{x^2}}, g(x) = e^{-\frac{1}{x^2}} + e^{-\frac{1}{1-x^2}}$, thus $h = f/g$. Consequently $f^{(n)}(0) = 0, f^{(n)}(1) = 0, g^{(n)}(0) = 0, g^{(n)}(1) = 0$ for all $n \geq 1$. As $g > 0$ always holds in $[0, 1]$ as shown in [Chen and Sridharan \(2025\)](#) and can be easily checked, we have $f = gh$. A straightforward induction gives $f^{(n)} = \sum_{k=0}^n \binom{n}{k} g^{(k)} h^{(n-k)}$ where $\binom{n}{k}$ is the binomial coefficient. We thus obtain $gh^{(n)} = f^{(n)} - \sum_{k=0}^{n-1} \binom{n}{k} g^{(k)} h^{(n-k)}$. For any $n \geq 1$,

taking $x = 0, 1$ in this expression for $h(x)$ and noting at least one of $k, n - k \geq 1$ for $0 \leq k \leq n - 1$ implies $g(0)h^{(n)}(0) = g(1)h^{(n)}(1) = 0$. Recalling $g(x) > 0$ on $[0, 1]$ proves $h^{(n)}(0) = h^{(n)}(1) = 0$ for $n \geq 1$, as requested.

- Finally, we check that $h'(x) > 0$ for all $x \in (0, 1)$. Consider any $x \in (0, 1)$. By a calculation in Lemma 11.5, [Chen and Sridharan \(2025\)](#), we have $h'(x) > 0$ if and only if $q(x) = \frac{2}{x^3} \left(e^{-\frac{1}{x^2}} + e^{-\frac{1}{1-x^2}} \right) + e^{\frac{-1}{x^2}} \cdot \frac{2}{x^3} + e^{-\frac{1}{1-x^2}} \cdot \frac{-2x}{(1-x^2)^2} > 0$. If $x \in [\frac{\sqrt{2}}{2}, 1)$, directly following the proof of Lemma 11.5 in [Chen and Sridharan \(2025\)](#) establishes that $q(x) > 0$. Otherwise if $x \in (0, \frac{\sqrt{2}}{2})$, note the strict inequality $\frac{1}{x^3} > \frac{x}{(1-x^2)^2}$, which in turn implies $q(x) > 0$.

By the above properties of h , it follows from the Chain Rule that for all $i \geq 0$, h_i satisfies the following properties:

- $h_i(t_i) = b_i$, $h_i(t_{i+1}) = b_{i+1}$, and $h_i(x) \in [b_i, b_{i+1}]$ for all $x \in [t_i, t_{i+1}]$.
- h_i is infinitely differentiable.
- $h'_i(x) > 0$ for $x \in (t_i, t_{i+1})$, and for all $x \in [t_i, t_{i+1}]$, $h'_i(x) \geq 0$.
- For all $n \geq 1$, $h_i^{(n)}(t_i) = h_i^{(n)}(t_{i+1}) = 0$, where again $h_i^{(n)}$ denotes the n -th derivative.

Finally, we check that \tilde{g} has the desired properties:

- \tilde{g} is well-defined: This follows because for all $i \geq 1$, we have $h_i(t_i) = h_{i-1}(t_i) = b_i$.
- \tilde{g} is strictly positive: This follows because $h_i(x) \in [b_i, b_{i+1}] \subseteq (0, \infty)$ for all $x \in [t_i, t_{i+1}]$.
- \tilde{g} is continuous, and moreover is infinitely differentiable: Continuity of \tilde{g} follows because each h_i is infinitely differentiable, and hence continuous, combined with the fact that for all $i \geq 1$, we have $h_i(t_i) = h_{i-1}(t_i) = b_i$. Infinite differentiability of \tilde{g} follows because each h_i is infinitely differentiable, and because for all $n \geq 1$ and all $i \geq 0$, $h_i^{(n)}(t_i) = h_i^{(n)}(t_{i+1}) = 0$.
- $\tilde{g}(x) \leq g(x)$ always holds for $x \in [1, \infty)$: Recall for all $i \geq 0$, we have $g(x) \geq b_{i+1}$ for all $x \in [t_i, t_{i+1}]$. Since we have $\tilde{g}(x) = h_i(x) \leq b_{i+1}$ for all $x \in [t_i, t_{i+1}]$, it follows that for all $x \in [t_i, t_{i+1}]$, $\tilde{g}(x) \leq g(x)$. The result follows upon recalling that $\bigcup_{i \geq 0} [t_i, t_{i+1})$ forms a disjoint union of $[1, \infty)$.
- \tilde{g} is strictly increasing: Consider any $x_1 < x_2, x_1, x_2 \in [1, \infty)$. Since $x_1 < x_2$, and recalling that $\bigcup_{i \geq 0} [t_i, t_{i+1})$ forms a disjoint union of $[1, \infty)$, it follows that for some $j \geq 0$, $(x_1, x_2) \cap (t_j, t_{j+1}) \neq \emptyset$. This intersection is open, and therefore contains some open interval $(a, b) \subseteq (t_j, t_{j+1})$. Let $c' = \inf_{x \in [\frac{2a+b}{3}, \frac{a+2b}{3}]} h'_j(x) > 0$, where the strict inequality follows as $[\frac{2a+b}{3}, \frac{a+2b}{3}] \subseteq (t_j, t_{j+1})$, and by continuity of h'_j on the compact $[\frac{2a+b}{3}, \frac{a+2b}{3}]$. Since we have $h'_i(x) \geq 0$ for all $x \in [t_i, t_{i+1}]$ for all $i \geq 0$, we obtain

$$\tilde{g}(x_2) \geq 0 + c' \cdot \frac{b-a}{3} + \tilde{g}(x_1) > \tilde{g}(x_1).$$

This proves that \tilde{g} is strictly increasing as claimed.

Thus, we have constructed a function \tilde{g} that satisfies the requested properties. \square

A.2 Comparison of Assumptions with Literature

Here, we establish that our regularity conditions are more general than those of literature.

Proposition A.1. *If $\|\nabla^2 F(\mathbf{w})\| \leq l(\nabla F(\mathbf{w}))$ for non-decreasing, differentiable sub-quadratic l (where sub-quadratic means that $\lim_{x \rightarrow \infty} \frac{l(x)}{x^2} = 0$), then our [Assumption 1.1](#) is satisfied for some non-decreasing $\rho_1(x)$. In this generality, $\rho_1(x)$ depends on $l(x)$, and can be found explicitly from the construction from [Lemma A.4](#).*

Furthermore, suppose F is (L_0, L_1) -smooth, that $\|\nabla^2 F(\mathbf{w})\| \leq L_0 + L_1 \|\nabla F(\mathbf{w})\|$ for $L_0, L_1 \geq 0$. Then [Assumption 1.1](#) is satisfied with $\rho_1(x) = \frac{3}{2}L_0 + 4L_1^2x$.

Proof. Essentially this follows from Lemma 3.5, Li et al. (2023a), where it is shown that these assumptions of Zhang et al. (2019), Li et al. (2023a) imply an upper bound on $\|\nabla F(\mathbf{w})\|$ in terms of an increasing function of $F(\mathbf{w})$; combining with the assumptions of Zhang et al. (2019); Li et al. (2023a) implies that $\|\nabla^2 F(\mathbf{w})\|$ is upper bounded in terms of an increasing function of $F(\mathbf{w})$.

Proof for general l : Consider any $\mathbf{w} \in \mathbb{R}^d$. By Lemma 3.5 of Li et al. (2023a),

$$\|\nabla F(\mathbf{w})\|^2 \leq 2\ell(2\|\nabla F(\mathbf{w})\|) \cdot F(\mathbf{w}).$$

This implies

$$\frac{4\|\nabla F(\mathbf{w})\|^2}{\ell(2\|\nabla F(\mathbf{w})\|)} \leq 8F(\mathbf{w}).$$

Let $2\|\nabla F(\mathbf{w})\| = t$. Consider when $t \geq 2$. Then the left hand side equals $\frac{t^2}{l(t)}$. Note that WLOG, we can add 1 to $l(\cdot)$ so that $l(t) \geq 1$ for $t \geq 1$. Thus $\frac{t^2}{l(t)}$ is continuous on $[1, \infty)$, and furthermore is positive on this interval. Now note $\lim_{x \rightarrow \infty} \frac{x^2}{l(x)} = \infty$ by the condition (including after adding 1 WLOG), and thus by Lemma A.4, $\frac{x^2}{l(x)}$ is lower bounded by some strictly increasing function $\tilde{g}(x)$ on $[2, \infty)$. Therefore, \tilde{g} is invertible and so we have

$$\tilde{g}(2\|\nabla F(\mathbf{w})\|) \leq \frac{4\|\nabla F(\mathbf{w})\|^2}{\ell(2\|\nabla F(\mathbf{w})\|)} \leq 8F(\mathbf{w}) \implies \|\nabla F(\mathbf{w})\| \leq \frac{1}{2}\tilde{g}^{-1}(8F(\mathbf{w})).$$

Then by the assumptions of Li et al. (2023b), it holds that

$$\|\nabla^2 F(\mathbf{w})\| \leq l\left(\frac{1}{2}\tilde{g}^{-1}(8F(\mathbf{w}))\right).$$

Else when $t < 2$, we have $\|\nabla F(\mathbf{w})\| \leq 1$, and by the assumptions of Li et al. (2023b), we have $\|\nabla^2 F(\mathbf{w})\| \leq l(1)$.

Thus the assumptions of Li et al. (2023b) imply that the following always holds:

$$\|\nabla^2 F(\mathbf{w})\| \leq l\left(\frac{1}{2}\tilde{g}^{-1}(8F(\mathbf{w}))\right) + l(1).$$

We thus can take $\rho_1(x) = l\left(\frac{1}{2}\tilde{g}^{-1}(8x)\right) + l(1)$, which is clearly non-negative. It remains to check that $l\left(\frac{1}{2}\tilde{g}^{-1}(8x)\right)$ is non-decreasing. As l is non-decreasing, as compositions of non-decreasing functions are non-decreasing, it remains to check that $\frac{1}{2}\tilde{g}^{-1}(8x)$ is non-decreasing. Since \tilde{g} is non-decreasing, \tilde{g}^{-1} is non-decreasing as well, and this completes the proof.

Proof for (L_0, L_1) -smoothness: First, when $L_1 = 0$ the result is immediate, so from here on out suppose $L_1 > 0$. By Lemma 3.5 from Li et al. (2023a) we have for all $\mathbf{w} \in \mathbb{R}^d$,

$$\|\nabla F(\mathbf{w})\|^2 \leq 2\ell(2\|\nabla F(\mathbf{w})\|) \cdot F(\mathbf{w}),$$

where $\ell(x) = L_0 + L_1(x)$ for $L_0, L_1 \geq 0$. We thus obtain:

$$\begin{aligned} \|\nabla F(\mathbf{w})\|^2 &\leq 2(L_0 + 2L_1\|\nabla F(\mathbf{w})\|) \cdot F(\mathbf{w}) \\ &= 2L_0F(\mathbf{w}) + 4L_1\|\nabla F(\mathbf{w})\|F(\mathbf{w}). \end{aligned}$$

Rewriting this inequality, we get

$$\|\nabla F(\mathbf{w})\|^2 - 4L_1\|\nabla F(\mathbf{w})\|F(\mathbf{w}) - 2L_0F(\mathbf{w}) \leq 0.$$

Consider the quadratic $x^2 - 4L_1F(\mathbf{w}) \cdot x - 2L_0F(\mathbf{w})$. The coefficient on the quadratic term is positive, and the quadratic is non-negative when $x = \|\nabla F(\mathbf{w})\|$. Thus $\|\nabla F(\mathbf{w})\|$ must be no larger than the largest root of $x^2 - 4L_1F(\mathbf{w}) \cdot x - 2L_0F(\mathbf{w})$, and we obtain

$$\begin{aligned} \|\nabla F(\mathbf{w})\| &\leq \frac{1}{2} \left(4L_1F(\mathbf{w}) + \sqrt{16L_1^2F(\mathbf{w})^2 + 8L_0F(\mathbf{w})} \right) \\ &\leq 2L_1F(\mathbf{w}) + \sqrt{(2L_1F(\mathbf{w}))^2 + 2L_0F(\mathbf{w})} \end{aligned} \tag{5}$$

If $F(\mathbf{w}) = 0$, the above immediately implies $\|\nabla F(\mathbf{w})\| = 0$. Otherwise, recall by shifting (in Notation) that $F(\mathbf{w}) \geq 0$ always holds, so suppose $F(\mathbf{w}) > 0$. Recall also from earlier that it suffices to show the result for $L_1 > 0$. Applying the inequality $\sqrt{a^2 + b} \leq a + \frac{b}{2a}$, valid for all $a > 0, b \geq 0$ with $a = 2L_1F(\mathbf{w}) > 0, b = 2L_0F(\mathbf{w}) \geq 0$, we obtain

$$\sqrt{(2L_1F(\mathbf{w}))^2 + 2L_0F(\mathbf{w})} \leq 2L_1F(\mathbf{w}) + \frac{L_0}{2L_1}.$$

Substituting into (5) gives that for all \mathbf{w} with $F(\mathbf{w}) > 0$, we have

$$\|\nabla F(\mathbf{w})\| \leq \frac{L_0}{2L_1} + 4L_1F(\mathbf{w}). \quad (6)$$

By the argument earlier, if $F(\mathbf{w}) = 0$, the above bound (6) holds too. Thus (6) holds for all $\mathbf{w} \in \mathbb{R}^d$. Now inserting (6) into the definition of (L_0, L_1) -smoothness gives

$$\|\nabla^2 F(\mathbf{w})\| \leq L_0 + L_1 \left(\frac{L_0}{2L_1} + 4L_1F(\mathbf{w}) \right) = \frac{3}{2}L_0 + 4L_1^2F(\mathbf{w}).$$

Hence Assumption 1.1 is satisfied with the increasing function $\rho_1(x) = \frac{3}{2}L_0 + 4L_1^2x$. \square

Proposition A.2. When F is (L_0, L_1) -smooth, letting $\rho_0(x) = 2L_0^{1/2}x^{1/2} + \frac{5L_1^2}{L_0^{1/2}}x^{3/2}$, we have $\|\nabla F(\mathbf{w})\| \leq \rho_0(F(\mathbf{w}))$.

Proof. By Proposition A.1, we can take $\rho_1(x) = \frac{3}{2}L_0 + 4L_1^2x$ in this case. As noted in Subsection 3.1, we need to show that $2L_0^{1/2}x^{1/2} + \frac{5L_1^2}{L_0^{1/2}}x^{3/2}$ is a pointwise upper bound on

$$\rho_1(x)\sqrt{2\theta(x)} \text{ where } \theta(x) = \int_0^x \frac{1}{\rho_1(v)} dv.$$

To this end note for each $x \geq 0$ that $\theta(x) \leq x \cdot \frac{1}{\frac{3}{2}L_0} = \frac{2}{3L_0}x$, thus for each $x \geq 0$,

$$\rho_1(x)\sqrt{2\theta(x)} \leq \left(\frac{3}{2}L_0 + 4L_1^2x \right) \sqrt{\frac{4}{3L_0}x} \leq 2L_0^{1/2}x^{1/2} + \frac{5L_1^2}{L_0^{1/2}}x^{3/2}.$$

This completes the proof. \square

Example 1. We now provide a natural example of a univariate function that satisfies our regularity assumptions but does not necessarily satisfy those of Li et al. (2023b) for non-convex optimization. Namely, consider the univariate function:

$$F(x) = 1 - \log(\cos(1+x)), 0 \leq x < \frac{\pi}{2} - 1.$$

The argument here is in radians. The first derivative is:

$$F'(x) = \tan(1+x).$$

The second derivative is:

$$F''(x) = \sec^2(1+x).$$

Thus as $\tan^2(\theta) + 1 = \sec^2(\theta)$, F satisfies the ODE:

$$F''(x) = F'(x)^2 + 1. \quad (7)$$

Suppose that F satisfied the conditions of Li et al. (2023b) for non-convex optimization on the relevant domain, thus for all $0 \leq x < \frac{\pi}{2} - 1$, we would have

$$F''(x) \leq \ell(F'(x)),$$

for some sub-quadratic $\ell(\cdot)$.

Then by (7) and noting $F'(x) > 0$ on the domain, we obtain for all $0 \leq x < \frac{\pi}{2} - 1$

$$1 \leq 1 + \frac{1}{F'(x)^2} = \frac{F'(x)^2 + 1}{F'(x)^2} = \frac{F''(x)}{F'(x)^2} \leq \frac{\ell(F'(x))}{F'(x)^2}.$$

As l is subquadratic, there exists $x' < \infty$ such that $l(x)/x^2 < 1$ for all $x > x'$. Noting $F'(x) \rightarrow \infty$ for $x \rightarrow \frac{\pi}{2} - 1$ yields a contradiction.

Consequently F does not satisfy the conditions of Li et al. (2023b) for non-convex optimization. However, we show that F satisfies Assumption 1.1. Rewriting $F''(x)$ in terms of $F(x)$, note that:

$$\cos(1+x) = e^{1-F(x)},$$

and thus:

$$F''(x) = \sec^2(1+x) = \frac{1}{\cos^2(1+x)} = e^{2(F(x)-1)}.$$

Hence we can define the increasing, non-negative function

$$\rho_1(t) = e^{2(t-1)},$$

which satisfies:

$$F''(x) \leq \rho_1(F(x)).$$

Thus F satisfies Assumption 1.1 (in the relevant domain).

We now discuss Assumption 1.2.

Example 2. First, we show that Assumption 1.2 captures several univariate functions of interest. Notice also if $F(\mathbf{w})$ is a sum of functions satisfying Assumption 1.2, Triangle Inequality implies that $F(\mathbf{w})$ also satisfies Assumption 1.2.

- **Polynomials:** Consider whenever $F(x)$ is a linear combination of monomials x^p for $p \geq 1$, combined with a constant term. We claim $F(x)$ satisfies Assumption 1.2. By linearity of derivative and Triangle Inequality, it suffices to prove this whenever $F(x) = x^p$ for $p \geq 1$ as the constant term vanishes, and then add up all the non-decreasing, non-negative functions on the right hand side to form ρ_1 and ρ_2 . To this end note $F''(x) = p(p-1)x^{p-2}$, thus

$$|F''(x)| = p(p-1)x^{p-2} \leq p(p-1)(x^p + 1) = p(p-1)(F(x) + 1).$$

Similarly, $F'''(x) = p(p-1)(p-2)x^{p-3}$, thus

$$|F'''(x)| = p(p-1)(p-2)x^{p-3} \leq p(p-1)(p-2)(1 + F(x)).$$

Noting $p(p-1)(1+t)$ and $p(p-1)(p-2)(1+t)$ are non-decreasing and non-negative for $t \geq 0$, combined with our earlier remarks that it suffices to prove this result when $F(x) = x^p$, completes the proof.

- **Single-exponential functions:** Consider when $F(x) = a^x = e^{x \ln a}$ for $a > 1$. Then $F''(x) = (\ln a)^2 e^{x \ln a}$, $F'''(x) = (\ln a)^3 e^{x \ln a}$, and so we can take $\rho_1(t) = (\ln a)^2 t$, $\rho_2(t) = (\ln a)^3 t$.
- **Doubly-exponential functions:** Consider when $F(x) = a^{b^x} = e^{\ln a e^{x \ln b}}$ for $a, b > 1$. Thus

$$F'(x) = e^{\ln a e^{x \ln b}} \cdot \ln a e^{x \ln b} \cdot \ln b = \ln a \ln b F(x) e^{x \ln b}.$$

It follows that

$$F''(x) = \ln a \ln b (F'(x) e^{x \ln b} + \ln b F(x) e^{x \ln b}) = (\ln a)(\ln b)^2 F(x) (e^{2x \ln b} \ln a + e^{x \ln b}).$$

This then implies

$$\begin{aligned} F'''(x) &= (\ln a)(\ln b)^2 F(x) (e^{2x \ln b} 2 \ln a \ln b + e^{x \ln b} \ln b) \\ &\quad + (\ln a)(\ln b)^2 (e^{2x \ln b} \ln a + e^{x \ln b}) \ln a \ln b F(x) e^{x \ln b} \\ &= (\ln a)(\ln b)^3 F(x) (2e^{2x \ln b} \ln a + e^{x \ln b} + e^{3x \ln b} (\ln a)^2 + e^{2x \ln b} \ln a). \end{aligned}$$

Notice

$$e^{x \ln b} \leq e^{\ln a e^{x \ln b}} - 1 < F(x),$$

therefore we have

$$\begin{aligned} F''(x) &\leq (\ln a)(\ln b)^2 F(x) (F(x)^2 \ln a + F(x)), \\ F'''(x) &\leq (\ln a)(\ln b)^3 F(x) (F(x)^3 (\ln a)^2 + 3F(x)^2 \ln a + F(x)). \end{aligned}$$

We thus can take

$$\begin{aligned} \rho_1(t) &= (\ln a)(\ln b)^2 t(t^2 \ln a + t), \\ \rho_2(t) &= (\ln a)(\ln b)^3 t(t^3 (\ln a)^2 + 3t^2 \ln a + t), \end{aligned}$$

which are clearly non-negative and non-decreasing on $[0, \infty)$.

- Next we highlight the natural example of any self-concordant function $F : \mathbb{R} \rightarrow \mathbb{R}$. Thus

$$|F'''(x)| \leq 2F''(x)^{3/2} \leq 2|F''(x)|^{3/2}.$$

Suppose F satisfies [Assumption 1.1](#). Then there exists a non-negative, non-decreasing ρ_1 such that $|F''(x)| \leq \rho_1(F(x))$. Thus,

$$|F'''(x)| \leq 2\rho_1(F(x))^{3/2}.$$

Since ρ_1 is non-negative and non-decreasing, $\rho_2(t) := 2\rho_1(t)^{3/2}$ is as well, and thus [Assumption 1.2](#) is satisfied.

Next, we show that the regularity assumptions Assumptions 1 and 3 of [Xie et al. \(2024\)](#), which they need for their guarantees finding SOSPs, are less general than [Assumption 1.2](#) when F is twice-differentiable. To do so we show they imply [Assumption 1.2](#), and are hence subsumed by [Assumption 1.2](#).

When F is twice-differentiable, their Assumption 1 implies (L_0, L_1) -smoothness. As shown in [Proposition A.2](#), this means that

$$\|\nabla F(\mathbf{w})\| \leq \rho_0(F(\mathbf{w})) \text{ where } \rho_0(x) = 2L_0^{1/2}x^{1/2} + \frac{5L_1^2}{L_0^{1/2}}x^{3/2}.$$

Their Assumption 3 implies for $M_0, M_1 \geq 0$ and some $\delta > 0$ that for all \mathbf{w}, \mathbf{w}' with $\|\mathbf{w} - \mathbf{w}'\| \leq \delta$,

$$\|\nabla^2 F(\mathbf{w}) - \nabla^2 F(\mathbf{w}')\|_{\text{op}} \leq \|\mathbf{w} - \mathbf{w}'\| (M_0 + M_1 \|\nabla F(\mathbf{w})\|).$$

Combining this with the earlier display gives for all \mathbf{w}, \mathbf{w}' with $\|\mathbf{w} - \mathbf{w}'\| \leq \delta$,

$$\|\nabla^2 F(\mathbf{w}) - \nabla^2 F(\mathbf{w}')\|_{\text{op}} \leq \|\mathbf{w} - \mathbf{w}'\| (M_0 + M_1 \rho_0(F(\mathbf{w}))),$$

where $\rho_0(x) = 2L_0^{1/2}x^{1/2} + \frac{5L_1^2}{L_0^{1/2}}x^{3/2}$. We thus see that F satisfies [Assumption 1.2](#) with the non-

decreasing, non-negative function $\rho_2(x) = M_0 + M_1 \left(2L_0^{1/2}x^{1/2} + \frac{5L_1^2}{L_0^{1/2}}x^{3/2} \right)$, where the latter two properties are evident as $\rho_0(\cdot)$ is non-decreasing and non-negative.

A.3 Proofs of Technical Results

Now, we prove general results used throughout our work. We prove [Corollary 1](#), which gives us control over the gradient:

Proof of [Corollary 1](#). Applying Lemma 11, [De Sa et al. \(2022\)](#) with Φ in place of F , we obtain

$$\|\nabla F(\mathbf{w})\| \leq \rho(F(\mathbf{w})) \sqrt{2\theta(F(\mathbf{w}))} = \rho_0(F(\mathbf{w})),$$

where $\theta(\cdot)$ is defined as in the statement of [Corollary 1](#). To prove $\rho_0(x)$ is increasing, simply note θ and thus $\sqrt{\theta}$ are clearly increasing, and are both non-negative. ρ_1 is non-decreasing and non-negative as well, thus ρ_0 is non-decreasing and non-negative. \square

We also prove the central [Lemma 3.1](#), which is very important to our results: it lets us control the change in function value under our regularity assumptions. We first state the following Lemma from [Li et al. \(2023a\)](#), a generalization of Gronwall's Inequality:

Lemma A.5 (Lemma A.3, Li et al. (2023a)). Let $\alpha : [a, b] \rightarrow [0, \infty)$ and $\beta : [0, \infty) \rightarrow [0, \infty)$ be two continuous functions. Suppose $\alpha'(t) \leq \beta(\alpha(t))$ almost everywhere over (a, b) . Let $\phi(u) = \int_0^u \frac{1}{\beta(v)} dv$. Then for all $t \in [a, b]$,

$$\phi(\alpha(t)) \leq \phi(\alpha(a)) - a + t.$$

This allows us to prove Lemma 3.1, which is an extension of Lemma A.4, Li et al. (2023a):

Proof of Lemma 3.1. The proof is essentially identical to the proof of Lemma A.4, Li et al. (2023a). Let $z(t) = (1-t)\mathbf{x} + t\mathbf{y}$, $\alpha(t) = F(z(t))$. Then for all $t \in (0, 1)$, we obtain

$$\begin{aligned} \alpha'(t) &= \lim_{s \rightarrow t} \frac{\alpha(s) - \alpha(t)}{s - t} \\ &\leq \lim_{s \rightarrow t} \frac{|F(z(s)) - F(z(t))|}{s - t} \\ &= \left| \lim_{s \rightarrow t} \frac{F(z(s)) - F(z(t))}{s - t} \right| \\ &= \left| \frac{d}{dt} F(z(t)) \right| \\ &= |\nabla F(z(t))^\top (\mathbf{y} - \mathbf{x})| \\ &\leq \rho_0(F(z(t))) \|\mathbf{y} - \mathbf{x}\|, \end{aligned}$$

the last step using $\|\nabla F(\mathbf{w})\| \leq \rho_0(F(\mathbf{w}))$. Let $\beta(x) = \|\mathbf{y} - \mathbf{x}\| \rho_0(x)$ and let $\phi(u) = \int_0^u \frac{1}{\beta(v)} dv$. Thus, $\alpha'(t) \leq \beta(\alpha(t))$ almost everywhere. Applying Lemma A.5 gives

$$\phi(F(\mathbf{y})) = \phi(\alpha(1)) \leq \phi(\alpha(0)) + 1 = \phi(F(\mathbf{x})) + 1.$$

Let $\psi(u) = \|\mathbf{y} - \mathbf{x}\| \phi(u) = \int_0^u \frac{1}{\rho_0(v)} dv$, which is clearly strictly increasing. Consequently we obtain from the above and assumption on \mathbf{y} that

$$\begin{aligned} \psi(F(\mathbf{y})) &\leq \psi(F(\mathbf{x})) + \|\mathbf{y} - \mathbf{x}\| \\ &\leq \psi(F(\mathbf{x})) + \frac{1}{\rho_0(F(\mathbf{x}) + 1)} \\ &\leq \int_0^{F(\mathbf{x})} \frac{1}{\rho_0(v)} dv + \int_{F(\mathbf{x})}^{F(\mathbf{x})+1} \frac{1}{\rho_0(v)} dv \\ &= \int_0^{F(\mathbf{x})+1} \frac{1}{\rho_0(v)} dv = \psi(F(\mathbf{x}) + 1). \end{aligned}$$

Since ψ is strictly increasing, taking inverses implies

$$F(\mathbf{y}) \leq F(\mathbf{x}) + 1,$$

as desired. \square

We also introduce the following Lemma, which lets us exploit Assumption 1.2 to control the Lipschitz constant of the Hessian of F .

Lemma A.6. Suppose F satisfies Assumption 1.2. Suppose $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ are such that $\|\mathbf{y} - \mathbf{x}\| \leq r$ for some $r > 0$. Then

$$\|\nabla^2 F(\mathbf{x}) - \nabla^2 F(\mathbf{y})\|_{\text{op}} \leq \|\mathbf{x} - \mathbf{y}\| \cdot \sup_{\mathbf{u} \in \mathbb{B}(\mathbf{x}, r)} \rho_2(F(\mathbf{u})).$$

In particular, we have

$$\|\nabla^2 F(\mathbf{x}) - \nabla^2 F(\mathbf{y})\|_{\text{op}} \leq \|\mathbf{x} - \mathbf{y}\| \cdot \sup_{\mathbf{u} \in \mathbb{B}(\mathbf{y}, r)} \rho_2(F(\mathbf{u})).$$

Proof. Consider $\delta > 0$, either from Assumption 1.2 if the second case of Assumption 1.2 holds, and otherwise set to some arbitrary positive real. Similar to the proof of Lemma 3.1, divide the line segment between \mathbf{x}, \mathbf{y} into $N = \frac{\|\mathbf{x} - \mathbf{y}\|}{\delta}$ equally spaced segments of length δ between points \mathbf{x}_i , where we define $\mathbf{x}_0 = \mathbf{x}, \mathbf{x}_1, \dots, \mathbf{x}_{N-1}, \mathbf{x}_N = \mathbf{y}$. Thus $\|\mathbf{x} - \mathbf{y}\| = N\delta$.

Suppose for all $\mathbf{u} \in \overline{\mathbf{x}\mathbf{y}}$ we have $\|\nabla^3 F(\mathbf{u})\|_{\text{op}} \leq L$. Consider any \mathbf{x}', \mathbf{y}' in the line segment $\overline{\mathbf{x}\mathbf{y}}$. Applying this for $\mathbf{x}' + t(\mathbf{y}' - \mathbf{x}')$ for $t \in [0, 1]$, which always lies in the line segment $\overline{\mathbf{x}\mathbf{y}}$, we obtain

$$\|\nabla^2 F(\mathbf{y}') - \nabla^2 F(\mathbf{x}')\|_{\text{op}} \leq \left\| \int_0^1 \langle \nabla^3 F(\mathbf{x}' + t(\mathbf{y}' - \mathbf{x}')), \mathbf{y}' - \mathbf{x}' \rangle dt \right\| \leq L \|\mathbf{y}' - \mathbf{x}'\|.$$

Consequently irrespective of which case of [Assumption 1.2](#) holds, because $\|\mathbf{x}_i - \mathbf{x}_{i-1}\| \leq \delta$, we have for each $i, 1 \leq i \leq N$ that

$$\|\nabla^2 F(\mathbf{x}_i) - \nabla^2 F(\mathbf{x}_{i-1})\|_{\text{op}} \leq \|\mathbf{x}_i - \mathbf{x}_{i-1}\| \sup_{\mathbf{u} \in \overline{\mathbf{x}\mathbf{y}}} \rho_2(F(\mathbf{u})).$$

Now Triangle Inequality gives

$$\begin{aligned} \|\nabla^2 F(\mathbf{x}) - \nabla^2 F(\mathbf{y})\|_{\text{op}} &\leq \sum_{i=1}^N \|\nabla^2 F(\mathbf{x}_i) - \nabla^2 F(\mathbf{x}_{i-1})\|_{\text{op}} \\ &\leq \sum_{i=1}^N \|\mathbf{x}_i - \mathbf{x}_{i-1}\| \sup_{\mathbf{u} \in \overline{\mathbf{x}\mathbf{y}}} \rho_2(F(\mathbf{u})) \\ &\leq N\delta \cdot \sup_{\mathbf{u} \in \overline{\mathbf{x}\mathbf{y}}} \rho_2(F(\mathbf{u})) \\ &= \|\mathbf{x} - \mathbf{y}\| \sup_{\mathbf{u} \in \overline{\mathbf{x}\mathbf{y}}} \rho_2(F(\mathbf{u})), \end{aligned}$$

as desired. \square

We will also generalize the proof of [Theorem 3.1](#) to show that GD, when initialized in the $F(\mathbf{w}_0)$ -sublevel set $\mathcal{L}_{F,F(\mathbf{w}_0)}$ with appropriate step size defined in terms of $F(\mathbf{w}_0)$, never increases function value.

Lemma A.7. *Consider any $\mathbf{w}_0 \in \mathbb{R}^d$, and consider iterates $\{\mathbf{u}_t\}_{t \geq 0}$ of GD initialized at any $\mathbf{u}_0 \in \mathcal{L}_{F,F(\mathbf{w}_0)}$, the $F(\mathbf{w}_0)$ -sublevel set. If the step size η of GD is at most $\frac{1}{L_1(\mathbf{w}_0)}$ where $L_1(\cdot)$ is defined as per (4), then $F(\mathbf{u}_t) \leq F(\mathbf{u}_0)$ for all $t \geq 0$.*

Proof. It suffices to prove this for $t = 1$; a simple inductive argument then establishes this for all $t \geq 0$. We have $\mathbf{u}_1 = \mathbf{u}_0 - \eta \nabla F(\mathbf{u}_0)$. By [Corollary 1](#) and because $\mathbf{u}_0 \in \mathcal{L}_{F,F(\mathbf{w}_0)}$, $\|\nabla F(\mathbf{u}_0)\| \leq \rho_0(F(\mathbf{u}_0)) \leq \rho_0(F(\mathbf{w}_0))$. Thus by choice of η and definition of $L_1(\mathbf{w}_0)$,

$$\|\mathbf{u}_1 - \mathbf{u}_0\| = \eta \|\nabla F(\mathbf{u}_0)\| \leq \eta \rho_0(F(\mathbf{w}_0)) \leq \frac{1}{\rho_0(F(\mathbf{w}_0) + 1)}.$$

By [Lemma 3.2](#), because $\mathbf{u}_0 \in \mathcal{L}_{F,F(\mathbf{w}_0)}$, for all \mathbf{p} in the line segment $\overline{\mathbf{u}_0 \mathbf{u}_1}$ we have $\|\nabla^2 F(\mathbf{p})\|_{\text{op}} \leq L_1(\mathbf{w}_0)$. By [Lemma A.1](#), it follows that

$$\begin{aligned} F(\mathbf{u}_1) &\leq F(\mathbf{u}_0) - \eta \|\nabla F(\mathbf{u}_0)\|^2 + \frac{L_1(\mathbf{w}_0)\eta^2}{2} \cdot \|\nabla F(\mathbf{u}_0)\|^2 \\ &\leq F(\mathbf{u}_0) + \|\nabla F(\mathbf{u}_0)\|^2 \cdot \left(-\eta + \frac{L_1(\mathbf{w}_0)\eta^2}{2} \right). \end{aligned}$$

Noting $-\eta + \frac{L_1(\mathbf{w}_0)\eta^2}{2} \leq 0$ for $\eta \in \left[0, \frac{2}{L_1(\mathbf{w}_0)}\right]$, the conclusion follows. \square

B Proof of Framework

Proof of Theorem 2.1. For convenience, for all $n \geq 0$, define $p_n := 1 - n \cdot \sup_{\mathbf{u} \in \mathcal{L}_{F,F(\mathbf{w}_0)}} \delta(\mathbf{u})$. Also let $T = \sup_{\mathbf{u} \in \mathcal{L}_{F,F(\mathbf{w}_0)}} \left\{ \frac{F(\mathbf{w}_0)}{\Delta(\mathbf{u})} \right\}$.

Lemma B.1. *For any $n \geq 0$, let \mathcal{E}_n be the event that the sequence of iterates $(\mathbf{w}_t)_{0 \leq t \leq n-1}$ satisfies either:*

1. The event $\mathcal{E}_{n,1}$: For all $0 \leq t \leq n-1$, $F(\mathcal{A}_1(\mathbf{w}_t)) < F(\mathbf{w}_t) - \Delta(\mathbf{w}_t)$.

2. The event $\mathcal{E}_{n,2}$: There exists $\mathbf{w}_t \in (\mathbf{w}_t)_{0 \leq t \leq n-1}$ such that $\mathcal{A}_2(\mathbf{w}_t) \cap \mathcal{S} \neq \{\}$, and for all \mathbf{w}_s with $0 \leq s < t$, we have $F(\mathcal{A}_1(\mathbf{w}_s)) < F(\mathbf{w}_s) - \Delta(\mathbf{w}_s)$.

That is, $\mathcal{E}_n = \mathcal{E}_{n,1} \cup \mathcal{E}_{n,2}$. Then over the randomness in \mathcal{A} , we have $\mathbb{P}(\mathcal{E}_n) \geq p_n$ for all $n \geq 0$.

Proof. We proceed by induction on n . The base case $n = 0$ is vacuously evident, and the case $n = 1$ follows immediately by the definition of a decrease procedure from [Definition 2.2](#) and hypotheses of [Theorem 2.1](#).

For the inductive step, suppose [Lemma B.1](#) is true for some $n \geq 1$; we show it is for $n + 1$. By the inductive hypothesis, we know that $\mathbb{P}(\mathcal{E}_n) \geq p_n$. We aim to show $\mathbb{P}(\mathcal{E}_{n+1}) \geq p_{n+1}$. If $p_n \leq 0$ there is nothing to prove, so suppose now that $p_n > 0$.

1. Let $p = \mathbb{P}(\mathcal{E}_{n,2} | \mathcal{E}_n)$. Note $\mathcal{E}_{n,2} \subseteq \mathcal{E}_{n+1,2} \subseteq \mathcal{E}_{n+1}$.
2. Let $\mathcal{B} := \mathcal{E}_{n,1} \cap \mathcal{E}_{n,2}^c$. Thus, if \mathcal{B} occurs, then all the $(\mathbf{w}_t)_{0 \leq t \leq n-1}$ are such that $F(\mathcal{A}_1(\mathbf{w}_t)) < F(\mathbf{w}_t) - \Delta(\mathbf{w}_t)$, but $\mathcal{E}_{n,2}$ did not occur. Note \mathcal{E}_n is the disjoint union $\mathcal{E}_{n,2} \sqcup \mathcal{B}$, so $\mathbb{P}(\mathcal{B} | \mathcal{E}_n) = 1 - p$.

Under \mathcal{B} , we know $\mathbf{w}_n = \mathcal{A}(\mathbf{w}_{n-1})$ is such that $F(\mathbf{w}_n) \leq F(\mathbf{w}_0)$. Hence $\mathbf{w}_n \in \mathcal{L}_{F,F(\mathbf{w}_0)}$. Therefore, conditioned on \mathcal{B} , by the hypotheses of [Theorem 2.1](#) we have with probability at least p_0 that either $F(\mathcal{A}_1(\mathbf{w}_n)) < F(\mathbf{w}_n) - \Delta(\mathbf{w}_n)$ or $\mathcal{A}_2(\mathbf{w}_n) \cap \mathcal{S} \neq \{\}$.

Let \mathcal{C} be the event that $F(\mathcal{A}_1(\mathbf{w}_n)) < F(\mathbf{w}_n) - \Delta(\mathbf{w}_n)$ occurs. Let \mathcal{D} be the event that $\mathcal{A}_2(\mathbf{w}_n) \cap \mathcal{S} \neq \{\}$ occurs but \mathcal{C} does not occur. Recall that $\mathbf{w}_n \in \mathcal{L}_{F,F(\mathbf{w}_0)}$ conditioned on \mathcal{B} . Furthermore recall that $\mathcal{A}(\mathbf{w}_n)$ is only a function of \mathbf{w}_n , and none of the $(\mathbf{w}_t)_{0 \leq t \leq n-1}$. Thus the definition of decrease procedure, [Definition 2.2](#), implies that

$$\mathbb{P}(\mathcal{C} \sqcup \mathcal{D} | \mathcal{B}) \geq p_0.$$

Now since $\mathbb{P}(\mathcal{B}) = \mathbb{P}(\mathcal{B} | \mathcal{E}_n) \mathbb{P}(\mathcal{E}_n) \geq (1 - p)p_n > 0$, Bayes' Rule implies

$$\begin{aligned} \mathbb{P}((\mathcal{B} \cap \mathcal{C}) \sqcup (\mathcal{B} \cap \mathcal{D}) | \mathcal{B}) &= \frac{\mathbb{P}(\mathcal{B} \cap ((\mathcal{B} \cap \mathcal{C}) \sqcup (\mathcal{B} \cap \mathcal{D})))}{\mathbb{P}(\mathcal{B})} \\ &= \frac{\mathbb{P}(\mathcal{B} \cap (\mathcal{C} \sqcup \mathcal{D}))}{\mathbb{P}(\mathcal{B})} = \mathbb{P}(\mathcal{C} \sqcup \mathcal{D} | \mathcal{B}) \geq p_0. \end{aligned}$$

Note $\mathcal{B} \cap \mathcal{C}$ implies that $\mathcal{E}_{n+1,1}$ occurs, since under $\mathcal{B} \cap \mathcal{C}$ we have $F(\mathcal{A}_1(\mathbf{w}_t)) < F(\mathbf{w}_t) - \Delta(\mathbf{w}_t)$ for all $0 \leq t \leq n$. Similarly, $\mathcal{B} \cap \mathcal{D}$ implies that $\mathcal{E}_{n+1,2}$ occurs, since under $\mathcal{B} \cap \mathcal{D}$ we have $F(\mathcal{A}_1(\mathbf{w}_t)) < F(\mathbf{w}_t) - \Delta(\mathbf{w}_t)$ for $0 \leq t \leq n - 1$ and $\mathcal{A}_2(\mathbf{w}_n) \cap \mathcal{S} \neq \{\}$.

Thus recalling $\mathcal{E}_{n,2}, \mathcal{B}$ are disjoint, we see that \mathcal{E}_{n+1} contains the following disjoint union of events:

$$\mathcal{E}_{n+1} \supseteq \mathcal{E}_{n,2} \sqcup (\mathcal{B} \cap \mathcal{C}) \sqcup (\mathcal{B} \cap \mathcal{D}).$$

The above observations imply via Bayes' Rule that

$$\begin{aligned} \mathbb{P}(\mathcal{E}_{n+1}) &\geq \mathbb{P}(\mathcal{E}_{n,2} \sqcup (\mathcal{B} \cap \mathcal{C}) \sqcup (\mathcal{B} \cap \mathcal{D})) \\ &= \mathbb{P}(\mathcal{E}_{n,2}) + \mathbb{P}((\mathcal{B} \cap \mathcal{C}) \sqcup (\mathcal{B} \cap \mathcal{D})) \\ &= \mathbb{P}(\mathcal{E}_{n,2} | \mathcal{E}_n) \mathbb{P}(\mathcal{E}_n) + \mathbb{P}((\mathcal{B} \cap \mathcal{C}) \sqcup (\mathcal{B} \cap \mathcal{D}) | \mathcal{B}) \mathbb{P}(\mathcal{B} | \mathcal{E}_n) \mathbb{P}(\mathcal{E}_n) \\ &= \mathbb{P}(\mathcal{E}_n) (p + \mathbb{P}((\mathcal{B} \cap \mathcal{C}) \sqcup (\mathcal{B} \cap \mathcal{D}) | \mathcal{B}) \cdot (1 - p)) \\ &\geq p_n (p + p_0 (1 - p)) \\ &\geq p_n (p_0 p + p_0 (1 - p)) = p_n p_0 \geq p_{n+1}. \end{aligned}$$

Here we used that $\mathbb{P}(\mathcal{E}_n) \geq p_n$, $p_n p_0 \geq p_{n+1}$ which follows immediately from the definition of p_n , $p_0 \leq 1$, and simple manipulations. The inductive step, and hence the proof, is thus complete. \square

Using [Lemma B.1](#) now readily proves the following:

Claim 3. Let \mathcal{E} be the event that there exists \mathbf{w}_t with $\mathbf{w}_t \in (\mathbf{w}_t)_{0 \leq t \leq T-1}$ such that $\mathcal{A}_2(\mathbf{w}_t) \cap \mathcal{S} \neq \{\}$, and for all \mathbf{w}_s with $0 \leq s < t$, we have $F(\mathcal{A}_1(\mathbf{w}_s)) < F(\mathbf{w}_s) - \Delta(\mathbf{w}_s)$. Then $\mathbb{P}(\mathcal{E}) \geq p_T$.

Proof of Claim 3. Apply [Lemma B.1](#) with $n = T$. Following the notation from there, we have that the event $\mathcal{E}_T = \mathcal{E}_{T,1} \sqcup \mathcal{E}_{T,2}$ has probability at least p_T .

Suppose that $\mathcal{E}_{T,1}$ occurs. Note $\mathcal{E}_{T,1}$ implies that $\mathbf{w}_t \in \mathcal{L}_{F,F(\mathbf{w}_0)}$ for all $0 \leq t \leq T$. Therefore

$$\Delta(\mathbf{w}_t) \geq \inf_{\mathbf{u} \in \mathcal{L}_{F,F(\mathbf{w}_0)}} \Delta(\mathbf{u}) \text{ for all } 0 \leq t \leq T. \quad (8)$$

Moreover, telescoping the direct implication of $\mathcal{E}_{T,1}$ gives that

$$F(\mathbf{w}_T) < F(\mathbf{w}_0) - \sum_{t=0}^{T-1} \Delta(\mathbf{w}_t). \quad (9)$$

Combining (8) and (9) and recalling that we shifted WLOG so F has minimum value 0 (see Notation) gives

$$T \inf_{\mathbf{u} \in \mathcal{L}_{F,F(\mathbf{w}_0)}} \Delta(\mathbf{u}) \leq \sum_{t=0}^{T-1} \Delta(\mathbf{w}_t) < F(\mathbf{w}_0) - F(\mathbf{w}_T) \leq F(\mathbf{w}_0).$$

This contradicts our choice of T .

Thus $\mathcal{E}_{T,1}$ cannot occur, and so $\mathcal{E}_{T,2}$ must occur, i.e. $\mathcal{E}_T = \mathcal{E}_{T,2}$. Note $\mathcal{E}_{T,2}$ is exactly the event \mathcal{E} . Thus

$$\mathbb{P}(\mathcal{E}) = \mathbb{P}(\mathcal{E}_{T,2}) = \mathbb{P}(\mathcal{E}_T) \geq p_T,$$

as desired. \square

Conditioning on the event \mathcal{E} from Claim 3, by Claim 3, we immediately recover the desired guarantee on the output, probability, and number of candidate vectors stated in Theorem 2.1. The only part remaining to prove Theorem 2.1 is to establish the bound $N = \frac{F(\mathbf{w}_0)}{\Delta} + \sup_{\mathbf{u} \in \mathcal{L}_{F,F(\mathbf{w}_0)}} t_{\text{oracle}}(\mathbf{u})$ on the number of oracle calls.

To this end, condition on \mathcal{E} from Claim 3 in all of the following, and follow the notation from there, in particular the definition of \mathbf{w}_t . Directly, we obtain that the number of oracle calls is at most $\sum_{i=0}^t t_{\text{oracle}}(\mathbf{w}_i)$ (the last term $t_{\text{oracle}}(\mathbf{w}_t)$ in the sum appears since computing $\mathcal{A}(\mathbf{w}_t)$ and $\mathcal{A}(\mathbf{w}_t)$ takes at most $t_{\text{oracle}}(\mathbf{w}_t)$ oracle calls). We now upper bound this sum.

As we are conditioning on \mathcal{E} and since we assumed WLOG by shifting that F has minimum value 0, we have

$$F(\mathbf{w}_{i+1}) - F(\mathbf{w}_i) < -\Delta(\mathbf{w}_i) < 0 \text{ for all } 0 \leq i \leq t-1 \implies \sum_{i=0}^{t-1} \Delta(\mathbf{w}_i) < F(\mathbf{w}_0) - F(\mathbf{w}_t) \leq F(\mathbf{w}_0). \quad (10)$$

The above also implies $F(\mathbf{w}_i) \leq F(\mathbf{w}_0)$, i.e. $\mathbf{w}_i \in \mathcal{L}_{F,F(\mathbf{w}_0)}$, for all $0 \leq i \leq t$. Therefore, $t_{\text{oracle}}(\mathbf{w}_i) \leq \sup_{\mathbf{u} \in \mathcal{L}_{F,F(\mathbf{w}_0)}} t_{\text{oracle}}(\mathbf{u})$ for all $0 \leq i \leq t$. Thus (10) gives

$$\frac{F(\mathbf{w}_0)}{\sum_{i=0}^{t-1} t_{\text{oracle}}(\mathbf{w}_i)} > \frac{\sum_{i=0}^{t-1} \Delta(\mathbf{w}_i)}{\sum_{i=0}^{t-1} t_{\text{oracle}}(\mathbf{w}_i)} \geq \min_{0 \leq i \leq t-1} \frac{\Delta(\mathbf{w}_i)}{t_{\text{oracle}}(\mathbf{w}_i)} \geq \bar{\Delta},$$

where the last inequality uses the elementary inequality $\frac{\sum_{i=1}^{k'} a_i}{\sum_{i=1}^{k'} b_i} \geq \min_i \frac{a_i}{b_i}$ for $a_i \geq 0, b_i > 0$, that $\mathbf{w}_i \in \mathcal{L}_{F,F(\mathbf{w}_0)}$ for all $0 \leq i \leq t-1$, and the definition of $\bar{\Delta}$. Rearranging and recalling $t_{\text{oracle}}(\mathbf{w}_t) \leq \sup_{\mathbf{u} \in \mathcal{L}_{F,F(\mathbf{w}_0)}} t_{\text{oracle}}(\mathbf{u})$ as justified above, we obtain

$$\sum_{i=0}^t t_{\text{oracle}}(\mathbf{w}_i) \leq \sup_{\mathbf{u} \in \mathcal{L}_{F,F(\mathbf{w}_0)}} t_{\text{oracle}}(\mathbf{u}) + \sum_{i=0}^{t-1} t_{\text{oracle}}(\mathbf{w}_i) \leq \sup_{\mathbf{u} \in \mathcal{L}_{F,F(\mathbf{w}_0)}} t_{\text{oracle}}(\mathbf{u}) + \frac{F(\mathbf{w}_0)}{\bar{\Delta}}.$$

This yields the desired conclusion on oracle complexity, completing the proof. \square

C First Order Convergence Proofs

C.1 Proofs for Adaptive GD

Proof. As with the proof of Theorem 3.1, we use Theorem 2.1. We again have $\mathcal{S} = \{\mathbf{w} : \|\nabla F(\mathbf{w})\| \leq \varepsilon\}$, and recall the choice of η from Theorem 3.2. Now we let $\mathcal{A}(\mathbf{u}_0) = (\mathbf{u}_0 - \eta_{\mathbf{u}_0} \nabla F(\mathbf{u}_0), \mathbf{u}_0)$. Thus $\mathcal{A}_1(\mathbf{u}_0) = \mathbf{u}_0 - \eta \nabla F(\mathbf{u}_0)$, $\mathcal{A}_2(\mathbf{u}_0) = \mathbf{u}_0$, and $t_{\text{oracle}}(\mathbf{u}_0) = 1$.

Claim 4. For any \mathbf{u}_0 in the $F(\mathbf{w}_0)$ -sublevel set $\mathcal{L}_{F,F(\mathbf{w}_0)}$, \mathcal{A} is a $(\mathcal{S}, 1, \min\left\{\frac{L'_1(\mathbf{w}_0)}{2\rho_0(F(\mathbf{w}_0)+1)^2}, \frac{\varepsilon^2}{2L'_1(\mathbf{w}_0)}\right\}, 0, \mathbf{u}_0)$ -decrease procedure.

To show this, analogously to the proof of [Theorem 3.1](#), for any $\mathbf{u}_0 \notin \mathcal{S}$ in the $F(\mathbf{w}_0)$ -sublevel set $\mathcal{L}_{F,F(\mathbf{w}_0)}$, we will show that the function will deterministically decrease by strictly greater than $\min\left\{\frac{L'_1(\mathbf{w}_0)}{\rho_0(F(\mathbf{w}_0)+1)^2}, \frac{\varepsilon^2}{2L'_1(\mathbf{w}_0)}\right\}$ at the next iterate. By definition of \mathcal{A}_2 , exactly as with the proof of [Theorem 3.1](#), we conclude via [Theorem 2.1](#) upon showing [Claim 4](#).

To show [Claim 4](#), by choice of step size, we have $\eta_{\mathbf{u}_0} \|\nabla F(\mathbf{u}_0)\| \leq \frac{1}{\rho_0(F(\mathbf{w}_0)+1)}$. Thus

$$\|\mathbf{u}_1 - \mathbf{u}_0\| \leq \frac{1}{\rho_0(F(\mathbf{w}_0) + 1)} \leq \frac{1}{\rho_0(F(\mathbf{u}_0) + 1)}.$$

Now combining [Lemma 3.1](#) with [Assumption 1.1](#), and because $\mathbf{u}_0 \in \mathcal{L}_{F,F(\mathbf{w}_0)}$, we see for all $\mathbf{p} \in \overline{\mathbf{u}_0 \mathbf{u}_1}$, $\|\nabla^2 F(\mathbf{p})\|_{\text{op}} \leq L'_1(\mathbf{w}_0)$ where $L'_1(\mathbf{w}_0)$ is defined as in the statement of [Theorem 3.2](#). We thus obtain by [Lemma A.1](#),

$$F(\mathbf{u}_1) \leq F(\mathbf{u}_0) - \eta \|\nabla F(\mathbf{u}_0)\|^2 + \frac{L'_1(\mathbf{w}_0)\eta^2}{2} \cdot \|\nabla F(\mathbf{u}_0)\|^2. \quad (11)$$

Recall that $\mathbf{u}_0 \notin \mathcal{S}$, so $\|\nabla F(\mathbf{u}_0)\| > \varepsilon$. We break into cases:

1. If $\|\nabla F(\mathbf{u}_0)\| > \frac{L'_1(\mathbf{w}_0)}{\rho_0(F(\mathbf{w}_0)+1)}$, then $\eta_{\mathbf{u}_0} = \frac{1}{\rho_0(F(\mathbf{w}_0)+1)\|\nabla F(\mathbf{u}_0)\|}$. In this case, substituting into (11) gives

$$\begin{aligned} F(\mathbf{u}_1) &\leq F(\mathbf{u}_0) - \eta \|\nabla F(\mathbf{u}_0)\|^2 + \frac{L'_1(\mathbf{w}_0)\eta^2}{2} \cdot \|\nabla F(\mathbf{u}_0)\|^2 \\ &= F(\mathbf{u}_0) - \frac{1}{\rho_0(F(\mathbf{w}_0) + 1)} \|\nabla F(\mathbf{u}_0)\| + \frac{L'_1(\mathbf{w}_0)}{2\rho_0(F(\mathbf{w}_0) + 1)^2} \\ &< F(\mathbf{u}_0) - \frac{1}{2} \cdot \frac{L'_1(\mathbf{w}_0)}{\rho_0(F(\mathbf{w}_0) + 1)^2}. \end{aligned}$$

2. Else if $\|\nabla F(\mathbf{u}_0)\| \leq L'_1(\mathbf{w}_0)$, then $\eta_{\mathbf{u}_0} = \frac{1}{L'_1(\mathbf{w}_0)}$. In this case, substituting into (11) gives

$$\begin{aligned} F(\mathbf{u}_1) &\leq F(\mathbf{u}_0) - \eta \|\nabla F(\mathbf{u}_0)\|^2 + \frac{L'_1(\mathbf{w}_0)\eta^2}{2} \cdot \|\nabla F(\mathbf{u}_0)\|^2 \\ &\leq F(\mathbf{u}_0) - \frac{\|\nabla F(\mathbf{u}_0)\|^2}{2L'_1(\mathbf{w}_0)} < F(\mathbf{u}_0) - \frac{\varepsilon^2}{2L'_1(\mathbf{w}_0)}, \end{aligned}$$

where we used that $\|\nabla F(\mathbf{u}_0)\| > \varepsilon$.

In either case, for $\|\nabla F(\mathbf{u}_0)\| > \varepsilon$ we have that

$$F(\mathbf{u}_1) < F(\mathbf{u}_0) - \min\left\{\frac{L'_1(\mathbf{w}_0)}{2\rho_0(F(\mathbf{w}_0) + 1)^2}, \frac{\varepsilon^2}{2L'_1(\mathbf{w}_0)}\right\}.$$

This proves [Claim 4](#). By our framework [Theorem 2.1](#), the proof is complete. \square

C.2 Proofs for SGD for FOSPs

Here, we prove [Theorem 3.3](#). We first introduce technical preliminaries, which will also be used in [Section E](#).

Theorem C.1 (Vector-Valued Azuma-Hoeffding, Theorem 3.5 in [Pinelis \(1994\)](#)). Let $\varepsilon_1, \dots, \varepsilon_K \in \mathbb{R}^d$ be such that for all k , $\mathbb{E}[\varepsilon_k | \mathcal{F}^{k-1}] = 0$, $\|\varepsilon_k\|^2 \leq \sigma_k^2$. Then for any $\lambda > 0$,

$$\mathbb{P}\left(\left\|\sum_{k=1}^K \varepsilon_k\right\| \geq \lambda\right) \leq 4 \exp\left(-\frac{\lambda^2}{4 \sum_{k=1}^K \sigma_k^2}\right).$$

Note the bound here is dimension free, so this result does not follow directly from standard Azuma-Hoeffding. Such a result can also be found in [Kallenberg and Sztencel \(1991\)](#); [Zhang \(2005\)](#); [Fang et al. \(2019\)](#).

Theorem C.2 (Data-Dependent Concentration Inequality, Lemma 3 in [Rakhlin et al. \(2012\)](#)). *Let $\varepsilon_1, \dots, \varepsilon_K \in \mathbb{R}$ be such that for all k , $\mathbb{E}[\varepsilon_k | \mathfrak{F}^{k-1}] = 0$, $\mathbb{E}[\varepsilon_k^2 | \mathfrak{F}^{k-1}] \leq \sigma_k^2$. Furthermore suppose that $\mathbb{P}(\varepsilon_k \leq b | \mathfrak{F}^{k-1}) = 1$. Letting $V_K = \sum_{k=1}^K \sigma_k^2$, for any $\delta < 1/e$, $K \geq 4$, we have*

$$\mathbb{P}\left(\sum_{k=1}^K \varepsilon_k > 2 \max\{2\sqrt{V_K}, b\sqrt{\log(1/\delta)}\} \sqrt{\log(1/\delta)}\right) \leq \delta \log(K).$$

Such a result is also presented in [Zhang \(2005\)](#); [Bartlett et al. \(2008\)](#); [Fang et al. \(2019\)](#).

We will first prove [Theorem 3.3](#) in the case where $\|\nabla f(\mathbf{w}; \boldsymbol{\zeta}) - \nabla F(\mathbf{w})\|$ is bounded by $\sigma(F(\mathbf{w}))$. As noted in [Fang et al. \(2019\)](#), these same inequalities hold when the martingale difference is not bounded or almost-surely bounded but rather the norms are sub-Gaussian with parameter σ_k . Thus after the proof, we remark how to straightforwardly generalize [Theorem 3.3](#) to the case when $\|\nabla f(\mathbf{w}; \boldsymbol{\zeta}) - \nabla F(\mathbf{w})\|$ is sub-Gaussian with parameter $\sigma(F(\mathbf{w}))$ in [Remark 7](#).

Now, we prove [Theorem 3.3](#).

Proof. We use our framework [Theorem 2.1](#) with $\mathcal{S} = \{\mathbf{w} : \|\nabla F(\mathbf{w})\| \leq \varepsilon\}$. Recall as per the discussion of SGD in our framework in [Subsection 2.3](#), we let $\mathbf{p}_0 = \mathbf{u}_0$, and define a sequence $(\mathbf{p}_i)_{0 \leq i \leq K_0}$ via

$$\mathbf{p}_i = \mathbf{p}_{i-1} - \eta \nabla f(\mathbf{p}_{i-1}; \boldsymbol{\zeta}_i),$$

where the $\boldsymbol{\zeta}_i$ are minibatch samples i.i.d. across different i . Note this sequence can be equivalently defined by repeated compositions of the function $\mathbf{u} \rightarrow \mathbf{u} - \eta \nabla f(\mathbf{u}; \boldsymbol{\zeta})$.

We now let $\mathcal{A}(\mathbf{u}_0) = (\mathbf{p}_{K_0}, (\mathbf{p}_i)_{0 \leq i \leq K_0-1})$, hence $\mathcal{A}_1(\mathbf{u}_0) = \mathbf{p}_{K_0}$, $\mathcal{A}_2(\mathbf{u}_0) = (\mathbf{p}_i)_{0 \leq i \leq K_0-1}$. Thus $t_{\text{oracle}}(\mathbf{u}_0) = K_0$. Also note the noise $\boldsymbol{\xi}_t$ used defining \mathcal{A} are independent across different t .

For appropriate $\eta = \tilde{\Theta}(\varepsilon^2)$, $K_0 = \tilde{\Theta}(\varepsilon^{-2})$ depending only on $\varepsilon, \delta, F(\mathbf{w}_0)$ and polylogarithmically in $1/\delta$, which we define below, we establish the following [Claim 5](#):

Claim 5. *For any \mathbf{u}_0 in the $F(\mathbf{w}_0)$ -sublevel set $\mathcal{L}_{F, F(\mathbf{w}_0)}$, \mathcal{A} is a $(\mathcal{S}, K_0, \frac{\eta K_0 \varepsilon^2}{4}, p, \mathbf{u}_0)$ -decrease procedure, where $p = \frac{\delta \eta K_0 \varepsilon^2}{4(F(\mathbf{w}_0)+1)}$.*

Then using [Theorem 2.1](#), we then directly conclude [Theorem 3.3](#).

To show [Claim 5](#), consider any \mathbf{u}_0 in the $F(\mathbf{w}_0)$ -sublevel set but not in \mathcal{S} . Following the notation from above, consider a ‘block’ of K_0 consecutive iterates of SGD starting at $\mathbf{p}_0 = \mathbf{u}_0$. We establish that with probability at least $1 - p$, if none of the iterates $\{\mathbf{p}_0 = \mathbf{u}_0, \dots, \mathbf{p}_{K_0-1}\}$ lie in \mathcal{S} , then $F(\mathbf{p}_{K_0}) < F(\mathbf{p}_0) - \Delta$ where $\Delta = \frac{\eta K_0 \varepsilon^2}{4}$. Then recalling the definitions of \mathcal{A}_2 , we immediately conclude [Claim 5](#).

Definitions and Parameters: For convenience, define

$$\begin{aligned} L_0(\mathbf{w}_0) &= \rho_0(F(\mathbf{w}_0) + 1), \\ L_1(\mathbf{w}_0) &= \rho_1(F(\mathbf{w}_0) + 1), \\ \sigma_1(\mathbf{w}_0) &= \sigma(F(\mathbf{w}_0) + 1), \\ B(\mathbf{w}_0) &= \sigma_1(\mathbf{w}_0)^2 + \frac{1}{8} \sigma_1(\mathbf{w}_0) L_0(\mathbf{w}_0). \end{aligned}$$

Also define

$$\boldsymbol{\xi}_{t+1} = \nabla f(\mathbf{p}_t; \boldsymbol{\zeta}_{t+1}) - \nabla F(\mathbf{p}_t),$$

where $\boldsymbol{\zeta}_{t+1}$ denotes the i.i.d. minibatch samples. Note by [Assumption 3.1](#) that $\mathbb{E}[\boldsymbol{\xi}_{t+1}] = 0$, where expectation is with respect to $\boldsymbol{\zeta}_{t+1}$.

In particular, we choose these parameters as follows:

$$\tilde{\eta} = \frac{\varepsilon^2}{\tilde{L}(\mathbf{w}_0) \log(1/\varepsilon)^6 \log(1/\delta)^6}$$

$$K_0 = \frac{C(\mathbf{w}_0)}{\varepsilon^2} \log(1/\tilde{\eta})^2 \log(1/\delta)^2 \log(1/\varepsilon)^2,$$

$$\eta = \frac{1}{\max\{1, \rho_0(F(\mathbf{w}_0) + 1)\}} \cdot \tilde{\eta},$$

where

$$C(\mathbf{w}_0) = 128B(\mathbf{w}_0) \vee 64(F(\mathbf{w}_0) + 1)^2,$$

$$\tilde{L}'(\mathbf{w}_0) = 8L_1(\mathbf{w}_0)(L_0(\mathbf{w}_0)^2 + \sigma_1(\mathbf{w}_0)^2) \vee 2L_0(\mathbf{w}_0) \vee 4\sigma_1(\mathbf{w}_0),$$

$$\tilde{L}(\mathbf{w}_0) = \tilde{L}'(\mathbf{w}_0)^2 C(\mathbf{w}_0)^2 \vee (3\sqrt{2} \log(\tilde{L}(\mathbf{w}_0)))^8 \vee (3\sqrt{2})^8.$$

Remark 6. Note that C, \tilde{L}', \tilde{L} depend only polynomially in terms of the self-bounding functions ρ_0, ρ_1, σ , and $F(\mathbf{w}_0)$.

Note we can assume WLOG that ε and the desired probability δ are at most some small enough *universal* constants in $(0, 1)$; by doing so, the result does not change up to universal constant, and hence is identical under the $O(\cdot)$. Consequently we may assume WLOG that $\tilde{\eta}$ and η are at most some small enough universal constant in $(0, 1)$ and that $K_0 \geq 4$.

Claim 6. For ε, δ small enough universal constants, the above choice of parameters satisfies the following properties:

$$\begin{aligned} & \max\{1, \rho_0(F(\mathbf{w}_0) + 1)\} \eta \\ &= \tilde{\eta} \leq \min \left\{ \frac{\varepsilon^2}{8L_1(\mathbf{w}_0)(L_0(\mathbf{w}_0)^2 + \sigma_1(\mathbf{w}_0)^2)}, \frac{1}{2K_0 L_0(\mathbf{w}_0)}, \frac{1}{4\sigma_1(\mathbf{w}_0)\sqrt{K_0 \log(4K_0/p)}} \right\}, \end{aligned} \quad (12)$$

$$K_0 \varepsilon^2 \geq 128B(\mathbf{w}_0) \log\left(\frac{2 \log K_0}{p}\right). \quad (13)$$

For the sake of brevity, we prove [Claim 6](#) after the our main proof. Checking this is a matter of elementary, albeit tedious, univariate inequalities.

Again, our plan is to apply [Theorem 2.1](#) by showing decrease with high probability for a block of K_0 iterates starting at \mathbf{p}_0 .

Notation: Let \mathfrak{F}^t denote the filtration of all information up through \mathbf{p}_t , but *not* including the mini-batch sample ζ_{t+1} . Let \mathcal{K} be a stopping time denoting the first t such that $\mathbf{p}_t \notin \mathbb{B}\left(\mathbf{p}_0, \frac{1}{\rho_0(F(\mathbf{w}_0)+1)}\right)$, i.e. the escape time of the iterates beginning at \mathbf{p}_0 from $\mathbb{B}\left(\mathbf{p}_0, \frac{1}{\rho_0(F(\mathbf{w}_0)+1)}\right) = \mathbb{B}\left(\mathbf{u}_0, \frac{1}{\rho_0(F(\mathbf{w}_0)+1)}\right)$.

We first detail two high probability events we will condition on for the remainder of the proof:

- By Vector-Valued Azuma Hoeffding [Theorem C.1](#), for a given $1 \leq t \leq K_0$ we have with probability at least $1 - \frac{p}{2K_0}$,

$$\left\| \eta \sum_{k=1}^t \xi_k \right\| \leq 2\eta \sqrt{\log(48K_0/p) \sum_{k=1}^t \sigma(F(\mathbf{p}_{k-1}))^2} = 2\eta \sqrt{\log(4K_0/p) \sum_{k=0}^{t-1} \sigma(F(\mathbf{p}_{k-1}))^2}.$$

This follows since each $\mathbb{E}[\xi_k | \mathfrak{F}^{k-1}] = 0$ as the stochastic gradient oracle is unbiased, and as $\|\xi_k\| \leq \sigma(F(\mathbf{p}_{k-1}))$ by [Assumption 3.1](#).

Thus by Union Bound, with probability at least $1 - p/2$, we have for all $1 \leq t \leq K_0$ that

$$\left\| \eta \sum_{k=1}^t \xi_k \right\| \leq 2\eta \sqrt{\log(4K_0/p) \sum_{k=0}^{t-1} \sigma(F(\mathbf{p}_k))^2}. \quad (14)$$

Denote this event by \mathcal{E}_1 , so $\mathbb{P}(\mathcal{E}_1) \geq 1 - p/2$.

- We define a stochastic process with the following trick to derive uniform bounds. Define the following sequence of real numbers:

$$Y_t := -\eta \langle \nabla F(\mathbf{p}_t), \boldsymbol{\xi}_{t+1} \rangle 1_{t < \mathcal{K}}.$$

Notice $1_{t < \mathcal{K}}$ is \mathfrak{F}^t -measurable, as $\{t < \mathcal{K}\}$ holds if and only if $\mathbf{p}_1, \dots, \mathbf{p}_t \in \mathbb{B}(\mathbf{p}_0, \frac{1}{\rho_0(F(\mathbf{w}_0)+1)})$.

Clearly $\nabla F(\mathbf{p}_t)$ is also \mathfrak{F}^t -measurable. Thus as the stochastic gradient oracle is unbiased (i.e. $\mathbb{E}[\boldsymbol{\xi}_{t+1} | \mathfrak{F}^t] = 0$),

$$\mathbb{E}[Y_t] = \mathbb{E}[\langle \nabla F(\mathbf{p}_t), \boldsymbol{\xi}_{t+1} \rangle 1_{t < \mathcal{K}} | \mathfrak{F}^t] = 0.$$

For $t \geq \mathcal{K}$ we have $Y_t \equiv 0$. For $t < \mathcal{K}$, we have $\mathbf{p}_t \in \mathbb{B}(\mathbf{p}_0, \frac{1}{\rho_0(F(\mathbf{w}_0)+1)})$. Consequently by [Lemma 3.1](#) and [Corollary 1](#) we have

$$|Y_t| \leq \eta |\langle \nabla F(\mathbf{p}_t), \boldsymbol{\xi}_{t+1} \rangle| \leq \eta \|\nabla F(\mathbf{p}_t)\| \|\boldsymbol{\xi}_{t+1}\| \leq \eta \rho_0(F(\mathbf{w}_0) + 1) \|\boldsymbol{\xi}_{t+1}\|.$$

Moreover by [Assumption 3.1](#) and [Lemma 3.1](#),

$$\|\boldsymbol{\xi}_{t+1}\| \leq \sigma(F(\mathbf{p}_t)) \leq \sigma(F(\mathbf{w}_0) + 1) = \sigma_1(\mathbf{w}_0).$$

In particular, recall that $\boldsymbol{\xi}_{t+1}$ is the difference between the gradient oracle and actual gradient at \mathbf{p}_t .

By the above arguments, both of the following inequalities hold deterministically:

$$\begin{aligned} |Y_t| &\leq \eta \|\nabla F(\mathbf{p}_t)\| \sigma_1(\mathbf{w}_0), \\ |Y_t| &\leq \eta \rho_0(F(\mathbf{w}_0) + 1) \sigma_1(\mathbf{w}_0) = \eta L_0(\mathbf{w}_0) \sigma_1(\mathbf{w}_0). \end{aligned}$$

We now apply both of these bounds in Data-Dependent Concentration Inequality, [Theorem C.2](#) (whose conditions hold because we can assume δ, ε are at most given universal constants, so $K_0 \geq 4, 2 \log K_0/p > e$). Consequently we obtain with probability at least $1 - \frac{p}{2}$ that

$$\begin{aligned} -\eta \sum_{t=0}^{K_0-1} \langle \nabla F(\mathbf{p}_t), \boldsymbol{\xi}_{t+1} \rangle 1_{t < \mathcal{K}} &\leq 2\eta L_0(\mathbf{w}_0) \sigma_1(\mathbf{w}_0) \log\left(\frac{2 \log K_0}{p}\right) \vee \\ &4 \sqrt{\eta^2 \sigma_1(\mathbf{w}_0)^2 \sum_{t=0}^{K_0-1} \|\nabla F(\mathbf{p}_t)\|^2} \sqrt{\log\left(\frac{2 \log K_0}{p}\right)}. \end{aligned} \quad (15)$$

Denote this event by \mathcal{E}_2 , so $\mathbb{P}(\mathcal{E}_2) \geq 1 - p/2$.

For the rest of this proof, we condition on $\mathcal{E}_1 \cap \mathcal{E}_2$. By the above, $\mathcal{E}_1 \cap \mathcal{E}$ occurs with probability at least $1 - p$. Denote $\mathcal{E} = \mathcal{E}_1 \cap \mathcal{E}_2$.

A-priori, these bounds are not particularly useful, especially in our more challenging setting under [Assumption 3.2](#) where noise can depend on function value. However conditioned on \mathcal{E} , we prove that SGD is sufficiently ‘local’, in particular that $\|\mathbf{p}_t - \mathbf{u}_0\| \leq 1$ for all $t, 1 \leq t \leq K_0$. This will then give us control over function value via [Lemma 3.1](#), which then allow us to make use of these bounds in a more standard way.

Lemma C.1. *Conditioned on \mathcal{E}_1 (and hence conditioned on \mathcal{E}), for all $t, 1 \leq t \leq K_0$, we have*

$$\|\mathbf{p}_t - \mathbf{p}_0\| = \|\mathbf{p}_t - \mathbf{u}_0\| \leq \frac{1}{\rho_0(F(\mathbf{w}_0) + 1)}.$$

Proof. We go by induction on t . Notice after t iterates,

$$\mathbf{p}_t = \mathbf{w}_0 - \eta \sum_{k=0}^{t-1} \nabla F(\mathbf{p}_k) - \eta \sum_{k=1}^t \boldsymbol{\xi}_k.$$

For the base case $t = 1$, we have from [Corollary 1](#) that $\|\nabla F(\mathbf{w}_0)\| \leq \rho_0(F(\mathbf{w}_0)) \leq L_0(\mathbf{w}_0)$. From the definition of the high-probability event \mathcal{E}_1 and properties of η from [Claim 6](#), and as $\sigma_1(\mathbf{w}_0) \geq \sigma(\mathbf{w}_0)$, it follows that

$$\|\eta \boldsymbol{\xi}_1\| \leq 2\eta \sigma(F(\mathbf{w}_0)) \sqrt{K_0 \log(4K_0/p)} \leq \frac{1}{2\rho_0(F(\mathbf{w}_0) + 1)}.$$

Consequently by properties of η from [Claim 6](#),

$$\|\mathbf{p}_1 - \mathbf{p}_0\| \leq \|\eta \nabla F(\mathbf{w}_0)\| + \|\eta \boldsymbol{\xi}_0\| \leq \frac{1}{\rho_0(F(\mathbf{w}_0) + 1)}.$$

This finishes the proof of the base case.

Now suppose [Lemma C.1](#) holds for all $1 \leq k \leq t-1$; we will show it for t . From [Lemma 3.1](#), for all $k \leq t-1$, we have

$$\|\nabla F(\mathbf{p}_k)\| \leq \rho_0(F(\mathbf{w}_0) + 1) \leq L_0(\mathbf{w}_0).$$

Thus for each k , we have

$$\sigma(F(\mathbf{p}_k)) \leq \sigma(F(\mathbf{w}_0) + 1) = \sigma_1(\mathbf{w}_0).$$

Thus conditioned on \mathcal{E}_1 we obtain

$$\begin{aligned} \|\mathbf{p}_t - \mathbf{p}_0\| &\leq \left\| \eta \sum_{k=0}^{t-1} \nabla F(\mathbf{p}_k) \right\| + \left\| \eta \sum_{k=1}^t \boldsymbol{\xi}_k \right\| \\ &\leq \eta K_0 L_0(\mathbf{w}_0) + 2\eta \sqrt{\log(4K_0/p) \sum_{k=0}^{K_0-1} \sigma_1(\mathbf{w}_0)^2} \\ &= \eta K_0 L_0(\mathbf{w}_0) + 2\eta \sigma_1(\mathbf{w}_0) \sqrt{K_0 \log(4K_0/p)} \\ &\leq \frac{1}{2\rho_0(F(\mathbf{w}_0) + 1)} + \frac{1}{2\rho_0(F(\mathbf{w}_0) + 1)} = \frac{1}{\rho_0(F(\mathbf{w}_0) + 1)}. \end{aligned}$$

Here we used the choice of η from [Claim 6](#) and the upper bound (14) on $\|\eta \sum_{k=1}^t \boldsymbol{\xi}_k\|$ implied by \mathcal{E}_1 . This completes the induction. \square

Now that we know the iterates of SGD are ‘sufficiently local’ for K_0 iterations via [Lemma C.1](#), the finish is straightforward. Condition on \mathcal{E} for the rest of the proof. Consider any $0 \leq t \leq K_0 - 1$. \mathcal{E} implies for all $\mathbf{p} \in \overline{\mathbf{p}_{t-1}\mathbf{p}_t}$, writing $\mathbf{p} = \theta \mathbf{p}_{t-1} + (1-\theta)\mathbf{p}_t$ for $\theta \in [0, 1]$, that we have

$$\|\mathbf{p} - \mathbf{p}_0\| \leq \theta \|\mathbf{p}_{t-1} - \mathbf{p}_0\| + (1-\theta) \|\mathbf{p}_t - \mathbf{p}_0\| \leq (1-\theta + \theta) \cdot \frac{1}{\rho_0(F(\mathbf{w}_0) + 1)} = \frac{1}{\rho_0(F(\mathbf{w}_0) + 1)}.$$

Consequently $F(\mathbf{p}) \leq \rho_0(F(\mathbf{w}_0) + 1)$, so the above combined with [Assumption 1.1](#) gives

$$\|\nabla^2 F(\mathbf{p})\| \leq L_1(\mathbf{w}_0). \quad (16)$$

We also obtain from [Lemma C.1](#) together with [Corollary 1](#) and [Assumption 3.1](#) that for all $0 \leq t \leq K_0$,

$$\begin{aligned} \|\boldsymbol{\xi}_t\| &\leq \sigma(F(\mathbf{w}_0) + 1) = \sigma_1(\mathbf{w}_0), \\ \|\nabla F(\mathbf{p}_t)\| &\leq \rho_0(F(\mathbf{w}_0) + 1) = L_0(\mathbf{w}_0). \end{aligned} \quad (17)$$

Now by [Lemma A.1](#) and (16),

$$\begin{aligned} F(\mathbf{p}_{t+1}) &\leq F(\mathbf{p}_t) - \eta \langle \nabla F(\mathbf{p}_t), \nabla f(\mathbf{p}_t; \boldsymbol{\zeta}_{t+1}) \rangle + \frac{\eta^2 L_1(\mathbf{w}_0)}{2} \|\nabla f(\mathbf{p}_t; \boldsymbol{\zeta}_{t+1})\|^2 \\ &\leq F(\mathbf{p}_t) - \eta \|\nabla F(\mathbf{p}_t)\|^2 - \eta \langle \nabla F(\mathbf{p}_t), \boldsymbol{\xi}_{t+1} \rangle + \eta^2 L_1(\mathbf{w}_0) (\|\nabla F(\mathbf{p}_t)\|^2 + \|\boldsymbol{\xi}_{t+1}\|^2). \end{aligned}$$

The last step uses the definition of $\boldsymbol{\xi}_{t+1}$ and Young’s Inequality.

Summing and telescoping the above for $0 \leq t \leq K_0 - 1$, and applying (17), gives

$$\begin{aligned} F(\mathbf{p}_{K_0}) &\leq F(\mathbf{p}_0) - \eta \sum_{t=0}^{K_0-1} \|\nabla F(\mathbf{p}_t)\|^2 - \eta \sum_{t=0}^{K_0-1} \langle \nabla F(\mathbf{p}_t), \boldsymbol{\xi}_{t+1} \rangle \\ &\quad + \eta^2 K_0 L_0(\mathbf{w}_0)^2 L_1(\mathbf{w}_0) + \eta^2 K_0 \sigma_1^2(\mathbf{w}_0) L_1(\mathbf{w}_0). \end{aligned} \quad (18)$$

Now, conditioned on \mathcal{E} , we upper bound

$$-\eta \sum_{t=0}^{K_0-1} \langle \nabla F(\mathbf{p}_t), \boldsymbol{\xi}_{t+1} \rangle$$

using (15). Under \mathcal{E} , by Lemma C.1 and Lemma 3.1, we have $\mathbf{p}_t \in \mathbb{B}(\mathbf{p}_0, \frac{1}{\rho_0(F(\mathbf{w}_0)+1)})$ for all $1 \leq t \leq K_0$, which implies that $t < \mathcal{K}$ for all $1 \leq t \leq K_0$. Therefore

$$-\eta \sum_{t=0}^{K_0-1} \langle \nabla F(\mathbf{p}_t), \boldsymbol{\xi}_{t+1} \rangle = -\eta \sum_{t=0}^{K_0-1} \langle \nabla F(\mathbf{p}_t), \boldsymbol{\xi}_{t+1} \rangle 1_{t < \mathcal{K}}.$$

Now AM-GM gives

$$\begin{aligned} & 4\sqrt{\eta^2 \sigma_1(\mathbf{w}_0)^2 \sum_{t=0}^{K_0-1} \|\nabla F(\mathbf{p}_t)\|^2} \sqrt{\log\left(\frac{2\log K_0}{p}\right)} \\ & \leq 2\eta \left(\frac{1}{4} \sum_{t=0}^{K_0-1} \|\nabla F(\mathbf{p}_t)\|^2 + 8\sigma_1(\mathbf{w}_0)^2 \log\left(\frac{2\log K_0}{p}\right) \right). \end{aligned}$$

Combining with (15), we obtain

$$\begin{aligned} -\eta \sum_{t=0}^{K_0-1} \langle \nabla F(\mathbf{p}_t), \boldsymbol{\xi}_{t+1} \rangle &= -\eta \sum_{t=0}^{K_0-1} \langle \nabla F(\mathbf{p}_t), \boldsymbol{\xi}_{t+1} \rangle 1_{t < \mathcal{K}} \\ &\leq \frac{\eta}{2} \sum_{t=0}^{K_0-1} \|\nabla F(\mathbf{p}_t)\|^2 + 16\eta B(\mathbf{w}_0) \log\left(\frac{2\log K_0}{p}\right). \end{aligned}$$

Combining with (18) gives

$$\begin{aligned} F(\mathbf{p}_{K_0}) &\leq F(\mathbf{p}_0) - \frac{\eta}{2} \sum_{t=0}^{K_0-1} \|\nabla F(\mathbf{p}_t)\|^2 + 16\eta B(\mathbf{w}_0) \log\left(\frac{2\log K_0}{p}\right) + \eta^2 K_0 L_0(\mathbf{w}_0)^2 L_1(\mathbf{w}_0) \\ &\quad + \eta^2 K_0 \sigma_1^2(\mathbf{w}_0) L_1(\mathbf{w}_0). \end{aligned} \tag{19}$$

Suppose that $\|\nabla F(\mathbf{p}_t)\| > \varepsilon$ for all $0 \leq t \leq K_0 - 1$. Then the above gives

$$\begin{aligned} F(\mathbf{p}_{K_0}) &< F(\mathbf{p}_0) - \frac{\eta K_0 \varepsilon^2}{2} + 16\eta B(\mathbf{w}_0) \log\left(\frac{2\log K_0}{p}\right) \\ &\quad + \eta^2 K_0 L_0(\mathbf{w}_0)^2 L_1(\mathbf{w}_0) + \eta^2 K_0 \sigma_1^2(\mathbf{w}_0) L_1(\mathbf{w}_0). \end{aligned}$$

To make use of this bound, by our choice of η , Claim 6 implies that

$$\eta^2 K_0 L_0(\mathbf{w}_0)^2 L_1(\mathbf{w}_0) + \eta^2 K_0 \sigma_1^2(\mathbf{w}_0) L_1(\mathbf{w}_0) \leq \frac{\eta K_0 \varepsilon^2}{8}.$$

By choice of K_0 , Claim 6 implies that

$$16\eta B(\mathbf{w}_0) \log\left(\frac{2\log K_0}{p}\right) \leq \frac{\eta K_0 \varepsilon^2}{8}.$$

The above was all conditioned on \mathcal{E} , which occurred with probability at least $1 - p$. Thus by (19), we obtain that with this same probability which is at least $1 - p$, if none of $\mathbf{p}_0, \dots, \mathbf{p}_{K_0-1}$ have gradient norm larger than ε , we have

$$F(\mathbf{p}_{K_0}) < F(\mathbf{p}_0) - \frac{\eta K_0 \varepsilon^2}{4} = F(\mathbf{u}_0) - \frac{\eta K_0 \varepsilon^2}{4}.$$

This establishes that \mathcal{A} is a $(\mathcal{S}, K_0 + 1, \frac{\eta K_0 \varepsilon^2}{4}, p, \mathbf{u}_0)$ -decrease procedure. Following our initial observations, we conclude via Theorem 2.1. \square

Now we prove Claim 6.

Proof of Claim 6. We first prove (13). Recall we chose

$$K_0 = \frac{C(\mathbf{w}_0)}{\varepsilon^2} \log(1/\tilde{\eta})^2 \log(1/\delta)^2 \log(1/\varepsilon)^2.$$

Furthermore recall $p = \frac{\delta \tilde{\eta} K_0 \varepsilon^2}{4(F(\mathbf{w}_0)+1)}$. Thus, (13) holds if and only if

$$C(\mathbf{w}_0) \log(1/\tilde{\eta})^2 \log(1/\delta)^2 \log(1/\varepsilon)^2 \geq 128B(\mathbf{w}_0) \log\left(\frac{8\log K_0 \cdot (F(\mathbf{w}_0) + 1)}{\delta \tilde{\eta} K_0 \varepsilon^2}\right).$$

As $C(\mathbf{w}_0) \geq 128B(\mathbf{w}_0) \vee 64(F(\mathbf{w}_0) + 1)^2$, again using the expression for K_0 , it suffices to prove

$$\log(1/\tilde{\eta})^2 \log(1/\delta)^2 \log(1/\varepsilon)^2 \geq \log\left(\frac{\log K_0}{C(\mathbf{w}_0)^{1/2} \delta \tilde{\eta} \log(1/\tilde{\eta})^2 \log(1/\delta)^2}\right).$$

As $\log(1/\delta)$, $\log(1/\tilde{\eta})$ are both larger than 1, it suffices to prove

$$\begin{aligned} & \log(1/\tilde{\eta})^2 \log(1/\delta)^2 \log(1/\varepsilon)^2 \\ & \geq \log(1/\tilde{\eta}) + \log(1/\delta) \\ & + \log\left(\frac{\log C(\mathbf{w}_0) + \log(1/\varepsilon^2) + 2 \log \log(1/\tilde{\eta}) + 2 \log \log(1/\delta) + 2 \log \log(1/\varepsilon)}{C(\mathbf{w}_0)^{1/2}}\right). \end{aligned}$$

Since $C(\mathbf{w}_0) \geq 64$, it satisfies $\log C(\mathbf{w}_0) < C(\mathbf{w}_0)^{1/2}$, so it suffices to prove

$$\begin{aligned} & \log(1/\tilde{\eta})^2 \log(1/\delta)^2 \log(1/\varepsilon)^2 \\ & \geq \log(1/\tilde{\eta}) + \log(1/\delta) \\ & + \log(1 + 2 \log(1/\varepsilon) + 2 \log \log(1/\tilde{\eta}) + 2 \log \log(1/\delta) + 2 \log \log(1/\varepsilon)). \end{aligned}$$

By comparing ‘degrees’, we conclude recalling we can assume WLOG that $\delta, \varepsilon, \tilde{\eta}$ are smaller than some universal constant.

Now we prove (12). We will prove that

$$\tilde{\eta} \leq \frac{1}{\tilde{L}'(\mathbf{w}_0) K_0 \sqrt{\log(4K_0/p)}}. \quad (20)$$

After proving (20), recalling our choice of $K_0 > 1/\varepsilon^2$ directly implies (12). To show (20), equivalently, we want to show

$$\tilde{\eta} \log(1/\tilde{\eta})^2 \sqrt{\log(4K_0/p)} \leq \frac{\varepsilon^2}{\tilde{L}'(\mathbf{w}_0) C(\mathbf{w}_0) \log(1/\delta)^2 \log(1/\varepsilon)^2}.$$

Recalling the definition of p , this holds if and only if

$$\tilde{\eta} \log(1/\tilde{\eta})^2 \sqrt{\log\left(\frac{16(F(\mathbf{w}_0) + 1)}{\delta \tilde{\eta} \varepsilon^2}\right)} \leq \frac{\varepsilon^2}{\tilde{L}'(\mathbf{w}_0) C(\mathbf{w}_0) \log(1/\delta)^2 \log(1/\varepsilon)^2}.$$

Now we explicitly recall our expression for $\tilde{\eta} = \frac{\varepsilon^2}{\tilde{L}(\mathbf{w}_0) \log(1/\varepsilon)^6 \log(1/\delta)^6}$. Plugging this in and recalling $\tilde{L}(\mathbf{w}_0) \geq \tilde{L}'(\mathbf{w}_0)^2 C(\mathbf{w}_0)^2$, it suffices to prove

$$\begin{aligned} & \frac{1}{\tilde{L}(\mathbf{w}_0)^{1/2} \log(1/\varepsilon)^6 \log(1/\delta)^6} \log\left(\frac{\tilde{L}(\mathbf{w}_0) \log(1/\varepsilon)^6 \log(1/\delta)^6}{\varepsilon^2}\right)^2 \\ & \cdot \sqrt{\log\left(\frac{16(F(\mathbf{w}_0) + 1) \tilde{L}(\mathbf{w}_0) \log(1/\varepsilon)^6 \log(1/\delta)^6}{\delta \varepsilon^4}\right)} \\ & \leq \frac{1}{\log(1/\delta)^2 \log(1/\varepsilon)^2}. \end{aligned}$$

Thus it suffices to prove:

$$\begin{aligned} & \frac{18}{\tilde{L}(\mathbf{w}_0)^{1/2}} \log\left(\frac{\tilde{L}(\mathbf{w}_0) \log(1/\varepsilon) \log(1/\delta)}{\varepsilon}\right)^2 \sqrt{\log\left(\frac{16(F(\mathbf{w}_0) + 1) \tilde{L}(\mathbf{w}_0) \log(1/\varepsilon) \log(1/\delta)}{\delta \varepsilon}\right)} \\ & \leq \log(1/\delta)^4 \log(1/\varepsilon)^4. \end{aligned}$$

Recall $\tilde{L}(\mathbf{w}_0)^{1/8} \geq 3\sqrt{2} \log(\tilde{L}(\mathbf{w}_0)) \vee 3\sqrt{2}$ and so

$$\frac{3\sqrt{2}}{\tilde{L}(\mathbf{w}_0)^{1/4}} \log\left(\frac{\tilde{L}(\mathbf{w}_0) \log(1/\varepsilon) \log(1/\delta)}{\varepsilon}\right)$$

$$\begin{aligned}
&\leq \frac{3\sqrt{2}}{\tilde{L}(\mathbf{w}_0)^{1/4}} (\log(1/\varepsilon) + \log \log(1/\varepsilon) + \log \log(1/\delta) + \log \tilde{L}(\mathbf{w}_0)) \\
&\leq \frac{1}{\tilde{L}(\mathbf{w}_0)^{1/8}} (1 + \log(1/\varepsilon) + \log \log(1/\varepsilon) + \log \log(1/\delta)).
\end{aligned}$$

Thus it suffices to show

$$\begin{aligned}
&\frac{1}{\tilde{L}(\mathbf{w}_0)^{1/4}} (1 + \log(1/\varepsilon) + \log \log(1/\varepsilon) + \log \log(1/\delta))^2 \\
&\cdot \sqrt{\log \left(\frac{16(F(\mathbf{w}_0) + 1) \tilde{L}(\mathbf{w}_0) \log(1/\varepsilon) \log(1/\delta)}{\delta \varepsilon} \right)} \\
&\leq \log(1/\delta)^4 \log(1/\varepsilon)^4.
\end{aligned}$$

To this end recall $\tilde{L}(\mathbf{w}_0)^{1/8} \geq \log(16(F(\mathbf{w}_0) + 1) \tilde{L}(\mathbf{w}_0))$, thus

$$\begin{aligned}
&\frac{1}{\tilde{L}(\mathbf{w}_0)^{1/8}} \log \left(\frac{16(F(\mathbf{w}_0) + 1) \tilde{L}(\mathbf{w}_0) \log(1/\varepsilon) \log(1/\delta)}{\delta \varepsilon} \right) \\
&= \frac{1}{\tilde{L}(\mathbf{w}_0)^{1/8}} (\log(16(F(\mathbf{w}_0) + 1) \tilde{L}(\mathbf{w}_0)) + \log(1/\delta) + \log(1/\varepsilon) + \log \log(1/\delta) + \log \log(1/\varepsilon)) \\
&\leq 1 + \log(1/\delta) + \log(1/\varepsilon) + \log \log(1/\delta) + \log \log(1/\varepsilon).
\end{aligned}$$

Therefore it suffices to show

$$\begin{aligned}
&(1 + \log(1/\varepsilon) + \log \log(1/\varepsilon) + \log \log(1/\delta))^2 \\
&\cdot (1 + \log(1/\delta) + \log(1/\varepsilon) + \log \log(1/\delta) + \log \log(1/\varepsilon))^{1/2} \\
&\leq \log(1/\delta)^4 \log(1/\varepsilon)^4.
\end{aligned}$$

Evidently the above holds for small enough universal constants δ, ε (compare ‘degrees’), so we conclude the proof. \square

Remark 7. We also discuss how to extend this result to when the $\|\xi_t\|$ has sub-Gaussianity parameter $\sigma(F(\mathbf{p}_t))$. The extension is straightforward. Again, we aim to prove [Claim 5](#). For the rest of this remark, follow the notation from the proof for SGD above. Besides applying [Theorem C.1](#), [Theorem C.2](#) when the relevant random variables are sub-Gaussian, which still hold true as mentioned in [Fang et al. \(2019\)](#), the only other time we used that $\|\xi_t\| \leq \sigma(F(\mathbf{p}_t))$ holds deterministically is to derive (18).

We apply [Theorem C.1](#), [Theorem C.2](#) identically to the proof earlier. This time, we have for $t < \mathcal{K}$ that ξ_{t+1} is sub-Gaussian with parameter $\sigma_1(\mathbf{w}_0)$, thanks to the same trick of multiplying with $1_{t < \mathcal{K}}$ when applying [Theorem C.2](#).

The only change is as follows: in the definition \mathcal{E} , add in the intersection the event \mathcal{E}_3 that for all $1 \leq t \leq K_0$, $\|\xi_t\|^2 \leq \sigma(F(\mathbf{p}_t))^2 \log(K_0/p)$, where p is defined the same as before. We control the probability of \mathcal{E}_3 via the following Lemma:

Lemma C.2 (Equivalent of Lemma 12, [De Sa et al. \(2022\)](#)). *With probability at least $1 - p$, we have for all $1 \leq t \leq K_0$,*

$$\|\xi_t\|^2 \leq \sigma(F(\mathbf{p}_t))^2 \log(K_0/p).$$

Proof. By [Assumption 3.1](#), with probability $1 - \frac{p}{K_0}$, we have

$$\frac{\|\xi_t\|^2}{\sigma(F(\mathbf{p}_t))^2} \leq \log(K_0/p).$$

A Union Bound finishes the proof. \square

Now we condition on $\mathcal{E} = \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$, which has probability at least $1 - 2p$ by combining

our earlier argument with [Lemma C.2](#). Note this only changes the resulting guarantee by a universal constant. We still have [Lemma C.1](#), which does not require an upper bound on *each* $\|\xi_t\|$ in its proof but simply uses concentration from event \mathcal{E}_1 .

Thus, conditioned on \mathcal{E} , we still have $F(\mathbf{p}_t) \leq F(\mathbf{w}_0) + 1$ by [Lemma C.1](#), [Lemma 3.1](#), and as $\mathbf{u}_0 \in \mathcal{L}_{F,F(\mathbf{w}_0)}$. Now conditioned on \mathcal{E} , by [Lemma C.2](#), we still have the following upper bound for all $1 \leq t \leq K_0$:

$$\|\xi_t\|^2 \leq \sigma(F(\mathbf{w}_0) + 1)^2 \log(K_0/p) = \sigma_1(\mathbf{w}_0)^2 \log(K_0/p).$$

Therefore conditioned on \mathcal{E} , we can still derive a bound analogous to (18). This resulting bound changes by only a $\log(K_0/p)$ factor (from [Lemma C.2](#), see the above display); moreover recall K_0, p depend polynomially in $\delta, 1/\varepsilon$. By adjusting η smaller by a $\text{polylog}(K_0/p)$ factor, the same proof as above goes through, up to changing quantities by polylogarithmic factors.

D Perturbed GD finding Second Order Stationary Points

D.1 Proof using the Framework

Here we prove [Theorem 3.4](#). We instantiate [Algorithm 1](#) formally here. The parameters of [Algorithm 1](#) will depend on $L_1(\mathbf{w}_0), L_2(\mathbf{w}_0)$, which are defined in (4), (21) respectively, and depend only on $\rho_1, \rho_2, F(\mathbf{w}_0)$. Given a desired success probability $1 - \delta$ for $\delta > 0$, a tolerance $\varepsilon > 0$, and $F(\mathbf{w}_0), L_1(\mathbf{w}_0), L_2(\mathbf{w}_0)$, the algorithm's other parameters are defined in terms of as follows:

1. $c \leq c_{\max}$ is a universal constant, where c_{\max} is a universal constant defined in [Lemma D.2](#).
2. $\tilde{\varepsilon} = \frac{\varepsilon}{L_2(\mathbf{w}_0)}$.
3. $\chi \leftarrow 4 \max\left\{\log\left(\frac{2dL_1(\mathbf{w}_0)^2 F(\mathbf{w}_0)}{c^2 \tilde{\varepsilon}^{2.5} \delta}\right), 5\right\}$.
4. $\eta \leftarrow \frac{c}{L_1(\mathbf{w}_0)}$.
5. $r \leftarrow \frac{\sqrt{c\tilde{\varepsilon}}}{\chi^2 L_1(\mathbf{w}_0)}$.
6. $g_{\text{thres}} \leftarrow \frac{\sqrt{c}}{\chi^2} \tilde{\varepsilon}$.
7. $f_{\text{thres}} \leftarrow \frac{c}{\chi^3} \sqrt{\frac{\tilde{\varepsilon}^3}{L_2(\mathbf{w}_0)}}$.
8. $t_{\text{thres}} \leftarrow \frac{\chi}{c^2} \frac{L_1(\mathbf{w}_0)}{\sqrt{L_2(\mathbf{w}_0) \tilde{\varepsilon}}}$.

Proof of Theorem 3.4 given Lemma D.2. We will first prove the following Lemma, which will define $L_2(\mathbf{w}_0)$ and explain its significance.

Lemma D.1. Define $L_1(\mathbf{w}_0)$ as in (4), and define

$$L_2(\mathbf{w}_0) = \max\{1, L_1(\mathbf{w}_0), \rho_2(F(\mathbf{w}_0) + 1)\}. \quad (21)$$

Then we have the following:

1. Suppose \mathbf{u} is such that $\|\mathbf{u} - \tilde{\mathbf{w}}\| \leq \frac{1}{\rho_0(F(\mathbf{w}_0)+1)}$, where $\tilde{\mathbf{w}} \in \mathcal{L}_{F,F(\mathbf{w}_0)}$, the $F(\mathbf{w}_0)$ -sublevel set. Then under [Assumption 1.1](#) (and in particular under [Assumption 1.2](#)),

$$\|\nabla^2 F(\mathbf{u})\|_{\text{op}} \leq L_1(\mathbf{w}_0).$$

2. Suppose that $\mathbf{u}_1, \mathbf{u}_2$ are such that $\|\mathbf{u}_1 - \tilde{\mathbf{w}}\|, \|\mathbf{u}_2 - \tilde{\mathbf{w}}\| \leq \frac{1}{\rho_0(F(\mathbf{w}_0)+1)}$, where $\tilde{\mathbf{w}} \in \mathcal{L}_{F,F(\mathbf{w}_0)}$. Then

$$\|\nabla^2 F(\mathbf{u}_1) - \nabla^2 F(\mathbf{u}_2)\|_{\text{op}} \leq L_2(\mathbf{w}_0) \|\mathbf{u}_1 - \mathbf{u}_2\|.$$

Remark 8. Note $L_1(\mathbf{w}_0), L_2(\mathbf{w}_0) \geq 1$, and $L_2(\mathbf{w}_0) \geq L_1(\mathbf{w}_0)$.

Proof of Lemma D.1. Recall by [Corollary 1](#) that $\|\nabla F(\mathbf{w})\| \leq \rho_0(F(\mathbf{w}))$. Now by [Lemma 3.1](#) and as $\tilde{\mathbf{w}} \in \mathcal{L}_{F,F(\mathbf{w}_0)}$, for any \mathbf{u}' with $\|\mathbf{u}' - \tilde{\mathbf{w}}\| \leq \frac{1}{\rho_0(F(\mathbf{w}_0)+1)} \leq \frac{1}{\rho_0(F(\tilde{\mathbf{w}})+1)}$, we have $F(\mathbf{u}') \leq F(\tilde{\mathbf{w}}) + 1$. The first part now directly follows by [Assumption 1.1](#).

Algorithm 1 Perturbed Gradient Descent, modified from Jin et al. (2017).

$\tilde{\varepsilon} = \frac{\varepsilon}{L_2(\mathbf{w}_0)}$, $\chi \leftarrow 4 \max\left\{\log\left(\frac{2dL_1(\mathbf{w}_0)^2 F(\mathbf{w}_0)}{c\tilde{\varepsilon}^{2.5}\delta}\right), 5\right\}$, $\eta \leftarrow \frac{c}{L_1(\mathbf{w}_0)}$, $r \leftarrow \frac{\sqrt{c}\tilde{\varepsilon}}{\chi^2 L_1(\mathbf{w}_0)}$, $g_{\text{thres}} \leftarrow \frac{\sqrt{c}}{\chi^2} \tilde{\varepsilon}$, $f_{\text{thres}} \leftarrow \frac{c}{\chi^3} \sqrt{\frac{\tilde{\varepsilon}^3}{L_2(\mathbf{w}_0)}}$, $t_{\text{thres}} \leftarrow \frac{\chi}{c^2} \frac{L_1(\mathbf{w}_0)}{\sqrt{L_2(\mathbf{w}_0)\tilde{\varepsilon}}}$. Here c refers to a small enough universal constant upper bounded by c_{max} in Lemma D.2.

```

while True do
  if  $\|\nabla F(\mathbf{w}_t)\| \leq g_{\text{thres}}$  then
     $\tilde{\mathbf{w}}_t \leftarrow \mathbf{w}_t$ ,  $t_{\text{noise}} \leftarrow t$ 
     $\mathbf{w}_t \leftarrow \tilde{\mathbf{w}}_t + \boldsymbol{\xi}_t$ ,  $\boldsymbol{\xi}_t$  uniform from  $\mathbb{B}(\tilde{\mathbf{0}}, r)$ 
     $s \leftarrow 0$ 
    while  $s < t_{\text{thres}}$  do
       $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla F(\mathbf{w}_t)$ ,  $s \leftarrow s + 1$ ,  $t \leftarrow t + 1$ 
    end while
    if  $F(\mathbf{w}_t) - F(\tilde{\mathbf{w}}_{t_{\text{noise}}}) > -f_{\text{thres}}$  then
      Return  $\tilde{\mathbf{w}}_{t_{\text{noise}}}$ 
    end if
  else
     $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla F(\mathbf{w}_t)$ ,  $t \leftarrow t + 1$ 
  end if
end while

```

The second part now follows by noting the line segment $\overline{\mathbf{u}_1 \mathbf{u}_2}$ is contained in $\mathbb{B}(\tilde{\mathbf{w}}, \frac{1}{\rho_0(F(\mathbf{w}_0)+1)})$ via Triangle Inequality, recalling $\tilde{\mathbf{w}} \in \mathcal{L}_{F, F(\mathbf{w}_0)}$, and then applying Lemma A.6 and Lemma 3.1. \square

We now prove Theorem 3.4 by instantiating our framework.

Define $\tilde{\varepsilon} = \frac{\varepsilon}{L_2(\mathbf{w}_0)}$ as we did earlier, and note $L_2(\mathbf{w}_0) \geq 1$. It suffices to show for $\tilde{\varepsilon} \leq 1$, that with probability at least $1 - \delta$, we will return \mathbf{w} such that $\|\nabla F(\mathbf{w})\| \leq \tilde{\varepsilon}$, $\nabla^2 F(\mathbf{w}) \geq -\sqrt{L_2(\mathbf{w}_0)\tilde{\varepsilon}}\mathbf{I}$ in $T = O\left(\frac{L_1(\mathbf{w}_0) \max\{F(\mathbf{w}_0), 1\} \chi^4}{\tilde{\varepsilon}^2}\right) = O\left(\frac{L_1(\mathbf{w}_0) L_2(\mathbf{w}_0)^2 \max\{F(\mathbf{w}_0), 1\} \chi^4}{\varepsilon^2}\right)$ oracle calls.⁷

Now let the set of interest

$$\mathcal{S} = \{\mathbf{w} : \|\nabla F(\mathbf{w})\| \leq g_{\text{thres}}, \nabla^2 F(\mathbf{w}) \geq -\sqrt{L_2(\mathbf{w}_0)\tilde{\varepsilon}}\mathbf{I}\}.$$

Note $g_{\text{thres}} \leq \tilde{\varepsilon}$, so $\mathbf{w} \in \mathcal{S}$ immediately implies $\|\nabla F(\mathbf{w})\| \leq \tilde{\varepsilon}$, $\nabla^2 F(\mathbf{w}) \geq -\sqrt{L_2(\mathbf{w}_0)\tilde{\varepsilon}}\mathbf{I}$. Also note it suffices to show the result for all $\tilde{\varepsilon} \leq \frac{1}{100L_2(\mathbf{w}_0)}$; otherwise for larger $\tilde{\varepsilon}$ we can just apply the result for $\tilde{\varepsilon} = \frac{1}{100L_2(\mathbf{w}_0)}$. Thus as $L_2(\mathbf{w}_0) \geq 1$, we can assume $\tilde{\varepsilon} \leq 1$. Clearly, we also can assume WLOG that $t_{\text{thres}} \geq 1$.

As in Subsection 2.3, we make the following definitions for Algorithm 1. For all $\mathbf{u}_0 \in \mathbb{R}^d$, if $\|\nabla F(\mathbf{u}_0)\| > g_{\text{thres}}$, we let

$$\mathcal{A}(\mathbf{u}_0) = (\mathbf{u}_0 - \eta \nabla F(\mathbf{u}_0), \mathbf{u}_0), \text{ hence } \mathcal{A}_1(\mathbf{u}_0) = \mathbf{u}_0 - \eta \nabla F(\mathbf{u}_0), \mathcal{A}_2(\mathbf{u}_0) = \mathbf{u}_0.$$

Otherwise if $\|\nabla F(\mathbf{u}_0)\| \leq g_{\text{thres}}$, we let $\mathbf{p}_0 = \mathbf{u}_0 + \boldsymbol{\xi}$ where $\boldsymbol{\xi}$ is uniform from $\mathbb{B}(\tilde{\mathbf{0}}, r)$, and define a sequence $(\mathbf{p}_i)_{0 \leq i \leq t_{\text{thres}}}$ via

$$\mathbf{p}_i = \mathbf{p}_{i-1} - \eta \nabla F(\mathbf{p}_{i-1}).$$

When then take

$$\mathcal{A}(\mathbf{u}_0) = (\mathbf{p}_{t_{\text{thres}}}, \mathbf{u}_0), \text{ hence } \mathcal{A}_1(\mathbf{u}_0) = \mathbf{p}_{t_{\text{thres}}}, \mathcal{A}_2(\mathbf{u}_0) = \mathbf{u}_0.$$

We then have

$$t_{\text{oracle}}(\mathbf{u}_0) = \begin{cases} t_{\text{thres}} & : \|\nabla F(\mathbf{u}_0)\| \leq g_{\text{thres}} \\ 1 & : \|\nabla F(\mathbf{u}_0)\| > g_{\text{thres}}. \end{cases}$$

We also define

$$\Delta(\mathbf{u}_0) = \begin{cases} f_{\text{thres}} & : \|\nabla F(\mathbf{u}_0)\| \leq g_{\text{thres}} \\ \frac{\eta}{2} \cdot g_{\text{thres}}^2 & : \|\nabla F(\mathbf{u}_0)\| > g_{\text{thres}}. \end{cases}$$

⁷The $\max\{1, F(\mathbf{w}_0)\}$ is a proof artifact.

We now establish the crucial [Claim 2](#): for all $\mathbf{u}_0 \in \mathcal{L}_{F,F(\mathbf{w}_0)}$, \mathcal{A} is a $(\mathcal{S}, t_{\text{oracle}}(\mathbf{u}_0), \Delta(\mathbf{u}_0), \frac{dL_1(\mathbf{w}_0)}{\sqrt{L_2(\mathbf{w}_0)}\tilde{\varepsilon}}e^{-\chi}, \mathbf{u}_0)$ -decrease procedure. (Recall $\tilde{\varepsilon} = \frac{\varepsilon}{L_2(\mathbf{w}_0)}$.)

To do this, we use the following crucial Lemma ensuring high-probability decrease around saddle points in the $F(\mathbf{w}_0)$ -sublevel set:

Lemma D.2 (Equivalent of Lemma 13, [Jin et al. \(2017\)](#)). *There exists a universal constant $c_{\max} \leq 1$ such that the following occurs. Suppose we start with a $\tilde{\mathbf{w}} \in \mathcal{L}_{F,F(\mathbf{w}_0)}$, that is in the $F(\mathbf{w}_0)$ -sublevel set, satisfying the following conditions:*

$$\|\nabla F(\tilde{\mathbf{w}})\| \leq g_{\text{thres}} \quad \text{and} \quad \lambda_{\min}(\nabla^2 F(\tilde{\mathbf{w}})) \leq -\sqrt{L_2(\mathbf{w}_0)}\tilde{\varepsilon}.$$

Now let $\mathbf{p}_0 = \tilde{\mathbf{w}} + \boldsymbol{\zeta}$, where $\boldsymbol{\zeta}$ is sampled uniformly from $\mathbb{B}(\vec{\mathbf{0}}, r)$ where r is defined in [Lemma D.3](#), and let $\{\mathbf{p}_t\}$ be the iterates of gradient descent starting from \mathbf{p}_0 . Then when the step size $\eta \leq \frac{c_{\max}}{L_1(\mathbf{w}_0)}$, with probability at least $1 - \frac{dL_1(\mathbf{w}_0)}{\sqrt{L_2(\mathbf{w}_0)}\tilde{\varepsilon}}e^{-\chi}$, we have:

$$F(\mathbf{p}_{t_{\text{thres}}}) - F(\tilde{\mathbf{w}}) < -f_{\text{thres}}.$$

The variables in the above are defined in [Algorithm 1](#). As noted earlier, because we work in the generalized smooth setting, the details require significant care compared to the proof of Lemma 13 in [Jin et al. \(2017\)](#).

With [Lemma D.2](#), we have the ingredients to prove [Theorem 3.4](#). First we establish [Claim 2](#).

Proof of Claim 2. We prove this by breaking into the following cases:

- Suppose $\|\nabla F(\mathbf{u}_0)\| > g_{\text{thres}}$. Then $\mathbf{u}_1 = \mathcal{A}_1(\mathbf{u}_0) = \mathbf{u}_0 - \eta \nabla F(\mathbf{u}_0)$.

Our condition on η implies that

$$\eta \leq \frac{1}{L_1(\mathbf{w}_0)} \leq \frac{1}{\rho_0(F(\mathbf{w}_0))\rho_0(F(\mathbf{w}_0) + 1)}.$$

As $\mathbf{u}_0 \in \mathcal{L}_{F,F(\mathbf{w}_0)}$, we have by [Corollary 1](#),

$$\|\mathbf{u}_1 - \mathbf{u}_0\| = \eta \|\nabla F(\mathbf{u}_0)\| \leq \eta \rho_0(F(\mathbf{u}_0)) \leq \eta \rho_0(F(\mathbf{w}_0)) \leq \frac{1}{\rho_0(F(\mathbf{w}_0) + 1)}.$$

Consequently, by [Lemma 3.1](#),

$$F(\mathbf{p}) \leq F(\mathbf{u}_0) + 1 \leq F(\mathbf{w}_0) + 1 \text{ for all } \mathbf{p} \in \overline{\mathbf{u}_0 \mathbf{u}_1}.$$

Now by [Lemma A.1](#) and [Assumption 1.1](#),

$$\begin{aligned} F(\mathbf{u}_1) &\leq F(\mathbf{u}_0) - \eta \|\nabla F(\mathbf{u}_0)\|^2 + \frac{L_1(\mathbf{w}_0)\eta^2}{2} \|\nabla F(\mathbf{u}_0)\|^2 \\ &\leq F(\mathbf{u}_0) - \frac{\eta}{2} \|\nabla F(\mathbf{u}_0)\|^2 \\ &< F(\mathbf{u}_0) - \frac{\eta}{2} \cdot g_{\text{thres}}^2 = F(\mathbf{u}_0) - \Delta(\mathbf{u}_0). \end{aligned}$$

- Else suppose $\|\nabla F(\mathbf{u}_0)\| \leq g_{\text{thres}}$. Then \mathbf{u}_0 is perturbed, and we consider the sequence of the next t_{thres} iterates $\mathbf{p}_0 = \mathbf{u}_0 + \boldsymbol{\xi}, \mathbf{p}_1, \dots, \mathbf{p}_{t_{\text{thres}}}$.

Consider the event \mathcal{E} from [Lemma D.2](#), which occurs with probability at least $1 - \frac{dL_1(\mathbf{w}_0)}{\sqrt{L_2(\mathbf{w}_0)}\tilde{\varepsilon}}e^{-\chi}$. Under \mathcal{E} , for such \mathbf{u}_0 , we have:

- Either

$$F(\mathbf{p}_{t_{\text{thres}}}) - F(\mathbf{u}_0) < -f_{\text{thres}},$$

that is

$$F(\mathbf{u}_1) = F(\mathbf{p}_{t_{\text{thres}}}) < F(\mathbf{u}_0) - f_{\text{thres}}.$$

- Or

$$\lambda_{\min}(\nabla^2 F(\mathbf{u}_0)) \geq -\sqrt{\tilde{\varepsilon}L_2(\mathbf{w}_0)}, \text{ hence } \mathbf{u}_0 \in \mathcal{S}.$$

In all cases, by definition of \mathcal{A} , we conclude that \mathcal{A} is a $(\mathcal{S}, t_{\text{oracle}}(\mathbf{u}_0), \Delta(\mathbf{u}_0), \frac{dL_1(\mathbf{w}_0)}{\sqrt{L_2(\mathbf{w}_0)}\tilde{\varepsilon}}e^{-\chi}, \mathbf{u}_0)$ decrease procedure for $\mathbf{u}_0 \in \mathcal{L}_{F, F(\mathbf{w}_0)}$. \square

Consider these two cases, and recall the definition of $\bar{\Delta}$ from [Theorem 2.1](#). Using the definition of $\eta, g_{\text{thres}}, f_{\text{thres}}$, we obtain for c a small enough universal constant,

$$\begin{aligned}\bar{\Delta} &\geq \frac{1}{2} \min \left\{ \frac{c^2 \tilde{\varepsilon}^2}{2L_1(\mathbf{w}_0)\chi^4}, \frac{c^3 \tilde{\varepsilon}^2}{\chi^4 L_1(\mathbf{w}_0)} \right\} \\ &\geq \frac{c^3 \tilde{\varepsilon}^2}{\chi^4 L_1(\mathbf{w}_0)}.\end{aligned}$$

Combining with [Theorem 2.1](#), and note $t_{\text{oracle}}(\mathbf{u}_0) \leq t_{\text{thres}} \leq \frac{\max\{1, F(\mathbf{w}_0)\}}{\bar{\Delta}}$ for $\tilde{\varepsilon} \leq 1$. We thus obtain the desired oracle complexity of $O\left(\frac{L_1(\mathbf{w}_0) \max\{F(\mathbf{w}_0), 1\} \chi^4}{\tilde{\varepsilon}^2}\right) = O\left(\frac{\bar{\Delta}}{L_1(\mathbf{w}_0) L_2(\mathbf{w}_0)^2 \max\{F(\mathbf{w}_0), 1\} \chi^4}\right)$ to obtain an iterate in \mathcal{S} .⁸

We finally show the desired probability of success. Through [Theorem 2.1](#), since $\chi \geq 18$ and by definition of χ , we can verify that the probability of failure is at most

$$\begin{aligned}&\frac{dL_1(\mathbf{w}_0)}{\sqrt{L_2(\mathbf{w}_0)}\tilde{\varepsilon}}e^{-\chi} \cdot \sup_{\mathbf{u} \in \mathcal{L}_{F, F(\mathbf{w}_0)}} \left\{ \frac{F(\mathbf{w}_0)}{\Delta(\mathbf{w})} \right\} \\ &\leq \frac{dL_1(\mathbf{w}_0)}{\sqrt{L_2(\mathbf{w}_0)}\tilde{\varepsilon}}e^{-\chi} \cdot \frac{F(\mathbf{w}_0)}{\frac{c^2 \tilde{\varepsilon}^2}{2\chi^4 L_1(\mathbf{w}_0) \sqrt{L_2(\mathbf{w}_0)}}} \\ &\leq \chi^4 e^{-\chi} \frac{2dL_1(\mathbf{w}_0)^2 F(\mathbf{w}_0)}{c^2 \tilde{\varepsilon}^{2.5}} \\ &\leq e^{-\chi/4} \cdot \frac{2F(\mathbf{w}_0) dL_1^2(\mathbf{w}_0)}{c \tilde{\varepsilon}^{2.5}} \\ &\leq \delta.\end{aligned}$$

This completes the proof, assuming [Lemma D.2](#). \square

D.2 Proving the key Lemma

We now prove [Lemma D.2](#) to complete the proof. The rest of the proof is similar to that of [Jin et al. \(2017\)](#), but hinges crucially on the fact that the analysis in [Jin et al. \(2017\)](#) is ‘local’.

Consider any $\gamma > 0$, and define the ‘units’ in a similar way as [Jin et al. \(2017\)](#), but now in terms of $L_1(\mathbf{w}_0), L_2(\mathbf{w}_0) > 0$ defined earlier. First let the new ‘condition number’ be $\kappa = \kappa(\mathbf{w}_0) := \frac{L_1(\mathbf{w}_0)}{L_2(\mathbf{w}_0)^\gamma}$ (note this is *not* the real condition number, but rather is the ‘effective condition number’ of $\nabla^2 F$ in $\mathcal{L}_{F, F(\mathbf{w}_0)}$). Now define the following positive reals:

$$\begin{aligned}\mathcal{F}_1 &= \eta L_1(\mathbf{w}_0) \frac{\gamma^3}{L_2(\mathbf{w}_0)^2} \log^{-3} \left(\frac{d\kappa}{\delta} \right), \\ \mathcal{F}_2 &= \frac{\log \left(\frac{d\kappa}{\delta} \right)}{\eta \gamma}, \\ \mathcal{G} &= \sqrt{\eta L_1(\mathbf{w}_0)} \frac{\gamma^2}{L_2(\mathbf{w}_0)} \log^{-2} \left(\frac{d\kappa}{\delta} \right), \\ \mathcal{L} &= \sqrt{\eta L_1(\mathbf{w}_0)} \frac{\gamma}{L_2(\mathbf{w}_0)} \log^{-1} \left(\frac{d\kappa}{\delta} \right).\end{aligned}$$

Our goal is to prove the following.

⁸Note t_{thres} generally does not decrease with $F(\mathbf{w}_0)$, and this is why the $\max\{1, F(\mathbf{w}_0)\}$ comes in.

Lemma D.3 (equivalent of Lemma 14 in Jin et al. (2017)). *There exists a universal constant c_{\max} such that the following holds. For any F satisfying the conditions of Theorem 3.4, for any $\delta \in (0, \frac{d\kappa}{e}]$, suppose we start with a point $\tilde{\mathbf{w}} \in \mathcal{L}_{F,F(\mathbf{w}_0)}$ satisfying the following conditions for some $\gamma > 0$, where \mathcal{G} is defined as above:*

$$\|\nabla F(\tilde{\mathbf{w}})\| \leq \mathcal{G} \quad \text{and} \quad \lambda_{\min}(\nabla^2 F(\tilde{\mathbf{w}})) \leq -\gamma.$$

Let $\mathbf{p}_0 = \tilde{\mathbf{w}} + \zeta$, where ζ is sampled from the uniform distribution over a ball with radius $\frac{\mathcal{L}}{\kappa \cdot \log(\frac{d\kappa}{\delta})} := r$ and where \mathcal{L} is defined as above. Let $\{\mathbf{p}_t\}$ be the iterates of gradient descent starting from \mathbf{p}_0 . Then, when the step size $\eta \leq \frac{c_{\max}}{L_1(\mathbf{w}_0)}$, with probability at least $1 - \delta$, we have the following for any $T \geq \frac{1}{c_{\max}} \mathcal{F}_2$:

$$F(\mathbf{p}_T) - F(\tilde{\mathbf{w}}) < -\mathcal{F}_1.$$

Plugging in $\gamma = \sqrt{L_2(\mathbf{w}_0)\varepsilon}$, $\eta = \frac{c_{\max}}{L_1(\mathbf{w}_0)}$, $\delta = \frac{dL_1(\mathbf{w}_0)}{\sqrt{L_2(\mathbf{w}_0)\varepsilon}} e^{-\chi}$ into the above expressions for $\mathcal{F}_1, \mathcal{F}_2, \mathcal{G}, \mathcal{L}$, using $c \leq c_{\max}$, and directly applying Lemma D.3, we immediately obtain Lemma D.2. The rest of Section D is thus devoted to proving Lemma D.3.

Remark 9. Note it suffices to prove Lemma D.3 for δ and γ smaller than universal constants, as the result Theorem 3.4 will remain identical under the $O(\cdot)$. Thus we can assume WLOG that $\log(d\kappa/\delta)$ is larger than some universal constant, and that $\gamma \leq \frac{1}{60}$. Also notice by our choice of step size $\eta \leq \frac{c_{\max}}{L_1(\mathbf{w}_0)}$ and the assumption $\gamma \leq \frac{1}{60}$, for $c \leq c_{\max} \leq \frac{1}{12100}$ we obtain

$$\kappa \geq 1, r \leq 1.$$

This in turn implies

$$\begin{aligned} \mathcal{G} &\leq \mathcal{L}, \\ \mathcal{F}_2 &\geq 40, \\ \mathcal{L} &\leq \sqrt{\eta L_1(\mathbf{w}_0)} \cdot \frac{\gamma}{L_2(\mathbf{w}_0)} \cdot \log^{-1}\left(\frac{d\kappa}{\delta}\right) \\ &\leq \frac{1}{6600} \cdot \min\left\{1, \frac{1}{\rho_0(F(\mathbf{w}_0) + 1)}, \frac{1}{\rho_0(F(\mathbf{w}_0))\rho_0(F(\mathbf{w}_0) + 1)}\right\}, \end{aligned}$$

where the second line uses that

$$L_2(\mathbf{w}_0) \geq L_1(\mathbf{w}_0) \geq \max\{1, \rho_0(F(\mathbf{w}_0) + 1), \rho_0(F(\mathbf{w}_0))\rho_0(F(\mathbf{w}_0) + 1)\}.$$

As these assumptions come with no loss of generality, we make these assumptions for the rest of the proof.

To show Lemma D.3, again as in Jin et al. (2017), we prove that the width of the stuck region is not too large.

Lemma D.4 (equivalent of Lemma 15 in Jin et al. (2017)). *There exists a universal constant c_{\max} such that the following occurs. For any $\delta \in (0, \frac{d\kappa}{e}]$, let F and $\tilde{\mathbf{w}}$ satisfy the conditions in Lemma D.3. Without loss of generality, by rotational symmetry, let \mathbf{e}_1 be the minimum eigenvector of $\nabla^2 F(\tilde{\mathbf{w}})$. Consider two gradient descent sequences $\{\mathbf{u}_t\}$ and $\{\mathbf{x}_t\}$ with initial points $\mathbf{u}_0, \mathbf{x}_0$ satisfying (again, denote the radius $r = \frac{\mathcal{L}}{\kappa \cdot \log(\frac{d\kappa}{\delta})}$):*

$$\|\mathbf{u}_0 - \tilde{\mathbf{w}}\| \leq r, \quad \mathbf{x}_0 = \mathbf{u}_0 \pm \mu \cdot r \cdot \mathbf{e}_1, \quad \mu \in \left[\frac{\delta}{2\sqrt{d}}, 1\right].$$

Then for any step size $\eta \leq \frac{c_{\max}}{L_1(\mathbf{w}_0)}$, and any $T \geq \frac{1}{c_{\max}} \mathcal{F}_2$, we have:

$$\min\{F(\mathbf{u}_T) - F(\mathbf{u}_0), F(\mathbf{x}_T) - F(\mathbf{x}_0)\} \leq -2.5\mathcal{F}_1.$$

Now, we prove Lemma D.3 given Lemma D.4.

Proof of Lemma D.3 given Lemma D.4. Recall as per Remark 9 that

$$\|\mathbf{p}_0 - \tilde{\mathbf{w}}\| \leq r \leq \mathcal{L} \leq \frac{1}{\rho_0(F(\mathbf{w}_0) + 1)}.$$

Also recall $\tilde{\mathbf{w}} \in \mathcal{L}_{F,F(\mathbf{w}_0)}$. Thus by [Lemma D.1](#) we obtain for all $\mathbf{u} \in \overline{\mathbf{p}_0 \tilde{\mathbf{w}}}$ that

$$\|\nabla^2 F(\mathbf{u})\|_{\text{op}} \leq L_1(\mathbf{w}_0).$$

Therefore by [Lemma A.1](#),

$$F(\mathbf{p}_0) \leq F(\tilde{\mathbf{w}}) + \|\nabla F(\tilde{\mathbf{w}})\| r + \frac{L_1(\mathbf{w}_0)}{2} r^2 \leq F(\tilde{\mathbf{w}}) + \mathcal{G}r + \frac{L_1(\mathbf{w}_0)}{2} r^2 = F(\tilde{\mathbf{w}}) + \mathcal{F}_1,$$

where we can readily verify from [Remark 9](#) that $\mathcal{G}r + \frac{L_1(\mathbf{w}_0)}{2} r^2 \leq \mathcal{F}_1$.

Now let the stuck region be the set of points \mathbf{p}_0 in $\mathbb{B}(\tilde{\mathbf{w}}, r)$ such that

$$F(\mathbf{p}_T) - F(\mathbf{p}_0) \geq -2.5\mathcal{F}_1.$$

Define the unstuck points by the complement of the stuck points.

We upper bound the volume of the stuck region as done in [Jin et al. \(2017\)](#); this step does not use gradient and Hessian Lipschitzness. Let $1_{\text{Stuck region}}(\cdot)$ be the indicator function of the stuck region. Write all $\mathbf{w} \in \mathbb{R}^d$ as $\mathbf{w} = (\mathbf{w}^{(1)}, \mathbf{w}^{(-1)})$, where $\mathbf{w}^{(1)}$ is the component of \mathbf{w} along \mathbf{e}_1 direction and $\mathbf{w}^{(-1)}$ is the component of \mathbf{w} along the orthogonal complement of \mathbf{e}_1 . By [Lemma D.4](#), for any $\mathbf{w} \in \mathbb{B}(\tilde{\mathbf{w}}, r)$,

$$\begin{aligned} 1_{\text{Stuck region}}(\mathbf{w}) d\mathbf{w} &= 1_{\text{Stuck region}}(\mathbf{w}) d\mathbf{w}^{(-1)} \int_{\tilde{\mathbf{w}} - \sqrt{r^2 - \|\tilde{\mathbf{w}}^{(-1)} - \mathbf{w}^{(-1)}\|^2}}^{\tilde{\mathbf{w}} + \sqrt{r^2 - \|\tilde{\mathbf{w}}^{(-1)} - \mathbf{w}^{(-1)}\|^2}} d\mathbf{w}^{(1)} \\ &\leq d\mathbf{w}^{(-1)} \cdot 2 \cdot \frac{\delta}{2\sqrt{d}} r. \end{aligned}$$

Using this, we have:

$$\begin{aligned} \text{Volume}(\text{Stuck region}) &= \int_{\mathbb{B}^d(\tilde{\mathbf{w}}, r)} 1_{\text{Stuck region}}(\mathbf{w}) d\mathbf{w} \\ &= \int_{\mathbb{B}^{d-1}(\tilde{\mathbf{w}}, r)} 1_{\text{Stuck region}}(\mathbf{w}) d\mathbf{w}^{(-1)} \int_{\tilde{\mathbf{w}} - \sqrt{r^2 - \|\tilde{\mathbf{w}}^{(-1)} - \mathbf{w}^{(-1)}\|^2}}^{\tilde{\mathbf{w}} + \sqrt{r^2 - \|\tilde{\mathbf{w}}^{(-1)} - \mathbf{w}^{(-1)}\|^2}} d\mathbf{w}^{(1)} \\ &\leq \int_{\mathbb{B}^{d-1}(\tilde{\mathbf{w}}, r)} d\mathbf{w}^{(-1)} \cdot 2 \cdot \frac{\delta}{2\sqrt{d}} r. \\ &= \text{Volume}(\mathbb{B}^{d-1}(\tilde{\mathbf{0}}, r)) \cdot \frac{\delta r}{\sqrt{d}}. \end{aligned}$$

Then letting $\Gamma(\cdot)$ denote the Gamma function, we have the following ratio:

$$\begin{aligned} \frac{\text{Volume}(\text{Stuck region})}{\text{Volume}(\mathbb{B}(\tilde{\mathbf{w}}, r))} &\leq \frac{\delta r}{\sqrt{d}} \cdot \frac{\text{Volume}(\mathbb{B}^{d-1}(\tilde{\mathbf{0}}, r))}{\text{Volume}(\mathbb{B}^d(\tilde{\mathbf{0}}, r))} \\ &= \frac{\delta}{\sqrt{\pi d}} \cdot \frac{\Gamma\left(\frac{d}{2} + 1\right)}{\Gamma\left(\frac{d}{2} + \frac{1}{2}\right)} \\ &\leq \frac{\delta}{\sqrt{\pi d}} \cdot \sqrt{\frac{d}{2} + \frac{1}{2}} \leq \delta. \end{aligned}$$

Here we use the following property of the Gamma function: for $x \geq 0$, $\frac{\Gamma(x+1)}{\Gamma(\frac{x}{2} + \frac{1}{2})} \leq \sqrt{x + \frac{1}{2}}$.

This directly implies that with probability at least $1 - \delta$, \mathbf{p}_0 is an unstuck point. Consequently with probability at least $1 - \delta$, for any $T \geq \frac{1}{c_{\max}} \mathcal{F}_2$, we have

$$F(\mathbf{p}_T) - F(\tilde{\mathbf{w}}) = F(\mathbf{p}_T) - F(\mathbf{p}_0) + F(\mathbf{p}_0) - F(\tilde{\mathbf{w}}) \leq -2.5\mathcal{F}_1 + \mathcal{F}_1 = -1.5\mathcal{F}_1 < -\mathcal{F}_1.$$

This proves [Lemma D.3](#). □

Now we prove [Lemma D.4](#), which we do with an analogous strategy as [Jin et al. \(2017\)](#) by coupling two gradient descent sequences. We have the following two Lemmas, analogous to

Lemmas 16, 17 in Jin et al. (2017). Again, the reason why they hold in our setting under generalized smoothness is because they all concern ‘local’ behavior around points in the sublevel set of $F(\mathbf{w}_0)$. Consequently Lemma 3.1 and Assumption 1.2 ensure we have the required ‘local’ smoothness properties.

Again define $H, \tilde{F}_{\mathbf{y}}(\mathbf{x})$ analogously to page 20, Jin et al. (2017), as follows:

$$H := \nabla^2 F(\tilde{\mathbf{w}}), \tilde{F}_{\mathbf{y}}(\mathbf{x}) := F(\mathbf{y}) + \langle \nabla F(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{1}{2}(\mathbf{x} - \mathbf{y})^\top H(\mathbf{x} - \mathbf{y}). \quad (22)$$

That is, $\tilde{F}_{\mathbf{y}}$ is a quadratic approximation of F , Taylor expanded about $\tilde{\mathbf{w}}$.

The aforementioned Lemmas are as follows:

Lemma D.5 (equivalent of Lemma 16 in Jin et al. (2017)). *Letting $\hat{c} = 11$, there exists a universal constant $c_{\max} \leq \frac{1}{12100}$ such that following holds. For any $\delta \in (0, \frac{d\kappa}{e}]$, consider $F, \tilde{\mathbf{w}}, r$ as in Lemma D.3. For any \mathbf{u}_0 with $\|\mathbf{u}_0 - \tilde{\mathbf{w}}\| \leq 2r = \frac{2\mathcal{L}}{\kappa \cdot \log(\frac{d\kappa}{\delta})}$, define*

$$T = \min \left\{ \inf_t \{t \mid \tilde{F}_{\mathbf{u}_0}(\mathbf{u}_t) - F(\mathbf{u}_0) \leq -3\mathcal{F}_1\}, \hat{c}\mathcal{F}_2 \right\}.$$

Then for any $\eta \leq \frac{c_{\max}}{L(\mathbf{w}_0)}$, we have for all $t < T$ that $\|\mathbf{u}_t - \tilde{\mathbf{w}}\| \leq 150\mathcal{L}\hat{c}$.

Lemma D.6 (equivalent of Lemma 17 in Jin et al. (2017)). *Letting $\hat{c} = 11$, there exists a universal constant $c_{\max} \leq \frac{1}{12100}$ such that the following holds. For any $\delta \in (0, \frac{d\kappa}{e}]$, consider $F, \tilde{\mathbf{w}}, r$ as in Lemma D.3, and sequences $\{\mathbf{u}_t\}, \{\mathbf{x}_t\}$ satisfying the conditions in Lemma D.4. Define:*

$$T = \min \left\{ \inf_t \{t \mid \tilde{F}_{\mathbf{x}_0}(\mathbf{x}_t) - F(\mathbf{x}_0) \leq -3\mathcal{F}_1\}, \hat{c}\mathcal{F}_2 \right\}.$$

Then, for any $\eta \leq \frac{c_{\max}}{L_1(\mathbf{w}_0)}$, if $\|\mathbf{u}_t - \tilde{\mathbf{w}}\| \leq 150\mathcal{L}\hat{c}$ for all $t < T$, we will have $T < \hat{c}\mathcal{F}_2$. Equivalently, this means that

$$\inf_t \{t : \tilde{F}_{\mathbf{x}_0}(\mathbf{x}_t) - F(\mathbf{x}_0) \leq -3\mathcal{F}_1\} < \hat{c}\mathcal{F}_2,$$

i.e. that we escaped the saddle point.

Proof of Lemma D.4 given Lemma D.5, Lemma D.6. Choosing c_{\max} to be the minimum of the c_{\max} from Lemma D.5, Lemma D.6, we can ensure both Lemmas hold. Clearly this preserves that $c_{\max} \leq \frac{1}{12100}$.

Define

$$T^* = \hat{c}\mathcal{F}_2, T' = \inf \{t : \tilde{F}_{\mathbf{u}_0}(\mathbf{u}_t) - F(\mathbf{u}_0) \leq -3\mathcal{F}_1\}.$$

We break into cases on T' versus T^* :

- $T' \leq T^*$: By Lemma D.5, $\|\mathbf{u}_{T'-1} - \tilde{\mathbf{w}}\| \leq 150\mathcal{L}\hat{c}$. Since $\mathcal{L} \leq \frac{1}{6600} \cdot \frac{1}{\rho_0(F(\mathbf{w}_0)+1)}$ from Remark 9 and $\hat{c} = 11$, this yields

$$\|\mathbf{u}_{T'-1} - \tilde{\mathbf{w}}\| \leq 150\mathcal{L}\hat{c} \leq \frac{1}{4} \cdot \frac{1}{\rho_0(F(\mathbf{w}_0)+1)}.$$

Thus because $\tilde{\mathbf{w}} \in \mathcal{L}_{F,F(\mathbf{w}_0)}$, by Lemma D.1, we have

$$\|\nabla^2 F(\mathbf{u})\| \leq L_1(\mathbf{w}_0) \text{ for all } \mathbf{u} \in \overline{\mathbf{u}_{T'-1}\tilde{\mathbf{w}}}.$$

Thus, recalling $\mathcal{G} \leq \mathcal{L}$ from Remark 9, we obtain

$$\begin{aligned} \|\nabla F(\mathbf{u}_{T'-1})\| &\leq \|\nabla F(\tilde{\mathbf{w}})\| + L_1(\mathbf{w}_0)\|\mathbf{u}_{T'-1} - \tilde{\mathbf{w}}\| \\ &\leq \mathcal{G} + 150\hat{c}L_1(\mathbf{w}_0)\mathcal{L} \leq \mathcal{L} + 150\hat{c}L_1(\mathbf{w}_0)\mathcal{L}. \end{aligned}$$

Therefore, as $\eta L_1(\mathbf{w}_0) \leq c_{\max} \leq 1$,

$$\begin{aligned} \|\mathbf{u}_{T'} - \tilde{\mathbf{w}}\| &\leq \|\mathbf{u}_{T'-1} - \tilde{\mathbf{w}}\| + \eta \|\nabla F(\mathbf{u}_{T'-1})\| \\ &\leq 150\mathcal{L}\hat{c} + \mathcal{L} + 150\hat{c} \cdot \eta L_1(\mathbf{w}_0)\mathcal{L} \leq (300\hat{c} + 1)\mathcal{L} \end{aligned} \quad (23)$$

Recalling $\kappa, \log(\frac{d\kappa}{\delta}) \geq 1$, the conditions of Lemma D.4 give

$$\|\mathbf{u}_0 - \tilde{\mathbf{w}}\| \leq r \leq \mathcal{L}. \quad (24)$$

Combining (23), (24) and applying Triangle Inequality gives

$$\|\mathbf{u}_{T'} - \mathbf{u}_0\| \leq (300\hat{c} + 2)\mathcal{L}. \quad (25)$$

Also by (24), we have $\|\mathbf{u}_0 - \tilde{\mathbf{w}}\| \leq \mathcal{L} \leq \frac{1}{\rho_0(F(\mathbf{w}_0)+1)}$. Thus as $\tilde{\mathbf{w}} \in \mathcal{L}_{F,F(\mathbf{w}_0)}$, by Lemma D.1 we obtain

$$\|\nabla^2 F(\mathbf{u}_0)\| \leq L_1(\mathbf{w}_0). \quad (26)$$

Moreover, by Triangle Inequality we obtain that for any $\mathbf{u} \in \overline{\mathbf{u}_0 \mathbf{u}_{T'}}$, we have

$$\|\mathbf{u} - \tilde{\mathbf{w}}\| \leq (300\hat{c} + 2)\mathcal{L} = 3302\mathcal{L} \leq \frac{1}{\rho_0(F(\mathbf{w}_0) + 1)}.$$

As $\tilde{\mathbf{w}} \in \mathcal{L}_{F,F(\mathbf{w}_0)}$, Lemma D.1 implies for all such $\mathbf{u}_1, \mathbf{u}_2 \in \overline{\mathbf{u}_0 \mathbf{u}_{T'}}$ that

$$\|\nabla^2 F(\mathbf{u}_1) - \nabla^2 F(\mathbf{u}_2)\|_{\text{op}} \leq \|\mathbf{u}_1 - \mathbf{u}_2\| L_2(\mathbf{w}_0).$$

Now applying Lemma A.2, and by choosing $\eta = \frac{c}{L(\mathbf{w}_0)}$ for a small enough universal constant c , we obtain:

$$\begin{aligned} & F(\mathbf{u}_{T'}) - F(\mathbf{u}_0) \\ & \leq \nabla F(\mathbf{u}_0)^\top (\mathbf{u}_{T'} - \mathbf{u}_0) + \frac{1}{2} (\mathbf{u}_{T'} - \mathbf{u}_0)^\top \nabla^2 F(\mathbf{u}_0) (\mathbf{u}_{T'} - \mathbf{u}_0) + \frac{L_2(\mathbf{w}_0)}{6} \|\mathbf{u}_{T'} - \mathbf{u}_0\|^3 \\ & \leq \tilde{F}_{\mathbf{u}_0}(\mathbf{u}_{T'}) - F(\mathbf{u}_0) + \frac{L_2(\mathbf{w}_0)}{2} \|\mathbf{u}_{T'} - \mathbf{u}_0\|^2 \|\mathbf{u}_0 - \tilde{\mathbf{w}}\| + \frac{L_2(\mathbf{w}_0)}{6} \|\mathbf{u}_{T'} - \mathbf{u}_0\|^3 \\ & \leq -3\mathcal{F}_1 + O(L_1(\mathbf{w}_0)\mathcal{L}^3) \\ & = -3\mathcal{F}_1 + O(\sqrt{\eta L_1(\mathbf{w}_0)}\mathcal{F}_1) \leq -2.5\mathcal{F}_1. \end{aligned}$$

Here we used (26), (24), (25), and that $\mathcal{L} \leq 1$ as per Remark 9. In the above, $O(\cdot)$ only hides universal constants as $\hat{c} = 11$ is a universal constant, and so these final inequalities can be made to hold by choosing c_{\max} a sufficiently small universal constant.

Since $\tilde{\mathbf{w}} \in \mathcal{L}_{F,F(\mathbf{w}_0)}$ and $\eta \leq \frac{2}{L_1(\mathbf{w}_0)}$, Lemma A.7 shows that gradient descent will not increase value (this is essentially the same as several steps the proof of Theorem 3.1, combined with induction). Thus for all $T \geq T'$ and hence for all $T \geq \frac{1}{c_{\max}}\mathcal{F}_2 \geq \hat{c}\mathcal{F}_2 \geq T'$ along this gradient descent trajectory, we have

$$F(\mathbf{u}_T) - F(\mathbf{u}_0) \leq F(\mathbf{u}_{T'}) - F(\mathbf{u}_0) \leq -2.5\mathcal{F}_1.$$

- $T' > T^*$: In this case, by Lemma D.5, we know $\|\mathbf{u}_t - \tilde{\mathbf{w}}\| \leq 150\mathcal{L}\hat{c}$ for all $t < T^* = \hat{c}\mathcal{F}_2$.

Define

$$T'' = \inf_t \{t \mid \tilde{F}_{\mathbf{x}_0}(\mathbf{x}_t) - F(\mathbf{x}_0) \leq -3\mathcal{F}_1\}.$$

Since $\|\mathbf{u}_t - \tilde{\mathbf{w}}\| \leq 150\mathcal{L}\hat{c}$ for all $t < T^* = \hat{c}\mathcal{F}_2$, it follows that $\|\mathbf{u}_t - \tilde{\mathbf{w}}\| \leq 150\mathcal{L}\hat{c}$ for all $t < \min\{T'', T^*\}$. Thus by Lemma D.6, we have that $\min\{T'', T^*\} < T^*$, and so $T'' < T^*$. Applying the same argument as in the first case to the $\{\mathbf{x}_t\}$, we have that for all $T \geq \frac{1}{c_{\max}}\mathcal{F}_2$ that

$$F(\mathbf{x}_T) - F(\mathbf{x}_0) \leq -2.5\mathcal{F}_1.$$

This proves Lemma D.4. □

Remark 10. Note that $\tilde{\mathbf{w}} \in \mathcal{L}_{F,F(\mathbf{w}_0)}$ is central to this argument, unlike the Lipschitz gradient and Hessian case from Jin et al. (2017).

D.3 Proof of Escaping Saddles Lemmas

Now we prove [Lemma D.5](#), [Lemma D.6](#).

Proof of Lemma D.5. We follow the proof of Lemma 16, [Jin et al. \(2017\)](#). Again, we aim to show that if the function value does not decrease, then all the iterates must remain constrained in a small ball. This is done by analyzing the dynamics of the iterates and decomposing the d -dimensional space into two subspaces: a subspace S , which is the span of the negative enough eigenvectors of the Hessian, and its orthogonal complement.

The main difference now is that now we cannot directly control relevant operator norms with global Lipschitz properties of the gradient and Hessian. However, it turns out that the proof of this Lemma will follow induction on the iterate \mathbf{u}_t , and consequently we will obtain that all of the prior iterates $\mathbf{u}_{t'}$ for $t' < t$ are close enough to $\tilde{\mathbf{w}}$. By a similar argument as in [Lemma D.3](#), since $\tilde{\mathbf{w}} \in \mathcal{L}_{F,F}(\mathbf{w}_0)$, this lets us upper bound the gradient of these points. By the Gradient Descent update rule, this in turn implies the current iterate is also close to $\tilde{\mathbf{w}}$, and thus we obtain bounds on the relevant derivatives in terms of $L_1(\mathbf{w}_0)$, $L_2(\mathbf{w}_0)$ for all points in the convex hull of the relevant iterates.

We begin the argument. Analogously to [Jin et al. \(2017\)](#), since $\delta \in (0, \frac{d\kappa}{e}]$, we always have $\log(\frac{d\kappa}{\delta}) \geq 1$. By the gradient descent update function, we have

$$\mathbf{u}_{t+1} = \mathbf{u}_t - \eta \nabla F(\mathbf{u}_t).$$

This can be expanded as:

$$\mathbf{u}_{t+1} = \mathbf{u}_t - \eta \nabla F(\mathbf{u}_0) - \eta \left(\int_0^1 \nabla^2 F(\theta(\mathbf{u}_t - \mathbf{u}_0) + \mathbf{u}_0) d\theta \right) (\mathbf{u}_t - \mathbf{u}_0).$$

Recall the definition $\mathbf{H} = \nabla^2 F(\tilde{\mathbf{w}})$. Let Δ_t be defined as:

$$\Delta_t := \int_0^1 \nabla^2 F(\theta(\mathbf{u}_t - \mathbf{u}_0) + \mathbf{u}_0) d\theta - \mathbf{H}.$$

Substituting, we obtain:

$$\mathbf{u}_{t+1} = (\mathbf{I} - \eta \mathbf{H} - \eta \Delta_t)(\mathbf{u}_t - \mathbf{u}_0) - \eta \nabla F(\mathbf{u}_0) + \mathbf{u}_0.$$

Note we do not immediately have an upper bound on the operator norm of Δ_t . In particular this is because t could diverge (logarithmically) in the dimension, only being upper bounded by \mathcal{F}_2 .

We now compute the projections of $\mathbf{u}_t - \mathbf{u}_0$ in different eigenspaces of \mathbf{H} . Define S as the subspace spanned by all eigenvectors of \mathbf{H} whose eigenvalues are less than $-\frac{\gamma}{\hat{c} \log(\frac{d\kappa}{\delta})}$. Let S^c denote the subspace of the remaining eigenvectors. Let α_t and β_t denote the projections of $\mathbf{u}_t - \mathbf{u}_0$ onto S and S^c respectively, i.e., $\alpha_t = \mathcal{P}_S(\mathbf{u}_t - \mathbf{u}_0)$, and $\beta_t = \mathcal{P}_{S^c}(\mathbf{u}_t - \mathbf{u}_0)$.

We can decompose the update equations for \mathbf{u}_{t+1} into:

$$\begin{aligned} \alpha_{t+1} &= (\mathbf{I} - \eta \mathbf{H})\alpha_t - \eta \mathcal{P}_S \Delta_t (\mathbf{u}_t - \mathbf{u}_0) - \eta \mathcal{P}_S \nabla F(\mathbf{u}_0), \\ \beta_{t+1} &= (\mathbf{I} - \eta \mathbf{H})\beta_t - \eta \mathcal{P}_{S^c} \Delta_t (\mathbf{u}_t - \mathbf{u}_0) - \eta \mathcal{P}_{S^c} \nabla F(\mathbf{u}_0). \end{aligned}$$

By the definition of T , we know for all $t < T$:

$$\begin{aligned} -3\mathcal{F}_1 &< \tilde{F}_{\mathbf{u}_0}(\mathbf{u}_t) - F(\mathbf{u}_0) = \nabla F(\mathbf{u}_0)^\top (\mathbf{u}_t - \mathbf{u}_0) - \frac{1}{2} (\mathbf{u}_t - \mathbf{u}_0)^\top \mathbf{H} (\mathbf{u}_t - \mathbf{u}_0) \\ &\leq \nabla F(\mathbf{u}_0)^\top (\mathbf{u}_t - \mathbf{u}_0) - \frac{\gamma}{2 \hat{c} \log(\frac{d\kappa}{\delta})} \|\alpha_t\|^2 + \frac{1}{2} \beta_t^\top \mathbf{H} \beta_t. \end{aligned}$$

Evidently we have $\|\mathbf{u}_t - \mathbf{u}_0\|^2 = \|\alpha_t\|^2 + \|\beta_t\|^2$, and thus the above rearranges to

$$\|\mathbf{u}_t - \mathbf{u}_0\|^2 \leq \frac{2\hat{c} \log(\frac{d\kappa}{\delta})}{\gamma} \left(3\mathcal{F}_1 + \nabla F(\mathbf{u}_0)^\top (\mathbf{u}_t - \mathbf{u}_0) + \frac{1}{2} \beta_t^\top \mathbf{H} \beta_t \right) + \|\beta_t\|^2. \quad (27)$$

Now we control $\|\nabla F(\mathbf{u}_0)\|$. We use the fact that $\tilde{\mathbf{w}} \in \mathcal{L}_{F,F}(\mathbf{w}_0)$ to give us the necessary control over this quantity. Similar ideas were used in the proof of [Lemma D.4](#), and will continue to be used in the rest of the proofs of [Lemma D.5](#), [Lemma D.6](#). In particular, recall as per [Remark 9](#) that

$$\|\mathbf{u}_0 - \tilde{\mathbf{w}}\| \leq 2r \leq 2\mathcal{L} \leq \frac{1}{\rho_0(F(\mathbf{w}_0) + 1)}.$$

Thus by Lemma D.1, as $\tilde{\mathbf{w}} \in cL_{F,F(\mathbf{w}_0)}$, we obtain

$$\|\nabla^2 F(\mathbf{u})\| \leq L_1(\mathbf{w}_0) \text{ for all } \mathbf{u} \in \overline{\mathbf{u}_0 \tilde{\mathbf{w}}}.$$

Consequently,

$$\|\nabla F(\mathbf{u}_0) - \nabla F(\tilde{\mathbf{w}})\| \leq L_1(\mathbf{w}_0)\|\mathbf{u}_0 - \tilde{\mathbf{w}}\| \leq 2rL_1(\mathbf{w}_0) = 2\mathcal{G},$$

which implies

$$\|\nabla F(\mathbf{u}_0)\| \leq \|\nabla F(\tilde{\mathbf{w}})\| + 2\mathcal{G} = 3\mathcal{G}. \quad (28)$$

This gives us an analogous bound on $\|\nabla F(\mathbf{u}_0)\|$ as in the proof of Lemma 16, Jin et al. (2017). Substituting this bound on $\|\nabla F(\mathbf{u}_0)\|$ into (27), we obtain

$$\|\mathbf{u}_t - \mathbf{u}_0\|^2 \leq 14 \max \left\{ \frac{\mathcal{G}\hat{c} \log\left(\frac{d\kappa}{\delta}\right)}{\gamma} \|\mathbf{u}_t - \mathbf{u}_0\|, \frac{\mathcal{F}_1 \hat{c} \log\left(\frac{d\kappa}{\delta}\right)}{\gamma}, \frac{\boldsymbol{\beta}_t^\top \mathbf{H} \boldsymbol{\beta}_t \hat{c} \log\left(\frac{d\kappa}{\delta}\right)}{\gamma}, \|\boldsymbol{\beta}_t\|^2 \right\}.$$

In turn this implies

$$\|\mathbf{u}_t - \mathbf{u}_0\| \leq 14 \max \left\{ \frac{\mathcal{G}\hat{c} \log\left(\frac{d\kappa}{\delta}\right)}{\gamma}, \sqrt{\frac{\mathcal{F}_1 \hat{c} \log\left(\frac{d\kappa}{\delta}\right)}{\gamma}}, \sqrt{\frac{\boldsymbol{\beta}_t^\top \mathbf{H} \boldsymbol{\beta}_t \hat{c} \log\left(\frac{d\kappa}{\delta}\right)}{\gamma}}, \|\boldsymbol{\beta}_t\| \right\}. \quad (29)$$

The key induction: Now, we induct on t to prove

$$\|\mathbf{u}_t - \mathbf{u}_0\| \leq 148\mathcal{L}\hat{c} \text{ for all } t < T. \quad (30)$$

Clearly this implies Lemma D.5, upon recalling $\|\mathbf{u}_0 - \tilde{\mathbf{w}}\| \leq 2r = 2\mathcal{L} \leq \hat{c}\mathcal{L}$ by our choice $\hat{c} = 11$.

The base case $t = 0$ is evident.

Now for the inductive step, suppose (30) is true for all $\tau \leq t$ such that $t + 1 < T$. We show it is true for $t + 1$.

Due to the above bound (29), it suffices to upper bound $\|\boldsymbol{\beta}_{t+1}\|, \boldsymbol{\beta}_{t+1}^\top \mathbf{H} \boldsymbol{\beta}_{t+1}$. We note as in the proof of Lemma 16 of Jin et al. (2017) that letting

$$\boldsymbol{\delta}_t := \mathcal{P}_{S^c}(\Delta_t(\mathbf{u}_t - \mathbf{u}_0) + \nabla F(\mathbf{u}_0)),$$

we have by the Triangle Inequality and properties of projections that

$$\|\boldsymbol{\delta}_t\| \leq \|\Delta_t\|_{\text{op}} \|\mathbf{u}_t - \mathbf{u}_0\| + \|\nabla F(\mathbf{u}_0)\|. \quad (31)$$

Furthermore, we have by definition of the update rule for $\boldsymbol{\beta}_{t+1}$ that

$$\boldsymbol{\beta}_{t+1} = (\mathbf{I} - \eta \mathbf{H})\boldsymbol{\beta}_t + \eta \boldsymbol{\delta}_t. \quad (32)$$

Thus,

$$\|\boldsymbol{\beta}_{t+1}\| \leq \|(\mathbf{I} - \eta \mathbf{H})\boldsymbol{\beta}_t\| + \eta \|\boldsymbol{\delta}_t\| \leq \|\boldsymbol{\beta}_t\| + \eta \|\mathbf{H}\boldsymbol{\beta}_t\| + \eta \|\boldsymbol{\delta}_t\|. \quad (33)$$

Now, consider any $\tau, 0 \leq \tau \leq t$. We upper bound $\|\Delta_\tau\|_{\text{op}}$. Rewrite

$$\Delta_\tau = \int_0^1 (\nabla^2 F(\theta(\mathbf{u}_\tau - \mathbf{u}_0) + \mathbf{u}_0) - \nabla^2 F(\mathbf{u}_0)) d\theta + \nabla^2 F(\mathbf{u}_0) - \nabla^2 F(\tilde{\mathbf{w}}).$$

Clearly, as per Remark 9,

$$\|\mathbf{u}_0 - \tilde{\mathbf{w}}\| \leq 2r \leq 2\mathcal{L} \leq \frac{1}{\rho_0(F(\mathbf{w}_0) + 1)}.$$

Recalling $\tilde{\mathbf{w}} \in \mathcal{L}_{F,F(\mathbf{w}_0)}$ and applying Lemma D.1 gives

$$\|\nabla^2 F(\mathbf{u}_0) - \nabla^2 F(\tilde{\mathbf{w}})\|_{\text{op}} \leq L_2(\mathbf{w}_0)\|\mathbf{u}_0 - \tilde{\mathbf{w}}\|. \quad (34)$$

Moreover by inductive hypothesis, we know that $\|\mathbf{u}_\tau - \mathbf{u}_0\| \leq 148\mathcal{L}\hat{c}$. Consequently as $\hat{c} = 11 \geq 1$ and following Remark 9, for all $\theta \in [0, 1]$, we have

$$\|(\theta(\mathbf{u}_\tau - \mathbf{u}_0) + \mathbf{u}_0) - \tilde{\mathbf{w}}\| \leq 2\mathcal{L} + 148\hat{c}\mathcal{L} \leq \frac{1}{\rho_0(F(\mathbf{w}_0) + 1)}.$$

Since $\tilde{\mathbf{w}} \in \mathcal{L}_{F,F}(\mathbf{w}_0)$, it follows by [Lemma D.1](#) that

$$\|\nabla^2 F(\theta(\mathbf{u}_\tau - \mathbf{u}_0) + \mathbf{u}_0) - \nabla^2 F(\mathbf{u}_0)\|_{\text{op}} \leq L_2(\mathbf{w}_0)\|\mathbf{u}_\tau - \mathbf{u}_0\| \text{ for all } \theta \in [0, 1]. \quad (35)$$

Hence by Triangle Inequality, from (34) and (35), we have

$$\|\Delta_t\|_{\text{op}} \leq L_2(\mathbf{w}_0)(\|\mathbf{u}_\tau - \mathbf{u}_0\| + \|\mathbf{u}_0 - \tilde{\mathbf{w}}\|) \leq L_2(\mathbf{w}_0)(148\mathcal{L}\hat{c} + \|\mathbf{u}_0 - \tilde{\mathbf{w}}\|). \quad (36)$$

Proceeding from here is now exactly the same as in [Jin et al. \(2017\)](#). We detail the argument for completeness.

Combining (31), (36), (28) and applying the inductive hypothesis and the condition of [Lemma D.3](#) that $\|\mathbf{u}_0 - \tilde{\mathbf{w}}\| \leq 2r$, gives

$$\begin{aligned} \|\delta_\tau\| &\leq L_2(\mathbf{w}_0)(148\mathcal{L}\hat{c} + \|\mathbf{u}_0 - \tilde{\mathbf{w}}\|)\|\mathbf{u}_\tau - \mathbf{u}_0\| + \|\nabla F(\mathbf{u}_0)\| \\ &\leq L_2(\mathbf{w}_0) \cdot 148\hat{c} \left(148\hat{c} + \frac{2}{\kappa \cdot \log\left(\frac{d\kappa}{\delta}\right)} \right) \mathcal{L}^2 + 3\mathcal{G}. \end{aligned}$$

Plugging in the choice of \mathcal{L} , and choosing a small enough constant $c_{\max} \leq \left(\frac{1}{2 \cdot 148\hat{c}(148\hat{c}+2)}\right)^2$ and choosing step size $\eta < \frac{c_{\max}}{L_1(\mathbf{w}_0)}$, gives for any $0 \leq \tau \leq t$:

$$\|\delta_\tau\| \leq \left\{ 148\hat{c} \left(148\hat{c} + \frac{2}{\kappa \cdot \log\left(\frac{d\kappa}{\delta}\right)} \right) \sqrt{\eta L_1(\mathbf{w}_0)} + 3 \right\} \mathcal{G} \leq 3.5\mathcal{G}. \quad (37)$$

We now bound $\|\beta_{t+1}\|, \beta_{t+1}^\top \mathbf{H} \beta_{t+1}$, which combining with (29) finishes the induction and thus the proof.

- In order to bound $\|\beta_{t+1}\|$, combining (33) with (37) and recalling the definition of \mathcal{S} and β_t gives:

$$\|\beta_{t+1}\| \leq \left(1 + \frac{\eta\gamma}{\hat{c} \log\left(\frac{d\kappa}{\delta}\right)} \right) \|\beta_t\| + 3.5\eta\mathcal{G}.$$

Since $\|\beta_0\| = 0$ and $t+1 \leq T$, by applying the above relation recursively, we have:

$$\|\beta_{t+1}\| \leq \sum_{\tau=0}^T 3.5 \left(1 + \frac{\eta\gamma}{\hat{c} \log\left(\frac{d\kappa}{\delta}\right)} \right)^\tau \eta\mathcal{G} \leq 3.5 \cdot 3 \cdot T\eta\mathcal{G} \leq 10.5\mathcal{L}\hat{c}. \quad (38)$$

In the above we used $T \leq \hat{c}\mathcal{F}$, which also implies $\left(1 + \frac{\eta\gamma}{\hat{c} \log\left(\frac{d\kappa}{\delta}\right)} \right)^T \leq \left(1 + \frac{\eta\gamma}{\hat{c} \log\left(\frac{d\kappa}{\delta}\right)} \right)^{\hat{c}\mathcal{F}} \leq 3$ (one can find an easy upper bound on \mathcal{F} based on its definition and check using $L_2(\mathbf{w}_0) \geq L_1(\mathbf{w}_0) \geq 1$ that this is the case).

- Now for bounding $\beta_{t+1}^\top \mathbf{H} \beta_{t+1}$, notice we can also write the update equation (32) for β_t as:

$$\beta_t = \eta \sum_{\tau=0}^{t-1} (\mathbf{I} - \eta\mathbf{H})^\tau \delta_{t-1-\tau}.$$

As \mathbf{H} is symmetric this gives:

$$\beta_{t+1}^\top \mathbf{H} \beta_{t+1} = \eta^2 \sum_{\tau_1=0}^t \sum_{\tau_2=0}^t \delta_{t-1-\tau_1}^\top (\mathbf{I} - \eta\mathbf{H})^{\tau_1} \mathbf{H} (\mathbf{I} - \eta\mathbf{H})^{\tau_2} \delta_{t-1-\tau_2}.$$

Thus we have:

$$\beta_{t+1}^\top \mathbf{H} \beta_{t+1} \leq \eta^2 \sum_{\tau_1=0}^t \sum_{\tau_2=0}^t \|\delta_{t-1-\tau_1}\| \|(\mathbf{I} - \eta\mathbf{H})^{\tau_1} \mathbf{H} (\mathbf{I} - \eta\mathbf{H})^{\tau_2}\| \|\delta_{t-1-\tau_2}\|.$$

Since for $0 \leq \tau_1, \tau_2 \leq t$ we have $\|\delta_{t-1-\tau_1}\|, \|\delta_{t-1-\tau_2}\| \leq 3.5\mathcal{G}$ as argued earlier, we have:

$$\beta_{t+1}^\top \mathbf{H} \beta_{t+1} \leq 3.5^2 \eta^2 \mathcal{G}^2 \sum_{\tau_1=0}^t \sum_{\tau_2=0}^t \|(\mathbf{I} - \eta\mathbf{H})^{\tau_1} \mathbf{H} (\mathbf{I} - \eta\mathbf{H})^{\tau_2}\|.$$

Let the eigenvalues of \mathbf{H} be $\{\lambda_i\}$. Thus for any $\tau_1, \tau_2 \geq 0$, the eigenvalues of $(\mathbf{I} - \eta\mathbf{H})^{\tau_1} \mathbf{H} (\mathbf{I} - \eta\mathbf{H})^{\tau_2}$ are $\{\lambda_i(1 - \eta\lambda_i)^{\tau_1 + \tau_2}\}$. We now detail a calculation from Jin et al. (2017). Letting $g_t(\lambda) := \lambda(1 - \eta\lambda)^t$ and setting its derivative to zero yields

$$\nabla g_t(\lambda) = (1 - \eta\lambda)^t - t\eta\lambda(1 - \eta\lambda)^{t-1} = 0.$$

It is easy to check that $\lambda_t^* = \frac{1}{(1+t)\eta}$ is the unique maximizer, and $g_t(\lambda)$ is monotonically increasing in $(-\infty, \lambda_t^*]$.

This gives:

$$\|(\mathbf{I} - \eta\mathbf{H})^{\tau_1} \mathbf{H} (\mathbf{I} - \eta\mathbf{H})^{\tau_2}\| = \max_i \lambda_i(1 - \eta\lambda_i)^{\tau_1 + \tau_2} \leq \hat{\lambda}(1 - \eta\hat{\lambda})^{\tau_1 + \tau_2} \leq \frac{1}{(1 + \tau_1 + \tau_2)\eta},$$

where $\hat{\lambda} = \min\{\ell, \lambda_{\tau_1 + \tau_2}^*\}$. Therefore, we have:

$$\beta_{t+1}^\top \mathbf{H} \beta_{t+1} \leq 3.5^2 \eta \mathcal{G}^2 \sum_{\tau_1=0}^t \sum_{\tau_2=0}^t \frac{1}{1 + \tau_1 + \tau_2}.$$

To bound the sum note:

$$\sum_{\tau_1=0}^t \sum_{\tau_2=0}^t \frac{1}{1 + \tau_1 + \tau_2} = \sum_{\tau=0}^{2t} \min\{1 + \tau, 2t + 1 - \tau\} \cdot \frac{1}{1 + \tau} \leq 2t + 1 < 2T.$$

Thus:

$$\beta_{t+1}^\top \mathbf{H} \beta_{t+1} \leq 2 \cdot 3.5^2 \eta T \mathcal{G}^2 \leq \frac{3.5^2 \mathcal{L}^2 \gamma \hat{c}}{\log\left(\frac{d\kappa}{\delta}\right)}. \quad (39)$$

Finally, substituting the previous upper bounds (38), (39) for $\|\beta_t\|$, $\beta_{t+1}^\top \mathbf{H} \beta_{t+1}$ into our prior display (29) for $\|\mathbf{u}_t - \mathbf{u}_0\|$, we obtain:

$$\|\mathbf{u}_t - \mathbf{u}_0\| \leq 14 \max \left\{ \frac{\mathcal{G} \hat{c} \log\left(\frac{d\kappa}{\delta}\right)}{\gamma}, \sqrt{\frac{\mathcal{F}_1 \hat{c} \log\left(\frac{d\kappa}{\delta}\right)}{\gamma}}, \sqrt{\frac{\beta_t^\top \mathbf{H} \beta_t \hat{c} \log\left(\frac{d\kappa}{\delta}\right)}{\gamma}}, \|\beta_t\| \right\} \leq 148 \mathcal{L} \hat{c}.$$

This finishes the induction, and hence the proof of the Lemma. \square

Proof of Lemma D.6. Again, we aim to show that if all iterates from \mathbf{u}_0 are contained in a small ball, then the iterates from \mathbf{x}_0 decrease function value. As with the proof of Lemma D.5, the proof combines the proof idea of Lemma 17, Jin et al. (2017) with the self-bounding framework. This time it goes through even easier, because the required new bounds that we need from the relevant iterates being ‘local’ hold not due to induction, but rather from a direct application of Lemma D.5.

Define $\mathbf{v}_t = \mathbf{x}_t - \mathbf{u}_t$. By the assumptions of this Lemma we have that $\mathbf{v}_0 = \pm \mu \left[\frac{\mathcal{L}}{\kappa \cdot \log\left(\frac{d\kappa}{\delta}\right)} \right] \mathbf{e}_1$ where $\mu \in \left[\frac{\delta}{2\sqrt{d}}, 1 \right]$. Consequently

$$\frac{\delta}{2\sqrt{d}} \cdot r \leq \|\mathbf{v}_0\| \leq r. \quad (40)$$

Recall the definition

$$\mathbf{H} = \nabla^2 F(\tilde{\mathbf{w}})$$

as per (22). Also define

$$\Delta'_t := \int_0^1 \nabla^2 F(\mathbf{u}_t + \theta \mathbf{v}_t) d\theta - \mathbf{H}.$$

Exactly as in the proof of Lemma 17, Jin et al. (2017), by directly writing the update equations, we have

$$\begin{aligned} \mathbf{u}_{t+1} + \mathbf{v}_{t+1} &= \mathbf{x}_{t+1} = \mathbf{x}_t - \eta \nabla F(\mathbf{x}_t) \\ &= \mathbf{u}_t + \mathbf{v}_t - \eta \nabla F(\mathbf{u}_t + \mathbf{v}_t) \end{aligned}$$

$$\begin{aligned}
&= \mathbf{u}_t + \mathbf{v}_t - \eta \nabla F(\mathbf{u}_t) - \eta \left(\int_0^1 \nabla^2 F(\mathbf{u}_t + \theta \mathbf{v}_t) d\theta \right) \mathbf{v}_t \\
&= \mathbf{u}_t + \mathbf{v}_t - \eta \nabla F(\mathbf{u}_t) - \eta (\mathbf{H} + \Delta'_t) \mathbf{v}_t \\
&= \mathbf{u}_t - \eta \nabla F(\mathbf{u}_t) + (\mathbf{I} - \eta \mathbf{H} - \eta \Delta'_t) \mathbf{v}_t.
\end{aligned}$$

Hence as $\mathbf{u}_{t+1} = \mathbf{u}_t - \eta \nabla F(\mathbf{u}_t)$, we obtain

$$\mathbf{v}_{t+1} = (\mathbf{I} - \eta \mathbf{H} - \eta \Delta'_t) \mathbf{v}_t. \quad (41)$$

The difference from the proof of Lemma 17, Jin et al. (2017) is now that we do not immediately have an upper bound on $\|\Delta'_t\|_{\text{op}}$ without global Lipschitzness of the gradient and Hessian. However, similarly as in the proof of Lemma D.5, we can obtain such a bound using the self-bounding framework, since the point $\tilde{\mathbf{w}}$ in question is in the $F(\mathbf{w}_0)$ -sublevel set $\mathcal{L}_{F, F(\mathbf{w}_0)}$.

Note by hypothesis on \mathbf{u}_0 from Lemma D.4 and as $\|\mathbf{v}_0\| \leq r$ by (40),

$$\|\mathbf{x}_0 - \tilde{\mathbf{w}}\| \leq \|\mathbf{u}_0 - \tilde{\mathbf{w}}\| + \|\mathbf{v}_0\| \leq r + r = 2r.$$

Applying Lemma D.5 directly to the $\{\mathbf{x}_t\}$ implies that

$$\|\mathbf{x}_t - \tilde{\mathbf{w}}\| \leq 150\mathcal{L}\hat{c} \text{ for all } t < T.$$

By assumption of this Lemma, we have

$$\|\mathbf{u}_t - \tilde{\mathbf{w}}\| \leq 150\mathcal{L}\hat{c} \text{ for all } t < T.$$

Triangle Inequality thus gives

$$\|\mathbf{v}_t\| \leq 300\mathcal{L}\hat{c}, \|\mathbf{u}_t - \mathbf{u}_0\| \leq 300\mathcal{L}\hat{c} \text{ for all } t < T.$$

Therefore for all $0 \leq \theta \leq 1$,

$$\mathbf{u}_t + \theta \mathbf{v}_t \in \mathbb{B}(\tilde{\mathbf{w}}, 600\mathcal{L}\hat{c}).$$

Note as per Remark 9,

$$600\mathcal{L}\hat{c} = 6600\mathcal{L} \leq \frac{1}{\rho_0(F(\mathbf{w}_0) + 1)}.$$

As $\tilde{\mathbf{w}} \in \mathcal{L}_{F, F(\mathbf{w}_0)}$, it follows from Lemma D.1 that

$$\|\nabla^2 F(\mathbf{u}_t + \theta \mathbf{v}_t) - \nabla^2 F(\mathbf{u}_t)\|_{\text{op}} \leq L_2(\mathbf{w}_0) \cdot \theta \|\mathbf{v}_t\| \text{ for all } \theta \in [0, 1]. \quad (42)$$

Similarly, by the above bound

$$\|\mathbf{u}_t - \tilde{\mathbf{w}}\| \leq 150\mathcal{L}\hat{c} \leq \frac{1}{\rho_0(F(\mathbf{w}_0) + 1)}$$

and as $\tilde{\mathbf{w}} \in \mathcal{L}_{F, F(\mathbf{w}_0)}$, Lemma D.1 proves that

$$\|\nabla^2 F(\mathbf{u}_t) - \nabla^2 F(\tilde{\mathbf{w}})\|_{\text{op}} \leq L_2(\mathbf{w}_0) \|\mathbf{u}_t - \tilde{\mathbf{w}}\|. \quad (43)$$

Now, rewrite

$$\Delta'_t = \int_0^1 (\nabla^2 F(\mathbf{u}_t + \theta \mathbf{v}_t) - \nabla^2 F(\mathbf{u}_t)) d\theta + \nabla^2 F(\mathbf{u}_t) - \nabla^2 F(\tilde{\mathbf{w}}).$$

By (42), (43), and the above bounds on $\|\mathbf{v}_t\|$, $\|\mathbf{u}_t - \tilde{\mathbf{w}}\|$, we obtain for all $\theta \in [0, 1]$ that

$$\|\Delta'_t\|_{\text{op}} \leq L_2(\mathbf{w}_0)(\theta \|\mathbf{v}_t\| + \|\mathbf{u}_t - \tilde{\mathbf{w}}\|) \leq L_2(\mathbf{w}_0)\mathcal{L}(450\hat{c} + 1). \quad (44)$$

From here, exactly the same proof as that of Lemma 17, Jin et al. (2017) lets us conclude. We detail it for completeness. Similar to the proof of Lemma 17, Jin et al. (2017), let S be the subspace corresponding to eigenvectors of \mathbf{H} with eigenvalues larger or equal in absolute value to γ , and let S^\perp be its orthogonal complement. Note $\mathbf{e}_1 \subseteq S$. Denote the norm of \mathbf{v}_t projected onto S by ψ_t , and the norm of \mathbf{v}_t projected onto S^\perp by ϕ_t .

Notice therefore from the assumptions of this Lemma that $\phi_0 = 0$ as \mathbf{v}_0 is a scalar multiple of \mathbf{e}_1 . Similarly, note $\psi_0 = \|\mathbf{v}_0\| \geq \frac{\delta}{2\sqrt{d}} \cdot r$ by (40).

Let

$$B := \eta L_2(\mathbf{w}_0)\mathcal{L}(450\hat{c} + 1).$$

Observe $B \leq 1$, as $\mathcal{L}L_2(\mathbf{w}_0) \leq 1$ and as $\eta \leq c_{\max} \leq \frac{1}{12100}$, $\hat{c} = 11$.

Combining (41) with (44) gives that

$$\psi_{t+1} \geq (1 + \gamma\eta)\psi_t - B\sqrt{\psi_t^2 + \phi_t^2}, \phi_{t+1} \leq (1 + \gamma\eta)\phi_t + B\sqrt{\psi_t^2 + \phi_t^2}. \quad (45)$$

The key induction: Now we induct on t to show that for all $t < T$,

$$\phi_t \leq 4Bt \cdot \psi_t.$$

For the base case, recall by hypotheses of the Lemma that \mathbf{v}_0 is a scalar multiple of \mathbf{e}_1 , thus $\phi_0 = 0$ and the base case holds.

Now, for the inductive step, assume that the inductive hypothesis holds true for all $\tau \leq t$ for some t such that $t + 1 \leq T$. Substituting the inequality (45) for ϕ_{t+1} and applying the inductive hypothesis $\phi_t \leq 4Bt \cdot \psi_t$, we obtain

$$\phi_{t+1} \leq 4Bt(1 + \gamma\eta)\psi_t + B\sqrt{\psi_t^2 + \phi_t^2}.$$

Also note (45) gives

$$4B(t+1)\psi_{t+1} \geq 4B(t+1)\left((1 + \gamma\eta)\psi_t - B\sqrt{\psi_t^2 + \phi_t^2}\right),$$

which rearranges to

$$4Bt(1 + \gamma\eta)\psi_t \leq 4B(t+1)\psi_{t+1} + 4B^2(t+1)\sqrt{\psi_t^2 + \phi_t^2} - 4B(1 + \gamma\eta)\psi_t.$$

Therefore,

$$\phi_{t+1} \leq 4B(t+1)\psi_{t+1} + \left(4B^2(t+1)\sqrt{\psi_t^2 + \phi_t^2} + B\sqrt{\psi_t^2 + \phi_t^2} - 4B(1 + \gamma\eta)\psi_t\right).$$

Thus, recalling $B \leq 1$, to complete the induction it suffices to show the following:

$$(1 + 4B^2(t+1))\sqrt{\psi_t^2 + \phi_t^2} \leq 4(1 + \gamma\eta)\psi_t.$$

Choosing $\sqrt{c_{\max}} \leq \frac{1}{450\hat{c}+1} \min\left\{\frac{1}{2\sqrt{2}}, \frac{1}{4\hat{c}}\right\}$ which is a universal constant, and choosing $\eta \leq \frac{c_{\max}}{L_1(\mathbf{w}_0)}$, we have:

$$4B(t+1) \leq 4BT \leq 4\eta L_2(\mathbf{w}_0)\mathcal{L}(450\hat{c}+1)\hat{c}\mathcal{F} = 4\sqrt{\eta L_1(\mathbf{w}_0)}(450\hat{c}+1)\hat{c} \leq 1.$$

By the inductive hypothesis, this gives $\phi_t \leq \psi_t$. In turn this implies that

$$4(1 + \gamma\eta)\psi_t \geq 4\psi_t \geq 2\sqrt{2}\psi_t \geq (1 + 4B(t+1))\sqrt{\psi_t^2 + \phi_t^2},$$

finishing the induction.

Finishing the proof from here: We thus obtain $\phi_t \leq 4Bt\psi_t \leq \psi_t$ for all t , where we use that $4BT \leq 1$ as proven above, which just follows from our choice of parameters. Therefore,

$$\psi_{t+1} \geq (1 + \gamma\eta)\psi_t - B\sqrt{2}\psi_t > \left(1 + \frac{\gamma\eta}{2}\right)\psi_t. \quad (46)$$

The last step follows upon noting $B \leq \eta L_2(\mathbf{w}_0)\mathcal{L}(450\hat{c}+1) \leq \sqrt{c_{\max}}(450\hat{c}+1)\gamma\eta \log^{-1}\left(\frac{d\kappa}{\delta}\right) < \frac{\gamma\eta}{2\sqrt{2}}$. The inequality is strict as $\gamma\eta > 0$.

Finally, recalling that $\|\mathbf{v}_t\| \leq 300\mathcal{L}\hat{c}$, $\psi_0 \geq \frac{\delta}{2\sqrt{d}} \cdot r$ and using (46), we have for all $t < T$:

$$\begin{aligned} 300(\mathcal{L} \cdot \hat{c}) &\geq \|\mathbf{v}_t\| \\ &\geq \psi_t \\ &> \left(1 + \frac{\gamma\eta}{2}\right)^t \psi_0 \\ &\geq \left(1 + \frac{\gamma\eta}{2}\right)^t \cdot \frac{\delta}{2\sqrt{d}} \cdot \frac{\mathcal{L}}{\kappa \cdot \log\left(\frac{d\kappa}{\delta}\right)}. \end{aligned} \quad (47)$$

Note that $\delta \in (0, \frac{d\kappa}{e}]$ implies $\log\left(\frac{d\kappa}{\delta}\right) \geq 1$. Applying (47) for $t = T - 1$ we obtain:

$$T < 1 + \log\left(600\kappa\sqrt{d}\delta^{-1} \cdot \hat{c} \log\left(\frac{d\kappa}{\delta}\right)\right) \cdot \log^{-1}\left(1 + \frac{\gamma\eta}{2}\right)$$

Algorithm 2 Restarted SGD, from Fang et al. (2019)

Initialize at \mathbf{w}_0 , and consider $K_0 = \tilde{\Theta}(\varepsilon^{-2})$, $\eta = \tilde{\Theta}(\varepsilon^{1.5})$, $B = \tilde{\Theta}(\varepsilon^{0.5})$, $\tilde{\sigma} = 2\sigma'_1(\mathbf{w}_0)$, all explicitly defined in Subsection E.1.
Let $t = 0$ (the total number of iterates), $k = 0$ (the restart counter), $\mathbf{x}^0 = \mathbf{w}_0$ (the point we consider the escape from).
while $k < K_0$ **do**
 Let $\mathbf{x}^{t+1} = \mathbf{x}^t - \eta(\nabla f(\mathbf{x}^t; \zeta_{t+1}) + \tilde{\sigma}\mathbf{\Lambda}^{t+1})$, where $\mathbf{\Lambda}^{t+1}$ is uniform from $\mathbb{B}(\vec{0}, 1)$ and independent of everything else, and ζ_{t+1} is an i.i.d. minibatch sample
 $t \leftarrow t + 1$, $k \leftarrow k + 1$
 if $\|\mathbf{x}^k - \mathbf{x}^0\| > B$ **then**
 $\mathbf{x}^0 \leftarrow \mathbf{x}^k$, $k \leftarrow 0$
 end if
end while
Return $\frac{1}{K_0} \sum_{k=0}^{K_0-1} \mathbf{x}^k$

$$\begin{aligned}
&\leq 1 + 2.01 \log \left(600\kappa\sqrt{d}\delta^{-1} \cdot \hat{c} \log \left(\frac{d\kappa}{\delta} \right) \right) \cdot \frac{1}{\gamma\eta} \\
&\leq 1 + 2.01(\log(600\hat{c}) + 1.01 \log(d\kappa/\delta)) \cdot \frac{1}{\gamma\eta} \\
&\leq \left(\frac{1}{40} + 1 + 2.0301 \right) \mathcal{F}_2 \leq \hat{c}\mathcal{F}_2.
\end{aligned}$$

These last steps follow by:

- Taking c_{\max} a small enough universal constant so that $\gamma\eta \leq \frac{1}{60} \cdot \frac{c_{\max}}{L_1(\mathbf{w}_0)} \leq \frac{c_{\max}}{60}$ satisfies $\frac{2.01}{x} > \log^{-1}(1 + x/2)$, which is valid for all $0 < x < 0.02$.
- Remark 9, which states that we can assume $\text{WLOG} \log(d\kappa/\delta)$ is larger than a universal constant. In particular we can assume WLOG that $\log(d\kappa/\delta)$ solves $\log x < x^{0.01}$ (hence $\log(\kappa\sqrt{d}\delta^{-1} \log(d\kappa/\delta)) \leq 1.01 \log(d\kappa/\delta)$), that $2.01 \log(600\hat{c}) = 2.01 \log(6600) \leq \log(d\kappa/\delta)$ (recall $\hat{c} = 11$), and that $\mathcal{F}_2 = \frac{\log(d\kappa/\delta)}{\gamma\eta} \geq 40$.

This completes the proof. \square

E Restarted SGD finding Second Order Stationary Points

Here, we formally prove Theorem 3.5. We formally instantiate Algorithm 2 here. One may notice a slight difference in Algorithm 2 vs the algorithm of Fang et al. (2019): we artificially inject bounded noise at a particular scale $\tilde{\sigma}$. This ensures we can escape saddle points that are in the $F(\mathbf{w}_0)$ -sublevel set $\mathcal{L}_{F,F(\mathbf{w}_0)}$. Note we may not be able to escape saddle points that are not in $\mathcal{L}_{F,F(\mathbf{w}_0)}$, but that does not matter thanks to our framework Theorem 2.1, which effectively lets us consider only behavior within $\mathcal{L}_{F,F(\mathbf{w}_0)}$. Also note a practitioner can find such a noise scaling $\tilde{\sigma}$ (depending on suboptimality at initialization $F(\mathbf{w}_0)$) via appropriate cross-validation.

The general proof strategy here is similar to the way we adapted the proof of Jin et al. (2017) in Section D. Namely, we use the self-bounding regularity conditions to control the derivatives of F in appropriate neighborhoods of the $F(\mathbf{w}_0)$ -sublevel set $\mathcal{L}_{F,F(\mathbf{w}_0)}$.

E.1 Notation and Parameters

We set the parameters of the algorithm as follows. We will highlight the significance of these parameters in Subsection E.3.

Noise Parameters: Define

$$\sigma'(\mathbf{w}_0) = \sigma(F(\mathbf{w}_0) + 1). \quad (48)$$

$$\tilde{\sigma} = 2\sigma'(\mathbf{w}_0). \quad (49)$$

$$\sigma_1(\mathbf{w}_0) = \max\{\sigma'(\mathbf{w}_0) + \tilde{\sigma}, 1\}. \quad (50)$$

Note this only depends on ρ_0 (and therefore only on ρ_1) and $F(\mathbf{w}_0)$. Note $\tilde{\sigma} \in [\sigma'(\mathbf{w}_0), 2\sigma'(\mathbf{w}_0)]$.⁹ Also note $\sigma_1(\mathbf{w}_0) \leq 3\sigma'(\mathbf{w}_0)$.

Update Rule: Define

$$\nabla \tilde{f}(\mathbf{x}^t; \boldsymbol{\zeta}_{t+1}) := \nabla f(\mathbf{x}^t; \boldsymbol{\zeta}_{t+1}) + \tilde{\sigma} \mathbf{\Lambda}^{t+1}.$$

Thus the SGD update rule in [Algorithm 2](#) (without considering the restarts) is $\mathbf{x}^{t+1} = \mathbf{x}^t - \eta \nabla \tilde{f}(\mathbf{x}^t; \boldsymbol{\zeta}_{t+1})$. Note the slight abuse of notation; $\nabla \tilde{f}(\mathbf{x}^t; \boldsymbol{\zeta}_{t+1})$ is not necessarily an actual gradient.¹⁰ This will not cause issues or ambiguity for the rest of this section.

Effective Smoothness Parameters in $F(\mathbf{w}_0)$ -sublevel set: We define the ‘local smoothness parameters’ as follows, slightly differently compared to the proof of [Theorem 3.4](#). Define

$$\begin{aligned} L_1(\mathbf{w}_0) &:= \max\{1, \rho_1(F(\mathbf{w}_0) + 1), \rho_3(\rho_0(F(\mathbf{w}_0) + 1) + \sigma'(\mathbf{w}_0), F(\mathbf{w}_0) + 1)\}, \\ L_2(\mathbf{w}_0) &:= \max\left\{1, \rho_2(F(\mathbf{w}_0) + 1), \rho_0(F(\mathbf{w}_0) + 1)^2 \max\left\{4, (\sigma_1(\mathbf{w}_0) + \rho_0(F(\mathbf{w}_0) + 1))^2\right\}\right\}. \end{aligned} \quad (51)$$

Note all of these parameters only depend on $F(\mathbf{w}_0)$, through $\rho_1(\cdot)$, $\rho_2(\cdot)$, $\rho_3(\cdot, \cdot)$ (recall $\rho_0(\cdot)$ can be defined in terms of $\rho_1(\cdot)$).

Parameters of [Algorithm 2](#): We define the remaining parameters of [Algorithm 2](#) as follows. Consider any $\varepsilon > 0$ and $p \in (0, 1)$. We choose:

$$\begin{aligned} \tilde{C}_1 &= 2 \lfloor \frac{\log(3/p)}{\log(0.8^{-1})} + 1 \rfloor \log\left(\frac{24\sqrt{d}}{\eta}\right), \\ \delta &= \sqrt{L_2(\mathbf{w}_0)\varepsilon}, \\ \delta_2 &= 16\delta, \\ B &= \frac{\delta}{L_2(\mathbf{w}_0)\tilde{C}_1}, \\ K_0 &= \tilde{C}_1 \eta^{-1} \delta_2^{-1}, \\ \eta &\leq \frac{B^2 \delta}{512 \max(\sigma_1(\mathbf{w}_0)^2, 1) \tilde{C}_1 \log(48K_0/p)} \cdot \frac{1}{3(1 + \log(K_0))}. \end{aligned} \quad (52)$$

Also define

$$K_o = 2 \log\left(\frac{24\sqrt{d}}{\eta}\right) \eta^{-1} \delta_2^{-1}, \text{ thus } K_0 = \lfloor \frac{\log(3/p)}{\log(0.8^{-1})} + 1 \rfloor K_o.$$

Remark 11. To choose η satisfying the above inequality, one can perform the same analysis as on footnote 4, page 7 of [Fang et al. \(2019\)](#). We first choose $\tilde{\eta}$ appropriately by setting

$$\tilde{\eta} = \frac{B^2 \delta}{4096 \max(\sigma_1(\mathbf{w}_0)^2, 1) \log(48/p) \log(p) \lfloor \frac{\log(3/p)}{\log(0.8^{-1})} + 1 \rfloor},$$

and then set $\eta = \tilde{\eta} \log^{-3}(1/\tilde{\eta})$.

Remark 12. Analogously to the proof of [Theorem 3.4](#), note it suffices to show the result for $\varepsilon \leq \frac{1}{L_2(\mathbf{w}_0)}$; for $\varepsilon > \frac{1}{L_2(\mathbf{w}_0)}$, we can just apply the result for $\varepsilon = \frac{1}{L_2(\mathbf{w}_0)}$, and the result remains the same up to $F(\mathbf{w}_0)$ -dependent parameters in the $O(\cdot)$. Thus we can suppose that δ_2 (and δ) are at most some universal constant. We also can take $L_1(\mathbf{w}_0), L_2(\mathbf{w}_0), \sigma_1(\mathbf{w}_0)$ to be the max between their currently definition and an appropriate universal constant. Thus due to the choice of parameters above, we may assume that

$$\tilde{C}_1, K_0 \geq 1,$$

⁹In fact, this is the only condition we need on $\tilde{\sigma}$. In practice, such a $\tilde{\sigma}$ by fine-enough cross validation in terms of only $F(\mathbf{w}_0)$.

¹⁰This choice of notation is made to demonstrate the artificial noise injections $\tilde{\sigma} \mathbf{\Lambda}^{t+1}$ are not fundamentally needed. They are not necessary if the stochastic gradient $\nabla f(\cdot; \cdot)$ enjoys suitable anticoncentration properties.

$$\begin{aligned}\log(K_0), \sigma_1(\mathbf{w}_0) &\geq 1, \\ B &\leq \min\left(1, \frac{\sigma_1(\mathbf{w}_0)}{L_1(\mathbf{w}_0)}, \frac{1}{L_1(\mathbf{w}_0)}, \frac{1}{L_2(\mathbf{w}_0)}\right), \\ \eta &\leq \min\left\{1, \frac{1}{\sigma_1(\mathbf{w}_0)^2}\right\}.\end{aligned}$$

From here note we have $\eta L_1(\mathbf{w}_0) \leq 1$. As *these assumptions come with no loss of generality*, we make these assumptions for the rest of the proof.

Notation: Consider a sequence of iterates $\mathbf{x}^0, \mathbf{x}^1, \dots$ beginning at \mathbf{x}^0 comprising an instance of the while loop in [Algorithm 2](#). For such a sequence, let \mathfrak{F}^k be the σ -algebra defined by all the prior iterates and the noise up through \mathbf{x}^k , namely $\sigma\{\mathbf{x}^0, \zeta_1, \Lambda^1, \mathbf{x}^1, \dots, \mathbf{x}^{k-1}, \zeta_k, \Lambda^k\}$. Let \mathcal{K}_0 be a stopping time given by

$$\mathcal{K}_0 = \inf_k \{k \geq 0 : \|\mathbf{x}^k - \mathbf{x}^0\| \geq B\}.$$

Note \mathbf{x}^k and $1_{\mathcal{K}_0 \geq k}, 1_{\mathcal{K}_0 > k}$ are \mathfrak{F}^k -measurable. Thus, $1_{\mathcal{K}_0 > k-1} \equiv 1_{\mathcal{K}_0 \geq k}$ is \mathfrak{F}^{k-1} -measurable.

E.2 Result

We now formally prove [Theorem 3.5](#). The following [Theorem E.1](#) can readily be seen to imply [Theorem 3.5](#).

Theorem E.1. *Suppose F satisfies [Assumption 1.2](#) and the stochastic gradient oracle satisfies [Assumption 3.1](#) and [Assumption 3.2](#). Run [Algorithm 2](#) initialized at \mathbf{w}_0 , run with parameters chosen as per [Subsection E.1](#).*

Consider any $p \in (0, 1)$. With probability at least $1 - \frac{7}{4}p \cdot \frac{(F(\mathbf{w}_0)+1)7\eta K_0}{B^2}$, upon making

$$K_0 + \frac{7\eta K_0^2 (F(\mathbf{w}_0) + 1)}{B^2} \text{ oracle calls to } \nabla f(\cdot; \cdot),$$

[Algorithm 2](#) will output $\tilde{O}\left(\frac{7\eta K_0^2 (F(\mathbf{w}_0)+1)}{B^2}\right)$ candidate vectors \mathbf{w} , one of which satisfies

$$\|\nabla F(\mathbf{w})\| \leq 18L_2(\mathbf{w}_0)B^2, \lambda_{\min}(\nabla^2 F(\mathbf{w})) \geq -17\delta.$$

Remark 13. Before proceeding, we justify why [Theorem E.1](#) implies [Theorem 3.5](#). Simply take $\varepsilon \leftarrow \frac{\varepsilon}{289L_2(\mathbf{w}_0)}$ in [Theorem E.1](#). Plugging this in, we obtain a result on finding a SOSF as per the definition in (2).¹¹ The oracle complexity has the desired dependence on ε and polylog dependence on d, p . The probability is at least $1 - p \cdot \tilde{\Theta}(\varepsilon^{-1.5})$, where the $\tilde{\Theta}$ are hiding polylog terms in $d, 1/\varepsilon, 1/p$ and dependence on $F(\mathbf{w}_0)$ (through $\rho_1(\cdot), \rho_2(\cdot), \rho_3(\cdot), \sigma(\cdot)$). This holds for any $p \in (0, 1)$.

Now consider the final desired success probability $1 - \tilde{\delta}$ governed in terms of $\tilde{\delta} \in (0, 1)$ in [Theorem 3.5](#). Let $p = \tilde{\delta}^{\varepsilon^{1.5}} \cdot \text{polylog}(d, 1/\varepsilon)$ in the guarantee from the above paragraph. This gives [Theorem 3.5](#), with the requested probability and oracle complexity.

We now prove [Theorem E.1](#) via our framework, [Theorem 2.1](#).

Proof of Theorem E.1 and thus Theorem 3.5. We again use our framework [Theorem 2.1](#). Consider any $p \in (0, 1)$, and choose parameters as per [Subsection E.1](#).

Let

$$\mathcal{S} = \{\mathbf{w} : \|\nabla F(\mathbf{w})\| \leq 18L_2(\mathbf{w}_0)B^2, \lambda_{\min}(\nabla^2 F(\mathbf{w})) \geq -17\delta\}.$$

Define \mathcal{A} as follows, identically to how we defined them for Restarted SGD in [Subsection 2.3](#). Consider any given $\mathbf{u}_0 \in \mathbb{R}^d$. Let $\mathbf{p}_0 = \mathbf{u}_0$. We define a sequence $(\mathbf{p}_i)_{0 \leq i \leq K_0}$ via $\mathbf{p}_i = \mathbf{p}_{i-1} - \eta(\nabla f(\mathbf{p}_{i-1}; \zeta_i) + \tilde{\sigma}\Lambda^i)$. Note this sequence can be equivalently defined by repeatedly composing the function $\mathbf{u} \rightarrow \mathbf{u} - \eta(\nabla f(\mathbf{u}; \zeta) + \tilde{\sigma}\Lambda)$.

¹¹Recall this definition refers to \mathbf{w} such that $\|\nabla F(\mathbf{w})\| \leq \varepsilon, \nabla^2 F(\mathbf{w}) \geq -\sqrt{\varepsilon}I$.

If it exists, let $i, 1 \leq i \leq K_0$ be the minimal index such that $\|\mathbf{p}_i - \mathbf{p}_0\| > B$. Otherwise let $i = K_0$. In either case, we define

$$\mathcal{A}(\mathbf{u}_0) = \left(\mathbf{p}_i, \frac{1}{i} \sum_{t=0}^{i-1} \mathbf{p}_t \right), \text{ hence } \mathcal{A}_1(\mathbf{u}_0) = \mathbf{p}_i, \mathcal{A}_2(\mathbf{u}_0) = \frac{1}{i} \sum_{t=0}^{i-1} \mathbf{p}_t.$$

We now let

$$t_{\text{oracle}}(\mathbf{u}_0) = K_0, \text{ and } \Delta = \frac{B^2}{7\eta K_0}.$$

Following the notation from [Algorithm 2](#), notice that $\mathcal{A}(\mathbf{u}_0)$ corresponds to next vector set to \mathbf{x}^0 in the while loop of [Algorithm 2](#), when the while loop begins at $\mathbf{x}^0 = \mathbf{u}_0$.

Crucial to this proof are the following two Lemmas. While inspired from [Fang et al. \(2019\)](#), a crucial difference is that *they hold only in the $F(\mathbf{w}_0)$ -sublevel set $\mathcal{L}_{F,F(\mathbf{w}_0)}$* .

Lemma E.1 (Equivalent of Proposition 10, [Fang et al. \(2019\)](#)). *Consider \mathbf{x}^0 in the while loop of [Algorithm 2](#). Suppose $\mathbf{x}^0 \in \mathcal{L}_{F,F(\mathbf{w}_0)}$. With probability at least $1 - p$, if \mathbf{x}^k does not move out of the ball $\mathbb{B}(\mathbf{x}^0, B)$ within the first K_0 iterations in the while loop of [Algorithm 2](#), letting $\bar{\mathbf{x}} = \frac{1}{K_0} \sum_{k=0}^{K_0-1} \mathbf{x}^k$, we have*

$$\|\nabla F(\bar{\mathbf{x}})\| \leq 18L_2(\mathbf{w}_0)B^2, \lambda_{\min}(\nabla^2 F(\bar{\mathbf{x}})) \geq -17\delta.$$

Lemma E.2 (Equivalent of Proposition 9, [Fang et al. \(2019\)](#)). *Consider \mathbf{x}^0 in the while loop of [Algorithm 2](#). Suppose $\mathbf{x}^0 \in \mathcal{L}_{F,F(\mathbf{w}_0)}$. With probability at least $1 - \frac{3}{4}p$, if \mathbf{x}^k moves out of $\mathbb{B}(\mathbf{x}^0, B)$ in K_0 iterations or fewer in the while loop of [Algorithm 2](#), we have*

$$F(\mathbf{x}^{K_0}) < F(\mathbf{x}^0) - \frac{B^2}{7\eta K_0}.$$

Finishing the proof: The main point is to prove the following Claim.

Claim 7. *For any $\mathbf{u}_0 \in \mathcal{L}_{F,F(\mathbf{w}_0)}$, \mathcal{A} is a $(\mathcal{S}, K_0, \Delta, \frac{7}{4}p, \mathbf{u}_0)$ -decrease procedure.*

Proof of Claim 7. Apply [Lemma E.1](#) and [Lemma E.2](#) to the sequence $(\mathbf{p}_i)_{0 \leq i \leq K_0}$, recalling that $\mathcal{A}(\mathbf{u}_0)$ corresponds to next vector set to \mathbf{x}^0 in the while loop of [Algorithm 2](#) when the while loop begins at $\mathbf{x}^0 = \mathbf{p}_0 = \mathbf{u}_0$. By a Union Bound over the events of [Lemma E.1](#) and [Lemma E.2](#), with probability at least $1 - \frac{7}{4}p$, we have the following:

- Suppose there exists $t < K_0$ such that $\mathbf{p}_t \notin \mathbb{B}(\mathbf{p}_0, B) = \mathbb{B}(\mathbf{u}_0, B)$. Let t' be the minimal such t . By [Lemma E.2](#), we have

$$F(\mathcal{A}_1(\mathbf{u}_0)) = F(\mathbf{p}_{t'}) \leq F(\mathbf{p}_0) - \frac{B^2}{7\eta K_0} = F(\mathbf{u}_0) - \Delta.$$

- Otherwise, we have $\mathcal{A}_2(\mathbf{u}_0) = \bar{\mathbf{p}}$ where $\bar{\mathbf{p}} = \frac{1}{K_0} \sum_{k=0}^{K_0-1} \mathbf{p}_k$. In this case, by [Lemma E.1](#), we have

$$\mathcal{A}_2(\mathbf{u}_0) = \bar{\mathbf{p}} \in \mathcal{S}.$$

Consequently, \mathcal{A} is a $(\mathcal{S}, K_0, \Delta, \frac{7}{4}p, \mathbf{u}_0)$ -decrease procedure. \square

Now with [Claim 7](#), directly applying [Theorem 2.1](#) and plugging in the relevant parameters, we obtain [Theorem E.1](#). \square

Remark 14. To sanity check these results, note the rate from [Lemma E.2](#) will get worse as η gets smaller because $K_0\eta = 2 \lfloor \frac{\log(3/p)}{\log(0.8^{-1})} + 1 \rfloor \log\left(\frac{24\sqrt{d}}{\eta}\right) \delta_2^{-1}$ will increase as η gets smaller.

The rest of [Section E](#) will now be devoted to the proofs of [Lemma E.1](#) and [Lemma E.2](#). For the rest of [Section E](#), we suppose F satisfies [Assumption 1.2](#) and the stochastic gradient oracle satisfies [Assumption 3.1](#) and [Assumption 3.2](#). These proofs are similar to that of [Fang et al. \(2019\)](#), but hinges crucially on the fact that the analysis in [Fang et al. \(2019\)](#) is ‘local’.

E.3 Preliminaries

We now establish useful properties of the parameters of the algorithm defined in [Subsection E.1](#), analogously to [Lemma D.1](#).

Locality of balls $\mathbb{B}(\mathbf{x}^0, B)$:

Lemma E.3. *We have $B \leq \frac{1}{2\rho_0(F(\mathbf{w}_0)+1)}$. In particular, for any $\mathbf{u} \in \mathbb{B}(\mathbf{w}, B)$ for $\mathbf{w} \in \mathcal{L}_{F,F(\mathbf{w}_0)}$, we have $\|\mathbf{u} - \mathbf{w}\| \leq \frac{1}{2\rho_0(F(\mathbf{w}_0)+1)} \leq \frac{1}{2\rho_0(F(\mathbf{w})+1)}$.*

Proof. As per [Remark 12](#), we have $\varepsilon \leq 1$. Thus by the choice of parameters in [\(52\)](#),

$$B \leq \frac{\delta}{L_2(\mathbf{w}_0)} \leq \frac{1}{\sqrt{L_2(\mathbf{w}_0)}} \leq \frac{1}{2\rho_0(F(\mathbf{w}_0) + 1)}.$$

This completes the proof. \square

Control over the stochastic gradient oracle:

Lemma E.4. *For all \mathbf{u} such that $\mathbf{u} \in \mathbb{B}\left(\mathbf{w}, \frac{1}{\rho_0(F(\mathbf{w}_0)+1)}\right)$ for $\mathbf{w} \in \mathcal{L}_{F,F(\mathbf{w}_0)}$, we have $\|\nabla f(\mathbf{u}; \zeta) - \nabla F(\mathbf{u})\| \leq \sigma'(\mathbf{w}_0)$ for all ζ .*

Proof. By [Assumption 3.1](#), we have

$$\|\nabla f(\mathbf{u}; \zeta) - \nabla F(\mathbf{u})\| \leq \sigma(F(\mathbf{u})).$$

Now as $\mathbf{w} \in \mathcal{L}_{F,F(\mathbf{w}_0)}$, we have

$$\frac{1}{\rho_0(F(\mathbf{w}_0) + 1)} \leq \frac{1}{\rho_0(F(\mathbf{w}) + 1)}.$$

Thus by [Lemma 3.1](#) and again as $\mathbf{w} \in \mathcal{L}_{F,F(\mathbf{w}_0)}$, we have

$$F(\mathbf{u}) \leq F(\mathbf{w}) + 1 \leq F(\mathbf{w}_0) + 1.$$

Combining these gives [Lemma E.4](#). \square

Lemma E.5. *For all \mathbf{u} such that $\mathbf{u} \in \mathbb{B}\left(\mathbf{w}, \frac{1}{\rho_0(F(\mathbf{w}_0)+1)}\right)$ for $\mathbf{w} \in \mathcal{L}_{F,F(\mathbf{w}_0)}$, $\|\nabla \tilde{f}(\mathbf{u}; \zeta) - \nabla F(\mathbf{u})\| \leq \sigma_1(\mathbf{w}_0)$ for all ζ .*

Proof. This immediately follows from [Lemma E.4](#) and the definition of $\nabla \tilde{f}(\mathbf{u}; \zeta)$, as $\|\tilde{\sigma}\Lambda^t\| \leq \tilde{\sigma}$. \square

Locality after one step of SGD:

Lemma E.6. *Consider any $\mathbf{u} \in \mathbb{B}(\mathbf{w}, B)$ for $\mathbf{w} \in \mathcal{L}_{F,F(\mathbf{w}_0)}$. Then for all points \mathbf{p} in the line segment between \mathbf{u} and $\mathbf{u} - \eta \nabla \tilde{f}(\mathbf{u}; \zeta)$ for any ζ , we have $\mathbf{p} \in \mathbb{B}\left(\mathbf{w}, \frac{1}{\rho_0(F(\mathbf{w}_0)+1)}\right)$.*

Proof. It suffices to show $\mathbf{u} - \eta \nabla \tilde{f}(\mathbf{u}; \zeta) \in \mathbb{B}\left(\mathbf{w}, \frac{1}{2(\rho_0(F(\mathbf{w}_0)+1))}\right)$; after establishing this, the result then follows by Triangle Inequality and [Lemma E.3](#). To this end, by Triangle Inequality, it suffices to show that

$$\eta \|\nabla \tilde{f}(\mathbf{u}; \zeta)\| \leq \frac{1}{2\rho_0(F(\mathbf{w}_0) + 1)}.$$

Indeed, the same reasoning as in the proof of [Lemma E.3](#) gives

$$F(\mathbf{u}) \leq F(\mathbf{w}_0) + 1.$$

Thus, [Assumption 3.2](#) gives

$$\|\nabla F(\mathbf{u})\| \leq \rho_0(F(\mathbf{w}_0) + 1),$$

and so [Lemma E.5](#) gives

$$\|\nabla \tilde{f}(\mathbf{u}; \boldsymbol{\zeta})\| \leq \sigma_1(\mathbf{w}_0) + \rho_0(F(\mathbf{w}_0) + 1).$$

As per [Remark 12](#), we have

$$\eta \leq \frac{1}{2} B^2 \delta \leq \frac{1}{2} \cdot \frac{\delta^3}{L_2(\mathbf{w}_0)^2} \leq \frac{1}{2L_2(\mathbf{w}_0)^{0.5}}.$$

Combining all the above gives

$$\begin{aligned} \eta \|\nabla \tilde{f}(\mathbf{u}; \boldsymbol{\zeta})\| &\leq \frac{1}{2L_2(\mathbf{w}_0)^{0.5}} \cdot (\sigma_1(\mathbf{w}_0) + \rho_0(F(\mathbf{w}_0) + 1)) \\ &\leq \frac{1}{2\rho_0(F(\mathbf{w}_0) + 1)(\sigma_1(\mathbf{w}_0) + \rho_0(F(\mathbf{w}_0) + 1))} \cdot (\sigma_1(\mathbf{w}_0) + \rho_0(F(\mathbf{w}_0) + 1)) \\ &\leq \frac{1}{2\rho_0(F(\mathbf{w}_0) + 1)}, \end{aligned}$$

which by our earlier remarks completes the proof. \square

Properties of the effective smoothness parameters:

Lemma E.7. Consider any $\mathbf{x}^0 \in \mathcal{L}_{F, F(\mathbf{w}_0)}$. Then we have $\|\nabla^2 F(\mathbf{u})\|_{\text{op}} \leq L_1(\mathbf{w}_0)$ for all \mathbf{u} such that either:

- $\mathbf{u} \in \mathbb{B}(\mathbf{x}^0, B)$,
- Or \mathbf{u} lies in the line segment between some $\mathbf{u}' \in \mathbb{B}(\mathbf{x}^0, B)$ and $\mathbf{u}' - \eta \nabla \tilde{f}(\mathbf{u}'; \boldsymbol{\zeta})$, for any $\boldsymbol{\zeta}$.

Proof. By [Lemma E.3](#) and [Lemma E.6](#), irrespective of which case for \mathbf{u} in the conditions of [Lemma E.7](#) holds, we have

$$\mathbf{u} \in \mathbb{B}\left(\mathbf{x}^0, \frac{1}{\rho_0(F(\mathbf{w}_0) + 1)}\right).$$

As $\mathbf{x}^0 \in \mathcal{L}_{F, F(\mathbf{w}_0)}$, this implies

$$\|\mathbf{u} - \mathbf{x}^0\| \leq \frac{1}{\rho_0(F(\mathbf{w}_0) + 1)} \leq \frac{1}{\rho_0(F(\mathbf{x}^0) + 1)}.$$

By [Lemma 3.1](#) and as $\mathbf{x}^0 \in \mathcal{L}_{F, F(\mathbf{w}_0)}$, it follows that

$$F(\mathbf{u}) \leq F(\mathbf{x}^0) + 1 \leq F(\mathbf{w}_0) + 1.$$

The conclusion now follows by [Assumption 1.1](#). \square

Lemma E.8. Consider any $\mathbf{x}^0 \in \mathcal{L}_{F, F(\mathbf{w}_0)}$. Consider any $\mathbf{u}_1, \mathbf{u}_2$ such that each \mathbf{u}_i , $i = 1, 2$ is such that either:

- $\mathbf{u}_i \in \mathbb{B}(\mathbf{x}^0, B)$,
- Or \mathbf{u}_i lies in the line segment between some $\mathbf{u}' \in \mathbb{B}(\mathbf{x}^0, B)$ and $\mathbf{u}' - \eta \nabla \tilde{f}(\mathbf{u}'; \boldsymbol{\zeta})$, for any $\boldsymbol{\zeta}$.

Then

$$\|\nabla^2 F(\mathbf{u}_1) - \nabla^2 F(\mathbf{u}_2)\|_{\text{op}} \leq L_2(\mathbf{w}_0) \|\mathbf{u}_1 - \mathbf{u}_2\|.$$

Proof. Irrespective of which condition applies to \mathbf{u}_i , By [Lemma E.3](#) and [Lemma E.6](#), we have

$$\mathbf{u}_i \in \mathbb{B}\left(\mathbf{x}^0, \frac{1}{\rho_0(F(\mathbf{w}_0) + 1)}\right)$$

for $i = 1, 2$. Thus the line segment $\overline{\mathbf{u}_1 \mathbf{u}_2}$ is contained in $\mathbb{B}\left(\tilde{\mathbf{w}}, \frac{1}{\rho_0(F(\mathbf{w}_0) + 1)}\right)$. As $\mathbf{x}^0 \in \mathcal{L}_{F, F(\mathbf{w}_0)}$, the result now follows from applying [Lemma A.6](#) and [Lemma 3.1](#). \square

Remark 15. The reason for the second case in the condition on \mathbf{u} or \mathbf{u}_i from [Lemma E.7](#), [Lemma E.8](#) will become clear in the proof of [Lemma E.2](#). In particular, to prove [Lemma E.2](#), we will consider $\mathbf{u} - \eta \nabla \tilde{f}(\mathbf{u}; \boldsymbol{\zeta})$ for $\mathbf{u} \in \mathbb{B}(\mathbf{x}^0, B)$ where $\mathbf{x}^0 \in \mathcal{L}_{F, F(\mathbf{w}_0)}$.

Lemma E.9. Consider any $\mathbf{x}^0 \in \mathcal{L}_{F, F(\mathbf{w}_0)}$. Then for any $\mathbf{u} \in \mathbb{B}(\mathbf{x}^0, B)$ and any $\boldsymbol{\zeta}$,

$$\|\nabla^2 f(\mathbf{u}; \boldsymbol{\zeta})\|_{\text{op}} \leq L_1(\mathbf{w}_0).$$

Proof. By [Lemma E.3](#), we have

$$\mathbf{u} \in \mathbb{B}\left(\mathbf{x}^0, \frac{1}{\rho_0(F(\mathbf{w}_0) + 1)}\right).$$

By [Lemma 3.1](#), because $\mathbf{x}^0 \in \mathcal{L}_{F, F(\mathbf{w}_0)}$, we have

$$F(\mathbf{u}) \leq F(\mathbf{w}_0) + 1.$$

Moreover, as $\mathbf{x}^0 \in \mathcal{L}_{F, F(\mathbf{w}_0)}$ and by [Lemma E.4](#) and [Corollary 1](#),

$$\|\nabla f(\mathbf{u}; \boldsymbol{\zeta})\| \leq \|\nabla F(\mathbf{u})\| + \sigma'(\mathbf{w}_0) \leq \rho_0(F(\mathbf{w}_0) + 1) + \sigma'(\mathbf{w}_0).$$

Thus the result follows from [Assumption 3.2](#). \square

Remark 16. While [Lemma E.9](#) is phrased as an upper bound on the operator norm of $\nabla^2 f(\cdot; \boldsymbol{\zeta})$, it can be easily phrased in terms of the local Lipschitz constant of $\nabla f(\cdot; \boldsymbol{\zeta})$, similar to one of the possibilities in [Assumption 1.2](#).

Enough noise to escape saddles: Now we verify that the noise scheme here gives us enough noise to escape saddle points in the $F(\mathbf{w}_0)$ -sublevel set $\mathcal{L}_{F, F(\mathbf{w}_0)}$.

Definition E.1 ((q^*, \mathbf{v}) -narrow property; Definition 2 in [Fang et al. \(2019\)](#)). A Borel set $\mathcal{A} \subset \mathbb{R}^d$ satisfies the (q^*, \mathbf{v}) -narrow property if for any $\mathbf{u} \in \mathcal{A}$, $q \geq q^*$, $\mathbf{u} + q\mathbf{v} \in \mathcal{A}^c$.

Immediately, we obtain the following properties of this definition, as also noted in [Fang et al. \(2019\)](#).

Lemma E.10. If \mathcal{A} satisfies the (q^*, \mathbf{v}) -narrow property, then for all $c_1 \in \mathbb{R}^d$, $c_2 \in \mathbb{R}$, $c_1 + c_2\mathcal{A}$ satisfies the $(|c_2|q^*, \mathbf{v})$ -narrow property.

We now introduce the following definition:

Definition E.2 (\mathbf{v} -dispersive Property; Equivalent of Definition 3 in [Fang et al. \(2019\)](#)). We say that a random vector $\tilde{\boldsymbol{\xi}}$ has the \mathbf{v} -dispersive property if for any \mathcal{A} satisfying the $\left(\frac{\sigma_1(\mathbf{w}_0)}{4\sqrt{d}}, \mathbf{v}\right)$ -narrow property, we have

$$\mathbb{P}(\tilde{\boldsymbol{\xi}} \in \mathcal{A}) \leq \frac{1}{2}.$$

Note the slight change of the constant $\frac{1}{2}$ rather than $\frac{1}{4}$ in the above definition compared to that of [Fang et al. \(2019\)](#); this subtle difference will appear in the following proofs, although this will not change too much conceptually.

Now we prove the following Lemma, which shows that our update rule contains enough noise to escape saddle points:

Lemma E.11 (Dispersive Noise; see also Algorithm 3, [Fang et al. \(2019\)](#)). The update $\nabla \tilde{f}(\mathbf{x}^t; \boldsymbol{\zeta}_{t+1})$ admits the \mathbf{v} -dispersive property for all unit vectors \mathbf{v} , for any \mathbf{x}^t .

Note this does not necessarily hold for the stochastic gradient oracle itself under our assumptions, hence the artificial noise injection of $\tilde{\sigma}\boldsymbol{\Lambda}^t$.

Proof of Lemma E.11. First, we prove that the random vector $\tilde{\sigma}\boldsymbol{\Lambda}^{t+1}$ satisfies the Dispersive Noise property for all unit vectors \mathbf{v} . Consider any \mathcal{A} satisfying the $\left(\frac{\sigma_1(\mathbf{w}_0)}{4\sqrt{d}}, \mathbf{v}\right)$ -narrow property. Note we have

$$\mathbb{P}(\tilde{\sigma}\boldsymbol{\Lambda}^{t+1} \in \mathcal{A}) = \mathbb{P}(\boldsymbol{\Lambda}^{t+1} \in \tilde{\sigma}^{-1}\mathcal{A})$$

$$\begin{aligned}
&\leq \frac{\sigma_1(\mathbf{w}_0)/4\sqrt{d}}{\tilde{\sigma}} \cdot \frac{\text{Vol}^{d-1}\mathbb{B}(\tilde{\mathbf{0}}, 1)}{\text{Vol}^d\mathbb{B}(\tilde{\mathbf{0}}, 1)} \\
&\leq \frac{\sigma_1(\mathbf{w}_0)/4\sqrt{d}}{\tilde{\sigma}} \cdot \sqrt{d} = \frac{\sigma_1(\mathbf{w}_0)}{4\tilde{\sigma}}.
\end{aligned}$$

Here, the inequality follows from an elementary calculation with multivariate calculus, analogous to the calculation in the proof of [Lemma D.3](#), which we detailed in full in this article. An analogous calculation can also be found in [Jin et al. \(2017\)](#), proof of Lemma 14, and in Appendix F, [Fang et al. \(2019\)](#).

Now, note as $\tilde{\sigma} \geq \sigma'(\mathbf{w}_0)$, we have

$$\frac{\sigma_1(\mathbf{w}_0)}{4\tilde{\sigma}} \leq \frac{\sigma'(\mathbf{w}_0) + \tilde{\sigma}}{4\tilde{\sigma}} \leq \frac{1}{2},$$

and so

$$\mathbb{P}(\tilde{\sigma}\mathbf{\Lambda}^{t+1} \in \mathcal{A}) \leq \frac{1}{2}.$$

Consequently the random vector $\tilde{\sigma}\mathbf{\Lambda}^t$ satisfies the Dispersive Noise property for all unit vectors \mathbf{v} .

Now, we show that $\nabla \tilde{f}(\mathbf{x}^t; \zeta_{t+1})$ satisfies the \mathbf{v} -dispersive property as wanted. The proof is analogous to part iii, Proposition 4 of [Fang et al. \(2019\)](#). Consider any unit vector \mathbf{v} . Recall that $\mathbf{\Lambda}^t$ and $\nabla f(\mathbf{x}^t; \zeta_{t+1})$ are independent. Since the (q^*, \mathbf{v}) -narrow property is evidently preserved with the same parameters by adding a fixed vector to \mathcal{A} , we obtain the following bound on the following conditional probability:

$$\begin{aligned}
\mathbb{P}(\nabla \tilde{f}(\mathbf{x}^t; \zeta_{t+1}) \in \mathcal{A} | \nabla f(\mathbf{x}^t; \zeta_{t+1})) &= \mathbb{P}(\nabla f(\mathbf{x}^t; \zeta_{t+1}) + \tilde{\sigma}\mathbf{\Lambda}^{t+1} \in \mathcal{A} | \nabla f(\mathbf{x}^t; \zeta_{t+1})) \\
&= \mathbb{P}(\tilde{\sigma}\mathbf{\Lambda}^{t+1} \in -\nabla f(\mathbf{x}^t; \zeta_{t+1}) + \mathcal{A} | \nabla f(\mathbf{x}^t; \zeta_{t+1})) \leq \frac{1}{2}.
\end{aligned}$$

This holds irrespective of conditioning, which implies that $\nabla \tilde{f}(\mathbf{x}^t; \zeta_{t+1})$ satisfies the \mathbf{v} -dispersive property. \square

E.4 Escaping Saddles

We first aim to prove that we can efficiently escape strict saddle points in the $F(\mathbf{w}_0)$ -sublevel set, similarly to [Fang et al. \(2019\)](#). In particular, we aim to prove the following [Lemma E.12](#). The contrapositive of [Lemma E.12](#) will in turn be used to prove [Lemma E.1](#), which establishes that [Algorithm 2](#) can find SOSPs.

Lemma E.12 (Equivalent of Proposition 7 in [Fang et al. \(2019\)](#)). *Consider a sequence of iterates $\mathbf{x}^0, \mathbf{x}^1, \dots$ beginning at \mathbf{x}^0 comprising an instance of the while loop in [Algorithm 2](#). Suppose $\mathbf{x}^0 \in \mathcal{L}_{F, F(\mathbf{w}_0)}$ and that $\lambda_{\min}(\nabla^2 F(\mathbf{x}^0)) \leq -\delta_2$ for $\delta_2 > 0$. Then when the while loop of [Algorithm 2](#) is initialized at \mathbf{x}^0 , with probability at least $1 - \frac{p}{3}$, we have*

$$\mathcal{K}_0 \leq K_0 = \lfloor \frac{\log(3/p)}{\log(0.8^{-1})} + 1 \rfloor K_o.$$

Remark 17. For δ_2 very small, note the guarantee from [Lemma E.12](#) will deteriorate because K_0 scales with δ_2^{-1} .

To prove [Lemma E.12](#), we use the same strategy as in [Fang et al. \(2019\)](#). However, as we do not have global Lipschitzness of the gradient and Hessian, we must be careful. We use that the strategy only requires control over points that are ‘local’, i.e. near \mathbf{x}^0 , since the proof strategy studies escape from the ball $\mathbb{B}(\mathbf{x}^0, B)$. We then appeal to control over F in $\mathbb{B}(\mathbf{x}^0, B)$ that we have by [Subsection E.3](#).

Remark 18. In this section [Subsection E.4](#), probability is over the samples ζ_k and the artificial noise injections $\mathbf{\Lambda}^k$.

Now we go into the details. As in [Fang et al. \(2019\)](#), let $\mathbf{w}^k(\mathbf{u})$ be the iterates of SGD starting from a given \mathbf{u} using the *same* stochastic samples as \mathbf{x}^k and the same noise additions $\tilde{\sigma}\mathbf{\Lambda}^k$. In particular

$$\mathbf{w}^k(\mathbf{u}) = \mathbf{w}^{k-1}(\mathbf{u}) - \eta \nabla \tilde{f}(\mathbf{w}^{k-1}(\mathbf{u}); \zeta_k).$$

Thus $\mathbf{x}^k = \mathbf{w}^k(\mathbf{x}^0)$.

Also for all \mathbf{u} , let $\mathcal{K}_{\text{exit}}(\mathbf{u})$ be the stopping time defined by

$$\mathcal{K}_{\text{exit}}(\mathbf{u}) := \inf\{k \geq 0 : \|\mathbf{w}^k(\mathbf{u}) - \mathbf{x}^0\| > B\}.$$

Thus $\mathcal{K}_0 = \mathcal{K}_{\text{exit}}(\mathbf{x}^0)$.

The high-level idea from Fang et al. (2019), similar to as in Jin et al. (2017), is to consider the ‘bad initialization region’ around $\mathbb{B}(\mathbf{x}^0, B)$ where iterates initialized in this bad region escape with low probability. We then prove that this bad initialization region is ‘narrow’, and consequently we can escape the saddle point efficiently.

In particular, define

$$\mathcal{S}_{K_o}^B(\mathbf{x}^0) = \{\mathbf{u} \in \mathbb{R}^d : \mathbb{P}(\mathcal{K}_{\text{exit}}(\mathbf{u}) < K_o) \leq 0.4\}.$$

Note by definition that $\mathcal{S}_{K_o}^B(\mathbf{x}^0) \subseteq \mathbb{B}(\mathbf{x}^0, B)$.

First let $q_0 = \frac{\sigma_1(\mathbf{w}_0)\eta}{4\sqrt{d}}$. We establish the following Lemma, which verifies that $\mathcal{S}_{K_o}^B(\mathbf{x}^0)$ is ‘narrow’ in a suitable sense.

Lemma E.13 (Equivalent of Lemma 8 in Fang et al. (2019); also similar to Lemma 15, Jin et al. (2017)). *Suppose the assumptions of Lemma E.12 hold. Let \mathbf{e}_1 be an arbitrary unit eigenvector of $\nabla^2 F(\mathbf{x}^0)$ corresponding to its smallest eigenvalue $-\delta_m \leq -\delta_2$. Then for any $q \geq q_0 = \frac{\sigma_1(\mathbf{w}_0)\eta}{4\sqrt{d}}$ and any $\mathbf{u}, \mathbf{u} + q\mathbf{e}_1 \in \mathbb{B}(\mathbf{x}^0, B)$, we have that*

$$\mathbb{P}((\mathcal{K}_{\text{exit}}(\mathbf{u}) \geq K_o) \text{ and } (\mathcal{K}_{\text{exit}}(\mathbf{u} + q\mathbf{e}_1) \geq K_o)) \leq 0.1.$$

Here probability is over the single sequence of samples used to compute stochastic gradients and the artificial noise injection.

Remark 19. The proof of Lemma E.13 crucially uses that $\nabla^2 F(\mathbf{x}^0)$ has a negative eigenvector, as one would expect.

Note we have, as in Fang et al. (2019), that

$$K_o = 2 \log\left(\frac{24\sqrt{d}}{\eta}\right) \eta^{-1} \delta_2^{-1} \geq \frac{\log(6/q_0)}{\log(1 + \eta\delta_2)} \geq \frac{\log(6B/q_0)}{\log(1 + \eta\delta_2)}. \quad (54)$$

This follows evidently from the choice of parameters and definition of q_0 , and Remark 12 which states that it is enough to show the result for $\eta\delta_2$ at most a universal constant, namely one satisfying $\log(1 + x) \geq \frac{x}{2}$. Now using Lemma E.13, we prove Lemma E.12:

Proof of Lemma E.12 given Lemma E.13. Given Lemma E.13, we first prove that the bad initialization region $\mathcal{S}_{K_o}^B(\mathbf{x}^0)$ satisfies the (q_0, \mathbf{e}_1) -narrow property, i.e. that there are no points $\mathbf{u}, \mathbf{u} + q\mathbf{e}_1 \in \mathcal{S}_{K_o}^B(\mathbf{x}^0)$ where $q \geq q_0 = \frac{\sigma_1(\mathbf{w}_0)\eta}{4\sqrt{d}}$. This part of the proof is identical to Proposition 7, Fang et al. (2019). If such points existed we would have

$$\mathbb{P}(\mathcal{K}_{\text{exit}}(\mathbf{u}) \geq K_o) \geq 0.6, \mathbb{P}(\mathcal{K}_{\text{exit}}(\mathbf{u} + q\mathbf{e}_1) \geq K_o) \geq 0.6.$$

This implies

$$\begin{aligned} \mathbb{P}((\mathcal{K}_{\text{exit}}(\mathbf{u}) \geq K_o) \text{ and } (\mathcal{K}_{\text{exit}}(\mathbf{u} + q\mathbf{e}_1) \geq K_o)) &\geq \mathbb{P}(\mathcal{K}_{\text{exit}}(\mathbf{u}) \geq K_o) + \mathbb{P}(\mathcal{K}_{\text{exit}}(\mathbf{u} + q\mathbf{e}_1) \geq K_o) - 1 \\ &\geq 0.2, \end{aligned}$$

which contradicts Lemma E.13.

From here, we prove Lemma E.12. For this rest of the proof of Lemma E.12, we only consider \mathbf{u} and do not consider the iterates from $\mathbf{u} + q\mathbf{e}_1$. Recall $\mathcal{S}_{K_o}^B$ satisfies the (q_0, \mathbf{e}_1) -narrow property with $q_0 = \frac{\eta\sigma_1(\mathbf{w}_0)}{4\sqrt{d}}$ as shown above. Thus we have for any $\mathbf{u} \in \mathbb{B}(\mathbf{x}^0, B)$,

$$\begin{aligned} \mathbb{P}(\mathbf{w}^1(\mathbf{u}) \in \mathcal{S}_{K_o}^B(\mathbf{x}^0)) &= \mathbb{P}(\mathbf{u} - \eta \nabla \tilde{f}(\mathbf{u}; \zeta_1) \in \mathcal{S}_{K_o}^B(\mathbf{x}^0)) \\ &= \mathbb{P}(\nabla \tilde{f}(\mathbf{u}; \zeta_1) \in \eta^{-1}(-\mathcal{S}_{K_o}^B(\mathbf{x}^0) + \mathbf{u})) \leq \frac{1}{2}. \end{aligned} \quad (55)$$

The last step follows from the definition of the $\mathbf{w}^k(\mathbf{u})$, the scale and translation properties of the (q_0, \mathbf{e}_1) -narrow property which implies that $\eta^{-1}(-S_{K_o}^B(\mathbf{x}^0) + \mathbf{u})$ satisfies the $(\frac{\sigma_1(\mathbf{w}_0)}{4\sqrt{d}}, \mathbf{e}_1)$ -narrow property, and that $\nabla \tilde{f}(\mathbf{u}; \zeta_1)$ satisfies the \mathbf{e}_1 -dispersive property by [Lemma E.11](#).

Note as events we have $\{\mathcal{K}_{\text{exit}}(\mathbf{w}^1(\mathbf{u})) < K_o\} \subseteq \{\mathcal{K}_{\text{exit}}(\mathbf{u}) \leq K_o\}$. Thus by Law of Total Expectation, for all $\mathbf{u} \in \mathbb{B}(\mathbf{x}^0, B)$,

$$\begin{aligned} \mathbb{P}(\mathcal{K}_{\text{exit}}(\mathbf{u}) \leq K_o) &\geq \mathbb{P}(\mathcal{K}_{\text{exit}}(\mathbf{w}^1(\mathbf{u})) < K_o) \\ &\geq \mathbb{E}[\mathbb{P}(\mathcal{K}_{\text{exit}}(\mathbf{w}^1(\mathbf{u})) < K_o | \mathfrak{F}^1) | \{\mathbf{w}^1(\mathbf{u}) \in (\mathcal{S}_{K_o}^B(\mathbf{x}^0))^c\}]. \end{aligned} \quad (56)$$

Conditioned on $\mathbf{w}^1(\mathbf{u}) \in (\mathcal{S}_{K_o}^B(\mathbf{x}^0))^c$, we have by definition of $\mathcal{S}_{K_o}^B(\mathbf{x}^0)$ that $\mathbb{P}(\mathcal{K}_{\text{exit}}(\mathbf{w}^1(\mathbf{u})) < K_o | \mathfrak{F}^1) \geq 0.4$. By [\(55\)](#), for all $\mathbf{u} \in \mathbb{B}(\mathbf{x}^0, B)$, we have

$$\mathbb{P}(\mathbf{w}^1(\mathbf{u}) \in \mathcal{S}_{K_o}^B(\mathbf{x}^0)^c) \geq \frac{1}{2}.$$

Thus combining with [\(56\)](#) implies for all $\mathbf{u} \in \mathbb{B}(\mathbf{x}^0, B)$,

$$\mathbb{P}(\mathcal{K}_{\text{exit}}(\mathbf{u}) \leq K_o) \geq 0.4 \cdot \frac{1}{2} = 0.2. \quad (57)$$

Now consider any $N' \geq 1$. Notice as events,

$$\begin{aligned} \{\mathcal{K}_{\text{exit}}(\mathbf{u}) > N'K_o\} &= \{\mathcal{K}_{\text{exit}}(\mathbf{w}^{(N'-1)K_o}(\mathbf{u})) > K_o\} \\ &= \{\mathcal{K}_{\text{exit}}(\mathbf{w}^{(N'-1)K_o}(\mathbf{u})) > K_o\} \cap \{\mathcal{K}_{\text{exit}}(\mathbf{u}) > (N'-1)K_o\}. \end{aligned}$$

Therefore,

$$\mathbb{P}(\mathcal{K}_{\text{exit}}(\mathbf{u}) > N'K_o) = \mathbb{E}[\mathbb{P}(\mathcal{K}_{\text{exit}}(\mathbf{w}^{(N'-1)K_o}(\mathbf{u})) > K_o | \mathfrak{F}^{K_o}) | \{\mathcal{K}_{\text{exit}}(\mathbf{u}) > (N'-1)K_o\}].$$

Note that conditioned on $\mathcal{K}_{\text{exit}}(\mathbf{u}) > (N'-1)K_o$, it follows that $\mathcal{K}_{\text{exit}}(\mathbf{w}^{(N'-1)K_o}(\mathbf{u})) \in \mathbb{B}(\mathbf{x}^0, B)$. Therefore $\mathbb{P}(\mathcal{K}_{\text{exit}}(\mathbf{w}^{(N'-1)K_o}(\mathbf{u})) > K_o | \mathfrak{F}^{K_o}) \leq \sup_{\mathbf{u}' \in \mathbb{B}(\mathbf{x}^0, B)} \mathbb{P}(\mathcal{K}_{\text{exit}}(\mathbf{u}') > K_o)$. Using [\(57\)](#), we can upper bound

$$\begin{aligned} \mathbb{P}(\mathcal{K}_{\text{exit}}(\mathbf{u}) > N'K_o) &\leq \mathbb{P}(\mathcal{K}_{\text{exit}}(\mathbf{u}) > (N'-1)K_o) \cdot \sup_{\mathbf{u}' \in \mathbb{B}(\mathbf{x}^0, B)} \mathbb{P}(\mathcal{K}_{\text{exit}}(\mathbf{u}') > K_o) \\ &\leq 0.8 \mathbb{P}(\mathcal{K}_{\text{exit}}(\mathbf{u}) > (N'-1)K_o). \end{aligned} \quad (58)$$

Recall that $K_0 = \lfloor \frac{\log(3/p)}{\log(0.8^{-1})} + 1 \rfloor K_o$. Let $N = \lfloor \frac{\log(3/p)}{\log(0.8^{-1})} + 1 \rfloor$. We obtain by repeatedly applying [\(58\)](#) for $N' = N, N-1, \dots$ that

$$\mathbb{P}(\mathcal{K}_{\text{exit}}(\mathbf{u}) > NK_o) \leq 0.8^N \leq p/3.$$

This gives the desired result. \square

Now we prove [Lemma E.13](#).

Proof of Lemma E.13. Again, we proceed similarly as the proof of Lemma 8, [Fang et al. \(2019\)](#). The main difference is we only have control over the relevant derivatives prior to the escape from $\mathbb{B}(\mathbf{x}^0, B)$ (recall $\mathbf{x}^0 \in \mathcal{L}_{F,F}(\mathbf{w}_0)$). However, it turns out that this is sufficient for the proof to go through.

Setup. Recall that we have $\mathbf{w}^0(\mathbf{u}) = \mathbf{u}$, and

$$\begin{aligned} \mathbf{w}^k(\mathbf{u}) &= \mathbf{w}^{k-1}(\mathbf{u}) - \eta \nabla \tilde{f}(\mathbf{w}^{k-1}(\mathbf{u}); \zeta_k), \\ \mathbf{w}^k(\mathbf{u} + q\mathbf{e}_1) &= \mathbf{w}^{k-1}(\mathbf{u} + q\mathbf{e}_1) - \eta \nabla \tilde{f}(\mathbf{w}^{k-1}(\mathbf{u} + q\mathbf{e}_1); \zeta_k). \end{aligned}$$

Now define the following stopping time:

$$\mathcal{K}_1 = \mathcal{K}_{\text{exit}}(\mathbf{u}) \wedge \mathcal{K}_{\text{exit}}(\mathbf{u} + q\mathbf{e}_1).$$

For solely the purpose of analysis, consider the following sequence:

$$\mathbf{z}^k = \begin{cases} \mathbf{w}^k(\mathbf{u} + q\mathbf{e}_1) - \mathbf{w}^k(\mathbf{u}) & : k \leq \mathcal{K}_1 \\ (\mathbf{I} - \eta \nabla^2 F(\mathbf{x}^0)) \mathbf{z}^{k-1} & : k > \mathcal{K}_1 \end{cases}. \quad (59)$$

Clearly the \mathbf{z}^k are \mathfrak{F}^k -measurable, because the event $\{k \leq \mathcal{K}_1\}$ is \mathfrak{F}^k -measurable.

Remark 20. Note unlike Fang et al. (2019), the first case holds when $k \leq \mathcal{K}_1$ rather than $k < \mathcal{K}_1$. That being said we expect that if one uses the exact same definition as in Fang et al. (2019) for the \mathbf{z}^k , the proof this generalized smooth setting will still work, with a slightly modified argument compared to the proof we present.

Notice by definition of $\mathbf{w}^0(\mathbf{u})$, $\mathbf{w}^0(\mathbf{u} + q\mathbf{e}_1)$ and assumption of Lemma E.13 that $\mathbf{u}, \mathbf{u} + q\mathbf{e}_1 \in \mathbb{B}(\mathbf{x}^0, B)$, we have $\mathcal{K}_1 > 0$. Thus,

$$\mathbf{z}^0 = q\mathbf{e}_1.$$

Controlling the \mathbf{z}^k . Let $\mathbf{H} = \nabla^2 F(\mathbf{x}^0)$. We have the following lemma to control the \mathbf{z}^k from (59).

For all k , define

$$\mathbf{D}^k := \nabla^2 F(\mathbf{x}^0) - \int_0^1 \nabla^2 F(\mathbf{w}^k(\mathbf{u}) + \theta(\mathbf{w}^k(\mathbf{u} + q\mathbf{e}_1) - \mathbf{w}^k(\mathbf{u}))) d\theta, \quad (60)$$

$$\boldsymbol{\xi}_d^k := (\nabla F(\mathbf{w}^{k-1}(\mathbf{u} + q\mathbf{e}_1)) - \nabla F(\mathbf{w}^{k-1}(\mathbf{u}))) - (\nabla \tilde{f}(\mathbf{w}^{k-1}(\mathbf{u} + q\mathbf{e}_1); \boldsymbol{\zeta}_k) - \nabla \tilde{f}(\mathbf{w}^{k-1}(\mathbf{u}); \boldsymbol{\zeta}_k)). \quad (61)$$

Recall by definition of $\mathbf{w}^k(\mathbf{u})$, we have

$$\begin{aligned} \nabla \tilde{f}(\mathbf{w}^{k-1}(\mathbf{u} + q\mathbf{e}_1); \boldsymbol{\zeta}_k) &= \nabla f(\mathbf{w}^{k-1}(\mathbf{u} + q\mathbf{e}_1); \boldsymbol{\zeta}_k) + \tilde{\sigma} \boldsymbol{\Lambda}^k, \\ \nabla \tilde{f}(\mathbf{w}^{k-1}(\mathbf{u}); \boldsymbol{\zeta}_k) &= \nabla f(\mathbf{w}^{k-1}(\mathbf{u}); \boldsymbol{\zeta}_k) + \tilde{\sigma} \boldsymbol{\Lambda}^k, \end{aligned}$$

for the same noise sequence $\boldsymbol{\Lambda}^k$. Thus we also have

$$\boldsymbol{\xi}_d^k = (\nabla F(\mathbf{w}^{k-1}(\mathbf{u} + q\mathbf{e}_1)) - \nabla F(\mathbf{w}^{k-1}(\mathbf{u}))) - (\nabla f(\mathbf{w}^{k-1}(\mathbf{u} + q\mathbf{e}_1); \boldsymbol{\zeta}_k) - \nabla f(\mathbf{w}^{k-1}(\mathbf{u}); \boldsymbol{\zeta}_k)). \quad (62)$$

Lemma E.14 (Equivalent of Lemma 13, Fang et al. (2019)). *We have that for all $k \leq \mathcal{K}_1$,*

$$\mathbf{z}^k = (\mathbf{I} - \eta \mathbf{H}) \mathbf{z}^{k-1} + \eta \mathbf{D}^{k-1} \mathbf{z}^{k-1} + \eta \boldsymbol{\xi}_d^k.$$

Furthermore, we have the following properties of the \mathbf{D}^k and $\boldsymbol{\xi}_d^k$ defined in (60), (61):

1. For all such $k \leq \mathcal{K}_1$, we have

$$\|\mathbf{D}^{k-1}\| \leq L_2(\mathbf{w}_0) \max(\|\mathbf{w}^{k-1}(\mathbf{u} + q\mathbf{e}_1) - \mathbf{x}^0\|, \|\mathbf{w}^{k-1}(\mathbf{u}) - \mathbf{x}^0\|) \leq L_2(\mathbf{w}_0)B.$$

2. For all k , we have

$$\mathbb{E}[\boldsymbol{\xi}_d^k | \mathfrak{F}^{k-1}] = 0.$$

3. For all $k \leq \mathcal{K}_1$, we have

$$\|\boldsymbol{\xi}_d^k\| \leq 2L_1(\mathbf{w}_0) \|\mathbf{z}^{k-1}\|.$$

Proof. We prove each part one at a time:

1. For $k \leq \mathcal{K}_1$, using the definition of \mathbf{z}^k , it follows that

$$\begin{aligned} \mathbf{z}^k &= \mathbf{w}^k(\mathbf{u} + q\mathbf{e}_1) - \mathbf{w}^k(\mathbf{u}) \\ &= \mathbf{w}^{k-1}(\mathbf{u} + q\mathbf{e}_1) - \mathbf{w}^{k-1}(\mathbf{u}) - \eta(\nabla \tilde{f}(\mathbf{w}^{k-1}(\mathbf{u} + q\mathbf{e}_1); \boldsymbol{\zeta}_k) - \nabla \tilde{f}(\mathbf{w}^{k-1}(\mathbf{u}); \boldsymbol{\zeta}_k)) \\ &= \mathbf{z}^{k-1} - \eta(\nabla F(\mathbf{w}^{k-1}(\mathbf{u} + q\mathbf{e}_1)) - \nabla F(\mathbf{w}^{k-1}(\mathbf{u}))) \\ &\quad + \eta[(\nabla F(\mathbf{w}^{k-1}(\mathbf{u} + q\mathbf{e}_1)) - \nabla F(\mathbf{w}^{k-1}(\mathbf{u}))) - (\nabla \tilde{f}(\mathbf{w}^{k-1}(\mathbf{u} + q\mathbf{e}_1); \boldsymbol{\zeta}_k) - \nabla \tilde{f}(\mathbf{w}^{k-1}(\mathbf{u}); \boldsymbol{\zeta}_k))] \\ &= \mathbf{z}^{k-1} - \eta \left[\int_0^1 \nabla^2 F(\mathbf{w}^{k-1}(\mathbf{u}) + \theta(\mathbf{w}^{k-1}(\mathbf{u} + q\mathbf{e}_1) - \mathbf{w}^{k-1}(\mathbf{u}))) d\theta \right] \mathbf{z}^{k-1} + \eta \boldsymbol{\xi}_d^k \\ &= \mathbf{z}^{k-1} - \eta(\mathbf{H} - \mathbf{D}^{k-1}) \mathbf{z}^{k-1} + \eta \boldsymbol{\xi}_d^k. \end{aligned}$$

This proves the desired property of the \mathbf{z}^k .

2. For the required properties of the D^{k-1} , consider any $k \leq \mathcal{K}_1$. First, notice $\mathbf{w}^{k-1}(\mathbf{u}) + \theta(\mathbf{w}^{k-1}(\mathbf{u} + q\mathbf{e}_1) - \mathbf{w}^{k-1}(\mathbf{u})) = \theta\mathbf{w}^{k-1}(\mathbf{u} + q\mathbf{e}_1) + (1 - \theta)\mathbf{w}^{k-1}(\mathbf{u})$ for any $\theta \in [0, 1]$. For $k \leq \mathcal{K}_1$, both $\mathbf{w}^{k-1}(\mathbf{u} + q\mathbf{e}_1), \mathbf{w}^{k-1}(\mathbf{u}) \in \mathbb{B}(\mathbf{x}^0, B)$. Note this still remains true for $k = \mathcal{K}_1$ because for $k - 1 = \mathcal{K}_1 - 1 < \mathcal{K}_1$, the definition of \mathcal{K}_1 implies that the iterates $\mathbf{w}^{k-1}(\mathbf{u} + q\mathbf{e}_1), \mathbf{w}^{k-1}(\mathbf{u}) \in \mathbb{B}(\mathbf{x}^0, B)$.

Thus for any $\theta \in [0, 1]$, $\mathbf{w}^{k-1}(\mathbf{u}) + \theta(\mathbf{w}^{k-1}(\mathbf{u} + q\mathbf{e}_1) - \mathbf{w}^{k-1}(\mathbf{u})) \in \mathbb{B}(\mathbf{x}^0, B)$, and so all points \mathbf{p} on the line segment between \mathbf{x}^0 and $\mathbf{w}^{k-1}(\mathbf{u}) + \theta(\mathbf{w}^{k-1}(\mathbf{u} + q\mathbf{e}_1) - \mathbf{w}^{k-1}(\mathbf{u}))$ lie in $\mathbb{B}(\mathbf{x}^0, B)$. Thus by [Lemma E.8](#),

$$\begin{aligned} \|D^{k-1}\| &= \left\| \nabla^2 F(\mathbf{x}^0) - \int_0^1 \nabla^2 F(\mathbf{w}^{k-1}(\mathbf{u}) + \theta(\mathbf{w}^{k-1}(\mathbf{u} + q\mathbf{e}_1) - \mathbf{w}^{k-1}(\mathbf{u}))) d\theta \right\| \\ &\leq \int_0^1 \left\| \nabla^2 F(\mathbf{x}^0) - \nabla^2 F(\mathbf{w}^{k-1}(\mathbf{u}) + \theta(\mathbf{w}^{k-1}(\mathbf{u} + q\mathbf{e}_1) - \mathbf{w}^{k-1}(\mathbf{u}))) \right\| d\theta \\ &\leq L_2(\mathbf{w}_0) \int_0^1 \left\| \theta(\mathbf{w}^{k-1}(\mathbf{u} + q\mathbf{e}_1) - \mathbf{x}^0) + (1 - \theta)(\mathbf{w}^{k-1}(\mathbf{u}) - \mathbf{x}^0) \right\| d\theta \\ &\leq L_2(\mathbf{w}_0) \max\{\|\mathbf{w}^{k-1}(\mathbf{u} + q\mathbf{e}_1) - \mathbf{x}^0\|, \|\mathbf{w}^{k-1}(\mathbf{u}) - \mathbf{x}^0\|\} \\ &\leq L_2(\mathbf{w}_0)B. \end{aligned}$$

The last line follows since $k \leq \mathcal{K}_1$, hence $k - 1 < \mathcal{K}_1$, thus $\mathbf{w}^{k-1}(\mathbf{u} + q\mathbf{e}_1), \mathbf{w}^{k-1}(\mathbf{u}) \in \mathbb{B}(\mathbf{x}^0, B)$.

3. Next as the stochastic gradient oracle $\nabla f(\cdot; \zeta)$ is unbiased, applying Linearity of Expectation on [\(62\)](#), it follows that $\mathbb{E}[\xi_d^k | \mathfrak{F}^{k-1}] = 0$ for all k .

For the bound on the magnitude of ξ_d^k , again recall by the above that for $k \leq \mathcal{K}_1$, we have

$$\mathbf{w}^{k-1}(\mathbf{u} + q\mathbf{e}_1), \mathbf{w}^{k-1}(\mathbf{u}) \in \mathbb{B}(\mathbf{x}^0, B).$$

Thus for all \mathbf{p} on the line segment between $\mathbf{w}^{k-1}(\mathbf{u} + q\mathbf{e}_1), \mathbf{w}^{k-1}(\mathbf{u})$, we have $\mathbf{p} \in \mathbb{B}(\mathbf{x}^0, B)$. Thus by [Lemma E.7](#), $\|\nabla^2 F(\mathbf{p})\| \leq L_1(\mathbf{w}_0)$. By [Lemma E.9](#), for any ζ , $\|\nabla^2 f(\mathbf{p}; \zeta)\| \leq L_1(\mathbf{w}_0)$. Recalling [\(62\)](#) gives

$$\begin{aligned} \|\xi_d^k\| &\leq \left\| \nabla F(\mathbf{w}^{k-1}(\mathbf{u} + q\mathbf{e}_1)) - \nabla F(\mathbf{w}^{k-1}(\mathbf{u})) \right\| + \left\| \nabla f(\mathbf{w}^{k-1}(\mathbf{u} + q\mathbf{e}_1); \zeta_k) - \nabla f(\mathbf{w}^{k-1}(\mathbf{u}); \zeta_k) \right\| \\ &\leq 2L_1(\mathbf{w}_0) \|\mathbf{w}^{k-1}(\mathbf{u} + q\mathbf{e}_1) - \mathbf{w}^{k-1}(\mathbf{u})\| \\ &= 2L_1(\mathbf{w}_0) \|\mathbf{z}^{k-1}\|. \end{aligned}$$

In the last step, we used the definition of \mathbf{z}^k for $k \leq \mathcal{K}_1$.

This proves all the desired parts of [Lemma E.14](#). \square

Controlling iterates under a high probability event. We now consider a rescaled iteration as considered in [Fang et al. \(2019\)](#). Recall the definition of $\delta_m \geq \delta_2$ in the statement of [Lemma E.13](#). For each $k = 0, 1, \dots$, we define:

$$\psi_k := q^{-1}(1 + \eta\delta_m)^{-k} \mathbf{z}_k.$$

Lemma E.15 (Equivalent of the first part of Lemma 14, [Fang et al. \(2019\)](#)). Define $\hat{D}_k := (1 + \eta\delta_m)^{-1} D_k$, and slightly overloading notation, define

$$\zeta_d^k := q^{-1}(1 + \eta\delta_m)^{-k} \xi_d^k.$$

Then for $k \leq \mathcal{K}_1$, we have $\psi^0 = \mathbf{e}_1$ and

$$\psi^k = \frac{\mathbf{I} - \eta H}{1 + \eta\delta_m} \psi^{k-1} + \eta \hat{D}^{k-1} \psi^{k-1} + \eta \zeta_d^k,$$

as well as the properties

$$\begin{aligned}\|\hat{\mathbf{D}}^k\| &\leq L_2(\mathbf{w}_0)B \text{ for all } 0 \leq k < \mathcal{K}_1, \\ \|\zeta_d^k\| &\leq 2L_1(\mathbf{w}_0)\|\psi^{k-1}\| \text{ for all } 1 \leq k \leq \mathcal{K}_1.\end{aligned}$$

Proof. We prove all the desired parts of [Lemma E.15](#).

- The fact that $\psi^0 = \mathbf{e}_1$ follows immediately, because $\mathbf{z}^0 = q\mathbf{e}_1$. For the general recursion for ψ^k , consider any $k \leq \mathcal{K}_1$. First note that by the recursion for the \mathbf{z}^k for $k \leq \mathcal{K}_1$ in [Lemma E.14](#), we have

$$\begin{aligned}\psi^k &= q^{-1}(1 + \eta\delta_m)^{-k}\mathbf{z}^k \\ &= \frac{\mathbf{I} - \eta\mathbf{H}}{1 + \eta\delta_m}q^{-1}(1 + \eta\delta_m)^{-(k-1)}\mathbf{z}^{k-1} \\ &\quad + \eta\frac{\mathbf{D}^{k-1}}{1 + \eta\delta_m}q^{-1}(1 + \eta\delta_m)^{-(k-1)}\mathbf{z}^{k-1} + \eta q^{-1}(1 + \eta\delta_m)^{-k}\xi_d^k \\ &= \frac{\mathbf{I} - \eta\mathbf{H}}{1 + \eta\delta_m}\psi^{k-1} + \eta\hat{\mathbf{D}}^{k-1}\psi^{k-1} + \eta\zeta_d^k.\end{aligned}$$

- Consider any $k \leq \mathcal{K}_1$. For the requisite properties of $\hat{\mathbf{D}}^k$ for $k < \mathcal{K}_1$, the upper bound on the norm of $\hat{\mathbf{D}}^k$ follows immediately from [Lemma E.14](#).

Next from the definition of ζ_d^k and [Lemma E.14](#), for $k \leq \mathcal{K}_1$ we have that

$$\begin{aligned}\|\zeta_d^k\| &\leq q^{-1}(1 + \eta\delta_m)^{-k}\|\xi_d^k\| \\ &\leq 2L_1(\mathbf{w}_0)q^{-1}\frac{(1 + \eta\delta_m)^{-(k-1)}}{1 + \eta\delta_m}\|\mathbf{z}^{k-1}\| \\ &\leq 2L_1(\mathbf{w}_0)\|\psi^{k-1}\|.\end{aligned}$$

This proves [Lemma E.15](#). □

Lemma E.16 (Equivalent of the rest of Lemma 14, [Fang et al. \(2019\)](#)). *With the step size η from (53), there exists an event \mathcal{H}_o (namely, from (66)) with probability at least 0.9, such that for all $k \leq \min(\mathcal{K}_1 - 1, K_0)$ we have*

$$\|\psi^k\|^2 \leq 4, \tag{63}$$

and

$$\mathbf{e}_1^\top \psi^k > \frac{1}{2}. \tag{64}$$

Proof. Define

$$\hat{\psi}^{k-1} = \frac{\mathbf{I} - \eta\mathbf{H}}{1 + \eta\delta_m}\psi^{k-1}.$$

Recall that $\mathbf{H} = \nabla^2 F(\mathbf{x}^0)$ and \mathbf{x}^0 is in the $F(\mathbf{w}_0)$ -sublevel set $\mathcal{L}_{F, F(\mathbf{w}_0)}$. Therefore, from [Assumption 1.1](#), $\|\mathbf{H}\| \leq L_1(\mathbf{w}_0)$. By definition of δ_m , it follows that

$$-\delta_m\mathbf{I} \leq \mathbf{H} \leq L_1(\mathbf{w}_0)\mathbf{I}.$$

Since $\eta L_1(\mathbf{w}_0) \leq 1$, it follows that the matrix $\mathbf{I} - \eta\mathbf{H}$ is symmetric and has all eigenvalues in $[0, 1 + \eta\delta_m]$. This implies

$$\|\hat{\psi}^{k-1}\| \leq \|\psi^{k-1}\|. \tag{65}$$

Note that $\hat{\psi}^{k-1}$ and ψ^{k-1} are measurable on \mathfrak{F}^{k-1} . This combined with [Lemma E.14](#) and [Lemma E.15](#) implies that for all $1 \leq k \leq \mathcal{K}_1$,

$$\mathbb{E}\left[(\hat{\psi}^{k-1})^\top \zeta_d^k \cdot 1_{\|\psi^{k-1}\| \leq 2} | \mathfrak{F}^{k-1}\right] = 1_{\|\psi^{k-1}\| \leq 2} \cdot \mathbb{E}[(\hat{\psi}^{k-1})^\top \zeta_d^k | \mathfrak{F}^{k-1}] = 0,$$

and

$$|(\hat{\psi}^{k-1})^\top \zeta_d^k \cdot 1_{\|\psi^{k-1}\| \leq 2}|^2 \leq 1_{\|\psi^{k-1}\| \leq 2} \cdot 4L_1^2(\mathbf{w}_0) \|\psi^{k-1}\|^4 \leq (8L_1(\mathbf{w}_0))^2.$$

Now define the following real-valued stochastic process:

$$Y_k = (\hat{\psi}^{k-1})^\top \zeta_d^k 1_{\|\psi^{k-1}\| \leq 2} 1_{k-1 < \mathcal{K}_1} = \begin{cases} (\hat{\psi}^{k-1})^\top \zeta_d^k \cdot 1_{\|\psi^{k-1}\| \leq 2} & : k \leq \mathcal{K}_1 \\ 0 & : k > \mathcal{K}_1. \end{cases}$$

Note Y_k is \mathfrak{F}_k -measurable, and that $(\hat{\psi}^{k-1})^\top, 1_{\|\psi^{k-1}\| \leq 2}, 1_{k-1 < \mathcal{K}_1} \equiv 1_{k \leq \mathcal{K}_1}$ are all \mathfrak{F}_{k-1} -measurable. Thus, by Lemma E.14 and the definition of ζ_d^k from Lemma E.15,

$$\mathbb{E}[Y_k | \mathfrak{F}_{k-1}] = 0.$$

Furthermore combining the above justification with the trivial case $k > \mathcal{K}_1$, we obtain

$$|Y_k| \leq 8L_1(\mathbf{w}_0).$$

By the (standard) Azuma's Inequality, with probability $1 - 0.1/(2K_0)$, for any given $l, 1 \leq l \leq K_0$:

$$\left| \sum_{k=1}^l Y_k \right| \leq 8L_1(\mathbf{w}_0) \sqrt{2l \log(40K_0)} \leq 8L_1(\mathbf{w}_0) \sqrt{2K_0 \log(40K_0)} \leq \frac{1}{\eta},$$

where the last inequality follows from the given choice of parameters.

Analogously, by Lemma E.14 and Lemma E.15, we also have for $1 \leq k \leq \mathcal{K}_1$:

$$\mathbb{E}[\mathbf{e}_1^\top \zeta_d^k \cdot 1_{\|\psi^{k-1}\| \leq 2} | \mathfrak{F}^{k-1}] = 0, |\mathbf{e}_1^\top \zeta_d^k \cdot 1_{\|\psi^{k-1}\| \leq 2}| \leq 4L_1(\mathbf{w}_0).$$

Define

$$Y'_k := \mathbf{e}_1^\top \zeta_d^k \cdot 1_{\|\psi^{k-1}\| \leq 2} 1_{k \leq \mathcal{K}_1}.$$

The (standard) Azuma's Inequality now implies that with probability at least $1 - 0.1/(2K_0)$, for any given $l, 1 \leq l \leq K_0$:

$$\left| \sum_{k=1}^l Y'_k \right| \leq 4L_1(\mathbf{w}_0) \sqrt{2l \log(40K_0)} \leq \frac{1}{4\eta}.$$

By the Union Bound, there exists an event \mathcal{H}_o happening with probability at least 0.9 such that the following inequalities hold for each $l = 1, 2, \dots, K_0$:

$$\left| \sum_{k=1}^l Y_k \right| \leq \frac{1}{\eta}, \left| \sum_{k=1}^l Y'_k \right| \leq \frac{1}{4\eta}. \quad (66)$$

In particular under the event \mathcal{H}_o , for any $l \leq \min(\mathcal{K}_1 - 1, K_0)$, using the definitions of Y_k, Y'_k we obtain

$$\left| \sum_{k=1}^l \hat{\psi}_{k-1}^\top \zeta_d^k \cdot 1_{\|\psi^{k-1}\| \leq 2} \right| \leq \frac{1}{\eta}, \left| \sum_{k=1}^l \mathbf{e}_1^\top \zeta_d^k \cdot 1_{\|\psi^{k-1}\| \leq 2} \right| \leq \frac{1}{4\eta}. \quad (67)$$

Now from Lemma E.15, it follows for all $k \leq \mathcal{K}_1$ that

$$\begin{aligned} \|\psi^k\|^2 &= \left\| \frac{\mathbf{I} - \eta \mathbf{H}}{1 + \eta \delta_m} \psi^{k-1} + \eta \hat{\mathbf{D}}^{k-1} \psi^{k-1} + \eta \zeta_d^k \right\|^2 \\ &= \left\| \hat{\psi}^{k-1} \right\|^2 + 2\eta (\hat{\psi}^{k-1})^\top \hat{\mathbf{D}}_{k-1} \psi^{k-1} + \eta^2 \left\| \hat{\mathbf{D}}_{k-1} \psi^{k-1} + \zeta_d^k \right\|^2 + 2\eta (\hat{\psi}^{k-1})^\top \zeta_d^k \\ &= \|\psi^{k-1}\|^2 + Q_{1,k} + Q_{2,k} + Q_{3,k} \end{aligned} \quad (68)$$

where we define

$$Q_{1,k} := 2\eta (\hat{\psi}^{k-1})^\top \hat{\mathbf{D}}^{k-1} \psi^{k-1}, Q_{2,k} := \eta^2 \left\| \hat{\mathbf{D}}_{k-1} \psi^{k-1} + \zeta_d^k \right\|^2, Q_{3,k} := 2\eta (\hat{\psi}^{k-1})^\top \zeta_d^k.$$

For $k \leq \mathcal{K}_1$, we have $k-1 < \mathcal{K}_1$. Thus by Lemma E.15 and (65), we have

$$Q_{1,k} \leq 2\eta L_2(\mathbf{w}_0) B \|\psi^{k-1}\|^2, \quad (69)$$

and

$$Q_{2,k} \leq 2\eta^2 \left\| \hat{\mathbf{D}}^{k-1} \psi^{k-1} \right\|^2 + 2\eta^2 \left\| \zeta_d^k \right\|^2$$

$$\begin{aligned}
&\leq 2\eta^2 \cdot L_2(\mathbf{w}_0)^2 B^2 \|\psi^{k-1}\|^2 + 8\eta^2 L_1(\mathbf{w}_0)^2 \|\psi^{k-1}\|^2 \\
&\leq 16\eta^2 L_1(\mathbf{w}_0)^2 \|\psi^{k-1}\|^2.
\end{aligned} \tag{70}$$

The last inequality above follows as per [Remark 12](#).

Now we complete the proof. Under the event \mathcal{H}_o from (66), we prove (63) by induction on k (recall our condition for k for [Lemma E.16](#) is that $0 \leq k \leq \min(\mathcal{K}_1 - 1, K_0)$).

When $k = 0$, by [Lemma E.15](#), $\psi^0 = \mathbf{e}_1$, so $\|\psi^0\| = \|\mathbf{e}_1\| = 1 \leq 2$ and $\mathbf{e}_1^\top \psi^0 = \|\mathbf{e}_1\|^2 = 1$ (recall \mathbf{e}_1 is a unit eigenvector), proving the base case.

Now for the inductive step, consider some $k \leq \min(\mathcal{K}_1 - 1, K_0)$. Suppose $\|\psi^l\| \leq 2$ holds for all $l, 0 \leq l \leq k - 1$. Then because $k < \mathcal{K}_1$, upon applying the above bounds (68), (69), (70) we have:

$$\begin{aligned}
\|\psi^k\|^2 &\leq \|\psi^0\|^2 + \sum_{s=1}^k Q_{1,s} + \sum_{s=1}^k Q_{2,s} + \sum_{s=1}^k Q_{3,s} \\
&\leq 1 + 2\eta \sum_{s=1}^k L_2(\mathbf{w}_0) B \|\psi^{s-1}\|^2 + 16\eta^2 L_1(\mathbf{w}_0)^2 \sum_{s=1}^k \|\psi^s\|^2 + 2\eta \sum_{s=1}^k (\hat{\psi}^{s-1})^\top \zeta_d^s \\
&\leq 1 + 2L_2(\mathbf{w}_0) B \cdot 4 \cdot \eta k + 16\eta^2 \cdot L_1(\mathbf{w}_0)^2 \cdot 4 \cdot k + 2\eta \sum_{s=1}^k (\hat{\psi}^{s-1})^\top \zeta_d^s \cdot 1_{\|\psi^{s-1}\| \leq 2} \\
&\leq 1 + 16L_2(\mathbf{w}_0) B \cdot \eta K_0 + 2\eta \sum_{s=1}^k \hat{\psi}_{s-1}^\top \zeta_d^s \cdot 1_{\|\psi^{s-1}\| \leq 2} \leq 1 + 1 + 2\eta \cdot \frac{1}{\eta} = 4.
\end{aligned}$$

To upper bound the above, we used our choice of step size $\eta \leq \frac{L_2(\mathbf{w}_0)B}{8L_1(\mathbf{w}_0)^2}$ and $B \leq \frac{1}{L_1(\mathbf{w}_0)}$ as per [Remark 12](#), our above upper bounds on $Q_{1,s}, Q_{2,s}$, and that the event \mathcal{H}_o implies (67).

This completes the induction and proves (63).

With (63), we prove (64). Namely note for $k \leq \min(\mathcal{K}_1 - 1, K_0)$, summing and telescoping the recursion for ψ^k from [Lemma E.15](#), we have:

$$\begin{aligned}
\mathbf{e}_1^\top \psi_k &= \mathbf{e}_1^\top \psi_0 + \sum_{s=0}^{k-1} \eta \mathbf{e}_1^\top \hat{D}_s \psi^s + \sum_{s=0}^{k-1} \eta \mathbf{e}_1^\top \zeta_d^s \\
&\geq 1 - \eta \sum_{s=0}^{k-1} 2L_2(\mathbf{w}_0) B \|\psi^s\| + \eta \sum_{s=0}^{k-1} \mathbf{e}_1^\top \zeta_d^s \cdot 1_{\|\psi^{s-1}\| \leq 2} \\
&\geq 1 - \eta \cdot K_0 \cdot 2L_2(\mathbf{w}_0) B \cdot 2 + \eta \sum_{s=0}^{k-1} \mathbf{e}_1^\top \zeta_d^s \cdot 1_{\|\psi^{s-1}\| \leq 2} \geq 1 - \frac{1}{8} - \frac{2}{8} \geq \frac{1}{2}.
\end{aligned}$$

Here to lower bound the final sum, we used that $\psi_0 = \mathbf{e}_1$ and the upper bound on $\|\hat{D}_s\|$ from [Lemma E.15](#), the fact that we have already established $\|\psi^s\| \leq 2$ for all $s < k$ as we showed (63), and that the event \mathcal{H}_o implies (67).

This proves all parts of [Lemma E.16](#). □

Finish. Now we prove [Lemma E.13](#) via the same high-level strategy as the proof of Lemma 8, [Fang et al. \(2019\)](#). Note on the event $\{\mathcal{K}_1 > K_o\}$, we have

$$z^{K_o} = \mathbf{w}^{K_o}(\mathbf{u} + q\mathbf{e}_1) - \mathbf{w}^{K_o}(\mathbf{u}) = (\mathbf{w}^{K_o}(\mathbf{u} + q\mathbf{e}_1) - \mathbf{x}^0) - (\mathbf{w}^{K_o}(\mathbf{u}) - \mathbf{x}^0).$$

Thus by definition of \mathcal{K}_1 , the event $\{\mathcal{K}_1 > K_o\}$ implies that

$$\|z^{K_o}\| \leq \|\mathbf{w}^{K_o}(\mathbf{u} + q\mathbf{e}_1) - \mathbf{x}^0\| + \|\mathbf{w}^{K_o}(\mathbf{u}) - \mathbf{x}^0\| \leq 2B.$$

That is,

$$\{\mathcal{K}_1 > K_o\} \subseteq \{\|z^{K_o}\| \leq 2B\}.$$

However, consider the event \mathcal{H}_o , (66) from Lemma E.16. On the event $\{\mathcal{K}_1 > K_o\} \cap \mathcal{H}_o$, we have $K_o \leq \min(\mathcal{K}_1 - 1, K_0)$, and so by Lemma E.16, we have

$$\mathbf{e}_1^\top \boldsymbol{\psi}^{K_o} > \frac{1}{2}.$$

Thus by definition of $\boldsymbol{\psi}^k$ and recalling $\delta_m \geq \delta_2 > 0$, on the event $\{\mathcal{K}_1 > K_o\} \cap \mathcal{H}_o$ we have

$$\|\mathbf{z}^{K_o}\| = q(1 + \eta\delta_m)^{K_o} \|\boldsymbol{\psi}^{K_o}\| \geq q_0(1 + \eta\delta_2)^{K_o} |\mathbf{e}_1^\top \boldsymbol{\psi}^{K_o}| > q_0 \cdot \frac{6B}{q_0} \cdot \frac{1}{2} = 3B,$$

where the last inequality uses (54). This means that

$$\{\mathcal{K}_1 > K_o\} \cap \mathcal{H}_o \subseteq \{\|\mathbf{z}^{K_o}\| \geq 3B\}.$$

Putting our work together, we see that

$$\{\mathcal{K}_1 > K_o\} \cap \mathcal{H}_o \subseteq \{\|\mathbf{z}^{K_o}\| \geq 3B\} \cap \{\|\mathbf{z}^{K_o}\| \leq 2B\} = \emptyset.$$

Therefore

$$\{\mathcal{K}_1 > K_o\} \subseteq \mathcal{H}_o^c \implies \mathbb{P}(\mathcal{K}_1 > K_o) \leq \mathbb{P}(\mathcal{H}_o^c) \leq 0.1.$$

Recalling the definition of \mathcal{K}_1 , we conclude Lemma E.13. \square

Remark 21. Note we only have $\mathbf{e}_1^\top \boldsymbol{\psi}^k > \frac{1}{2}$ for $k < \mathcal{K}_1$ due to the lack of global Lipschitz bounds on the gradient and Hessian of F , unlike in the proof of Lemma 8, Fang et al. (2019).

E.5 Faster Descent

Setup: As in Subsection E.4, let \mathcal{K}_0 denote the escape time of $\mathbb{B}(\mathbf{x}^0, B)$ for while loop of Algorithm 2 when the while loop begins at \mathbf{x}^0 . In this section, we aim to prove Lemma E.2.

As in Subsection E.4, the difference between Lemma E.2 and Proposition 9 of Fang et al. (2019) is that *this result only holds at points in the $F(\mathbf{w}_0)$ -sublevel set $\mathcal{L}_{F, F(\mathbf{w}_0)}$* . For the rest of this section, we work under the assumptions of Lemma E.2; thus for the rest of this section, \mathbf{x}^0 is in the $F(\mathbf{w}_0)$ -sublevel set $\mathcal{L}_{F, F(\mathbf{w}_0)}$.

The idea here is similar to that of Subsection E.4. At a high level, we have the requisite control over the gradient and Hessian since the iterates we consider are in a neighborhood of a point $\mathbf{x}^0 \in \mathcal{L}_{F, F(\mathbf{w}_0)}$. As in the previous part and as in Fang et al. (2019), we let

$$\mathbf{H} := \nabla^2 F(\mathbf{x}^0),$$

and let

$$\boldsymbol{\xi}^{k+1} := \nabla \tilde{f}(\mathbf{x}^k; \boldsymbol{\zeta}_{k+1}) - \nabla F(\mathbf{x}^k), \quad k \geq 0. \quad (71)$$

Note as $\boldsymbol{\Lambda}^{k+1}$ has mean 0 and as the stochastic gradient oracle is unbiased, we have that for all $k \geq 0$,

$$\mathbb{E}[\boldsymbol{\xi}^{k+1} | \mathcal{F}^k] = \mathbf{0}.$$

Let \mathcal{S} be the subspace spanned by all eigenvectors of $\nabla^2 F(\mathbf{x}^0)$ whose eigenvalue is greater than 0, and \mathcal{S}^\perp denotes the complement space. Also, let $\mathcal{P}_{\mathcal{S}} \in \mathbb{R}^{d \times d}$ and $\mathcal{P}_{\mathcal{S}^\perp} \in \mathbb{R}^{d \times d}$ denote the projection matrices onto the spaces \mathcal{S} and \mathcal{S}^\perp , respectively. Let $\mathbf{u}^k = \mathcal{P}_{\mathcal{S}}(\mathbf{x}^k - \mathbf{x}^0)$, and $\mathbf{v}^k = \mathcal{P}_{\mathcal{S}^\perp}(\mathbf{x}^k - \mathbf{x}^0)$. We can decompose the update equation of SGD as:

$$\mathbf{u}^{k+1} = \mathbf{u}^k - \eta \mathcal{P}_{\mathcal{S}} \nabla F(\mathbf{x}^k) - \eta \mathcal{P}_{\mathcal{S}} \boldsymbol{\xi}^{k+1},$$

$$\mathbf{v}^{k+1} = \mathbf{v}^k - \eta \mathcal{P}_{\mathcal{S}^\perp} \nabla F(\mathbf{x}^k) - \eta \mathcal{P}_{\mathcal{S}^\perp} \boldsymbol{\xi}^{k+1},$$

for $k \geq 0$. Clearly $\mathbf{u}^0 = \mathbf{0}$, $\mathbf{v}^0 = \mathbf{0}$.

Now decompose $\mathbf{H} = \mathbf{U} \boldsymbol{\Lambda} \mathbf{U}^T$ by the Spectral Theorem where $\mathbf{U} \in \mathbb{R}^{d \times d}$ is unitary and $\boldsymbol{\Lambda} \in \mathbb{R}^{d \times d}$ is diagonal. Let $\boldsymbol{\Lambda}_{>0}$ denote the diagonal matrix with diagonal entries equal to the positive (diagonal) entries of $\boldsymbol{\Lambda}$. Let $\boldsymbol{\Lambda}_{\leq 0}$ denote the diagonal matrix with diagonal entries equal to the zero or negative (diagonal) entries of $\boldsymbol{\Lambda}$. Now define

$$\mathbf{H}_{\mathcal{S}} := \mathbf{U} \boldsymbol{\Lambda}_{>0} \mathbf{U}^T, \mathbf{H}_{\mathcal{S}^\perp} := \mathbf{U} \boldsymbol{\Lambda}_{\leq 0} \mathbf{U}^T.$$

Thus $\mathbf{H}_{\mathcal{S}}$ has range in \mathcal{S} , and $\mathbf{H}_{\mathcal{S}^\perp}$ has range in \mathcal{S}^\perp . Note $\mathbf{H}_{\mathcal{S}}, \mathbf{H}_{\mathcal{S}^\perp}$ are both symmetric.

From here, define the following quadratic approximations:

$$G_{\mathcal{S}}(\mathbf{u}) := [\mathcal{P}_{\mathcal{S}} \nabla F(\mathbf{x}^0)]^\top \mathbf{u} + \frac{1}{2} \mathbf{u}^\top \mathbf{H}_{\mathcal{S}} \mathbf{u}, G_{\mathcal{S}^\perp}(\mathbf{v}) := [\mathcal{P}_{\mathcal{S}^\perp} \nabla F(\mathbf{x}^0)]^\top \mathbf{v} + \frac{1}{2} \mathbf{v}^\top \mathbf{H}_{\mathcal{S}^\perp} \mathbf{v}.$$

Now define the quadratic approximation

$$G(\mathbf{x}) = G_{\mathcal{S}}(\mathbf{u}) + G_{\mathcal{S}^\perp}(\mathbf{v}) \text{ where } \mathbf{u} = \mathcal{P}_{\mathcal{S}}(\mathbf{x} - \mathbf{x}^0), \mathbf{v} = \mathcal{P}_{\mathcal{S}^\perp}(\mathbf{x} - \mathbf{x}^0).$$

It is easy to see that

$$G(\mathbf{x}) = [\nabla F(\mathbf{x}^0)]^\top (\mathbf{x} - \mathbf{x}^0) + \frac{1}{2} (\mathbf{x} - \mathbf{x}^0)^\top \mathbf{H} (\mathbf{x} - \mathbf{x}^0).$$

For convenience, let

$$\nabla_{\mathbf{u}} F(\mathbf{x}^k) = \mathcal{P}_{\mathcal{S}} \nabla F(\mathbf{x}^k), \nabla_{\mathbf{v}} F(\mathbf{x}^k) = \mathcal{P}_{\mathcal{S}^\perp} \nabla F(\mathbf{x}^k).$$

Similarly, let

$$\boldsymbol{\xi}_{\mathbf{u}}^k = \mathcal{P}_{\mathcal{S}} \boldsymbol{\xi}^k, \boldsymbol{\xi}_{\mathbf{v}}^k = \mathcal{P}_{\mathcal{S}^\perp} \boldsymbol{\xi}^k.$$

Also denote the stopping time

$$\mathcal{K} = \mathcal{K}_0 \wedge K_0.$$

Due to its ‘local’ nature around the \mathbf{x}^0 in the $F(\mathbf{w}_0)$ -sublevel set, we still have the following result from Fang et al. (2019):

Lemma E.17 (Equivalent of Lemma 15, Fang et al. (2019)). *Consider any $\mathbf{u} \in \mathcal{L}_{F, F(\mathbf{w}_0)}$, and consider any $\mathbf{x} \in \mathbb{B}(\mathbf{u}, B)$. Then we have*

$$\|\nabla F(\mathbf{x}) - \nabla G(\mathbf{x})\| \leq \frac{L_2(\mathbf{w}_0)B^2}{2}.$$

Furthermore, for any symmetric matrix \mathbf{A} , with $0 < a \leq \frac{1}{\|\mathbf{A}\|_2}$, for any $i = 0, 1, \dots$, and $j = 0, 1, \dots$, we have

$$\|(I - a\mathbf{A})^i \mathbf{A} (I - a\mathbf{A})^j\|_2 \leq \frac{1}{a(i+j+1)}.$$

Proof. Notice that for all $0 \leq \theta \leq 1$, $\theta \mathbf{x} + (1 - \theta)\mathbf{u} \in \mathbb{B}(\mathbf{u}, B)$. Thus as $\mathbf{u} \in \mathcal{L}_{F, F(\mathbf{w}_0)}$, by Lemma E.8, we have

$$\|\nabla^2 F(\theta \mathbf{x} + (1 - \theta)\mathbf{u}) - \nabla^2 F(\mathbf{u})\| \leq L_2(\mathbf{w}_0) \cdot \theta \|\mathbf{x} - \mathbf{u}\| \text{ for all } 0 \leq \theta \leq 1.$$

Thus we have

$$\begin{aligned} \|\nabla F(\mathbf{x}) - \nabla G(\mathbf{x})\| &= \|\nabla F(\mathbf{x}) - \nabla F(\mathbf{x}^0) - \nabla^2 F(\mathbf{u})(\mathbf{x} - \mathbf{u})\| \\ &= \left\| \left\{ \int_0^1 (\nabla^2 F(\mathbf{x}^0 + \theta(\mathbf{x} - \mathbf{u})) - \nabla^2 F(\mathbf{u})) d\theta \right\} (\mathbf{x} - \mathbf{u}) \right\| \\ &\leq \left\| \int_0^1 \{L_2(\mathbf{w}_0) \cdot \theta \|\mathbf{x} - \mathbf{u}\|\} d\theta \right\| \cdot \|\mathbf{x} - \mathbf{u}\| \\ &\leq \frac{L_2(\mathbf{w}_0)B^2}{2}. \end{aligned}$$

The second part of the Lemma follows from the exact same proof of Lemma D.5 in Section D. It is also proved in the proofs of Lemma 15, Fang et al. (2019), and in the proof of Lemma 16 of Jin et al. (2017). For more detail, let the eigenvalues of \mathbf{A} be $\{\lambda_k\}$. Thus for any $i, j \geq 0$, the eigenvalues of $(I - a\mathbf{A})^i \mathbf{A} (I - a\mathbf{A})^j$ are $\{\lambda_k(1 - a\lambda_k)^{i+j}\}$. We now detail a calculation from Jin et al. (2017). Letting $g_t(\lambda) := \lambda(1 - a\lambda)^t$ and setting its derivative to zero yields

$$\nabla g_t(\lambda) = (1 - a\lambda)^t - ta\lambda(1 - a\lambda)^{t-1} = 0.$$

It is easy to check that $\lambda_t^* = \frac{1}{(1+t)a}$ is the unique maximizer, and $g_t(\lambda)$ is monotonically increasing in $(-\infty, \lambda_t^*]$.

This gives:

$$\|(I - a\mathbf{A})^i \mathbf{A} (I - a\mathbf{A})^j\| = \max_k \lambda_i(1 - a\lambda_k)^{i+j} \leq \hat{\lambda}(1 - a\hat{\lambda})^{i+j} \leq \frac{1}{(1+i+j)a},$$

where $\hat{\lambda} = \min\{\ell, \lambda_{i+j}^*\}$. □

Lemma E.18. For any $k \leq \mathcal{K}_0$, we have

$$\|\xi^k\| \leq \sigma_1(\mathbf{w}_0).$$

Proof. Note for $k \leq \mathcal{K}_0$, we have $k-1 < \mathcal{K}_0$ and so $\mathbf{x}^{k-1} \in \mathbb{B}(\mathbf{x}^0, B)$. Recall furthermore that $\mathbf{x}^0 \in \mathcal{L}_{F,F}(\mathbf{w}_0)$. Thus, by Lemma E.5 and Lemma E.3,

$$\|\xi^k\| = \|\nabla \tilde{f}(\mathbf{x}^{k-1}; \zeta_k) - \nabla F(\mathbf{x}^{k-1})\| \leq \sigma_1(\mathbf{w}_0),$$

as desired. \square

Analyzing the Quadratic Approximation: We now analyze the quadratic approximation $G(\mathbf{x})$ as done in Fang et al. (2019). First we analyze the part in \mathcal{S} :

Lemma E.19 (Equivalent of Lemma 16, Fang et al. (2019)). Set hyperparameters from (8). With probability at least $1 - p/4$, we have

$$\begin{aligned} & G_{\mathcal{S}}(\mathbf{u}^{\mathcal{K}}) - G_{\mathcal{S}}(\mathbf{u}^0) \\ & \leq -\frac{25\eta}{32} \sum_{k=0}^{\mathcal{K}-1} \|\nabla G_{\mathcal{S}}(\mathbf{y}^k)\|^2 + 4\eta\sigma_1(\mathbf{w}_0)^2(\log(K_0) + 3) \log\left(\frac{48K_0}{p}\right) + \eta L_2(\mathbf{w}_0)^2 B^4 K_0 \\ & = -\frac{25\eta}{32} \sum_{k=0}^{\mathcal{K}-1} \|\nabla G_{\mathcal{S}}(\mathbf{y}^k)\|^2 + \tilde{O}(\varepsilon^{1.5}). \end{aligned}$$

Proof. We follow a similar strategy as before of combining the proof of Fang et al. (2019) with our self-bounding framework. To analyze $G_{\mathcal{S}}(\cdot)$ we first consider an auxiliary Gradient Descent trajectory, which performs the update:

$$\mathbf{y}^{k+1} = \mathbf{y}^k - \eta \nabla G_{\mathcal{S}}(\mathbf{y}^k), \quad k \geq 0,$$

and $\mathbf{y}^0 = \mathbf{u}^0$. \mathbf{y}^k performs Gradient Descent on $G_{\mathcal{S}}(\cdot)$, which is deterministic given \mathbf{x}^0 .

Noting $G_{\mathcal{S}}$ has Hessian $\mathbf{H}_{\mathcal{S}}$, and that \mathbf{H} is the Hessian of F at the point $\mathbf{x}^0 \in \mathcal{L}_{F,F}(\mathbf{w}_0)$, we obtain from Assumption 1.1 that

$$\|\mathbf{H}_{\mathcal{S}}\| \leq \|\mathbf{H}\| \leq L_1(\mathbf{w}_0).$$

Since the following only concern $G_{\mathcal{S}}$, then identically to the proof of Lemma 16, Fang et al. (2019), we obtain the following:

- By $L_1(\mathbf{w}_0)$ -smoothness of $G_{\mathcal{S}}$ (recall $G_{\mathcal{S}}$ has Hessian $\mathbf{H}_{\mathcal{S}}$), we obtain the so-called ‘Descent Lemma’:

$$\begin{aligned} G_{\mathcal{S}}(\mathbf{y}^{k+1}) & \leq G_{\mathcal{S}}(\mathbf{y}^k) + \langle \nabla G_{\mathcal{S}}(\mathbf{y}^k), \mathbf{y}^{k+1} - \mathbf{y}^k \rangle + \frac{L_1(\mathbf{w}_0)}{2} \|\mathbf{y}^{k+1} - \mathbf{y}^k\|^2 \\ & = G_{\mathcal{S}}(\mathbf{y}^k) - \eta \left(1 - \frac{L_1(\mathbf{w}_0)\eta}{2}\right) \|\nabla G_{\mathcal{S}}(\mathbf{y}^k)\|^2. \end{aligned}$$

- Telescoping the above for $0 \leq k \leq \mathcal{K}-1$, and by our choice of η which satisfies $\eta L_1(\mathbf{w}_0) \leq \frac{1}{16}$ as per Remark 12, we obtain

$$G_{\mathcal{S}}(\mathbf{y}^{\mathcal{K}}) \leq G_{\mathcal{S}}(\mathbf{y}^0) - \frac{31\eta}{32} \sum_{k=0}^{\mathcal{K}-1} \|\nabla G_{\mathcal{S}}(\mathbf{y}^k)\|^2. \quad (72)$$

To obtain Lemma E.19, we upper bound the difference between $\mathbf{u}^{\mathcal{K}}$ and $\mathbf{y}^{\mathcal{K}}$. For all $k \geq 0$, define

$$\mathbf{z}^k := \mathbf{u}^k - \mathbf{y}^k.$$

We aim to upper bound $\mathbf{z}^{\mathcal{K}}$ (in an appropriate sense) using the concentration argument of Fang et al. (2019):

Lemma E.20 (Equivalent of Lemma 17, Fang et al. (2019)). *With probability at least $1 - p/6$, we have*

$$\|\mathbf{z}^k\| \leq \frac{3B}{32} \approx \tilde{\Theta}(\varepsilon^{0.5}), \quad (73)$$

and

$$\mathbf{z}^{k^\top} \mathbf{H}_S \mathbf{z}^k \leq 8\sigma_1(\mathbf{w}_0)^2 \eta (\log(K_0) + 1) \log\left(\frac{48K_0}{p}\right) + \eta L_2(\mathbf{w}_0)^2 B^4 K_0 \approx \tilde{\Theta}(\varepsilon^{0.5}). \quad (74)$$

Here $\tilde{\Theta}(\cdot)$ hides $F(\mathbf{w}_0)$ -dependence.

Proof of Lemma E.20. Clearly $\mathbf{z}^0 = \mathbf{0}$. From the definitions of $\mathbf{u}^k, \mathbf{y}^k$, we have

$$\begin{aligned} \mathbf{z}^{k+1} &= \mathbf{z}^k - \eta(\nabla G_S(\mathbf{u}^k) - \nabla G_S(\mathbf{y}^k)) - \eta(\nabla_{\mathbf{u}} F(\mathbf{x}^k) - \nabla G_S(\mathbf{u}^k)) - \eta \xi_{\mathbf{u}}^{k+1} \\ &= (\mathbf{I} - \eta \mathbf{H}_S) \mathbf{z}^k - \eta(\nabla_{\mathbf{u}} F(\mathbf{x}^k) - \nabla G_S(\mathbf{u}^k)) - \eta \xi_{\mathbf{u}}^{k+1}, \quad k \geq 0. \end{aligned} \quad (75)$$

Unraveling the above recursion gives:

$$\mathbf{z}^k = - \sum_{j=1}^k \eta (\mathbf{I} - \eta \mathbf{H}_S)^{k-j} \xi_{\mathbf{u}}^j - \eta \sum_{j=0}^{k-1} (\mathbf{I} - \eta \mathbf{H}_S)^{k-1-j} (\nabla_{\mathbf{u}} F(\mathbf{x}^j) - \nabla G_S(\mathbf{u}^j)), \quad k \geq 0. \quad (76)$$

Setting $k = \mathcal{K}$, Triangle Inequality gives

$$\|\mathbf{z}^{\mathcal{K}}\| \leq \left\| \sum_{j=1}^{\mathcal{K}} \eta (\mathbf{I} - \eta \mathbf{H}_S)^{\mathcal{K}-j} \xi_{\mathbf{u}}^j \right\| + \left\| \eta \sum_{j=0}^{\mathcal{K}-1} (\mathbf{I} - \eta \mathbf{H}_S)^{\mathcal{K}-1-j} (\nabla_{\mathbf{u}} F(\mathbf{x}^j) - \nabla G_S(\mathbf{u}^j)) \right\|.$$

We separately bound these two terms:

- For the first term, for any fixed l from 1 to K_0 , and any j from 1 to $\min(l, \mathcal{K}_0)$, we have

$$\mathbb{E}[\eta (\mathbf{I} - \eta \mathbf{H}_S)^{l-j} \xi_{\mathbf{u}}^j | \mathfrak{F}^{j-1}] = 0, \quad \|\eta (\mathbf{I} - \eta \mathbf{H}_S)^{l-j} \xi_{\mathbf{u}}^j\| \leq \eta \sigma_1(\mathbf{w}_0).$$

The first equality uses $\|\xi_{\mathbf{u}}^j\| = \|\mathcal{P}_S \xi^j\|$ and that the stochastic gradient oracle is unbiased. The inequality uses that \mathcal{P} is a projection matrix, $\|\xi_{\mathbf{u}}^j\| = \|\mathcal{P}_S \xi^j\| \leq \sigma_1(\mathbf{w}_0)$ which follows as $j \leq \mathcal{K}_0$ and Lemma E.18, and $\|(\mathbf{I} - \eta \mathbf{H}_S)^{l-j}\| \leq 1$ which follows as $l \geq j$ and $\mathbf{H}_S \geq 0$. (Note the importance that $j \leq \mathcal{K}_0$, which gives us enough control over the noise term $\xi_{\mathbf{u}}^j$.)

Now to deal with the fact that the above control only applies for certain j , we define a stochastic process as follows, analogously to our proof of Lemma E.13. For all fixed $1 \leq l \leq K_0$, define a stochastic process $Y_{l,j}$ over all $1 \leq j \leq l$ by:

$$Y_{l,j} = \eta (\mathbf{I} - \eta \mathbf{H}_S)^{l-j} \xi_{\mathbf{u}}^j 1_{j-1 < \mathcal{K}} = \begin{cases} \eta (\mathbf{I} - \eta \mathbf{H}_S)^{l-j} \xi_{\mathbf{u}}^j & : j \leq \mathcal{K} \\ 0 & : j > \mathcal{K}. \end{cases}$$

Recalling $\mathcal{K} = \mathcal{K}_0 \wedge K_0$, it's easy to check that for any fixed l , $Y_{l,j}$ is \mathfrak{F}^j -measurable. Furthermore, $\eta (\mathbf{I} - \eta \mathbf{H}_S)^{l-j} 1_{j-1 < \mathcal{K}}$ are both \mathfrak{F}^{j-1} -measurable. Thus combining with the earlier observations, we obtain that

$$\mathbb{E}[Y_{l,j} | \mathfrak{F}^{j-1}] = 0, \quad \|Y_{l,j}\| \leq \eta \sigma_1(\mathbf{w}_0).$$

Thus, by the Vector-Martingale Concentration Inequality Theorem C.1, we have with probability $1 - p/(12K_0)$,

$$\left\| \sum_{j=1}^l Y_{l,j} \right\| \leq 2\eta \sigma_1(\mathbf{w}_0) \sqrt{l \log\left(\frac{48K_0}{p}\right)} \leq 2\eta \sigma_1(\mathbf{w}_0) \sqrt{K_0 \log\left(\frac{48K_0}{p}\right)} \leq \frac{B}{16}. \quad (77)$$

The last inequality uses our choice of parameters.

By a Union Bound, with probability at least $1 - p/12$, (77) holds for all l from 1 to K_0 . In particular, with probability at least $1 - p/12$ we have for \mathcal{K} (recall $\mathcal{K} \leq K_0$) that

$$\left\| \sum_{j=1}^{\mathcal{K}} \eta (\mathbf{I} - \eta \mathbf{H}_S)^{\mathcal{K}-j} \xi_{\mathbf{u}}^j \right\| = \left\| \sum_{j=1}^{\mathcal{K}} Y_{\mathcal{K},j} \right\| \leq \frac{B}{16},$$

where we define $Y_{\mathcal{K},j}$ the obvious way. This holds because with probability at least $1 - p/12$, we have the bound (77) on $\left\| \sum_{j=1}^l Y_{l,j} \right\|$ irrespective of which value of $1 \leq l \leq K_0$ that \mathcal{K} takes on. The first equality holds by our definition of $Y_{l,j}$ for $j \leq l = \mathcal{K}$.

- For the second term, we have

$$\begin{aligned}
\left\| \eta \sum_{j=0}^{\mathcal{K}-1} (\mathbf{I} - \eta \mathbf{H}_S)^{\mathcal{K}-1-j} (\nabla_{\mathbf{u}} F(\mathbf{x}^j) - \nabla G_S(\mathbf{u}^j)) \right\| &\leq \eta \sum_{j=0}^{\mathcal{K}-1} \|\nabla_{\mathbf{u}} F(\mathbf{x}^j) - \nabla G_S(\mathbf{u}^j)\| \\
&\leq \eta \sum_{j=0}^{\mathcal{K}-1} \|\nabla F(\mathbf{x}^j) - \nabla G(\mathbf{x}^j)\| \\
&\leq \frac{\eta L_2(\mathbf{w}_0) B^2 K_0}{2} \leq \frac{B}{32}.
\end{aligned}$$

The first inequality uses the Triangle Inequality and that $\|(\mathbf{I} - \eta \mathbf{H}_S)^{\mathcal{K}-1-j}\|_2 \leq 1$ for j from 0 to $\mathcal{K} - 1$; this follows because $\|\mathbf{H}_S\| \leq L_1(\mathbf{w}_0)$ and as $\eta \leq \frac{1}{L_1(\mathbf{w}_0)}$. The second inequality uses $\|\mathcal{P}_S(\nabla F(\mathbf{x}) - \nabla G(\mathbf{x}))\| \leq \|\nabla F(\mathbf{x}) - \nabla G(\mathbf{x})\|$ because \mathcal{P}_S is a projection matrix. The third inequality follows from Lemma E.17, and the fact that for all $j \leq \mathcal{K} - 1$, $\mathbf{x}^j \in \mathbb{B}(\mathbf{x}^0, B)$. The last inequality uses the choice of parameters.

Combining the above gives (73), the first part of Lemma E.20.

Now prove the second part of Lemma E.20, namely (74). Using the fact that $(\mathbf{a} + \mathbf{b})^\top \mathbf{A}(\mathbf{a} + \mathbf{b}) \leq 2\mathbf{a}^\top \mathbf{A} \mathbf{a} + 2\mathbf{b}^\top \mathbf{A} \mathbf{b}$ for any symmetric positive definite matrix \mathbf{A} and the recursion (76) for \mathbf{z}^k , we have

$$\begin{aligned}
&(\mathbf{z}^{\mathcal{K}})^\top \mathbf{H}_S \mathbf{z}^{\mathcal{K}} \\
&\leq 2\eta^2 \left(\sum_{j=1}^{\mathcal{K}} (\mathbf{I} - \eta \mathbf{H}_S)^{\mathcal{K}-j-1} \right)^\top \mathbf{H}_S \left(\sum_{j=1}^{\mathcal{K}} (\mathbf{I} - \eta \mathbf{H}_S)^{\mathcal{K}-j} \boldsymbol{\xi}_{\mathbf{u}}^j \right) \\
&+ 2\eta^2 \left(\sum_{j=0}^{\mathcal{K}-1} (\mathbf{I} - \eta \mathbf{H}_S)^{\mathcal{K}-1-j} (\nabla_{\mathbf{u}} F(\mathbf{x}^j) - \nabla G_S(\mathbf{u}^j)) \right)^\top \mathbf{H}_S \left(\sum_{j=0}^{\mathcal{K}-1} (\mathbf{I} - \eta \mathbf{H}_S)^{\mathcal{K}-1-j} (\nabla_{\mathbf{u}} F(\mathbf{x}^j) - \nabla G_S(\mathbf{u}^j)) \right) \\
&= 2 \left\| \eta \sum_{j=1}^{\mathcal{K}} \mathbf{H}_S^{1/2} (\mathbf{I} - \eta \mathbf{H}_S)^{\mathcal{K}-j} \boldsymbol{\xi}_{\mathbf{u}}^j \right\|^2 \\
&+ 2\eta^2 \sum_{j=0}^{\mathcal{K}-1} \sum_{l=0}^{\mathcal{K}-1} (\nabla_{\mathbf{u}} F(\mathbf{x}^j) - \nabla G_S(\mathbf{u}^j))^\top (\mathbf{I} - \eta \mathbf{H}_S)^{\mathcal{K}-1-j} \mathbf{H}_S (\mathbf{I} - \eta \mathbf{H}_S)^{\mathcal{K}-1-l} (\nabla_{\mathbf{u}} F(\mathbf{x}^l) - \nabla G_S(\mathbf{u}^l)) \\
&\leq 2 \left\| \eta \sum_{j=1}^{\mathcal{K}} \mathbf{H}_S^{1/2} (\mathbf{I} - \eta \mathbf{H}_S)^{\mathcal{K}-j} \boldsymbol{\xi}_{\mathbf{u}}^j \right\|^2 + 2\eta^2 \frac{L_2(\mathbf{w}_0)^2 B^4}{4} \sum_{j=0}^{\mathcal{K}-1} \sum_{l=0}^{\mathcal{K}-1} \|(\mathbf{I} - \eta \mathbf{H}_S)^{\mathcal{K}-1-j} \mathbf{H}_S (\mathbf{I} - \eta \mathbf{H}_S)^{\mathcal{K}-1-l}\|.
\end{aligned}$$

The last inequality follows by properties of projection matrices and by Lemma E.17, recalling that for $j \leq \mathcal{K} - 1$, $\mathbf{x}^j \in \mathbb{B}(\mathbf{x}^0, B)$.

Now we bound each of these two terms separately:

- For the first term, for any fixed $l, 1 \leq l \leq K_0$, again we define a stochastic process for any $j, 1 \leq j \leq l$ by:

$$Y_{l,j} = \eta \left(\mathbf{H}_S^{1/2} (\mathbf{I} - \eta \mathbf{H}_S)^{l-j} \boldsymbol{\xi}_{\mathbf{u}}^j \right) 1_{j-1 < \mathcal{K}} = \begin{cases} \eta \left(\mathbf{H}_S^{1/2} (\mathbf{I} - \eta \mathbf{H}_S)^{l-j} \boldsymbol{\xi}_{\mathbf{u}}^j \right) & : j \leq \mathcal{K} \\ 0 & : j > \mathcal{K}. \end{cases}$$

Analogously to earlier, recalling $\mathcal{K} \leq K_0$, for fixed l , it is evident that $Y_{l,j}$ is \mathfrak{F}^j -measurable, $\eta \mathbf{H}_S^{1/2} (\mathbf{I} - \eta \mathbf{H}_S)^{l-j} 1_{j-1 < \mathcal{K}}$ is \mathfrak{F}^{j-1} -measurable, and thus

$$\mathbb{E}[Y_{l,j} | \mathfrak{F}^{j-1}] = 0.$$

We furthermore have

$$\|Y_{l,j}\|^2 \leq \frac{\eta \sigma_1(\mathbf{w}_0)^2}{1 + 2(l-j)},$$

which follows by noting for any $1 \leq l \leq K_0$ and $j \leq \mathcal{K} \leq K_0$,

$$\left\| \eta (\mathbf{H}_S^{1/2} (\mathbf{I} - \eta \mathbf{H}_S)^{l-j} \boldsymbol{\xi}_{\mathbf{u}}^j) \right\|^2 \leq \eta^2 \|\boldsymbol{\xi}_{\mathbf{u}}^j\|^2 \left\| \mathbf{H}_S^{1/2} (\mathbf{I} - \eta \mathbf{H}_S)^{l-j} \mathbf{H}_S (\mathbf{I} - \eta \mathbf{H}_S)^{l-j} \right\| \|\boldsymbol{\xi}_{\mathbf{u}}^j\|^2$$

$$\leq \frac{\eta\sigma_1(\mathbf{w}_0)^2}{1+2(l-j)}.$$

This uses the second part of [Lemma E.17](#), that $\|\mathbf{H}_S\| \leq L_1(\mathbf{w}_0)$, that $j \leq K_0$ which gives $\|\xi_u^j\| \leq \sigma_1(\mathbf{w}_0)$ by [Lemma E.18](#), and our choice of η (which cancels one of the $\sigma_1(\mathbf{w}_0)^2$ factors).

For a given l , by the Vector-Martingale Concentration Inequality [Theorem C.1](#), we have with probability $1 - p/(12K_0)$ that

$$\begin{aligned} \left\| \sum_{j=1}^l Y_{l,j} \right\|^2 &\leq 4\eta\sigma_1(\mathbf{w}_0)^2 \log\left(\frac{48K_0}{p}\right) \sum_{j=1}^l \frac{1}{1+2(l-j)} \\ &\leq 4\eta\sigma_1(\mathbf{w}_0)^2 (\log(K_0) + 1) \log\left(\frac{48K_0}{p}\right). \end{aligned} \quad (78)$$

The last step above uses $l \leq K_0$, $\sum_{j=1}^l \frac{1}{1+j} \leq \log(K_0) + 1$.

By the Union Bound, with probability at least $1 - \frac{p}{12}$, (78) holds for all l from 1 to K_0 . Because $1 \leq \mathcal{K} \leq K_0$, using the definition of $Y_{l,j}$ for $l \leq \mathcal{K}$, we obtain with probability at least $1 - \frac{p}{12}$ that

$$\eta \left\| \sum_{j=1}^{\mathcal{K}} \mathbf{H}_S^{1/2} (\mathbf{I} - \eta \mathbf{H}_S)^{K-j} \xi_u^j \right\|^2 = \left\| \sum_{j=1}^{\mathcal{K}} Y_{\mathcal{K},j} \right\|^2 \leq 4\eta\sigma_1(\mathbf{w}_0)^2 (\log(K_0) + 1) \log\left(\frac{48K_0}{p}\right).$$

- For the second term, using the second part of [Lemma E.17](#) and that $\mathcal{K} \leq K_0$, and then rearranging order of the sum and performing explicit calculation yields

$$\begin{aligned} &\eta^2 \frac{L_2(\mathbf{w}_0)^2 B^4}{4} \sum_{j=0}^{\mathcal{K}-1} \sum_{l=0}^{\mathcal{K}-1} \|(\mathbf{I} - \eta \mathbf{H}_S)^{\mathcal{K}-1-j} \mathbf{H}_S (\mathbf{I} - \eta \mathbf{H}_S)^{\mathcal{K}-1-l}\| \\ &\leq \eta \frac{L_2(\mathbf{w}_0)^2 B^4}{4} \sum_{j=0}^{K_0-1} \sum_{l=0}^{K_0-1} \frac{1}{1+j+l} \\ &\leq \eta \frac{L_2(\mathbf{w}_0)^2 B^4}{4} \sum_{l=0}^{2(K_0-1)} \frac{\min(1+j, 2K_0-1-j)}{1+j} \\ &\leq \frac{\eta L_2(\mathbf{w}_0)^2 B^4 K_0}{2}. \end{aligned}$$

Combining the above two bounds proves (74), the second part of [Lemma E.20](#). \square

We introduce one more Lemma, an intermediate step in the proof of [Fang et al. \(2019\)](#).

Lemma E.21. *We have with probability at least $1 - p/12$ that*

$$\langle \nabla G_S(\mathbf{y}^{\mathcal{K}}), \mathbf{u}^{\mathcal{K}} - \mathbf{y}^{\mathcal{K}} \rangle \leq \frac{3\eta \sum_{k=0}^{\mathcal{K}} \|\nabla G_S(\mathbf{y}^k)\|^2}{16} + 8\eta\sigma_1(\mathbf{w}_0)^2 \log(48K_0/p) + \eta L_2(\mathbf{w}_0)^2 B^4 K_0/2.$$

Proof of Lemma E.21. Let $\mathbf{y}^* = \arg \min_{\mathbf{y}} G_S(\mathbf{y})$; this exists as G is convex in the subspace \mathcal{S} , by the definition of \mathcal{S} . By the optimality condition of \mathbf{y}^* , we have:

$$\nabla_{\mathbf{u}} F(\mathbf{x}^0) = -\mathbf{H}_S \mathbf{y}^*. \quad (79)$$

Let $\tilde{\mathbf{y}}^k = \mathbf{y}^k - \mathbf{y}^*$. From the update rule of \mathbf{y}^k and the optimality condition (79), we obtain:

$$\mathbf{H}_S \tilde{\mathbf{y}}^k = \nabla G_S(\mathbf{y}^k), \tilde{\mathbf{y}}^{k+1} = \tilde{\mathbf{y}}^k - \eta \mathbf{H}_S \tilde{\mathbf{y}}^k. \quad (80)$$

Consequently, using (80) and (76), we have:

$$\langle \nabla G_S(\mathbf{y}^{\mathcal{K}}), \mathbf{u}^{\mathcal{K}} - \mathbf{y}^{\mathcal{K}} \rangle$$

$$\begin{aligned}
&= \langle \tilde{\mathbf{y}}^{\mathcal{K}}, \mathbf{z}^{\mathcal{K}} \rangle_{\mathbf{H}_S} \\
&= \eta \sum_{k=1}^{\mathcal{K}} \langle \tilde{\mathbf{y}}^{k-1}, \boldsymbol{\xi}_u^k \rangle_{\mathbf{H}_S(\mathbf{I}-\eta\mathbf{H}_S)^{\mathcal{K}-k+1}} - \eta \sum_{k=0}^{\mathcal{K}-1} \langle \tilde{\mathbf{y}}^k, \nabla_{\mathbf{u}} F(\mathbf{x}^k) - \nabla G_S(\mathbf{u}^k) \rangle_{\mathbf{H}_S(\mathbf{I}-\eta\mathbf{H}_S)^{\mathcal{K}-k}}.
\end{aligned}$$

Now we bound both of these sums in a manner similar to the proof of [Lemma E.20](#):

- For the first term: For any fixed l , $1 \leq l \leq K_0$, define a real-valued stochastic process for any k , $1 \leq k \leq \min(l, \mathcal{K}_0)$ by:

$$Y_{l,k} = \langle \tilde{\mathbf{y}}^{k-1}, \boldsymbol{\xi}_u^k \rangle_{\mathbf{H}_S(\mathbf{I}-\eta\mathbf{H}_S)^{l-k+1}} \mathbf{1}_{k-1 < \mathcal{K}} = \begin{cases} \langle \tilde{\mathbf{y}}^{k-1}, \boldsymbol{\xi}_u^k \rangle_{\mathbf{H}_S(\mathbf{I}-\eta\mathbf{H}_S)^{l-k+1}} & : k \leq \mathcal{K} \\ 0 & : k > \mathcal{K}. \end{cases}$$

Analogously to earlier, recalling $\mathcal{K} \leq \mathcal{K}_0$, it's easy to check that for any fixed l , $Y_{l,k}$ is \mathfrak{F}^k measurable, and that all terms defining $Y_{l,k}$ are \mathfrak{F}^{k-1} measurable except $\boldsymbol{\xi}_u^k$. Thus,

$$\mathbb{E}[Y_{l,k} | \mathfrak{F}^{k-1}] = 0.$$

We furthermore have for any fixed l , $1 \leq l \leq K_0$ and k , $1 \leq k \leq l$,

$$\|Y_{l,k}\|^2 \leq \sigma_1(\mathbf{w}_0)^2 \|\nabla G_S(\mathbf{y}^{k-1})\|^2.$$

To justify why the above holds, clearly this is evident for $k > \mathcal{K}$. For $k \leq \mathcal{K} \leq \mathcal{K}_0$, note that

$$\begin{aligned}
|Y_{l,k}|^2 &= \|\langle \tilde{\mathbf{y}}^{k-1}, \boldsymbol{\xi}_u^k \rangle_{\mathbf{H}_S(\mathbf{I}-\eta\mathbf{H}_S)^{l-k+1}}\|^2 = \|\langle \mathbf{H}_S \tilde{\mathbf{y}}^{k-1}, \boldsymbol{\xi}_u^k \rangle_{(\mathbf{I}-\eta\mathbf{H}_S)^{l-k+1}}\|^2 \\
&= \|\langle \nabla G_S(\mathbf{y}^{k-1}), \boldsymbol{\xi}_u^k \rangle_{(\mathbf{I}-\eta\mathbf{H}_S)^{l-k+1}}\|^2 \\
&\leq \sigma_1(\mathbf{w}_0)^2 \|\nabla G_S(\mathbf{y}^{k-1})\|^2 \|(\mathbf{I}-\eta\mathbf{H}_S)^{l-k+1}\|^2 \\
&\leq \sigma_1(\mathbf{w}_0)^2 \|\nabla G_S(\mathbf{y}^{k-1})\|^2.
\end{aligned}$$

Here we used that \mathbf{H}_S is symmetric, that (80), that $\|\mathbf{I}-\eta\mathbf{H}_S^{l-k+1}\| \leq 1$ which we have argued earlier in the proof of [Lemma E.20](#), and that $\|\boldsymbol{\xi}_u^k\| \leq \sigma_1(\mathbf{w}_0)$ as $k \leq l \leq \mathcal{K}_0$ by [Lemma E.18](#) and properties of projection matrices.

Now for any l , $1 \leq l \leq K_0$, by the Azuma–Hoeffding inequality, we have with probability at least $1 - p/(12K_0)$ that

$$\left| \eta \sum_{k=1}^l Y_{l,k} \right| \leq \sqrt{2\eta^2 \sigma_1(\mathbf{w}_0)^2 \log(24K_0/p) \sum_{k=0}^{l-1} \|\nabla G_S(\mathbf{y}^k)\|^2}.$$

Taking a Union Bound, it follows that with probability at least $1 - p/12$, the above holds for all l with $1 \leq l \leq K_0$.

Because $1 \leq \mathcal{K} \leq K_0$ always holds, using the definition of $Y_{l,k}$ for $k \leq \mathcal{K}$, we obtain with probability at least $1 - \frac{p}{12}$ that

$$\begin{aligned}
\left| \eta \sum_{k=1}^{\mathcal{K}} \langle \tilde{\mathbf{y}}_{k-1}, \boldsymbol{\xi}_u^k \rangle_{\mathbf{H}_S(\mathbf{I}-\eta\mathbf{H}_S)^{\mathcal{K}-k+1}} \right| &= \left| \eta \sum_{k=1}^{\mathcal{K}} Y_{\mathcal{K},k} \right| \\
&\leq \sqrt{2\eta^2 \sigma_1(\mathbf{w}_0)^2 \log(24K_0/p) \sum_{k=0}^{\mathcal{K}-1} \|\nabla G_S(\mathbf{y}^k)\|^2} \\
&\leq \frac{\eta}{16} + 8\eta \sigma_1(\mathbf{w}_0)^2 \log(48K_0/p)
\end{aligned}$$

where we used AM-GM in the last step. This holds because we have this upper bound on $|\sum_{k=1}^l Y_{l,k}|$ irrespective of which value of l , $1 \leq l \leq K_0$ that \mathcal{K} takes on. The first equality holds by our definition of $Y_{l,k}$ for $k \leq \mathcal{K}$.

- For the second term: note

$$\eta \sum_{k=0}^{\mathcal{K}-1} \langle \tilde{\mathbf{y}}^k, \nabla_{\mathbf{u}} F(\mathbf{x}^k) - \nabla G_S(\mathbf{u}^k) \rangle_{\mathbf{H}_S(\mathbf{I}-\eta\mathbf{H}_S)^{\mathcal{K}-k}}$$

$$\begin{aligned}
&= \eta \sum_{k=0}^{\mathcal{K}-1} \langle \nabla G_{\mathcal{S}}(\mathbf{y}^{\mathcal{K}}), \nabla_{\mathbf{u}} F(\mathbf{x}^k) - \nabla G_{\mathcal{S}}(\mathbf{u}^k) \rangle_{(\mathbf{I} - \eta \mathbf{H}_{\mathcal{S}})^{\mathcal{K}-k}} \\
&\leq \eta \sum_{k=0}^{\mathcal{K}-1} \|\nabla G_{\mathcal{S}}(\mathbf{y}^{\mathcal{K}})\| \|\nabla_{\mathbf{u}} F(\mathbf{x}^k) - \nabla G_{\mathcal{S}}(\mathbf{u}^k)\| \\
&\leq \frac{\eta \sum_{k=0}^{\mathcal{K}-1} \|\nabla G_{\mathcal{S}}(\mathbf{y}^{\mathcal{K}})\|^2}{8} + 2\eta \sum_{k=0}^{\mathcal{K}-1} \|\nabla_{\mathbf{u}} F(\mathbf{x}^k) - \nabla G_{\mathcal{S}}(\mathbf{u}^k)\|^2 \\
&\leq \frac{\eta \sum_{k=0}^{\mathcal{K}-1} \|\nabla G_{\mathcal{S}}(\mathbf{y}^{\mathcal{K}})\|^2}{8} + \frac{1}{2} \eta L_2(\mathbf{w}_0)^2 B^4 K_0.
\end{aligned}$$

The first step above uses that $\mathbf{H}_{\mathcal{S}}$ is symmetric and (80). The second step uses that $k \leq \mathcal{K}$ and that $\|\mathbf{I} - \eta \mathbf{H}_{\mathcal{S}}\| \leq 1$, as argued in the proof of Lemma E.20. The third step uses AM-GM.

The last step uses that $\mathcal{K} \leq K_0$ and Lemma E.17; for $k < \mathcal{K}$, we have $\mathbf{x}^k \in \mathbb{B}(\mathbf{x}^0, B)$.

Combining these above two bounds proves Lemma E.21. \square

Now we finish the proof of Lemma E.19. As done in Fang et al. (2019), we combine Lemma E.20, Lemma E.21 with (72) to prove Lemma E.19 as follows. In particular, taking a Union Bound over the events from Lemma E.20 and Lemma E.21, we obtain with probability at least $1 - p/4$ that

$$\begin{aligned}
G_{\mathcal{S}}(\mathbf{u}^{\mathcal{K}}) &= G_{\mathcal{S}}(\mathbf{y}^{\mathcal{K}}) + \langle \nabla G_{\mathcal{S}}(\mathbf{y}^{\mathcal{K}}), \mathbf{u}^{\mathcal{K}} - \mathbf{y}^{\mathcal{K}} \rangle + \frac{1}{2} (\mathbf{u}^{\mathcal{K}} - \mathbf{y}^{\mathcal{K}})^{\top} \mathbf{H} (\mathbf{u}^{\mathcal{K}} - \mathbf{y}^{\mathcal{K}}) \\
&\leq G_{\mathcal{S}}(\mathbf{y}^{\mathcal{K}}) + \langle \nabla G_{\mathcal{S}}(\mathbf{y}^{\mathcal{K}}), \mathbf{u}^{\mathcal{K}} - \mathbf{y}^{\mathcal{K}} \rangle + \frac{1}{2} (\mathbf{u}^{\mathcal{K}} - \mathbf{y}^{\mathcal{K}})^{\top} \mathbf{H}_{\mathcal{S}} (\mathbf{u}^{\mathcal{K}} - \mathbf{y}^{\mathcal{K}}) \\
&\leq G_{\mathcal{S}}(\mathbf{y}^{\mathcal{K}}) + \frac{3\eta}{16} \sum_{k=0}^{\mathcal{K}-1} \|\nabla G_{\mathcal{S}}(\mathbf{y}^k)\|^2 \\
&\quad + 4\eta\sigma_1(\mathbf{w}_0)^2 (\log(K_0) + 3) \log(48K_0/p) + L_2(\mathbf{w}_0)^2 \eta B^4 K_0.
\end{aligned}$$

Here the first two lines used the definition of $G_{\mathcal{S}}$ and \mathcal{S} . The last line above applied Lemma E.21 together with the second part of Lemma E.20.

Now combining the above with (72), we obtain

$$\begin{aligned}
G_{\mathcal{S}}(\mathbf{u}^{\mathcal{K}}) &\leq G_{\mathcal{S}}(\mathbf{y}^{\mathcal{K}}) + \frac{3\eta}{16} \sum_{k=0}^{\mathcal{K}-1} \|\nabla G_{\mathcal{S}}(\mathbf{y}^k)\|^2 \\
&\quad + 4\eta\sigma_1(\mathbf{w}_0)^2 (\log(K_0) + 3) \log(48K_0/p) + L_2(\mathbf{w}_0)^2 \eta B^4 K_0 \\
&\leq G_{\mathcal{S}}(\mathbf{u}^0) - \frac{25}{32} \sum_{k=0}^{\mathcal{K}-1} \|\nabla G_{\mathcal{S}}(\mathbf{y}^k)\|^2 \\
&\quad + 4\eta\sigma_1(\mathbf{w}_0)^2 (\log(K_0) + 3) \log(48K_0/p) + \eta L_2(\mathbf{w}_0)^2 B^4 K_0,
\end{aligned}$$

where we also used $\mathbf{y}^0 = \mathbf{u}^0$. This proves Lemma E.19. \square

We now analyze the orthogonal complement of \mathcal{S} , \mathcal{S}^{\perp} as in Fang et al. (2019), where the analysis again goes through since the iterates are ‘local’, being prior to the escape time \mathcal{K} :

Lemma E.22 (Equivalent of Lemma 18, Fang et al. (2019)). *Deterministically, we have:*

$$G_{\mathcal{S}^{\perp}}(\mathbf{v}^{\mathcal{K}}) \leq G_{\mathcal{S}^{\perp}}(\mathbf{v}^0) - \sum_{k=1}^{\mathcal{K}} \eta \langle \nabla G_{\mathcal{S}^{\perp}}(\mathbf{v}_{\mathcal{K}-1}), \boldsymbol{\xi}_{\mathbf{v}}^k \rangle - \frac{7\eta}{8} \sum_{k=0}^{\mathcal{K}-1} \|\nabla G_{\mathcal{S}^{\perp}}(\mathbf{x}^k)\|^2 + L_2(\mathbf{w}_0)^2 B^4 \eta K_0^2.$$

Note by choice of parameters that $L_2(\mathbf{w}_0)^2 B^4 \eta K_0^2 = \tilde{O}(\varepsilon^{1.5})$, where again the $\tilde{O}(\cdot)$ hides $F(\mathbf{w}_0)$ -dependence.

Proof. By definition of $G_{\mathcal{S}^{\perp}}$, and using definition of \mathcal{S}^{\perp} which implies $\mathbf{H}_{\mathcal{S}^{\perp}} \leq 0$, we obtain

$$G_{\mathcal{S}^{\perp}}(\mathbf{v}^{k+1}) = G_{\mathcal{S}^{\perp}}(\mathbf{v}^k) + \langle \nabla G_{\mathcal{S}^{\perp}}(\mathbf{v}^k), \mathbf{v}^{k+1} - \mathbf{v}^k \rangle + \frac{1}{2} (\mathbf{v}^{k+1} - \mathbf{v}^k)^{\top} \mathbf{H}_{\mathcal{S}^{\perp}} (\mathbf{v}^{k+1} - \mathbf{v}^k)$$

$$\begin{aligned}
&\leq G_{\mathcal{S}^\perp}(\mathbf{v}^k) + \langle \nabla G_{\mathcal{S}^\perp}(\mathbf{v}^k), \mathbf{v}^{k+1} - \mathbf{v}^k \rangle \\
&= G_{\mathcal{S}^\perp}(\mathbf{v}^k) - \eta \langle \nabla G_{\mathcal{S}^\perp}(\mathbf{v}^k), \nabla_{\mathbf{v}} F(\mathbf{x}^k) + \boldsymbol{\xi}_{\mathbf{v}}^{k+1} \rangle \\
&= G_{\mathcal{S}^\perp}(\mathbf{v}^k) - \eta \|\nabla G_{\mathcal{S}^\perp}(\mathbf{v}^k)\|^2 - \langle \eta \nabla G_{\mathcal{S}^\perp}(\mathbf{v}^k), \nabla_{\mathbf{v}} F(\mathbf{x}^k) - \nabla G_{\mathcal{S}^\perp}(\mathbf{v}^k) \rangle \\
&\quad - \eta \langle \nabla G_{\mathcal{S}^\perp}(\mathbf{v}^k), \boldsymbol{\xi}_{\mathbf{v}}^{k+1} \rangle \\
&\leq G_{\mathcal{S}^\perp}(\mathbf{v}^k) - \eta \langle \nabla G_{\mathcal{S}^\perp}(\mathbf{v}^k), \boldsymbol{\xi}_{\mathbf{v}}^{k+1} \rangle - \frac{7\eta}{8} \|\nabla G_{\mathcal{S}^\perp}(\mathbf{v}^k)\|^2 + 2\eta \|\nabla_{\mathbf{v}} F(\mathbf{x}^k) - \nabla G_{\mathcal{S}^\perp}(\mathbf{v}^k)\|^2.
\end{aligned}$$

The last step uses AM-GM.

Substituting and telescoping the above for k from 0 to $\mathcal{K} - 1$, we have:

$$\begin{aligned}
&G_{\mathcal{S}^\perp}(\mathbf{v}^{\mathcal{K}}) \\
&\leq G_{\mathcal{S}^\perp}(\mathbf{v}^0) - \sum_{k=1}^{\mathcal{K}} \eta \langle \nabla G_{\mathcal{S}^\perp}(\mathbf{v}^{k-1}), \boldsymbol{\xi}_{\mathbf{v}}^k \rangle - \frac{7\eta}{8} \sum_{k=0}^{\mathcal{K}-1} \|\nabla G_{\mathcal{S}^\perp}(\mathbf{x}^k)\|^2 + 2\eta \sum_{k=0}^{\mathcal{K}-1} \|\nabla_{\mathbf{v}} F(\mathbf{x}^k) - \nabla G_{\mathcal{S}^\perp}(\mathbf{v}^k)\|^2 \\
&\leq G_{\mathcal{S}^\perp}(\mathbf{v}^0) - \sum_{k=1}^{\mathcal{K}} \eta \langle \nabla G_{\mathcal{S}^\perp}(\mathbf{v}^{k-1}), \boldsymbol{\xi}_{\mathbf{v}}^k \rangle - \frac{7\eta}{8} \sum_{k=0}^{\mathcal{K}-1} \|\nabla G_{\mathcal{S}^\perp}(\mathbf{x}^k)\|^2 + \frac{L_2(\mathbf{w}_0)^2 B^4 \eta \mathcal{K}_0}{2}.
\end{aligned}$$

Here, the second inequality uses that by [Lemma E.17](#), for all $k \leq \mathcal{K} - 1$, we have $\mathbf{x}^k \in \mathbb{B}(\mathbf{x}^0, B)$ and so

$$\|\nabla_{\mathbf{v}} F(\mathbf{x}^k) - \nabla G_{\mathcal{S}^\perp}(\mathbf{v}^k)\| = \|\mathcal{P}_{\mathcal{S}^\perp}(\nabla F(\mathbf{x}^k) - \nabla G(\mathbf{x}^k))\| \leq \|\nabla F(\mathbf{x}^k) - \nabla G(\mathbf{x}^k)\| \leq \frac{L_2(\mathbf{w}_0)B^2}{2}.$$

This completes the proof. \square

Completing the Proof: Now we have all the ingredients in hand to prove [Lemma E.2](#).

Proof of Lemma E.2. Again, we follow the strategy of [Fang et al. \(2019\)](#) and adapt it to our setting here where we do not have global bounds on the Lipschitz constants of the gradient and Hessian. With [Lemma E.19](#) and [Lemma E.22](#) in hand, the idea will be to show

$$\sum_{k=0}^{\mathcal{K}_0-1} \|\nabla G_{\mathcal{S}^\perp}(\mathbf{v}^k)\|^2 + \sum_{k=0}^{\mathcal{K}_0-1} \|\nabla G_{\mathcal{S}}(\mathbf{y}^k)\|^2 = \tilde{\Omega}(1),$$

and to bound the noise term

$$-\sum_{k=1}^{\mathcal{K}} \eta \langle \nabla G_{\mathcal{S}^\perp}(\mathbf{v}^{k-1}), \boldsymbol{\xi}_{\mathbf{v}}^k \rangle.$$

We break the proof of [Lemma E.2](#) into two cases:

1. $\|\nabla F(\mathbf{x}^0)\| > 5\sigma_1(\mathbf{w}_0)$.
2. $\|\nabla F(\mathbf{x}^0)\| \leq 5\sigma_1(\mathbf{w}_0)$.

Case 1: This case is more straightforward as the gradient is large, and will not use the quadratic approximation we developed earlier.

Consider any $k, 0 \leq k \leq \mathcal{K} - 1$. Thus $\mathbf{x}^k \in \mathbb{B}(\mathbf{x}^0, B)$, and so $\mathbf{u} \in \mathbb{B}(\mathbf{x}^0, B)$ for all $\mathbf{u} \in \overline{\mathbf{x}^0 \mathbf{x}^k}$. By [Lemma E.7](#), as $\mathbf{x}^0 \in \mathcal{L}_{F, F(\mathbf{w}_0)}$, we have $\|\nabla^2 F(\mathbf{u})\| \leq L_1(\mathbf{w}_0)$ for all such \mathbf{u} . Thus as $\|\nabla F(\mathbf{x}^0)\| > 5\sigma_1(\mathbf{w}_0)$ and by our choice of parameters,

$$\|\nabla F(\mathbf{x}^k)\| \geq \|\nabla F(\mathbf{x}^0)\| - \|\nabla F(\mathbf{x}^k) - \nabla F(\mathbf{x}^0)\| \geq 5\sigma_1(\mathbf{w}_0) - L_1(\mathbf{w}_0)B \geq \frac{9}{2}\sigma_1(\mathbf{w}_0). \quad (81)$$

Similarly, as $\mathbf{x}^{k+1} = \mathbf{x}^k - \eta \nabla \tilde{f}(\mathbf{x}^k; \zeta_{k+1})$ and again as $\mathbf{x}^0 \in \mathcal{L}_{F, F(\mathbf{w}_0)}$, we have $\|\nabla^2 F(\mathbf{u})\| \leq L_1(\mathbf{w}_0)$ for all $\mathbf{u} \in \overline{\mathbf{x}^k \mathbf{x}^{k+1}}$ by [Lemma E.7](#). Applying [Lemma A.1](#), for all $0 \leq k \leq \mathcal{K} - 1$, we obtain:

$$F(\mathbf{x}^{k+1}) - F(\mathbf{x}^k) \leq \langle \nabla F(\mathbf{x}^k), \mathbf{x}^{k+1} - \mathbf{x}^k \rangle + \frac{L_1(\mathbf{w}_0)}{2} \|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2$$

$$\begin{aligned}
&= -\eta \|\nabla F(\mathbf{x}^k)\|^2 - \eta \langle \nabla F(\mathbf{x}^k), \boldsymbol{\xi}^{k+1} \rangle + \frac{L_1(\mathbf{w}_0)\eta^2}{2} \|\nabla F(\mathbf{x}^k) + \boldsymbol{\xi}^{k+1}\|^2 \\
&\leq -\eta \|\nabla F(\mathbf{x}^k)\|^2 - \eta \langle \nabla F(\mathbf{x}^k), \boldsymbol{\xi}^{k+1} \rangle + L_1(\mathbf{w}_0)\eta^2 \|\nabla F(\mathbf{x}^k)\|^2 + L_1(\mathbf{w}_0)\eta^2 \|\boldsymbol{\xi}^{k+1}\|^2 \\
&\leq \eta \left(-\frac{15}{16} + \frac{5}{32} \right) \|\nabla F(\mathbf{x}^k)\|^2 + \frac{8}{5} \eta \sigma_1(\mathbf{w}_0)^2 + L_1(\mathbf{w}_0)\eta^2 \sigma_1(\mathbf{w}_0)^2 \\
&\leq -\frac{25\eta}{32} \|\nabla F(\mathbf{x}^k)\|^2 + 2\eta\sigma^2 \\
&\leq -\eta \left(\frac{25}{32} - \frac{8}{81} \right) \|\nabla F(\mathbf{x}^k)\|^2.
\end{aligned}$$

Note here that we need to consider a bound on the Lipschitz constant of the gradient between \mathbf{x}^{K-1} and \mathbf{x}^K ; see [Remark 15](#). Here, we used the update rule of SGD, AM-GM and Young's Inequality, that $L_1(\mathbf{w}_0)\eta \leq \frac{1}{16}$ by our choice of hyperparameters, [Lemma E.18](#), and finally (81) in the last step.

Telescoping the above inequality from $k = 0$ to $K - 1$, we get:

$$F(\mathbf{x}^K) - F(\mathbf{x}^0) \leq -\eta \left(\frac{25}{32} - \frac{8}{81} \right) \sum_{k=0}^{K-1} \|\nabla F(\mathbf{x}^k)\|^2. \quad (82)$$

To upper bound the right hand side above, note by Triangle Inequality that

$$\begin{aligned}
\left\| \eta \sum_{k=0}^{K-1} \nabla F(\mathbf{x}^k) \right\| &= \left\| -\eta \sum_{k=0}^{K-1} \nabla F(\mathbf{x}^k) \right\| \\
&= \left\| \mathbf{x}^K - \mathbf{x}^0 + \eta \sum_{k=1}^K \boldsymbol{\xi}^k \right\| \\
&\geq \|\mathbf{x}^K - \mathbf{x}^0\| - \left\| \eta \sum_{k=1}^K \boldsymbol{\xi}^k \right\|.
\end{aligned} \quad (83)$$

By the Vector-Martingale Concentration Inequality [Theorem C.1](#) and the bound $\|\boldsymbol{\xi}^k\| \leq \sigma_1(\mathbf{w}_0)$ for all $k \leq K$ by [Lemma E.5](#), we obtain with probability at least $1 - p/12$:

$$\left\| \eta \sum_{k=1}^K \boldsymbol{\xi}^k \right\| = \left\| \eta \sum_{k=1}^{K_0} \boldsymbol{\xi}^k \mathbf{1}_{k \leq K} \right\| \leq 2\eta\sigma_1(\mathbf{w}_0) \sqrt{K_0 \log(48/p)} \leq \frac{B}{16}. \quad (84)$$

Here, we used the fact that $\mathbf{1}_{k \leq K} \equiv \mathbf{1}_{k-1 < K}$ and consequently $\mathbf{1}_{k \leq K}$ is \mathcal{F}^{k-1} -measurable, and that $\mathbb{E}[\boldsymbol{\xi}^k | \mathcal{F}^{k-1}] = 0$, $\|\boldsymbol{\xi}^k\| \leq \sigma_1(\mathbf{w}_0)$ for all $k \leq K$.

Suppose the above event implying (84) occurs, which has probability at least $1 - \frac{p}{12}$. Under this event, suppose that \mathbf{x}^k is able to leave the ball $\mathbb{B}(\mathbf{x}^0, B)$ in K_0 iterations or less. If this is the case, then we have $K = K_0 \leq K_0$, and so $\|\mathbf{x}^K - \mathbf{x}^0\| \geq B$. Thus conditioned on the aforementioned event implying (84), if \mathbf{x}^k is able to leave the ball $\mathbb{B}(\mathbf{x}^0, B)$ in K_0 iterations or less, we obtain

$$\eta \sum_{k=0}^{K-1} \|\nabla F(\mathbf{x}^k)\|^2 \geq \frac{1}{\eta K} \left\| \sum_{k=0}^{K-1} \eta \nabla F(\mathbf{x}^k) \right\|^2 \geq \frac{1}{\eta K} \left(B - \frac{1}{16} B \right)^2 \geq \frac{15^2 B^2}{16^2 \eta K} \geq \frac{15^2 B^2}{16^2 \eta K_0},$$

where we combined (83), (84) to lower bound $\left\| \sum_{k=0}^{K-1} \eta \nabla F(\mathbf{x}^k) \right\|$. Here the first step holds by the elementary inequality $\left\| \sum_{i=0}^l \mathbf{a}_i \right\|^2 \leq l \sum_{i=0}^l \|\mathbf{a}_i\|^2$, and the last step uses $K_0 \geq K$.

Consequently by combining with (82), with probability at least $1 - \frac{p}{12}$, if \mathbf{x}^k is able to leave the ball $\mathbb{B}(\mathbf{x}^0, B)$ in K_0 iterations or less, we have

$$F(\mathbf{x}^K) \leq F(\mathbf{x}^0) - \left(\frac{25}{32} - \frac{8}{81} \right) \cdot \frac{15^2 B^2}{16^2 \eta K_0} < F(\mathbf{x}^0) - \frac{B^2}{7\eta K_0}.$$

Case 2: Suppose $\|\nabla F(\mathbf{x}^0)\| \leq 5\sigma_1(\mathbf{w}_0)$. To obtain the desired result, we first define and prove the following Lemmas. Proving these Lemmas in turn utilizes the Lemmas on quadratic approximation we have established earlier.

Lemma E.23. For all $0 \leq k \leq \mathcal{K} - 1$, we have

$$\|\nabla G_{\mathcal{S}^\perp}(\mathbf{v}^k)\| \leq \frac{11}{2}\sigma_1(\mathbf{w}_0).$$

Proof. By the condition in this case, properties of projection matrices, and as $\mathbf{v}^0 = 0$,

$$\|\nabla G_{\mathcal{S}^\perp}(\mathbf{v}^0)\| = \|\nabla_{\mathbf{v}} F(\mathbf{x}^0)\| \leq \|\nabla F(\mathbf{x}^0)\| \leq 5\sigma_1(\mathbf{w}_0).$$

Note for $k \leq \mathcal{K} - 1$, we have

$$\|\mathbf{v}^k - \mathbf{v}^0\| = \|\mathcal{P}_{\mathcal{S}^\perp}(\mathbf{x}^k - \mathbf{x}^0)\| \leq B.$$

Thus

$$\begin{aligned} \|\nabla G_{\mathcal{S}^\perp}(\mathbf{v}^k)\| &\leq \|\nabla G_{\mathcal{S}^\perp}(\mathbf{v}^0)\| + \|\nabla G_{\mathcal{S}^\perp}(\mathbf{v}^k) - \nabla G_{\mathcal{S}^\perp}(\mathbf{v}^0)\| \\ &\leq 5\sigma_1(\mathbf{w}_0) + L_1(\mathbf{w}_0)B \\ &\leq \frac{11}{2}\sigma_1(\mathbf{w}_0). \end{aligned}$$

The above uses our choice of hyperparameters, and that

$$\|\nabla G_{\mathcal{S}^\perp}(\mathbf{v}^k) - \nabla G_{\mathcal{S}^\perp}(\mathbf{v}^0)\| = \|\mathbf{H}_{\mathcal{S}^\perp}(\mathbf{v}^k - \mathbf{v}^0)\| \leq \|\mathbf{H}\| \|\mathbf{v}^k - \mathbf{v}^0\| \leq L_1(\mathbf{w}_0) \|\mathbf{v}^k - \mathbf{v}^0\|,$$

which in turn follows because $\mathbf{x}^0 \in \mathcal{L}_{F,F}(\mathbf{w}_0)$ and by [Assumption 1.1](#). \square

The next Lemma is obtained by combining [Lemma E.19](#) and [Lemma E.22](#), and it gives us a way to upper bound $F(\mathbf{x}^k) - F(\mathbf{x}^0)$.

Lemma E.24 (Equivalent of Lemma 19 in [Fang et al. \(2019\)](#)). If $\|\nabla F(\mathbf{x}^0)\| \leq 5\sigma_1(\mathbf{w}_0)$, with probability $1 - \frac{\rho}{4}$, we have

$$\begin{aligned} F(\mathbf{x}^\mathcal{K}) \leq & F(\mathbf{x}^0) - \eta \sum_{k=1}^{\mathcal{K}} \langle \nabla G_{\mathcal{S}^\perp}(\mathbf{v}^{k-1}), \boldsymbol{\xi}_{\mathbf{v}}^k \rangle + \left(\frac{3}{256} + \frac{1}{80} \right) \frac{B^2}{\eta K_0} \\ & - \frac{7\eta}{8} \sum_{k=0}^{\mathcal{K}-1} \|\nabla G_{\mathcal{S}^\perp}(\mathbf{v}^k)\|^2 - \frac{25\eta}{32} \sum_{k=0}^{\mathcal{K}-1} \|\nabla G_{\mathcal{S}}(\mathbf{y}^k)\|^2. \end{aligned}$$

Proof. For $k \leq \mathcal{K} - 1$, we have $\mathbf{x}^k \in \mathbb{B}(\mathbf{x}^0, B)$. Consequently the entire line segment $\overline{\mathbf{x}^0 \mathbf{x}^k}$ lies in $\mathbb{B}(\mathbf{x}^0, B)$. As $\mathbf{x}^0 \in \mathcal{L}_{F,F}(\mathbf{w}_0)$, by [Lemma E.7](#), we have

$$\|\nabla F(\mathbf{x}^k) - \nabla F(\mathbf{x}^0)\| \leq L_1(\mathbf{w}_0) \|\mathbf{x}^k - \mathbf{x}^0\| \leq L_1(\mathbf{w}_0)B.$$

Thus by our choice of parameters, as per [Remark 12](#),

$$\|\nabla F(\mathbf{x}^k)\| \leq \|\nabla F(\mathbf{x}^0)\| + \|\nabla F(\mathbf{x}^k) - \nabla F(\mathbf{x}^0)\| \leq 5\sigma_1(\mathbf{w}_0) + L_1(\mathbf{w}_0)B \leq \frac{11}{2}\sigma_1(\mathbf{w}_0).$$

Recalling $\|\boldsymbol{\xi}^\mathcal{K}\| \leq \sigma_1(\mathbf{w}_0)$ by [Lemma E.18](#), we obtain from our choice of parameters as per [Remark 12](#) that

$$\|\mathbf{x}^\mathcal{K} - \mathbf{x}^0\| \leq \|\mathbf{x}^0 - \mathbf{x}^{\mathcal{K}-1}\| + \eta \|\nabla F(\mathbf{x}^{\mathcal{K}-1}) + \boldsymbol{\xi}^\mathcal{K}\| \leq B + \frac{13}{2}\eta\sigma_1(\mathbf{w}_0) \leq B + \frac{B}{100}. \quad (85)$$

Using this, we then bound the difference between $F(\mathbf{x}^\mathcal{K})$ and $G(\mathbf{x}^\mathcal{K})$. As $\mathbf{x}^\mathcal{K} = \mathbf{x}^{\mathcal{K}-1} - \eta \nabla \tilde{f}(\mathbf{x}^{\mathcal{K}-1}; \boldsymbol{\zeta}_\mathcal{K})$, as $\mathbf{x}^{\mathcal{K}-1} \in \mathbb{B}(\mathbf{x}^0, B)$, and as $\mathbf{x}^0 \in \mathcal{L}_{F,F}(\mathbf{w}_0)$, we have $\|\nabla^2 F(\mathbf{u}) - \nabla^2 F(\mathbf{x}^0)\| \leq L_2(\mathbf{w}_0) \|\mathbf{u} - \mathbf{x}^0\|$ for all $\mathbf{u} \in \overline{\mathbf{x}^{\mathcal{K}-1} \mathbf{x}^\mathcal{K}}$ by [Lemma E.8](#). Applying [Lemma A.2](#) and recalling that $G_{\mathcal{S}}(\mathbf{u}^\mathcal{K}) + G_{\mathcal{S}^\perp}(\mathbf{v}^\mathcal{K}) = G(\mathbf{x}^\mathcal{K} - \mathbf{x}^0)$, we obtain

$$F(\mathbf{x}^\mathcal{K}) - F(\mathbf{x}^0) - G_{\mathcal{S}}(\mathbf{u}^\mathcal{K}) - G_{\mathcal{S}^\perp}(\mathbf{v}^\mathcal{K}) \leq \frac{L_2(\mathbf{w}_0)}{6} \|\mathbf{x}^\mathcal{K} - \mathbf{x}^0\|^3 \leq \frac{L_2(\mathbf{w}_0)B^3}{5}. \quad (86)$$

Here, we used (85) in the last step. Note here that we need to consider a bound on the Lipschitz constant of the Hessian between $\mathbf{x}^{\mathcal{K}-1}$ and $\mathbf{x}^\mathcal{K}$; see [Remark 15](#).

Now, take a Union Bound over [Lemma E.19](#) and [Lemma E.22](#). We now add the bounds from [Lemma E.19](#) and [Lemma E.22](#) to upper bound $G_S(\mathbf{u}^\mathcal{K}) + G_{S^\perp}(\mathbf{v}^\mathcal{K})$ and use that $G_S(\mathbf{u}^0) + G_{S^\perp}(\mathbf{v}^0) = 0$. Combining with (86), we obtain with probability at least $1 - p/4$ that

$$\begin{aligned} F(\mathbf{x}^\mathcal{K}) &\leq F(\mathbf{x}^0) - \eta \sum_{k=1}^{\mathcal{K}} \langle \nabla G_{S^\perp}(\mathbf{v}_{k-1}), \boldsymbol{\xi}_v^k \rangle + 4\eta\sigma_1(\mathbf{w}_0)^2(1 + 3\log(K_0))\log\left(\frac{48}{p}\right) \\ &\quad - \frac{7\eta}{8} \sum_{k=0}^{\mathcal{K}-1} \|\nabla G_{S^\perp}(\mathbf{v}^k)\|^2 - \frac{25\eta}{32} \sum_{k=0}^{\mathcal{K}-1} \|\nabla G_S(\mathbf{y}^k)\|^2 + \frac{3L_2(\mathbf{w}_0)B^4\eta K_0}{2} + \frac{L_2(\mathbf{w}_0)B^3}{5}. \end{aligned} \quad (87)$$

Note by our choice of hyperparameters (analogous to the choice of hyperparameters from [Fang et al. \(2019\)](#)), we have the following bounds: $4\eta\sigma_1(\mathbf{w}_0)^2(1 + 3\log(K_0))\log\left(\frac{48}{p}\right) \leq \frac{B^2}{256\eta K_0}$, $\frac{3L_2(\mathbf{w}_0)B^4\eta K_0}{2} \leq \frac{B^2}{128\eta K_0}$, $\frac{L_2(\mathbf{w}_0)B^3}{5} \leq \frac{B^2}{80\eta K_0}$.

Combining these above inequalities with (87), with probability at least $1 - p/4$, we obtain

$$\begin{aligned} F(\mathbf{x}^\mathcal{K}) &\leq F(\mathbf{x}^0) - \eta \sum_{k=1}^{\mathcal{K}} \langle \nabla G_{S^\perp}(\mathbf{v}_{k-1}), \boldsymbol{\xi}_v^k \rangle + \left(\frac{3}{256} + \frac{1}{80}\right) \frac{B^2}{\eta K_0} \\ &\quad - \frac{7\eta}{8} \sum_{k=0}^{\mathcal{K}-1} \|\nabla G_{S^\perp}(\mathbf{v}^k)\|^2 - \frac{25\eta}{32} \sum_{k=0}^{\mathcal{K}-1} \|\nabla G_S(\mathbf{y}^k)\|^2. \end{aligned}$$

This implies [Lemma E.24](#). \square

By [Lemma E.24](#), we want to lower bound the gradient norm of G_{S^\perp}, G_S . We do this in the following Lemma, assuming \mathbf{x}^k leaves the ball $\mathbb{B}(\mathbf{x}^0, B)$ in K_0 iterations.

Lemma E.25 (Equivalent of Lemma 20 in [Fang et al. \(2019\)](#)). *With probability $1 - \frac{p}{6}$, if \mathbf{x}^k exits $\mathbb{B}(\mathbf{x}^0, B)$ in K_0 iterations (i.e. $\mathcal{K} = \mathcal{K}_0 \leq K_0$), we have*

$$\eta \sum_{k=0}^{\mathcal{K}-1} \|\nabla G_{S^\perp}(\mathbf{v}^k)\|^2 + \eta \sum_{k=0}^{\mathcal{K}-1} \|\nabla G_S(\mathbf{y}^k)\|^2 \geq \frac{169B^2}{512\eta K_0}.$$

Proof. At a high level, the proof idea is similar to the proof of Case 1 earlier. Telescoping the recursions $\mathbf{v}^k = \mathbf{v}^{k-1} - \eta \boldsymbol{\xi}_v^k - \eta \nabla_{\mathbf{v}} F(\mathbf{x}^k)$ and $\mathbf{y}^k = \mathbf{y}^{k-1} - \eta \nabla G_S(\mathbf{y}^k)$, we obtain

$$\begin{aligned} \left\| \eta \sum_{k=0}^{\mathcal{K}-1} (\nabla G_{S^\perp}(\mathbf{v}^k) + \nabla G_S(\mathbf{y}^k)) \right\| &= \left\| -\eta \sum_{k=0}^{\mathcal{K}-1} (\nabla G_{S^\perp}(\mathbf{v}^k) + \nabla G_S(\mathbf{y}^k)) \right\| \\ &= \left\| \mathbf{v}^\mathcal{K} - \mathbf{v}^0 + \eta \sum_{k=0}^{\mathcal{K}-1} (\boldsymbol{\xi}_v^{k+1} - \nabla G_{S^\perp}(\mathbf{v}^k) + \nabla_{\mathbf{v}} F(\mathbf{x}^k)) + \mathbf{y}^\mathcal{K} - \mathbf{y}^0 \right\| \\ &\geq \left\| \mathbf{v}^\mathcal{K} - \mathbf{v}^0 + \eta \sum_{k=0}^{\mathcal{K}-1} \boldsymbol{\xi}_v^{k+1} + (\mathbf{u}^\mathcal{K} - \mathbf{u}^0) - (\mathbf{z}^\mathcal{K} - \mathbf{z}^0) \right\| \\ &\quad - \left\| \eta \sum_{k=0}^{\mathcal{K}-1} (\nabla G_{S^\perp}(\mathbf{v}^k) - \nabla_{\mathbf{v}} F(\mathbf{x}^k)) \right\|. \end{aligned}$$

Here, we used that $\mathbf{z}^k = \mathbf{u}^k - \mathbf{y}^k$ and the Triangle Inequality.

Next, recall $\mathbf{x}^k - \mathbf{x}^0 = \mathbf{u}^k + \mathbf{v}^k$ for all $k \geq 0$, and $\mathbf{u}^0 = \mathbf{v}^0 = 0$. Thus $\mathbf{x}^k - \mathbf{x}^0 = \mathbf{v}^k - \mathbf{v}^0 + \mathbf{u}^k - \mathbf{u}^0$. Furthermore notice

$$\nabla G_{S^\perp}(\mathbf{v}^k) - \nabla_{\mathbf{v}} F(\mathbf{x}^k) = \mathbf{H}_{S^\perp}(\nabla G(\mathbf{x}^k) - \nabla F(\mathbf{x}^k)).$$

For all $k \leq \mathcal{K} - 1$ we have $\mathbf{x}^k \in \mathbb{B}(\mathbf{x}^0, B)$, so as $\mathbf{x}^0 \in \mathcal{L}_{F,F(\mathbf{w}_0)}$, [Lemma E.17](#) gives

$$\left\| \eta \sum_{k=0}^{\mathcal{K}-1} (\nabla G_{S^\perp}(\mathbf{v}^k) - \nabla_{\mathbf{v}} F(\mathbf{x}^k)) \right\| \leq \eta K_0 \cdot \frac{L_2(\mathbf{w}_0)B^2}{2}.$$

Applying these observations and Triangle Inequality again, we obtain

$$\left\| \eta \sum_{k=0}^{\mathcal{K}-1} (\nabla G_{S^\perp}(\mathbf{v}^k) + \nabla G_S(\mathbf{y}^k)) \right\| \geq \|\mathbf{x}^\mathcal{K} - \mathbf{x}^0\| - \|\mathbf{z}^\mathcal{K} - \mathbf{z}^0\| - \eta \left\| \sum_{k=1}^{\mathcal{K}} \boldsymbol{\xi}_v^k \right\| - \frac{\eta K_0 L_2(\mathbf{w}_0)B^2}{2}$$

$$\geq \|\mathbf{x}^\mathcal{K} - \mathbf{x}^0\| - \|\mathbf{z}^\mathcal{K} - \mathbf{z}^0\| - \frac{B}{32} - \eta \left\| \sum_{k=1}^{\mathcal{K}} \boldsymbol{\xi}_v^k \right\|. \quad (88)$$

and Lemma E.17 combined with the fact that projection matrices do not increase norm and that $\mathbf{x}^k \in \mathbb{B}(\mathbf{x}^0, B)$ for $k < \mathcal{K}$, and the final statement is by the choice of hyperparameters.

Using Lemma E.20 and that $\mathbf{z}^0 = 0$, we obtain with probability at least $1 - \frac{p}{12}$ that

$$\|\mathbf{z}^\mathcal{K} - \mathbf{z}^0\| \leq \frac{3B}{32}. \quad (89)$$

Now recall that $1_{k \leq \mathcal{K}} \equiv 1_{k-1 < \mathcal{K}}$ is \mathfrak{F}^{k-1} -measurable, which implies

$$\mathbb{E}[\boldsymbol{\xi}_v^k 1_{\{k \leq \mathcal{K}\}} | \mathfrak{F}^{k-1}] = \mathbf{0},$$

as the stochastic gradient oracle is unbiased. Furthermore, recall $\|\boldsymbol{\xi}^k\| \leq \sigma_1(\mathbf{w}_0)$ for $k \leq \mathcal{K}$, and projection matrices do not increase norm. Thus by the Vector-Martingale Concentration Inequality Theorem C.1, with probability at least $1 - \frac{p}{12}$, we have

$$\left\| \eta \sum_{k=1}^{\mathcal{K}} \boldsymbol{\xi}_v^k \right\| = \left\| \eta \sum_{k=1}^{K_0} \boldsymbol{\xi}_v^k 1_{\{k \leq \mathcal{K}\}} \right\| \leq 2\eta\sigma_1(\mathbf{w}_0) \sqrt{K_0 \log\left(\frac{48}{p}\right)} \leq \frac{B}{16}. \quad (90)$$

Thus taking a Union Bound over the events implying (89), (89) and combining with the earlier display (88), with probability at least $1 - \frac{p}{6}$, we have

$$\left\| \eta \sum_{k=0}^{\mathcal{K}-1} \nabla G_{\mathcal{S}^\perp}(\mathbf{v}^k) + \nabla G_{\mathcal{S}}(\mathbf{y}^k) \right\| \geq \|\mathbf{x}^\mathcal{K} - \mathbf{x}^0\| - \frac{3B}{16}.$$

Thus with probability at least $1 - \frac{p}{6}$, if \mathbf{x}^k exits $\mathbb{B}(\mathbf{x}^0, B)$ in K_0 iterations (that is, if we have $K_0 \geq \mathcal{K}$), we have

$$\left\| \eta \sum_{k=0}^{\mathcal{K}-1} \nabla G_{\mathcal{S}^\perp}(\mathbf{v}^k) + \nabla G_{\mathcal{S}}(\mathbf{y}^k) \right\| \geq \|\mathbf{x}^\mathcal{K} - \mathbf{x}^0\| - \frac{3B}{16} \geq B - \frac{3B}{16},$$

and so

$$\begin{aligned} \eta \sum_{k=0}^{\mathcal{K}-1} \|\nabla G_{\mathcal{S}^\perp}(\mathbf{v}^k)\|^2 + \eta \sum_{k=0}^{\mathcal{K}-1} \|\nabla G_{\mathcal{S}}(\mathbf{y}^k)\|^2 &\geq \frac{1}{2\eta\mathcal{K}} \left\| \eta \sum_{k=0}^{\mathcal{K}-1} (\nabla G_{\mathcal{S}^\perp}(\mathbf{v}^k) + \nabla G_{\mathcal{S}}(\mathbf{y}^k)) \right\|^2 \\ &\geq \frac{1}{2\eta\mathcal{K}} \left(B - \frac{3B}{16} \right)^2 = \frac{169B^2}{512\eta\mathcal{K}} \geq \frac{169B^2}{512\eta K_0}. \end{aligned}$$

In the first step above we used the elementary inequality $\|\sum_{i=1}^l \mathbf{a}_i\|^2 \leq l \sum_{i=1}^l \|\mathbf{a}_i\|^2$ and Young's Inequality. This proves Lemma E.25. \square

We now combine Lemma E.24, Lemma E.25 to prove Lemma E.2. First recall by Lemma E.24, with probability $1 - p/4$, we have

$$\begin{aligned} F(\mathbf{x}^\mathcal{K}) &\leq F(\mathbf{x}^0) - \eta \sum_{k=1}^{\mathcal{K}} \langle \nabla G_{\mathcal{S}^\perp}(\mathbf{v}^{k-1}), \boldsymbol{\xi}_v^k \rangle + \left(\frac{3}{256} + \frac{1}{80} \right) \frac{B^2}{\eta K_0} \\ &\quad - \frac{7\eta}{8} \sum_{k=0}^{\mathcal{K}-1} \|\nabla G_{\mathcal{S}^\perp}(\mathbf{v}^k)\|^2 - \frac{25\eta}{32} \sum_{k=0}^{\mathcal{K}-1} \|\nabla G_{\mathcal{S}}(\mathbf{y}^k)\|^2. \end{aligned} \quad (91)$$

We first control $\sum_{k=1}^{\mathcal{K}} \langle \nabla G_{\mathcal{S}^\perp}(\mathbf{v}^{k-1}), \boldsymbol{\xi}_v^k \rangle$ by concentration. For all k from 1 to K_0 , note

$$\mathbb{E}[\eta \langle \nabla G_{\mathcal{S}^\perp}(\mathbf{v}^{k-1}), \boldsymbol{\xi}_v^k \rangle 1_{k \leq \mathcal{K}} | \mathfrak{F}_{k-1}] = 0,$$

because $1_{k \leq \mathcal{K}} \equiv 1_{k-1 \leq \mathcal{K}}$, so all terms in $\eta \langle \nabla G_{\mathcal{S}^\perp}(\mathbf{v}^{k-1}), \boldsymbol{\xi}_v^k \rangle 1_{k \leq \mathcal{K}}$ except $\boldsymbol{\xi}_v^k$ are \mathfrak{F}^{k-1} -measurable.

Furthermore, by Lemma E.23 and Lemma E.18, for all $k \leq \mathcal{K}$, we have

$$\left\| \eta \langle \nabla G_{\mathcal{S}^\perp}(\mathbf{v}^{k-1}), \boldsymbol{\xi}_v^k \rangle 1_{k \leq \mathcal{K}} \right\| \leq \frac{11\eta\sigma_1(\mathbf{w}_0)^2}{2},$$

and

$$\mathbb{E}\left[\left\{\eta\langle\nabla G_{S^\perp}(\mathbf{v}^{k-1}), \boldsymbol{\xi}_v^k\rangle 1_{k \leq \mathcal{K}}\right\}^2 \middle| \mathfrak{F}^{k-1}\right] \leq \eta^2 \sigma_1(\mathbf{w}_0)^2 1_{k \leq K} \|\nabla G_{S^\perp}(\mathbf{v}^k)\|^2.$$

Taking $\delta = \frac{p}{3 \log(K_0)}$ in the Data-Dependent Bernstein Inequality [Theorem C.2](#), we obtain with probability at least $1 - \frac{p}{3}$,

$$\begin{aligned} & \sum_{k=1}^{\mathcal{K}} -\eta \langle \nabla G_{S^\perp}(\mathbf{v}^{k-1}), \boldsymbol{\xi}_v^k \rangle \\ &= \sum_{k=1}^{K_0} -\eta \langle \nabla G_{S^\perp}(\mathbf{v}^{k-1}), \boldsymbol{\xi}_v^k \rangle 1_{k \leq \mathcal{K}} \\ &\leq \max \left\{ 11\eta \sigma_1(\mathbf{w}_0)^2 \log \left(\frac{3 \log(K_0)}{p} \right), 4 \sqrt{\eta^2 \sigma_1(\mathbf{w}_0)^2 \sum_{k=0}^{\mathcal{K}-1} \|\nabla G_{S^\perp}(\mathbf{v}^k)\|^2 \log \left(\frac{3 \log(K_0)}{p} \right)} \right\}. \end{aligned} \tag{92}$$

We upper bound each of these terms in the maximum. With our choice of parameters and one application of AM-GM, we have

$$11\eta \sigma_1(\mathbf{w}_0)^2 \log \left(\frac{3 \log(K_0)}{p} \right) \leq \frac{B^2}{100\eta K_0},$$

and

$$\begin{aligned} 4 \sqrt{\eta^2 \sigma^2 \sum_{k=0}^{K-1} \|\nabla G_{S^\perp}(\mathbf{v}^k)\|^2 \log \left(\frac{3 \log(K_0)}{p} \right)} &\leq 32 \log \left(\frac{3 \log(K_0)}{p} \right) \eta \sigma_1(\mathbf{w}_0)^2 + \frac{\eta}{8} \sum_{k=0}^{\mathcal{K}-1} \|\nabla G_{S^\perp}(\mathbf{v}^k)\|^2 \\ &\leq \frac{B^2}{32\eta K_0} + \frac{\eta}{8} \sum_{k=0}^{\mathcal{K}-1} \|\nabla G_{S^\perp}(\mathbf{v}^k)\|^2. \end{aligned}$$

Consequently the second upper bound dominates the maximum from (92). Substituting the above into (92), with probability at least $1 - \frac{p}{3}$, we obtain

$$\sum_{k=1}^{\mathcal{K}} -\eta \langle \nabla G_{S^\perp}(\mathbf{v}^{k-1}), \boldsymbol{\xi}_v^k \rangle \leq \frac{B^2}{32\eta K_0} + \frac{\eta}{8} \sum_{k=0}^{\mathcal{K}-1} \|\nabla G_{S^\perp}(\mathbf{v}^k)\|^2.$$

Combining with (91), we obtain with probability at least $1 - \frac{7p}{12}$ that

$$F(\mathbf{x}^\mathcal{K}) - F(\mathbf{x}^0) \leq \left(\frac{3}{256} + \frac{1}{80} + \frac{1}{32} \right) \frac{B^2}{\eta K_0} - \frac{3\eta}{4} \sum_{k=0}^{\mathcal{K}-1} \|\nabla G_{S^\perp}(\mathbf{v}^k)\|^2 - \frac{3\eta}{4} \sum_{k=0}^{\mathcal{K}-1} \|\nabla G_S(\mathbf{y}^k)\|^2$$

Taking a Union Bound with the event from [Lemma E.25](#), we obtain with probability at least $1 - \frac{3}{4}p$, if \mathbf{x}^k moves out of the ball $B(\mathbf{x}^0, B)$ within K_0 iterations (i.e. $\mathcal{K} = \mathcal{K}_0 \leq K_0$), then

$$F(\mathbf{x}^{\mathcal{K}_0}) - F(\mathbf{x}^0) = F(\mathbf{x}^\mathcal{K}) - F(\mathbf{x}^0) \leq -\left(\frac{3}{4} \cdot \frac{169}{512} - \frac{3}{256} - \frac{1}{80} - \frac{1}{32} \right) \frac{B^2}{\eta K_0} < -\frac{B^2}{7\eta K_0}.$$

This proves [Lemma E.2](#) in Case 2.

Combining Case 1 and Case 2, we obtain [Lemma E.2](#). \square

E.6 Finding Second Order Stationary Points

Here, we finish the proof by showing with high probability, if the algorithm does not escape $B(\mathbf{x}^0, B)$ in K_0 iterates, then the average of the K_0 iterates is a SOSF. In particular, we aim to prove [Lemma E.1](#). Here is where [Lemma E.12](#) is used. In the following, we define $\boldsymbol{\xi}^k$ as in (71). Furthermore, note the proofs of [Lemma E.17](#) and [Lemma E.18](#) still go through under the conditions of [Lemma E.1](#), so we may apply those Lemmas in our proof here.

Proof. We adopt the proof strategy of [Fang et al. \(2019\)](#) in a similar way as we have thus far.

- By Lemma E.12, with probability $1 - \frac{p}{3}$ (namely if the event (66) from Lemma E.12 occurs), then if $\lambda_{\min}(\nabla^2 F(\bar{\mathbf{x}})) \leq -\delta_2$, \mathbf{x}^k will move out of the ball $\mathbb{B}(\mathbf{x}^0, B)$ within K_0 iterations. By taking the contrapositive, we see that with probability $1 - \frac{p}{3}$, if \mathbf{x}^k does not move out of the ball $\mathbb{B}(\mathbf{x}^0, B)$ in K_0 iterations, then $\lambda_{\min}(\nabla^2 F(\mathbf{x}^0)) \geq -\delta_2$. In this case, we have $\mathbf{x}^k \in \mathbb{B}(\mathbf{x}^0, B)$ for all $1 \leq k \leq K_0$, so $\bar{\mathbf{x}} \in \mathbb{B}(\mathbf{x}^0, B)$. Thus by Lemma E.8 and as $\mathbf{x}^0 \in \mathcal{L}_{F,F}(\mathbf{w}_0)$,

$$\lambda_{\min}(\nabla^2 F(\bar{\mathbf{x}})) \geq \lambda_{\min}(\nabla^2 F(\mathbf{x}^0)) - L_2(\mathbf{w}_0) \|\bar{\mathbf{x}} - \mathbf{x}^0\| \geq -\delta_2 - L_2(\mathbf{w}_0)B \geq -17\delta,$$

where the final inequality follows from our choice of parameters. That is, with probability $1 - \frac{p}{3}$, if \mathbf{x}^k does not move out of the ball $\mathbb{B}(\mathbf{x}^0, B)$ in K_0 iterations, then $\lambda_{\min}(\nabla^2 F(\bar{\mathbf{x}})) \geq -17\delta$.

- To complete the proof and show $\bar{\mathbf{x}}$ is a SOSp, we will show that $\|\nabla F(\bar{\mathbf{x}})\|$ is small. To this end, we upper bound $\frac{1}{K_0} \left\| \sum_{k=1}^{K_0} \boldsymbol{\xi}^k \right\|$ using concentration. In deriving this bound we do *not* yet suppose that \mathbf{x}^k does not move out of $\mathbb{B}(\mathbf{x}^0, B)$ in its first K_0 iterations. Consider

$$\left\| \sum_{k=1}^{K_0} \boldsymbol{\xi}^k 1_{k \leq \mathcal{K}_0} \right\| = \left\| \sum_{k=1}^{K_0} \boldsymbol{\xi}^k 1_{k-1 < \mathcal{K}_0} \right\|.$$

As $1_{k-1 < \mathcal{K}_0}$ is \mathfrak{F}^{k-1} -measurable,

$$\mathbb{E}[\boldsymbol{\xi}^k 1_{k \leq \mathcal{K}_0} | \mathfrak{F}^{k-1}] = \mathbf{0}.$$

Furthermore by Lemma E.18, for $k \leq \mathcal{K}_0$ we have

$$\|\boldsymbol{\xi}^k 1_{k \leq \mathcal{K}_0}\| \leq \sigma_1(\mathbf{w}_0).$$

Thus the Vector-Martingale Concentration Inequality Theorem C.1 gives with probability at least $1 - 2p/3$ that

$$\frac{1}{K_0} \left\| \sum_{k=1}^{K_0} \boldsymbol{\xi}^k 1_{k \leq \mathcal{K}_0} \right\| \leq \frac{2\sigma_1(\mathbf{w}_0) \sqrt{K_0 \log(6/p)}}{K_0} \leq L_2(\mathbf{w}_0) B^2. \quad (93)$$

The last inequality follows from our choice of parameters.

Now conditioning on the above event implying (93) which occurs with probability at least $1 - 2p/3$, suppose \mathbf{x}^k does not move out of the ball $\mathbb{B}(\mathbf{x}^0, B)$ in K_0 iterations. Then we have $\mathcal{K}_0 > K_0$, and so from (93), we have

$$\frac{1}{K_0} \left\| \sum_{k=1}^{K_0} \boldsymbol{\xi}^k \right\| = \frac{1}{K_0} \left\| \sum_{k=1}^{K_0} \boldsymbol{\xi}^k 1_{k \leq \mathcal{K}_0} \right\| \leq L_2(\mathbf{w}_0) B^2.$$

Furthermore, if \mathbf{x}^k does not move out of the ball $\mathbb{B}(\mathbf{x}^0, B)$ in K_0 iterations, then we have $\bar{\mathbf{x}} \in \mathbb{B}(\mathbf{x}^0, B)$. We find an upper bound $\|\nabla F(\bar{\mathbf{x}})\|^2$. We again consider the quadratic approximation $G(\mathbf{x})$ at \mathbf{x}^0 defined in Subsection E.5, and follow the notation from there. Noting $G(\cdot)$ is a quadratic and so its gradient is a linear map, we obtain

$$\begin{aligned} \|G(\bar{\mathbf{x}})\| &= \left\| \frac{1}{K_0} \sum_{k=0}^{K_0-1} \nabla G(\mathbf{x}^k) \right\| \\ &\leq \left\| \frac{1}{K_0} \sum_{k=0}^{K_0-1} \nabla F(\mathbf{x}^k) \right\| + \left\| \frac{1}{K_0} \sum_{k=0}^{K_0-1} \nabla G(\mathbf{x}^k) - \nabla F(\mathbf{x}^k) \right\| \\ &= \frac{1}{K_0 \eta} \left\| \mathbf{x}^{K_0-1} - \mathbf{x}^0 - \eta \sum_{k=1}^{K_0} \boldsymbol{\xi}^k \right\| + \left\| \frac{1}{K_0} \sum_{k=0}^{K_0-1} \nabla G(\mathbf{x}^k) - \nabla F(\mathbf{x}^k) \right\| \\ &\leq \frac{B}{K_0 \eta} + \frac{1}{K_0} \left\| \sum_{k=1}^{K_0} \boldsymbol{\xi}^k \right\| + \frac{1}{K_0} \cdot K_0 \cdot \frac{L_2(\mathbf{w}_0) B^2}{2} \\ &\leq \left(\frac{16}{\tilde{C}_1} + \frac{1}{2} \right) L_2(\mathbf{w}_0) B^2 + \frac{1}{K_0} \left\| \sum_{k=1}^{K_0} \boldsymbol{\xi}^k \right\|. \end{aligned}$$

Here we used the choice of parameters, that $\mathbf{x}^k \in \mathbb{B}(\mathbf{x}^0, B)$ for all $0 \leq k \leq K_0$ combined with [Lemma E.17](#) and that $\mathbf{x}^0 \in \mathcal{L}_{F,F}(\mathbf{w}_0)$, and Triangle Inequality repeatedly.

Note because $\mathbf{x}^0 \in \mathcal{L}_{F,F}(\mathbf{w}_0)$ and as $\bar{\mathbf{x}} \in \mathbb{B}(\mathbf{x}^0, B)$, by [Lemma E.17](#), the above implies

$$\|\nabla F(\bar{\mathbf{x}})\| \leq \|\nabla G(\bar{\mathbf{x}})\| + \frac{L_2(\mathbf{w}_0)B^2}{2} \leq 17L_2(\mathbf{w}_0)B^2 + \frac{1}{K_0} \left\| \sum_{k=1}^{K_0} \xi^k \right\| \leq 18L_2(\mathbf{w}_0)B^2.$$

Consequently, with probability at least $1 - 2p/3$, if \mathbf{x}^k does not move out of the ball $\mathbb{B}(\mathbf{x}^0, B)$ within K_0 iterations, then

$$\|\nabla F(\bar{\mathbf{x}})\| \leq 18L_2(\mathbf{w}_0)B^2.$$

Taking a Union Bound, it follows that with probability at least $1 - p$, if \mathbf{x}^k does not escape $\mathbb{B}(\mathbf{x}^0, B)$ within the first K_0 iterations, we have both

$$\|\nabla F(\bar{\mathbf{x}})\| \leq 18L_2(\mathbf{w}_0)B^2, \lambda_{\min}(\nabla^2 F(\bar{\mathbf{x}})) \geq -17\delta.$$

This proves [Lemma E.1](#). □

F Examples

F.1 Phase Retrieval

By [Theorem 3.4](#) and [Theorem 3.5](#), it suffices to show that 1) F_{pr} satisfies [Assumption 1.2](#) and 2) F_{pr} is a strict saddle problem (that is, all SOSP are near-optima in a suitable sense). In the rest of this subsection, denote F_{pr} by F for short. As shown in [Candes et al. \(2015\)](#); [De Sa et al. \(2022\)](#), Section 2.3 and Lemma 16 part a respectively, direct calculation shows $F(\mathbf{w})$ takes the form

$$F(\mathbf{w}) = \mathbf{w}^\top (\mathbf{I} - (\mathbf{w}^*)(\mathbf{w}^*)^\top) \mathbf{w} + \frac{3}{4}(\|\mathbf{w}\|^2 - 1)^2. \quad (94)$$

As $\|\mathbf{w}^*\| = 1$, we have $F(\mathbf{w}) \geq 0$. Furthermore, we have $\inf_{\mathbf{w} \in \mathbb{R}^d} F(\mathbf{w}) = 0$, attained for example at $\mathbf{w} = \pm \mathbf{w}^*$. Also note for any fixed \mathbf{w} , F is absolutely continuous on a compact neighborhood of \mathbf{w} .

F satisfies [Assumption 1.2](#): By [De Sa et al. \(2022\)](#), Lemma 20, we have that

$$\|\nabla^2 F(\mathbf{w})\| \leq \rho_1(F(\mathbf{w}))$$

for $\rho_1(x) = 9\sqrt{x} + 10$. It remains to show that

$$\|\nabla^3 F(\mathbf{w})\| \leq \rho_2(F(\mathbf{w}))$$

for some increasing, non-negative ρ_2 , where $\|\nabla^3 F(\mathbf{w})\|$ refers to operator norm of the third order tensor. Equivalently, we will show that for any \mathbf{w} and any unit vector \mathbf{u} , we have

$$\lim_{\delta \rightarrow 0} \frac{\|\nabla^2 F(\mathbf{w} + \delta \mathbf{u}) - \nabla^2 F(\mathbf{w})\|_{\text{op}}}{\delta \|\mathbf{u}\|} \leq \rho_2(F(\mathbf{w})).$$

As shown in the proof of Lemma 20, [De Sa et al. \(2022\)](#), we obtain from direct calculation that

$$\nabla^2 F(\mathbf{w}) = 2\mathbf{I} - 2(\mathbf{w}^*)(\mathbf{w}^*)^\top + 3(\|\mathbf{w}\|^2 - 1)\mathbf{I} + 6\mathbf{w}\mathbf{w}^\top. \quad (95)$$

Thus, by repeatedly applying Triangle Inequality and [Lemma A.3](#) and as $\|\mathbf{u}\| = 1$,

$$\begin{aligned} & \|\nabla^2 F(\mathbf{w} + \delta \mathbf{u}) - \nabla^2 F(\mathbf{w})\|_{\text{op}} \\ &= \left\| 3(\|\mathbf{w} + \delta \mathbf{u}\|^2 - \|\mathbf{w}\|^2)\mathbf{I} + 6(\mathbf{w} + \delta \mathbf{u})(\mathbf{w} + \delta \mathbf{u})^\top - 6\mathbf{w}\mathbf{w}^\top \right\|_{\text{op}} \\ &\leq 3\|\mathbf{w} + \delta \mathbf{u}\| - \|\mathbf{w}\| \cdot (\|\mathbf{w} + \delta \mathbf{u}\| + \|\mathbf{w}\|) \\ &\quad + 6\|(\mathbf{w} + \delta \mathbf{u})(\mathbf{w} + \delta \mathbf{u})^\top - \mathbf{w}(\mathbf{w} + \delta \mathbf{u})^\top + \mathbf{w}(\mathbf{w} + \delta \mathbf{u})^\top - \mathbf{w}\mathbf{w}^\top\|_{\text{op}} \\ &\leq 3\delta\|\mathbf{u}\|(2\|\mathbf{w}\| + \delta) + 6\left(\|\delta \mathbf{u}(\mathbf{w} + \delta \mathbf{u})^\top\|_{\text{op}} + \|\mathbf{w}(\delta \mathbf{u})^\top\|_{\text{op}}\right) \end{aligned}$$

$$\begin{aligned} &\leq \delta \|\mathbf{u}\| (3(2\|\mathbf{w}\| + \delta) + 6\|\mathbf{w} + \delta\mathbf{u}\| + 6\|\mathbf{w}\|) \\ &\leq \delta \|\mathbf{u}\| (18\|\mathbf{w}\| + 9\delta). \end{aligned}$$

Here, we used the inequality $\|\mathbf{x} + \mathbf{y}\| - \|\mathbf{x}\| \leq \|\mathbf{y}\|$.

Consequently,

$$\lim_{\delta \rightarrow 0} \frac{\|\nabla^2 F(\mathbf{w} + \delta\mathbf{u}) - \nabla^2 F(\mathbf{w})\|_{\text{op}}}{\delta \|\mathbf{u}\|} \leq \lim_{\delta \rightarrow 0} 18\|\mathbf{w}\| + 9\delta \leq 18\|\mathbf{w}\| + 1.$$

By Lemma 16 part d, [De Sa et al. \(2022\)](#), using Jensen's Inequality we have

$$F(\mathbf{w}) \geq (\|\mathbf{w}\|^2 - 1)^2.$$

Note for $\|\mathbf{w}\| \geq 2$, this implies

$$18\|\mathbf{w}\| + 1 \leq 18(\|\mathbf{w}\| + 1)^2 (\|\mathbf{w}\| - 1)^2 \leq 18F(\mathbf{w}).$$

Combining with the case $\|\mathbf{w}\| < 2$, we obtain

$$\lim_{\delta \rightarrow 0} \frac{\|\nabla^2 F(\mathbf{w} + \delta\mathbf{u}) - \nabla^2 F(\mathbf{w})\|_{\text{op}}}{\delta \|\mathbf{u}\|} \leq 18\|\mathbf{w}\| + 1 \leq 18F(\mathbf{w}) + 37,$$

so we can just take $\rho_2(x) = 18x + 37$.

Next, we check that F is a strict saddle problem: We check this here. Similar results, in slightly different of a setting where we solve phase retrieval from samples from data, are shown in [Sun et al. \(2018\)](#).

Suppose $\|\nabla F(\mathbf{w})\| \leq \delta$ for $\delta \leq (\frac{1}{20})^4$. Note by Lemma 16 part b, [De Sa et al. \(2022\)](#), $\langle \mathbf{w}^*, \nabla F(\mathbf{w}) \rangle = 3(\|\mathbf{w}\|^2 - 1)\langle \mathbf{w}, \mathbf{w}^* \rangle$. By Cauchy-Schwartz and recalling \mathbf{w}^* is a unit vector, this gives

$$\delta \geq \|\mathbf{w}^*\| \|\nabla F(\mathbf{w})\| \geq |\langle \mathbf{w}^*, \nabla F(\mathbf{w}) \rangle| = 3|\|\mathbf{w}\|^2 - 1| \cdot |\langle \mathbf{w}, \mathbf{w}^* \rangle|. \quad (96)$$

- Suppose $|\langle \mathbf{w}, \mathbf{w}^* \rangle| \geq \sqrt{\delta}$. Combining this with (96) gives

$$|\|\mathbf{w}\|^2 - 1| \leq \frac{\sqrt{\delta}}{3}.$$

By Lemma 16 part c, [De Sa et al. \(2022\)](#),

$$\begin{aligned} \|\nabla F(\mathbf{w})\|^2 &= 12\|\mathbf{w}\|^2 F(\mathbf{w}) - 8(\|\mathbf{w}\|^2 - \langle \mathbf{w}, \mathbf{w}^* \rangle^2) \\ &= (12\|\mathbf{w}\|^2 - 8)F(\mathbf{w}) + 6(\|\mathbf{w}\|^2 - 1)^2, \end{aligned}$$

where the last equality follows from the explicit form $F(\mathbf{w})$ from (94). Thus using $|\|\mathbf{w}\|^2 - 1| \leq \frac{\sqrt{\delta}}{3}$, we obtain

$$\delta^2 \geq \|\nabla F(\mathbf{w})\|^2 = (12\|\mathbf{w}\|^2 - 8)F(\mathbf{w}) + 6(\|\mathbf{w}\|^2 - 1)^2 \geq (4 - 4\sqrt{\delta})F(\mathbf{w}).$$

For $\delta \leq \frac{1}{4}$, this gives

$$F(\mathbf{w}) \leq \frac{\delta^2}{4 - 4\sqrt{\delta}} \leq \frac{\delta^2}{2}.$$

- Otherwise, suppose $|\langle \mathbf{w}, \mathbf{w}^* \rangle| \leq \sqrt{\delta}$. Note by differentiating (94), as shown in the proof of Lemma 16 part b, [De Sa et al. \(2022\)](#),

$$\nabla F(\mathbf{w}) = 2\mathbf{w} - 2\langle \mathbf{w}, \mathbf{w}^* \rangle \mathbf{w}^* + 3(\|\mathbf{w}\|^2 - 1)\mathbf{w} = -2\langle \mathbf{w}, \mathbf{w}^* \rangle \mathbf{w}^* + (3\|\mathbf{w}\|^2 - 1)\mathbf{w}.$$

Thus by Triangle Inequality,

$$|3\|\mathbf{w}\|^2 - 1| \cdot \|\mathbf{w}\| \leq \|\nabla F(\mathbf{w})\| + 2|\langle \mathbf{w}, \mathbf{w}^* \rangle| \|\mathbf{w}^*\| \leq \delta + 2\sqrt{\delta} \leq 4\sqrt{\delta}.$$

Consequently either $\|\mathbf{w}\| \leq 2\delta^{1/4}$ or $|3\|\mathbf{w}\|^2 - 1| \leq 2\delta^{1/4}$.

In the first case, by Cauchy Schwartz and (95), notice for any unit vector \mathbf{u} that

$$\begin{aligned}\mathbf{u}^\top \nabla^2 F(\mathbf{w}) \mathbf{u} &= \mathbf{u}^\top \left(2\mathbf{I} - 2(\mathbf{w}^*)(\mathbf{w}^*)^\top + 3(\|\mathbf{w}\|^2 - 1)\mathbf{I} + 6\mathbf{w}\mathbf{w}^\top \right) \mathbf{u} \\ &\leq -\|\mathbf{u}\|^2 + 3\|\mathbf{u}\|^2 \cdot (2\delta^{1/4})^2 + 6\|\mathbf{u}\|^2 \cdot (2\delta^{1/4})^2 \\ &\leq -1 + 36\delta^{1/2} \leq -\frac{9}{10},\end{aligned}$$

since $\delta \leq (\frac{1}{20})^4$.

In the second case, using (95), notice as $\|\mathbf{w}^*\| = 1$, we have

$$\begin{aligned}\mathbf{w}^{*\top} \nabla^2 F(\mathbf{w}) \mathbf{w}^* &= \mathbf{w}^{*\top} (3\|\mathbf{w}\|^2 - 1)\mathbf{w}^* - 2\|\mathbf{w}^*\|^2 + 6|\langle \mathbf{w}, \mathbf{w}^* \rangle|^2 \\ &\leq 2\delta^{1/4} - 2 + 6\delta \leq -\frac{9}{5}.\end{aligned}$$

Consequently in either case, $\nabla^2 F(\mathbf{w})$ has at least one negative eigenvalue with value at most $-\frac{9}{10}$.

Consider ε smaller than a universal constant, and take $\delta = \sqrt{\varepsilon}$ in the above result. It follows from the analysis here that if we find an SOSF to tolerance ε as per the definition (2), we obtain \mathbf{w} with $F(\mathbf{w}) \leq \frac{\varepsilon}{2}$.

Thus, it follows that running Perturbed GD or Restarted SGD as described in Theorem 3.4 or Theorem 3.5 respectively, we will obtain \mathbf{w} with suboptimality $F(\mathbf{w}) \leq \varepsilon$, where the number of oracle calls depends on $1/\varepsilon, d, F(\mathbf{w}_0)$ in the same way as in Theorem 3.4 or Theorem 3.5 respectively.

F.2 Matrix PCA

Again by Theorem 3.4, Theorem 3.5, it suffices to show that 1) F_{pca} satisfies Assumption 1.2 and 2) is a strict saddle problem (that is, all SOSFs are near-optima in a suitable sense). We will show this, with the parameters governing the strict saddle property depending on the spectral gap $\lambda_1(\mathbf{M}) - \lambda_2(\mathbf{M})$.¹² In the rest of this subsection, denote F_{pca} by F for short. Recall the loss function for PCA takes the form

$$F(\mathbf{w}) = \frac{1}{2} \|\mathbf{w}\mathbf{w}^\top - \mathbf{M}\|_F^2,$$

where \mathbf{M} is a symmetric PD matrix. Note for any fixed \mathbf{w} , F is absolutely continuous on a compact neighborhood of \mathbf{w} . Note $F(\mathbf{w}) \geq 0$ always holds. While it is not true that $\inf_{\mathbf{w} \in \mathbb{R}^d} F(\mathbf{w}) = 0$, to enforce this, we can consider the shifted function $G := F - \inf_{\mathbf{w} \in \mathbb{R}^d} F(\mathbf{w})$. The derivatives of G are identical to those of F , and furthermore $G(\mathbf{x}) - G(\mathbf{y}) = F(\mathbf{x}) - F(\mathbf{y})$ for all \mathbf{x}, \mathbf{y} . Thus to apply Theorem 3.4, Theorem 3.5 and show that Perturbed GD or Restarted SGD can globally optimize G and therefore F by finding SOSFs, it remains to show F satisfies Assumption 1.2 and is strict saddle.

F satisfies Assumption 1.2: Direct calculation, also in Jin et al. (2021a), yields

$$\nabla F(\mathbf{w}) = (\mathbf{w}\mathbf{w}^\top - \mathbf{M})\mathbf{w}, \nabla^2 F(\mathbf{w}) = \|\mathbf{w}\|^2 \mathbf{I} + 2\mathbf{w}\mathbf{w}^\top - \mathbf{M}. \quad (97)$$

We now check self-bounding regularity for the Hessian and third order derivative tensor. First observe

$$\mathbf{w}^\top (\mathbf{w}\mathbf{w}^\top) \mathbf{w} = \|\mathbf{w}\|^4.$$

Combining with Lemma A.3, we obtain

$$\begin{aligned}\|\mathbf{w}\| &= \|\mathbf{w}\mathbf{w}^\top\|_{\text{op}}^{1/2} \\ &\leq \left(\|\mathbf{w}\mathbf{w}^\top - \mathbf{M}\|_{\text{op}} + \|\mathbf{M}\|_{\text{op}} \right)^{1/2} \\ &\leq \|\mathbf{w}\mathbf{w}^\top - \mathbf{M}\|_F^{1/2} + \|\mathbf{M}\|_{\text{op}}^{1/2} \\ &\leq 2F(\mathbf{w})^{1/4} + \|\mathbf{M}\|_{\text{op}}^{1/2}.\end{aligned} \quad (98)$$

¹²Thus our result will be vacuous when the spectral gap is 0.

Now we check the self bounding conditions. For the Hessian, note from (97) and (98) and using Lemma A.3,

$$\|\nabla^2 F(\mathbf{w})\|_{\text{op}} \leq 3\|\mathbf{w}\|^2 + \|\mathbf{M}\|_{\text{op}} \leq 3(2F(\mathbf{w})^{1/4} + \|\mathbf{M}\|_{\text{op}}^{1/2})^2 + \|\mathbf{M}\|_{\text{op}}.$$

Thus we can take $\rho_1(x) = 3(2x^{1/4} + \|\mathbf{M}\|_{\text{op}}^{1/2})^2 + \|\mathbf{M}\|_{\text{op}}$.

For the third order derivative tensor, following the strategy in Subsection F.1, we will show that for any \mathbf{w} and any unit vector \mathbf{u} , we have

$$\lim_{\delta \rightarrow 0} \frac{\|\nabla^2 F(\mathbf{w} + \delta \mathbf{u}) - \nabla^2 F(\mathbf{w})\|_{\text{op}}}{\delta \|\mathbf{u}\|} \leq \rho_3(F(\mathbf{w})).$$

Applying (97) and Lemma A.3 and note

$$\begin{aligned} (\mathbf{w} + \delta \mathbf{u})(\mathbf{w} + \delta \mathbf{u})^\top - \mathbf{w}\mathbf{w}^\top &= (\mathbf{w} + \delta \mathbf{u})(\mathbf{w} + \delta \mathbf{u})^\top - (\mathbf{w} + \delta \mathbf{u})\mathbf{w}^\top + (\mathbf{w} + \delta \mathbf{u})\mathbf{w}^\top - \mathbf{w}\mathbf{w}^\top \\ &= (\mathbf{w} + \delta \mathbf{u})(\delta \mathbf{u})^\top + \delta \mathbf{u}\mathbf{w}^\top. \end{aligned}$$

This gives

$$\begin{aligned} &\lim_{\delta \rightarrow 0} \frac{\|\nabla^2 F(\mathbf{w} + \delta \mathbf{u}) - \nabla^2 F(\mathbf{w})\|_{\text{op}}}{\delta \|\mathbf{u}\|} \\ &= \lim_{\delta \rightarrow 0} \frac{(\|\mathbf{w} + \delta \mathbf{u}\|^2 - \|\mathbf{w}\|^2) + 2\|(\mathbf{w} + \delta \mathbf{u})(\mathbf{w} + \delta \mathbf{u})^\top - \mathbf{w}\mathbf{w}^\top\|_{\text{op}}}{\delta \|\mathbf{u}\|} \\ &\leq \lim_{\delta \rightarrow 0} \frac{\|\mathbf{w} + \delta \mathbf{u}\| - \|\mathbf{w}\| \cdot (2\|\mathbf{w}\| + \delta \|\mathbf{u}\|) + \delta \|\mathbf{u}\|(2\|\mathbf{w}\| + \delta \|\mathbf{u}\|)}{\delta \|\mathbf{u}\|} \\ &\leq \lim_{\delta \rightarrow 0} \frac{\delta \|\mathbf{u}\|(2\|\mathbf{w}\| + \delta \|\mathbf{u}\|) + \delta \|\mathbf{u}\|(2\|\mathbf{w}\| + \delta \|\mathbf{u}\|)}{\delta \|\mathbf{u}\|} \\ &= \lim_{\delta \rightarrow 0} 4\|\mathbf{w}\| + 2\delta \|\mathbf{u}\| \\ &= 4\|\mathbf{w}\| \\ &\leq 8F(\mathbf{w})^{1/4} + 4\|\mathbf{M}\|_{\text{op}}^{1/2}. \end{aligned}$$

Here we used the inequality $\|\mathbf{x} + \mathbf{y}\| - \|\mathbf{x}\| \leq \|\mathbf{y}\|$. The last step used (98). Thus we can take $\rho_2(x) = 8x^{1/4} + 4\|\mathbf{M}\|_{\text{op}}^{1/2}$.

Next, we check F is a strict saddle problem: We check this here. A similar verification is done in Ge et al. (2017).

Let $\mathbf{v}_1, \dots, \mathbf{v}_d$ be the (unit) eigenvectors of \mathbf{M} corresponding to $\lambda_1(\mathbf{M}) \geq \lambda_2(\mathbf{M}) \geq \dots \geq \lambda_d(\mathbf{M}) > 0$ respectively (recall \mathbf{M} is assumed to be PD). Thus the \mathbf{v}_i form an orthonormal basis of \mathbb{R}^d . Furthermore for convenience let $\lambda_i := \lambda_i(\mathbf{M})$ for all $1 \leq i \leq d$. As \mathbf{M} is symmetric and PD, by the Spectral Theorem, we can write

$$\mathbf{M} = \sum_{i=1}^d \lambda_i \mathbf{v}_i \mathbf{v}_i^\top.$$

Suppose \mathbf{w} is a SOSF to tolerance ε for $\varepsilon < \min\left\{1, \frac{(\lambda_1 - \lambda_2)^2}{16}, \frac{3}{8}(\lambda_1 - \lambda_2)^{5/2}\right\}$. Note the minimizers of F are $\mathbf{w} = \pm \sqrt{\lambda_1} \mathbf{v}_1$. We will show that \mathbf{w} is close to these minimizers: in particular, that $\min\left\{\|\mathbf{w} - \sqrt{\lambda_1} \mathbf{v}_1\|^2, \|\mathbf{w} + \sqrt{\lambda_1} \mathbf{v}_1\|^2\right\} \leq \varepsilon$.

Write $\mathbf{w} = c_1 \mathbf{v}_1 + \dots + c_d \mathbf{v}_d$. Thus, our goal is to show that $|(c_1^2 + \dots + c_d^2) - \lambda_1| < \sqrt{\varepsilon}$. By (97), we have

$$\varepsilon \geq \|\nabla F(\mathbf{w})\| = \|\mathbf{M}\mathbf{w} - \|\mathbf{w}\|^2 \mathbf{w}\| = \left\| \sum_{i=1}^d ((c_1^2 + \dots + c_d^2) - \lambda_i) c_i \mathbf{v}_i \right\|.$$

That is, we have

$$\sum_{i=1}^d c_i^2 ((c_1^2 + \dots + c_d^2) - \lambda_i)^2 \leq \varepsilon^2. \quad (99)$$

Furthermore by (97), we have

$$\nabla^2 F(\mathbf{w}) = (c_1^2 + \dots + c_d^2)\mathbf{I} + 2 \sum_{i,j} c_i c_j \mathbf{v}_i \mathbf{v}_j^\top - \sum_{i=1}^d \lambda_i \mathbf{v}_i \mathbf{v}_i^\top.$$

Since \mathbf{w} is a SOSp, for all $\mathbf{v}_k, 1 \leq k \leq d$, we have

$$-\sqrt{\varepsilon} \leq \mathbf{v}_k^\top \nabla^2 F(\mathbf{w}) \mathbf{v}_k = (c_1^2 + \dots + c_d^2) + 2c_k^2 - \lambda_k. \quad (100)$$

We now break into cases:

- Suppose for all i , we have $|(c_1^2 + \dots + c_d^2) - \lambda_i| \geq \sqrt{\varepsilon}$. From (99), this gives $\sum_{i=1}^d c_i^2 \leq \varepsilon$. Taking $k = 1$ in (100), we obtain

$$-\sqrt{\varepsilon} \leq 3 \sum_{i=1}^d c_i^2 - \lambda_1 \leq 3\varepsilon - \lambda_1 \implies \lambda_1 \leq \sqrt{\varepsilon} + 3\varepsilon,$$

contradicting that $\varepsilon < \min\left\{1, \frac{(\lambda_1 - \lambda_2)^2}{16}\right\}$.

- Else, suppose there exists i such that $|(c_1^2 + \dots + c_d^2) - \lambda_i| < \sqrt{\varepsilon}$. Suppose that $i \geq 2$. Then taking $k = 1$ in (100), we obtain

$$-\sqrt{\varepsilon} \leq \lambda_i + \sqrt{\varepsilon} + 2c_1^2 - \lambda_1 \implies c_1^2 \geq \frac{\lambda_1 - \lambda_i}{2} - \sqrt{\varepsilon} \geq \frac{\lambda_1 - \lambda_2}{4},$$

where the last inequality uses $\lambda_i \leq \lambda_2$ and $\varepsilon < \left(\frac{\lambda_1 - \lambda_2}{4}\right)^2$.

Note furthermore that as $\varepsilon \leq \left(\frac{\lambda_1 - \lambda_2}{4}\right)^2$, as $|(c_1^2 + \dots + c_d^2) - \lambda_i| < \sqrt{\varepsilon}$, and as $\lambda_i \leq \lambda_2 < \lambda_1$, we have $|(c_1^2 + \dots + c_d^2) - \lambda_1| > \frac{3(\lambda_1 - \lambda_2)}{4}$. Thus (99) implies

$$\varepsilon^2 > 0 + \frac{\lambda_1 - \lambda_2}{4} \cdot \frac{9}{16} (\lambda_1 - \lambda_2)^2,$$

contradicting that $\varepsilon < \frac{3}{8}(\lambda_1 - \lambda_2)^{5/2}$.

Therefore, we must have $i = 1$ in the second case above. That is, $|(c_1^2 + \dots + c_d^2) - \lambda_1| < \sqrt{\varepsilon}$, as desired.

Thus, it follows that running Perturbed GD or Restarted SGD as described in Theorem 3.4 or Theorem 3.5 respectively, we will obtain \mathbf{w} that is distance at most $\sqrt{\varepsilon}$ from a global minimizer of F for $\varepsilon < \min\left\{1, \frac{(\lambda_1 - \lambda_2)^2}{16}, \frac{3}{8}(\lambda_1 - \lambda_2)^{5/2}\right\}$. Here the number of oracle calls depends on $1/\varepsilon, d, F(\mathbf{w}_0)$ the same way as in Theorem 3.4 or Theorem 3.5 respectively. For $\varepsilon \geq \min\left\{1, \frac{(\lambda_1 - \lambda_2)^2}{16}, \frac{3}{8}(\lambda_1 - \lambda_2)^{5/2}\right\}$, we can replace ε by any real strictly smaller than $\min\left\{1, \frac{(\lambda_1 - \lambda_2)^2}{16}, \frac{3}{8}(\lambda_1 - \lambda_2)^{5/2}\right\}$ in the guarantees from Theorem 3.4 or Theorem 3.5.

G Simulations

Our algorithmic results Theorem 3.1, Theorem 3.2, Theorem 3.3, Theorem 3.4, and Theorem 3.5 have strong practical implications. They directly suggest that under generalized smoothness, the step sizes η that lead to convergence/successful optimization become smaller for larger initialization $F(\mathbf{w}_0)$ and larger self-bounding functions $\rho_1(\cdot), \rho_2(\cdot)$. For example in Theorem 3.1, we set $\eta = \frac{1}{L_1(\mathbf{w}_0)}$ where $L_1(\mathbf{w}_0) = \max\{1, \rho_0(F(\mathbf{w}_0) + 1), \rho_0(F(\mathbf{w}_0))\rho_0(F(\mathbf{w}_0) + 1), \rho_1(F(\mathbf{w}_0) + 1)\}$ was defined in (4).

That is, our work suggests that larger suboptimality at initialization and larger self-bounding functions shrink the ‘window’ for choosing a working η in practice, when the loss function satisfies generalized smoothness. This has strong practical implications: it implies that for losses with non-Lipschitz gradient/Hessian, one should tune η based on suboptimality at initialization. This contrasts sharply with the Lipschitz gradient/Hessian case, see e.g. (Bubeck et al., 2015; Jin et al., 2017; Fang et al., 2019), where the range of working η is fixed in terms of the Lipschitz constant of the gradient and/or Hessian, and does not depend on the initialization.

In this section, we empirically validate this implication of our work.

G.1 Synthetic Simulations with GD

Simulation Details: We consider $F(\mathbf{w}) = \|\mathbf{A}\mathbf{w}\|^p$ for $p = 2, 3, 4, 5, 6$, where $\mathbf{A} = \text{diag}(\frac{1}{20}, \frac{1}{19}, \dots, \frac{1}{2}, 1)$. When $p = 2$, $F(\mathbf{w})$ is smooth. When $p \geq 3$, $F(\mathbf{w})$ is not smooth, but it is straightforward to verify that it satisfies [Assumption 1.1](#), similar to our verifications in [Subsection A.2](#). One can furthermore verify that as p increases, the corresponding self-bounding function $\rho_1(\cdot)$ from [Assumption 1.1](#) increase. This choice of generalized smooth function was motivated by [Gaash et al. \(2025\)](#), who used $\|\mathbf{A}\mathbf{w}\|^4$ with the exact same \mathbf{A} in their experiments to study optimization with first-order methods under generalized smoothness.

For each $p = 2, 3, 4, 5, 6$, we consider the following settings for GD:

- Step sizes: We consider 30 step sizes $\{\eta_i\}_{i=1}^{30}$, $\eta_1 < \dots < \eta_{30}$ evenly spaced on a log scale between 10^{-8} and 10^1 , inclusive.
- Initialization: For each step size η_i , we initialize GD at 4 distributions $\pi_j = \mathcal{N}(\vec{\mathbf{0}}, c_j \mathbf{I}_{20})$ for $c_j \in \{2.5, 5, 7.5, 10\}$. For each of these 4 distributions π_j , we draw 100 points $\mathbf{w}_0 \sim \pi_j$ to use as our initialization.
- Number of steps: For each η_i and each $\mathbf{w}_0 \sim \pi_j$, we run GD initialized at \mathbf{w}_0 with step size η_i for $T = 1000$ iterations. Here as F is known, we analytically compute the gradient.

For each p and initialization π_j , we consider all 30 possible η_i , which we plot on the x -axis. For each η_i , we consider all 100 initializations $\mathbf{w}_0 \sim \pi_j$. For each initialization \mathbf{w}_0 , letting $\{\mathbf{w}_t\}$ be the resulting sequence of iterates of GD, we compute $\frac{\|\nabla F(\mathbf{w}_T)\|}{F(\mathbf{w}_0)}$ for $T = 1000$. For η_i that led to faithful convergence of GD, on the y -axis, we then plot the mean of $\frac{\|\nabla F(\mathbf{w}_T)\|}{F(\mathbf{w}_0)}$ over those 100 initializations as a blue dot, with blue vertical error bars indicating ± 2 standard deviations. We considered the ratio $\frac{\|\nabla F(\mathbf{w}_T)\|}{F(\mathbf{w}_0)}$ because for L -smooth functions, established optimization theory predicts that this converges at a rate independent of $F(\mathbf{w}_0)$ and only depending on T and L ([Bubeck et al., 2015](#)).

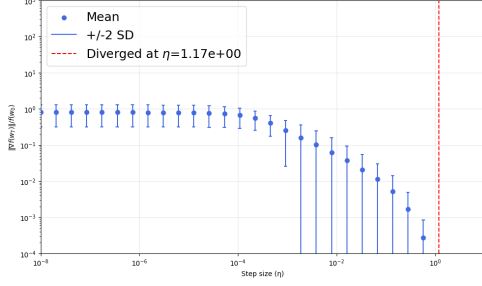
The simulations for [Subsection G.1](#) were run on a Jupyter notebook in Python in Google Colab Pro, connected to a single NVIDIA T4 GPU. Our code can be found in the attached files.

Divergence of GD and working step sizes: We observe that for some η_i larger than some threshold depending on p and π_j , the iterates of GD diverge. In particular, the resulting ratio $\frac{\|\nabla F(\mathbf{w}_T)\|}{F(\mathbf{w}_0)}$ becomes massive, often on the order of 10^5 or more, indicating that η_i was too large for GD to converge. To identify the smallest η_i where this first occurs, or equivalently find the largest working step size among $\{\eta_i\}_{i=1}^{30}$, for a given π_j and η_i , we computed the average $\frac{\|\nabla F(\mathbf{w}_T)\|}{F(\mathbf{w}_0)}$ over the 100 initializations. If this average was 100 or more times larger than this average for η_{i-1} , we took this as an indication that the iterates of GD with this step size η_i or larger step sizes diverge, and for this p and π_j , we stopped considering any larger $\eta_{i'}, i' > i$. We then save this η_i to indicate the smallest η_i for which divergence occurred. This η_i is indicated with a red line in the following plots.

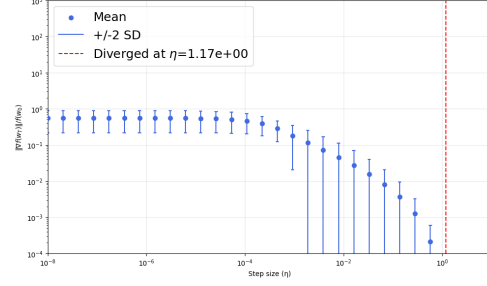
This smallest η_i for which divergence occurred plays a crucial role in validating our theoretical claims. Established optimization theory predicts that for smooth functions (here, when $p = 2$), this η_i is identical across different initializations ([Bubeck et al., 2015](#)). Meanwhile for generalized smooth functions, as per our remarks earlier and from [Subsection 3.6](#), we predict that as $F(\mathbf{w}_0)$ increases, the range of working step sizes, and consequently also the smallest η_i for which divergence occurs, will *decrease*. Note as c_j increases (recall $\pi_j \sim \mathcal{N}(\vec{\mathbf{0}}, c_j \mathbf{I}_{20})$ and $c_j \in \{2.5, 5, 7.5, 10\}$), we expect $F(\mathbf{w}_0)$ to increase, at least on average or with high probability over the 100 initializations $\mathbf{w}_0 \sim \pi_j$.

Results: Our simulations validate this theory very accurately. Note in the following figures that the y -axis is normalized, as we plot $\frac{\|\nabla F(\mathbf{w}_T)\|}{F(\mathbf{w}_0)}$ where $T = 1000$. Thus larger c_j lead to comparable values on the y -axis.

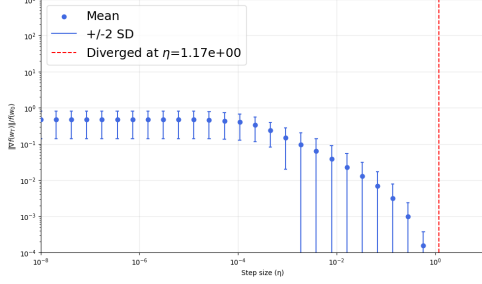
- When $p = 2$: In [Figure 1](#), we plot the results in the manner described above for all 4 initializations π_j . As is predicted by established optimization theory for smooth functions ([Bubeck et al., 2015](#)), the first step size leading to divergence η_i is identical across all the π_j .
- When $p = 3, 4, 5, 6$: We plot the results in the manner described above for all 4 initializations π_j in [Figure 2](#), [Figure 3](#), [Figure 4](#), [Figure 5](#) respectively. Unlike the $p = 2$ case, in all of



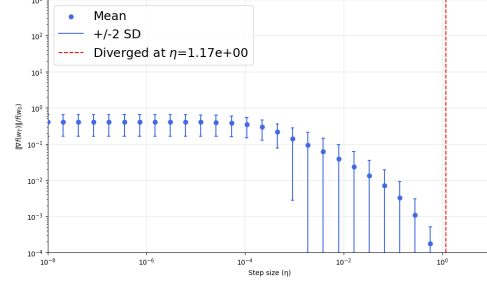
(a) $\pi_j = \mathcal{N}(\vec{0}, 2.5\mathbf{I}_{20})$. The first divergence is at $\eta_i \approx 1.17$.



(b) $\pi_j = \mathcal{N}(\vec{0}, 5.0\mathbf{I}_{20})$. The first divergence is at $\eta_i \approx 1.17$.



(c) $\pi_j = \mathcal{N}(\vec{0}, 7.5\mathbf{I}_{20})$. The first divergence is at $\eta_i \approx 1.17$.



(d) $\pi_j = \mathcal{N}(\vec{0}, 10\mathbf{I}_{20})$. The first divergence is at $\eta_i \approx 1.17$.

Figure 1: GD simulation results for $p = 2$. For all π_j , the smallest η_i leading to divergence is ≈ 1.17 .

these cases, the first step size leading to divergence η_i generally decreases as the covariance $c_j\mathbf{I}_{20}$ of π_j increases from 2.5 to 10.

We also notice the following, both in line with our theoretical claims:

- For a given p , consider how this first step size η_i leading to divergence decreases as the covariance $c_j\mathbf{I}_{20}$ of π_j increases from 2.5 to 10. We find that the rate of this decrease increases as p increases. The ratio of the first η_i leading to divergence for π_1 vs π_4 is approximately 4.18, 4.18, 8.53, 17.43 for $p = 3, 4, 5, 6$ respectively.

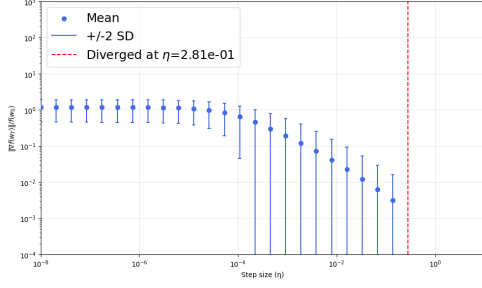
As remarked earlier, for larger p , the corresponding self-bounding function $\rho_1(\cdot)$ is larger for $F(\mathbf{w}) = \|\mathbf{A}\mathbf{w}\|^p$ (see Subsection A.2 for a similar verification). Thus this behavior is consistent with our results, as the step size from all of our results depends on $F(\mathbf{w}_0)$ through $\rho_1(\cdot)$.

- Fixing π_j and comparing across p , we see that the first step size leading to divergence η_i decreases as p increases. Again this is not a surprise considering our theoretical results, as for larger p , both $F(\mathbf{w}_0)$ for $\mathbf{w}_0 \sim \pi_j$ and the self-bounding function $\rho_1(\cdot)$ become larger.

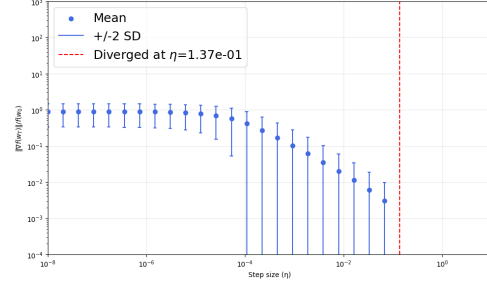
For each $p \in \{2, 3, 4, 5, 6\}$ and π_j , we also record the smallest η_i for which divergence occurred in Table 1 on page 92, which highlights the aforementioned trends.

	$\pi_j = \mathcal{N}(\vec{0}, 2.5\mathbf{I}_{20})$	$\pi_j = \mathcal{N}(\vec{0}, 5.0\mathbf{I}_{20})$	$\pi_j = \mathcal{N}(\vec{0}, 7.5\mathbf{I}_{20})$	$\pi_j = \mathcal{N}(\vec{0}, 10\mathbf{I}_{20})$
$p = 2$	$1.17 \cdot 10^0$	$1.17 \cdot 10^0$	$1.17 \cdot 10^0$	$1.17 \cdot 10^0$
$p = 3$	$2.81 \cdot 10^{-1}$	$1.37 \cdot 10^{-1}$	$1.37 \cdot 10^{-1}$	$6.72 \cdot 10^{-2}$
$p = 4$	$3.29 \cdot 10^{-2}$	$3.29 \cdot 10^{-2}$	$1.61 \cdot 10^{-2}$	$7.88 \cdot 10^{-3}$
$p = 5$	$7.88 \cdot 10^{-3}$	$3.86 \cdot 10^{-3}$	$9.24 \cdot 10^{-4}$	$9.24 \cdot 10^{-4}$
$p = 6$	$9.24 \cdot 10^{-4}$	$4.52 \cdot 10^{-4}$	$5.30 \cdot 10^{-5}$	$5.30 \cdot 10^{-5}$

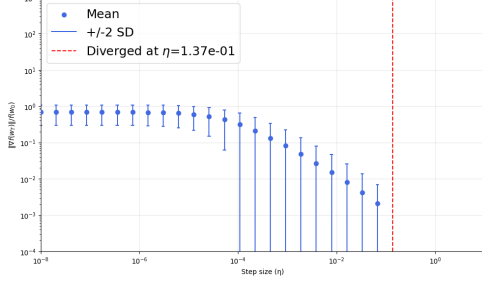
Table 1: The smallest η_i leading to divergence for a given p and initialization π_j .



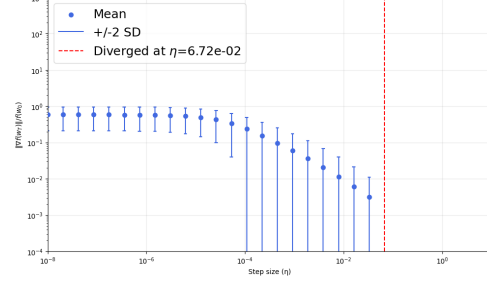
(a) $\pi_j = \mathcal{N}(\vec{0}, 2.5\mathbf{I}_{20})$. The first divergence is at $\eta_i \approx 0.281$.



(b) $\pi_j = \mathcal{N}(\vec{0}, 5.0\mathbf{I}_{20})$. The first divergence is at $\eta_i \approx 0.137$.

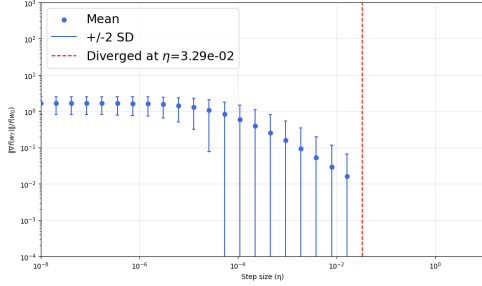


(c) $\pi_j = \mathcal{N}(\vec{0}, 7.5\mathbf{I}_{20})$. The first divergence is at $\eta_i \approx 0.137$.

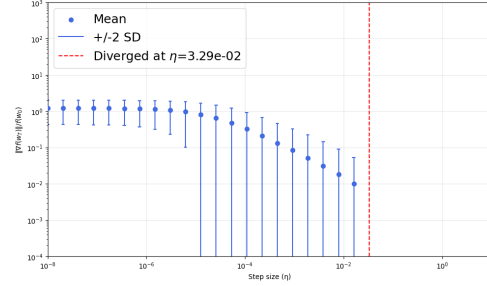


(d) $\pi_j = \mathcal{N}(\vec{0}, 10\mathbf{I}_{20})$. The first divergence is at $\eta_i \approx 0.0672$.

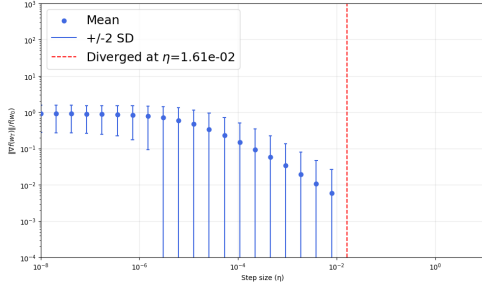
Figure 2: GD simulation results for $p = 3$. For $\pi_j = \mathcal{N}(\vec{0}, 2.5\mathbf{I}_{20})$, the first divergence is at $\eta_i \approx 0.281$. For $\pi_j = \mathcal{N}(\vec{0}, 5\mathbf{I}_{20}), \mathcal{N}(\vec{0}, 7.5\mathbf{I}_{20})$, the first divergence is at $\eta_i \approx 0.137$. For $\pi_j = \mathcal{N}(\vec{0}, 10\mathbf{I}_{20})$, the first divergence is at $\eta_i \approx 0.0672$.



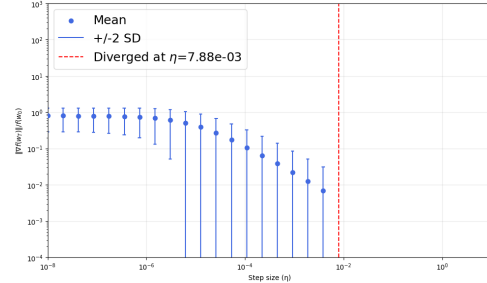
(a) $\pi_j = \mathcal{N}(\vec{0}, 2.5\mathbf{I}_{20})$. The first divergence is at $\eta_i \approx 3.29 \cdot 10^{-2}$.



(b) $\pi_j = \mathcal{N}(\vec{0}, 5.0\mathbf{I}_{20})$. The first divergence is at $\eta_i \approx 3.29 \cdot 10^{-2}$.

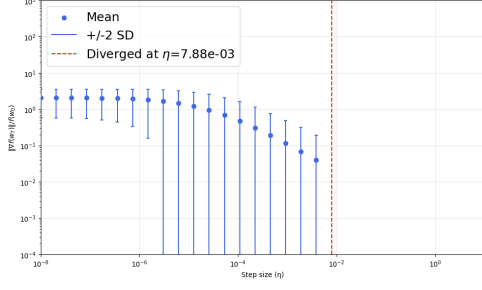


(c) $\pi_j = \mathcal{N}(\vec{0}, 7.5\mathbf{I}_{20})$. The first divergence is at $\eta_i \approx 1.61 \cdot 10^{-2}$.

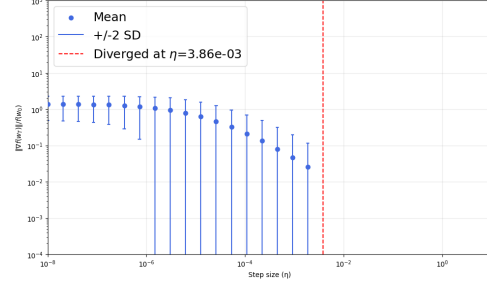


(d) $\pi_j = \mathcal{N}(\vec{0}, 10\mathbf{I}_{20})$. The first divergence is at $\eta_i \approx 7.88 \cdot 10^{-3}$.

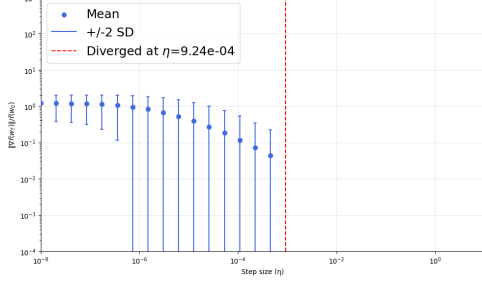
Figure 3: GD simulation results for $p = 4$. For $\pi_j = \mathcal{N}(\vec{0}, 2.5\mathbf{I}_{20}), \mathcal{N}(\vec{0}, 5\mathbf{I}_{20})$, the first divergence is at $\eta_i \approx 3.29 \cdot 10^{-2}$. For $\pi_j = \mathcal{N}(\vec{0}, 7.5\mathbf{I}_{20})$, the first divergence is at $\eta_i \approx 1.61 \cdot 10^{-2}$. For $\pi_j = \mathcal{N}(\vec{0}, 10\mathbf{I}_{20})$, the first divergence is at $\eta_i \approx 7.88 \cdot 10^{-3}$.



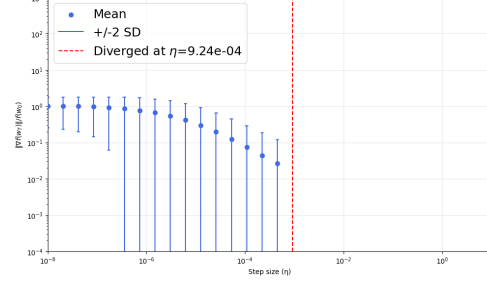
(a) $\pi_j = \mathcal{N}(\vec{0}, 2.5\mathbf{I}_{20})$. The first divergence is at $\eta_i \approx 7.88 \cdot 10^{-3}$.



(b) $\pi_j = \mathcal{N}(\vec{0}, 5.0\mathbf{I}_{20})$. The first divergence is at $\eta_i \approx 3.86 \cdot 10^{-3}$.

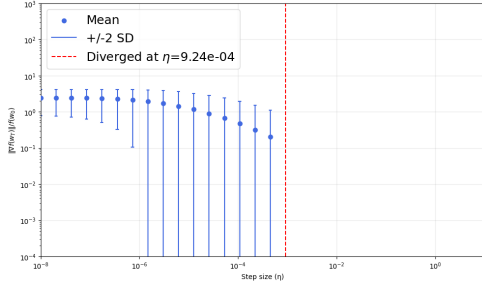


(c) $\pi_j = \mathcal{N}(\vec{0}, 7.5\mathbf{I}_{20})$. The first divergence is at $\eta_i \approx 9.24 \cdot 10^{-4}$.

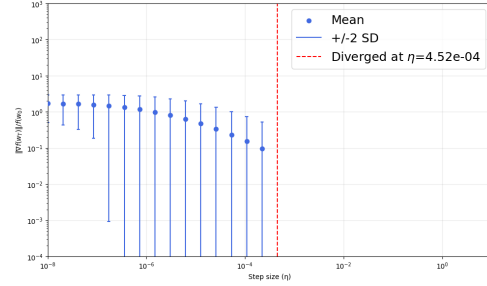


(d) $\pi_j = \mathcal{N}(\vec{0}, 10\mathbf{I}_{20})$. The first divergence is at $\eta_i \approx 9.24 \cdot 10^{-4}$.

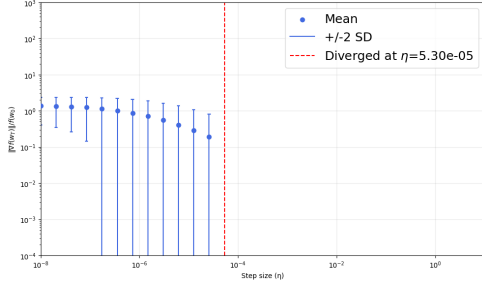
Figure 4: GD simulation results for $p = 5$. For $\pi_j = \mathcal{N}(\vec{0}, 2.5\mathbf{I}_{20})$, the first divergence is at $\eta_i \approx 7.88 \cdot 10^{-3}$. For $\pi_j = \mathcal{N}(\vec{0}, 5\mathbf{I}_{20})$, the first divergence is at $\eta_i \approx 3.86 \cdot 10^{-3}$. For $\pi_j = \mathcal{N}(\vec{0}, 7.5\mathbf{I}_{20}), \mathcal{N}(\vec{0}, 10\mathbf{I}_{20})$, the first divergence is at $\eta_i \approx 9.24 \cdot 10^{-4}$.



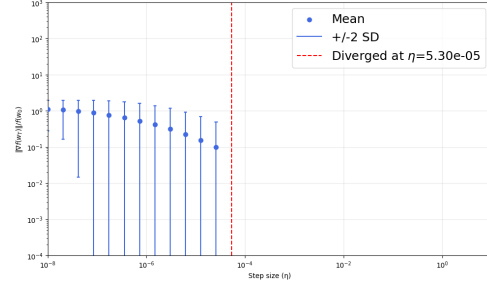
(a) $\pi_j = \mathcal{N}(\vec{0}, 2.5\mathbf{I}_{20})$. The first divergence is at $\eta_i \approx 9.24 \cdot 10^{-4}$.



(b) $\pi_j = \mathcal{N}(\vec{0}, 5.0\mathbf{I}_{20})$. The first divergence is at $\eta_i \approx 4.52 \cdot 10^{-4}$.



(c) $\pi_j = \mathcal{N}(\vec{0}, 7.5\mathbf{I}_{20})$. The first divergence is at $\eta_i \approx 5.30 \cdot 10^{-5}$.



(d) $\pi_j = \mathcal{N}(\vec{0}, 10\mathbf{I}_{20})$. The first divergence is at $\eta_i \approx 5.30 \cdot 10^{-5}$.

Figure 5: GD simulation results for $p = 6$. For $\pi_j = \mathcal{N}(\vec{0}, 2.5\mathbf{I}_{20})$, the first divergence is at $\eta_i \approx 9.24 \cdot 10^{-4}$. For $\pi_j = \mathcal{N}(\vec{0}, 5\mathbf{I}_{20})$, the first divergence is at $\eta_i \approx 4.52 \cdot 10^{-4}$. For $\pi_j = \mathcal{N}(\vec{0}, 7.5\mathbf{I}_{20}), \mathcal{N}(\vec{0}, 10\mathbf{I}_{20})$, the first divergence is at $\eta_i \approx 5.30 \cdot 10^{-5}$.

G.2 Synthetic Simulations with SGD

Simulation Details: We adopt the exact same settings as in [Subsection G.1](#). The only difference is that we study SGD rather than GD, and hence we simulate stochastic gradients. We do so similarly to [Gaash et al. \(2025\)](#): we artificially add $\mathcal{N}(\vec{0}, 0.01\mathbf{I}_{20})$ to ∇F at each iteration of SGD.¹³ The simulations for [Subsection G.2](#) were again run on a Jupyter notebook in Python in Google Colab Pro, connected to a single NVIDIA T4 GPU. Our code is in the attached files.

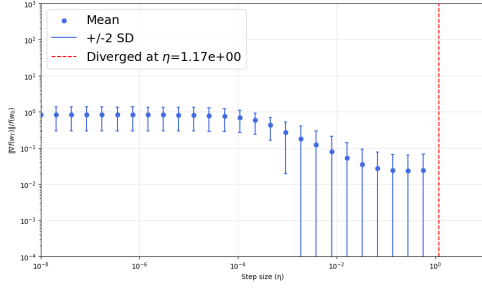
Results: Our conclusions are similar to those from [Subsection G.1](#). When $p = 2$, as predicted by established optimization theory for smooth functions, the first step size leading to divergence η_i is identical across the π_j (see [Figure 6](#)). In contrast for $p = 3, 4, 5, 6$, this η_i generally decreases as the covariance $c_j\mathbf{I}_{20}$ of π_j increases from 2.5 to 10 (see [Figure 7](#), [Figure 8](#), [Figure 9](#), [Figure 10](#)). We note that while the general trends are similar to those from [Subsection G.1](#), we can clearly see the presence of the stochastic gradients in these plots. In many of these plots, $\frac{\|\nabla F(\mathbf{w}_T)\|}{F(\mathbf{w}_0)}$ becomes roughly constant for η large enough such that $T = 1000$ yields reasonable convergence; for such η , by $T = 1000$, the true gradients are small enough and the noise from the stochastic gradients takes over.

Once more, consider how the first step size leading to divergence η_i decreases as the covariance $c_j\mathbf{I}_{20}$ of π_j increases from 2.5 to 10. We find that the rate of this decrease generally increases as p increases. We also again see that fixing π_j and comparing across p , the first step size leading to divergence η_i decreases as p increases. As discussed in [Subsection G.1](#), both of these phenomena are consistent with our theoretical results. For each $p \in \{2, 3, 4, 5, 6\}$ and π_j , we again record the smallest η_i for which divergence occurred in [Table 2](#) on [page 95](#).

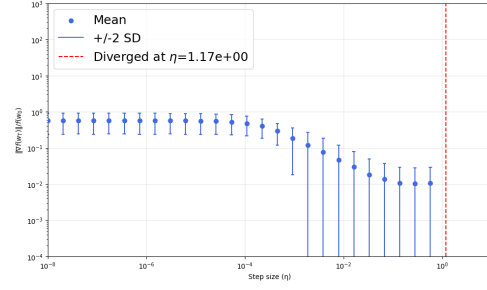
	$\pi_j = \mathcal{N}(\vec{0}, 2.5\mathbf{I}_{20})$	$\pi_j = \mathcal{N}(\vec{0}, 5.0\mathbf{I}_{20})$	$\pi_j = \mathcal{N}(\vec{0}, 7.5\mathbf{I}_{20})$	$\pi_j = \mathcal{N}(\vec{0}, 10\mathbf{I}_{20})$
$p = 2$	$1.17 \cdot 10^0$	$1.17 \cdot 10^0$	$1.17 \cdot 10^0$	$1.17 \cdot 10^0$
$p = 3$	$2.81 \cdot 10^{-1}$	$1.37 \cdot 10^{-1}$	$6.72 \cdot 10^{-2}$	$1.37 \cdot 10^{-1}$
$p = 4$	$3.29 \cdot 10^{-2}$	$3.29 \cdot 10^{-2}$	$1.61 \cdot 10^{-2}$	$7.88 \cdot 10^{-3}$
$p = 5$	$7.88 \cdot 10^{-3}$	$1.89 \cdot 10^{-3}$	$9.24 \cdot 10^{-4}$	$4.52 \cdot 10^{-4}$
$p = 6$	$4.52 \cdot 10^{-4}$	$4.52 \cdot 10^{-4}$	$1.08 \cdot 10^{-4}$	$5.30 \cdot 10^{-5}$

Table 2: Smallest η_i leading to divergence for a given p and initialization π_j .

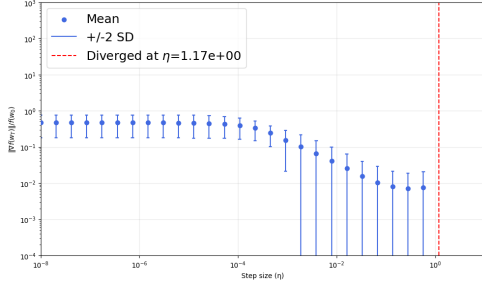
¹³Note our result for convergence of SGD to FOSPs, [Theorem 3.3](#), applies for Gaussian noise as per [Remark 7](#).



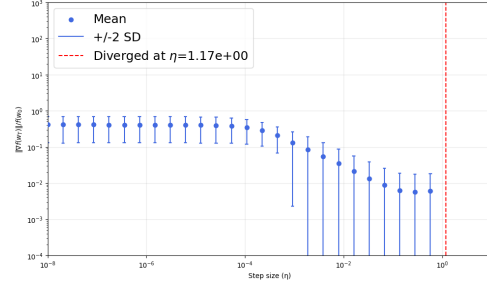
(a) $\pi_j = \mathcal{N}(\vec{0}, 2.5\mathbf{I}_{20})$. The first divergence is at $\eta_i \approx 1.17$.



(b) $\pi_j = \mathcal{N}(\vec{0}, 5.0\mathbf{I}_{20})$. The first divergence is at $\eta_i \approx 1.17$.

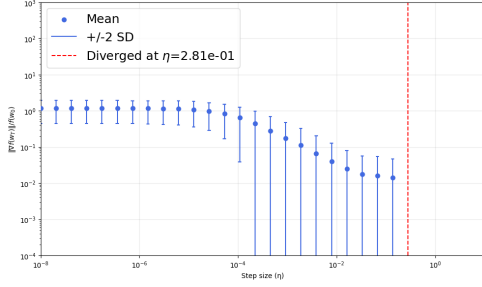


(c) $\pi_j = \mathcal{N}(\vec{0}, 7.5\mathbf{I}_{20})$. The first divergence is at $\eta_i \approx 1.17$.

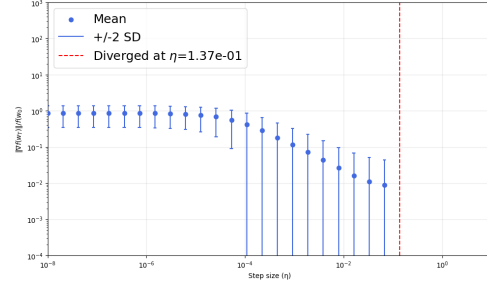


(d) $\pi_j = \mathcal{N}(\vec{0}, 10\mathbf{I}_{20})$. The first divergence is at $\eta_i \approx 1.17$.

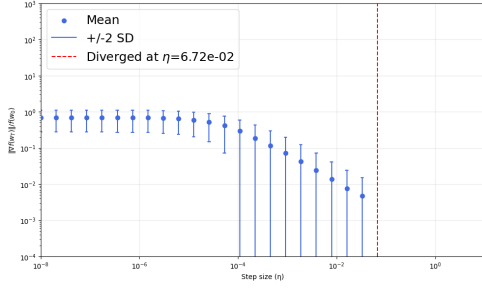
Figure 6: SGD simulation results for $p = 2$. For all π_j , the smallest η_i leading to divergence is ≈ 1.17 .



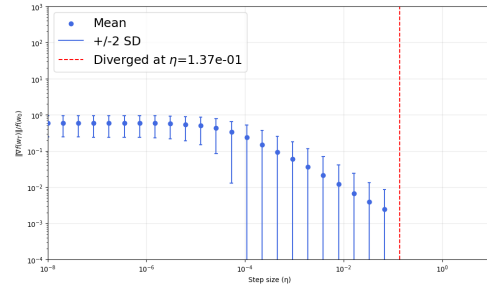
(a) $\pi_j = \mathcal{N}(\vec{0}, 2.5\mathbf{I}_{20})$. The first divergence is at $\eta_i \approx 0.281$.



(b) $\pi_j = \mathcal{N}(\vec{0}, 5.0\mathbf{I}_{20})$. The first divergence is at $\eta_i \approx 0.137$.

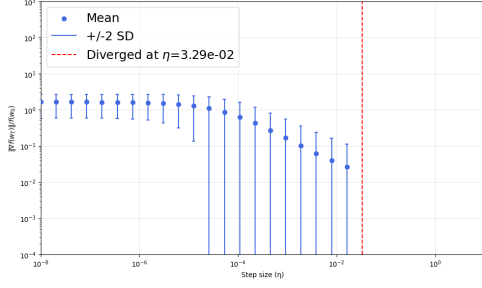


(c) $\pi_j = \mathcal{N}(\vec{0}, 7.5\mathbf{I}_{20})$. The first divergence is at $\eta_i \approx 6.72 \cdot 10^{-2}$.

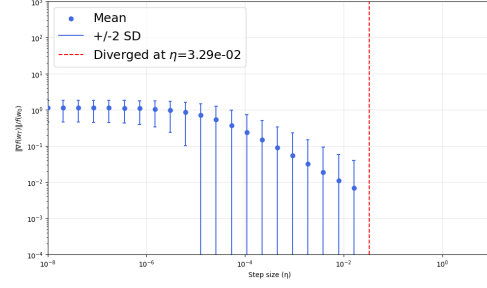


(d) $\pi_j = \mathcal{N}(\vec{0}, 10\mathbf{I}_{20})$. The first divergence is at $\eta_i \approx 0.137$.

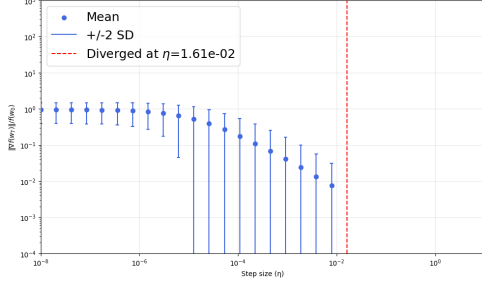
Figure 7: SGD simulation results for $p = 3$. For $\pi_j = \mathcal{N}(\vec{0}, 2.5\mathbf{I}_{20})$, the first divergence is at $\eta_i \approx 0.281$. For $\pi_j = \mathcal{N}(\vec{0}, 7.5\mathbf{I}_{20})$, the first divergence is at $\eta_i \approx 6.72 \cdot 10^{-2}$. For the other π_j , the first divergence is at $\eta_i \approx 0.137$.



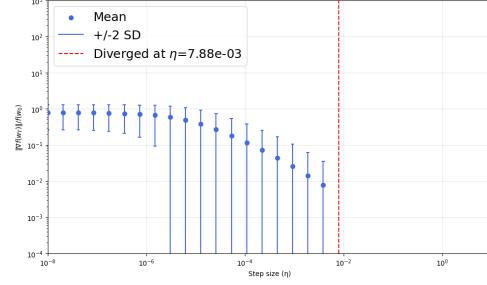
(a) $\pi_j = \mathcal{N}(\vec{0}, 2.5\mathbf{I}_{20})$. The first divergence is at $\eta_i \approx 3.29 \cdot 10^{-2}$.



(b) $\pi_j = \mathcal{N}(\vec{0}, 5.0\mathbf{I}_{20})$. The first divergence is at $\eta_i \approx 3.29 \cdot 10^{-2}$.

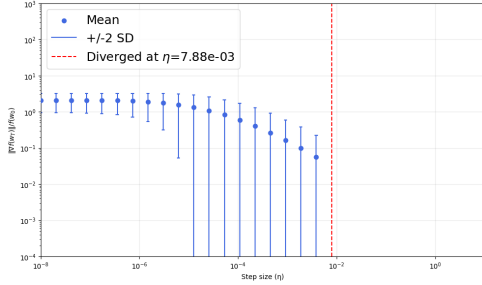


(c) $\pi_j = \mathcal{N}(\vec{0}, 7.5\mathbf{I}_{20})$. The first divergence is at $\eta_i \approx 1.61 \cdot 10^{-2}$.

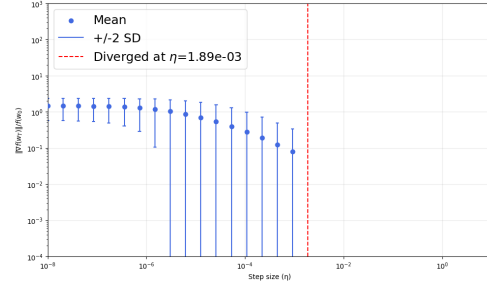


(d) $\pi_j = \mathcal{N}(\vec{0}, 10\mathbf{I}_{20})$. The first divergence is at $\eta_i \approx 7.88 \cdot 10^{-3}$.

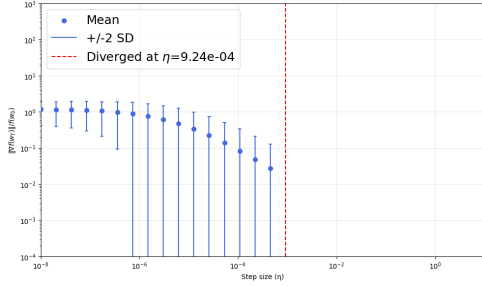
Figure 8: SGD simulation results for $p = 4$. For $\pi_j = \mathcal{N}(\vec{0}, 2.5\mathbf{I}_{20}), \mathcal{N}(\vec{0}, 5.0\mathbf{I}_{20})$, the first divergence is at $\eta_i \approx 3.29 \cdot 10^{-2}$. For $\pi_j = \mathcal{N}(\vec{0}, 7.5\mathbf{I}_{20})$, the first divergence is at $\eta_i \approx 1.61 \cdot 10^{-2}$. For $\pi_j \sim \mathcal{N}(\vec{0}, 10\mathbf{I}_{20})$, the first divergence is at $\eta_i \approx 7.88 \cdot 10^{-3}$.



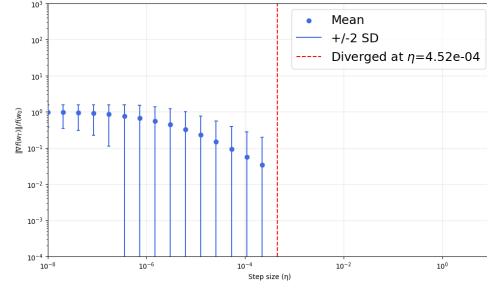
(a) $\pi_j = \mathcal{N}(\vec{0}, 2.5\mathbf{I}_{20})$. The first divergence is at $\eta_i \approx 7.88 \cdot 10^{-3}$.



(b) $\pi_j = \mathcal{N}(\vec{0}, 5.0\mathbf{I}_{20})$. The first divergence is at $\eta_i \approx 1.89 \cdot 10^{-3}$.

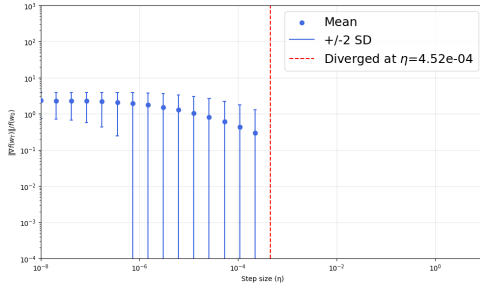


(c) $\pi_j = \mathcal{N}(\vec{0}, 7.5\mathbf{I}_{20})$. The first divergence is at $\eta_i \approx 9.24 \cdot 10^{-4}$.

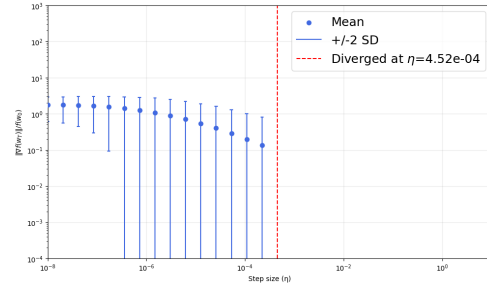


(d) $\pi_j = \mathcal{N}(\vec{0}, 10\mathbf{I}_{20})$. The first divergence is at $\eta_i \approx 4.52 \cdot 10^{-4}$.

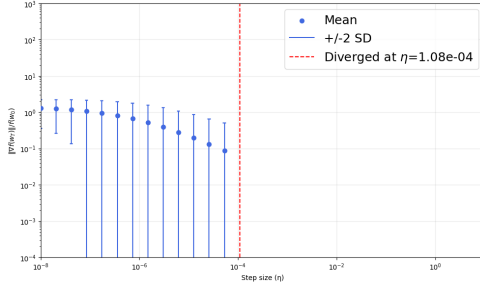
Figure 9: SGD simulation results for $p = 5$. For $\pi_j = \mathcal{N}(\vec{0}, 2.5\mathbf{I}_{20}), \mathcal{N}(\vec{0}, 5.0\mathbf{I}_{20}), \mathcal{N}(\vec{0}, 7.5\mathbf{I}_{20}), \mathcal{N}(\vec{0}, 10\mathbf{I}_{20})$, the first divergence is at $\eta_i \approx 7.88 \cdot 10^{-3}, 1.89 \cdot 10^{-3}, 9.24 \cdot 10^{-4}, 4.52 \cdot 10^{-4}$ respectively.



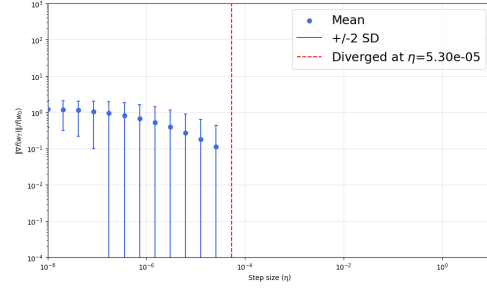
(a) $\pi_j = \mathcal{N}(\vec{0}, 2.5\mathbf{I}_{20})$. The first divergence is at $\eta_i \approx 4.52 \cdot 10^{-4}$.



(b) $\pi_j = \mathcal{N}(\vec{0}, 5.0\mathbf{I}_{20})$. The first divergence is at $\eta_i \approx 4.52 \cdot 10^{-4}$.



(c) $\pi_j = \mathcal{N}(\vec{0}, 7.5\mathbf{I}_{20})$. The first divergence is at $\eta_i \approx 1.08 \cdot 10^{-4}$.



(d) $\pi_j = \mathcal{N}(\vec{0}, 10\mathbf{I}_{20})$. The first divergence is at $\eta_i \approx 5.30 \cdot 10^{-5}$.

Figure 10: SGD simulation results for $p = 6$. For $\pi_j = \mathcal{N}(\vec{0}, 2.5\mathbf{I}_{20}), \mathcal{N}(\vec{0}, 5.0\mathbf{I}_{20}), \mathcal{N}(\vec{0}, 7.5\mathbf{I}_{20}), \mathcal{N}(\vec{0}, 10\mathbf{I}_{20})$, the first divergence are at $\eta_i \approx 4.52 \cdot 10^{-4}, 4.52 \cdot 10^{-4}, 1.08 \cdot 10^{-4}, 5.30 \cdot 10^{-5}$ respectively.