A. Proof of Lemmas in Preliminaries

A.1. Proof of Lemma 3

Proof. Let $g(t) = f(x + t(y - x))$ for $t \in [0, 1]$, then g is $L\|y - x\|$-Lipschitz implying that g is absolutely continuous. Thus from the fundamental theorem of calculus (Lebesgue), g has a derivative g' almost everywhere, and the derivative is Lebesgue integrable such that

$$g(t) = g(0) + \int_0^t g'(s)ds.$$

Moreover, if g is differentiable at t, then

$$g'(t) = \lim_{\delta t \to 0} \frac{g(t + \delta t) - g(t)}{\delta t} = \lim_{\delta t \to 0} \frac{f(x + (t + \delta t)(y - x)) - f(x + t(y - x))}{\delta t} = f'(x + t(y - x), y - x).$$

Since this equality holds almost everywhere, we have

$$f(y) - f(x) = g(1) - g(0) = \int_0^1 g'(t)dt = \int_0^1 f'(x + t(y - x), y - x)dt.$$

□

A.2. Proof of Lemma 4

Proof. For any $\varphi(t) = x + td$ as given in Definition 3, let $t_k \to 0$. Denote $x_k = \varphi(t_k), \delta_k = \|x_k - x\| \to 0$. By Proposition 1.6, we know that there exists $g_{k,j} \in \cup_{y \in x + \delta_k B} \partial f(y)$ such that

$$f(x_k) - f(x) = \langle g_{k,j}, x_k - x \rangle.$$

By the existence of directional derivative, we know that

$$\lim_{k \to \infty} \langle g_{k,j}, d \rangle = \lim_{k \to \infty} \frac{\langle g_{k,j}, t_k d \rangle}{t_k} = f'(x, d)$$

$g_{k,j}$ is in a bounded set with norm less than L. The Lemma follows by the fact that any accumulation point of $g_{k,j}$ is in $\partial f(x)$ due to upper-semicontinuity of $\partial f(x)$. □

B. Proof of Lemmas in Algorithm Complexity

B.1. Proof of Theorem 5

Our proof strategy is similar to Theorem 1.1.2 in (Nesterov, 2018), where we use the resisting strategy to prove lower bound. Given a one dimensional function f, let $x_k, k \in [1, K]$ be the sequence of points queried in ascending order instead of query order. We assume without loss of generality that the initial point is queried and is an element of $\{x_k\}_{k=0}^K$ (otherwise, query the initial point first before proceeding with the algorithm).

Then we define the resisting strategy: always return

$$f(x) = 0, \text{and} \quad \nabla f(x) = L.$$

If we can prove that for any set of points $x_k, k \in [1, K]$, there exists two functions such that they satisfy the resisting strategy $f(x_k) = 0, \text{and} \quad \nabla f(x_k) = L, k \in [1, K]$, and that the two functions do not share any common stationary points, then we know no randomized/deterministic can return an $\epsilon-$stationary points with probability more than $1/2$ for both functions simultaneously. In other word, no algorithm that query K points can distinguish these two functions. Hence we proved the theorem following the definition of complexity in (5) with $\delta = 0$.

All we need to do is to show that such two functions exist in the Lemma below.

Lemma 12. *Given a finite sequence of real numbers $\{x_k\}_{k \in [1,K]} \in \mathbb{R}$, there is a family of functions $f_\theta \in \mathcal{F}(\Delta, L)$ such that for any $k \in [1, K]$,*

$$f_\theta(x_k) = 0 \quad \textit{and} \quad \nabla f_\theta(x_k) = L$$

and for ϵ sufficiently small, the set of ϵ-stationary points of f_θ are all disjoint, i.e $\{\epsilon$-stationary points of $f_{\theta_1}\} \cap \{\epsilon$-stationary points of $f_{\theta_2}\} = \emptyset$ for any $\theta_1 \neq \theta_2$.

Proof. Up to a permutation of the indices, we could reorder the sequence in the increasing order. WLOG, we assume x_k is increasing. Let $\delta = \min\{\min_{x_i \neq x_j}\{|x_i - x_j|\}, \frac{\Delta}{L}\}$. For any $0 < \theta < 1/2$, we define f_θ by

$$
\begin{aligned}
&f_\theta(x) = -L(x - x_1 + 2\theta\delta) \quad \text{for} \quad x \in (-\infty, x_1 - \theta\delta] \\
&f_\theta(x) = L(x - x_k) \quad \text{for} \quad x \in \left[x_k - \theta\delta, \frac{x_k + x_{k+1}}{2} - \theta\delta\right] \\
&f_\theta(x) = -L(x - x_{k+1} + 2\theta\delta) \quad \text{for} \quad x \in \left[\frac{x_k + x_{k+1}}{2} - \theta\delta, x_{k+1} - \theta\delta\right] \\
&f_\theta(x) = L(x - x_K) \quad x \in [x_K + \theta\delta, +\infty).
\end{aligned}
$$

It is clear that f_θ is directional differentiable at all point and $\nabla f_\theta(x_k) = L$. Moreover, the minimum $f_\theta^* = -L\theta\delta \geq -\Delta$. This implies that $f_\theta \in \mathcal{F}(\Delta, L)$. Note that $\nabla f_\theta = L$ or $-L$ except at the local extremum. Therefore, for any $\epsilon < L$ the set of ϵ-stationary points of f_θ are exactly

$$
\{\epsilon\text{-stationary points of } f_\theta\} = \{x_k - \theta\delta \mid k \in [1, K]\} \cup \left\{\frac{x_k + x_{k+1}}{2} - \theta\delta \mid k \in [1, K-1]\right\},
$$

which is clearly distinct for different choice of θ. □

B.2. Proof of Proposition 6

Proof. When x is $(\frac{\epsilon}{3L}, \frac{\epsilon}{3})$ stationary, we have $d(0, \partial f(x + \frac{\epsilon}{3L}B)) \leq \frac{\epsilon}{3}$. By definition, we could find $g \in conv(\cup_{y \in x + \frac{\epsilon}{3L}B} \nabla f(y))$ such that $\|g\| \leq 2\epsilon/3$. This means, there exists $x_1, \cdots, x_k \in x + \frac{\epsilon}{3L}B$, and $\alpha_1, \cdots, \alpha_k \in [0, 1]$ such that $\alpha_1 + \cdots + \alpha_k = 1$ and

$$
g = \sum_{i=1}^{k} \alpha_i \nabla f(x_i)
$$

Therefore

$$
\begin{aligned}
\|\nabla f(x)\| &\leq \|g\| + \|\nabla f(x) - g\| \\
&\leq \frac{2\epsilon}{3} + \sum_{i=1}^{k} \alpha_i \|\nabla f(x) - \nabla f(x_k)\| \\
&\leq \frac{2\epsilon}{3} + \sum_{i=1}^{k} \alpha_i L \|x - x_k\| \\
&\leq \frac{2\epsilon}{3} + \sum_{i=1}^{k} \alpha_i L \frac{\epsilon}{3L} = \epsilon.
\end{aligned}
$$

Therefore, x is an ϵ-stationary point in the standard sense. □

B.3. Proof of Lemma 7

Proof. First, we show that the limit exists. By Lipschitzness and Jenson inequality, we know that $\partial f(x + \delta_{k+1}B)$ lies in a bounded ball with radius L. For any sequence of $\{\delta_k\}$ with $\delta_k \downarrow 0$, we know that $\partial f(x + \delta_{k+1}B) \subseteq \partial f(x + \delta_k B)$. Therefore, the limit exists by the monotone convergence theorem.

Next, we show that $\lim_{\delta \downarrow 0} \partial f(x + \delta B) = \partial f(x)$. For one direction, we show that $\partial f(x) \subseteq \lim_{\delta \downarrow 0} \partial f(x + \delta B)$. This follows by proposition 1.5 and the fact that

$$
\cup_{y \in x + \delta B} \partial f(y) \subseteq \text{conv}(\cup_{y \in x + \delta B} \partial f(y)) = \partial f(x + \delta B).
$$

Next, we show the other direction $\lim_{\delta \downarrow 0} \partial f(x + \delta B) \subseteq \partial f(x)$. By upper semicontinuity, we know that for any $\epsilon > 0$, there exists $\delta > 0$ such that

$$
\cup_{y \in x + \delta B} \partial f(y) \subseteq \partial f(x) + \epsilon B.
$$

Then by convexity of $\partial f(x)$ and ϵB, we know that their Minkowski sum $\partial f(x) + \epsilon B$ is convex. Therefore, we conclude that for any $\epsilon > 0$, there exists $\delta > 0$ such that

$$\partial f(x + \delta B) = \text{conv}(\cup_{y \in x+\delta B} \partial f(y)) \subseteq \partial f(x) + \epsilon B.$$

□

C. Proof of Theorem 8

Before we prove the theorem, we first analyze how many times the algorithm iterates in the while loop.

Lemma 13. *Let $K = \frac{48L^2}{\epsilon^2}$. Given $t \in [1, T]$,*

$$\mathbb{E}[\|m_{t,K}\|^2] \leq \frac{\epsilon^2}{16}.$$

where for convenience of analysis, we define $m_{t,k} = 0$ for all $k > k_0$ if the k-loop breaks at (t, k_0). Consequently, for any $\gamma < 1$, with probability $1 - \gamma$, there are at most $log(1/\gamma)$ restarts of the while loop at the t-th iteration.

Proof. Let $\mathfrak{F}_{t,k} = \sigma(y_{t,1}, \cdots, y_{t,k+1})$, then $x_{t,k}, m_{t,k} \in \mathfrak{F}_{t,k}$. We denote $D_{t,k}$ as the event that k-loop does not break at $x_{t,k}$, i.e. $\|m_{t,k}\| > \epsilon$ and $f(x_{t,k}) - f(x_t) > -\frac{\delta\|m_{t,k}\|}{4}$. It is clear that $D_{t,k} \in \mathfrak{F}_{t,k}$.

Let $\gamma(\lambda) = (1-\lambda)x_t + \lambda x_{t,k},\ \lambda \in [0,1]$. Note that $\gamma'(\lambda) = x_{t,k} - x_t = -\delta\frac{m_{t,k}}{\|m_{t,k}\|}$. Since $y_{t,k+1}$ is uniformly sampled from line segment $[x_t, x_{t,k}]$, we know

$$\mathbb{E}[\langle g_{t,k+1}, x_{t,k} - x_t\rangle | \mathfrak{F}_{t,k}] = \int_0^1 f'(\gamma(t), x_{t,k} - x_t)dt = f(x_{t,k}) - f(x_t)$$

where the second equality comes from directional differentiability. Since $x_{k+1} - x_k = -\delta\frac{m_{t,k}}{\|m_{t,k}\|}$, we know that

$$\mathbb{E}[\langle g_{t,k+1}, m_{t,k}\rangle | \mathfrak{F}_{t,k}] = -\frac{\|m_{t,k}\|}{\delta}(f(x_{t,k}) - f(x_t)). \tag{7}$$

By construction $m_{t,k+1} = \beta m_{t,k} + (1-\beta)g_{t,k+1}$ under $D_{t,k} \cap \cdots \cap D_{t,1}$, and $m_{t,k+1} = 0$ otherwise. Therefore,

$$\begin{aligned}
&\mathbb{E}[\|m_{t,k+1}\|^2 | \mathfrak{F}_{t,k}] \\
=&\mathbb{E}[\|\beta m_{t,k} + (1-\beta)g_{t,k+1}\|^2 \mathbb{1}_{D_{t,k}\cap\cdots\cap D_{t,1}} | \mathfrak{F}_{t,k}] \\
\leq& \left(\beta^2\|m_{t,k}\|^2 + (1-\beta)^2L^2 + 2\beta(1-\beta)\mathbb{E}[\langle g_{t,k+1}, m_{t,k}\rangle | \mathfrak{F}_{t,k}]\right)\mathbb{1}_{D_{t,k}\cap\cdots\cap D_{t,1}} \\
\leq& \beta^2\|m_{t,k}\|^2 + (1-\beta)^2L^2 - 2\beta(1-\beta)\frac{\|m_{t,k}\|}{\delta}(f(x_{t,k}) - f(x_t))\mathbb{1}_{D_{t,k}\cap\cdots\cap D_{t,1}} \\
\leq& \beta^2\|m_{t,k}\|^2 + (1-\beta)^2L^2 + 2\beta(1-\beta)\frac{\|m_{t,k}\|^2}{4}
\end{aligned}$$

where in the third line, we use the fact $\beta, D_{t,k} \cap \cdots \cap D_{t,1} \in \mathfrak{F}_{t,k}$; in the fourth line we use the fact under $D_{t,k}$, $f(x_{t,k}) - f(x_t) \geq -\frac{\delta\|m_{t,k}\|}{4}$. The last equation is a quadratic function with respect to β, which could be rewritten as

$$h(\beta) = \beta^2(\frac{\|m_{t,k}\|^2}{2} + L^2) - 2\beta(L^2 - \frac{\|m_{t,k}\|^2}{4}) + L^2.$$

It achieves the minimum at $\beta = \frac{4L^2 - \|m_{t,k}\|^2}{4L^2 + 2\|m_{t,k}\|^2}$, which belongs to $\mathfrak{F}_{t,k}$. Since $\|m_{t,k}\| \leq L$, we have

$$h^* = \frac{L^2}{L^2 + \frac{\|m_{t,k}\|^2}{2}}\|m_{t,k}\|^2 \leq \left(1 - \frac{\|m_{t,k}\|^2}{3L^2}\right)\|m_{t,k}\|^2$$

Therefore,

$$\begin{aligned}&\mathbb{E}[\|m_{t,k+1}\|^2]\\=&\mathbb{E}[\mathbb{E}[\|m_{t,k+1}\|^2|\mathfrak{F}_{t,k}]]\\\leq&\mathbb{E}\left[\left(1-\frac{\|m_{t,k}\|^2}{3L^2}\right)\|m_{t,k}\|^2\right]\\\leq&\left(1-\frac{\mathbb{E}[\|m_{t,k}\|^2]}{3L^2}\right)\mathbb{E}[\|m_{t,k}\|^2]\end{aligned}$$

where the last inequality follows from Jensen's inequality under the fact that the function $x \to (1 - x/3L^2)x$ is concave. Now consider the sequence $v_k = \mathbb{E}[\|m_{t,k}\|^2]/L^2 \in [0, 1]$, we get

$$v_{k+1} \leq v_k - v_k^2/3 \quad \Longrightarrow \quad \frac{1}{v_{k+1}} \geq \frac{1}{v_k - v_k^2/3} \geq \frac{1}{v_k} + \frac{1}{3}.$$

Knowing that $v_1 \leq 1$, we therefore have

$$v_k \leq \frac{3}{k+2}.$$

When $K > \frac{48L^2}{\epsilon^2}$, we have $\mathbb{E}[\|m_{t,K}\|^2] \leq \frac{\epsilon^2}{16}$. Therefore, by Markov inequality, $\mathcal{P}\{\|m_{t,K}\| \geq \epsilon\} \leq 1/4$. In other word, the while-loop restart with probability at most $1/4$. Therefore, with probability $1 - \gamma$, there are at most $\log(1/\gamma)$ restarts. □

Now we are ready to prove the main theorem.

Proof of Theorem 8. We notice that $m_{t,k}$ is always a convex combinations of generalized gradients within the δ ball of x_k, i.e.

$$m_{t,k} \in \partial f(x_t + \delta B) = \text{conv}(\cup_{y \in x_t + \delta B} \partial f(y))$$

Therefore, if at any t, k, $\|m_{t,k}\| \leq \epsilon$, then the corresponding x_t is a (δ, ϵ) approximate stationary point. To show that our algorithm always find a $\|m_{t,k}\| \leq \epsilon$, we need to control the number of times the descent condition is satisfied, which breaks the while-loop without satisfying $\|m_{t,k}\| \leq \epsilon$. Indeed, when the descent condition holds, we have

$$f(x_{t,k}) - f(x_t) \leq -\frac{\delta\|m_{t,k}\|}{4} < -\frac{\delta\epsilon}{4},$$

where we use the fact $\|m_{t,k}\| > \epsilon$, otherwise, the algorithm already terminates. Consequently, there are at most $\frac{4\Delta}{\delta\epsilon} - 1 = T - 1$ iterations that the descent condition holds. As a result, for at least one t , the while-loop ends providing a (δ, ϵ) approximate stationary point.

By Lemma 13, we know that with probability $1 - \frac{\gamma\delta\epsilon}{4\Delta}$, the t-th iteration terminates in $\log(\frac{4\Delta}{\gamma\delta\epsilon})$ restarts. Consequently, with probability $1 - \gamma$, the algorithm returns a (δ, ϵ) approximate stationary point using

$$\frac{192\Delta L^2}{\epsilon^3\delta}\log\left(\frac{4\Delta}{\gamma\delta\epsilon}\right) \quad \text{oracle calls.}$$

□

D. Proof of Theorem 10

Stochastic INGD has convergence guarantee as stated in the next theorem.

Theorem 14. *Under the stochastic Assumption 1, the Stochastic INGD algorithm in Algorithm 2 with parameters* $\beta = 1 - \frac{\epsilon^2}{64G^2}$, $p = \frac{64G^2\ln(16G/\epsilon)}{\delta\epsilon^2}$, $q = 4Gp$, $T = \frac{2^{16}G^3\Delta\ ln(16G/\epsilon)}{\epsilon^4\delta}\max\{1, \frac{G\delta}{8\Delta}\}$, $K = p\delta$ *has algorithm complexity upper bounded by*

$$\frac{2^{16}G^3\Delta\ ln(16G/\epsilon)}{\epsilon^4\delta}\max\{1, \frac{G\delta}{8\Delta}\} = \tilde{O}\left(\frac{G^3\Delta}{\epsilon^4\delta}\right).$$

Proof. First, we are going to show that

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|m_t\|] \le \epsilon/4. \tag{8}$$

From construction of the descent direction, we have

$$\|m_{t+1}\|^2 = (1-\beta)^2\|g(y_{t+1})\|^2 + 2\beta(1-\beta)\langle g(y_{t+1}), m_t\rangle + \beta^2\|m_t\|^2. \tag{9}$$

Multiply both side by η_t and sum over t, we get

$$0 = (1-\beta)^2\underbrace{\sum_{t=1}^{T}\eta_t\|g(y_{t+1})\|^2}_{\text{i}} + 2\beta(1-\beta)\underbrace{\sum_{t=1}^{T}\langle g(y_{t+1}), \eta_t m_t\rangle}_{\text{ii}} + \underbrace{\sum_{t=1}^{T}\eta_t(-\|m_{t+1}\|^2 + \beta^2\|m_t\|^2)}_{\text{iii}}. \tag{10}$$

We remark that at each iteration, we have two randomized/stochastic procedure: first we draw y_{t+1} randomly between the segment $[x_t, x_{t+1}]$, second we draw a stochastic gradient at y_{t+1}. For convenience of analysis, we denote $\mathcal{G}_t$ as the sigma field generated by $g(y_t)$, and $\mathcal{Y}_t$ as the sigma field generated by y_t. Clearly, $\mathcal{G}_t \subset \mathcal{Y}_{t+1} \subset \mathcal{G}_{t+1}$. By definition η_t is determined by m_t, which is further determined by g_t. Hence, the vectors m_t, η_t and x_{t+1} are $\mathcal{G}_t$-measurable.

Now we analyze each term one by one.

Term i: This term could be easily bound by

$$\mathbb{E}[\eta_t\|g(y_{t+1})\|^2] \le \frac{1}{q}\mathbb{E}[\|g(y_{t+1})\|^2] = \frac{1}{q}\mathbb{E}[\mathbb{E}[\|g(y_{t+1})\|^2|\mathcal{Y}_{t+1}]] \le \frac{G^2}{q} \tag{11}$$

Term ii: Note that $\eta_t m_t = x_t - x_{t+1}$, we have

$$\begin{aligned}
\mathbb{E}[\langle g(y_{t+1}), \eta_t m_t\rangle \mid \mathcal{G}_t] &= \mathbb{E}[\mathbb{E}[\langle g(y_{t+1}), x_t - x_{t+1}\rangle \mid \mathcal{Y}_{t+1}]|\ \mathcal{G}_t] \\
&= \mathbb{E}[f'(y_{t+1}; x_t - x_{t+1})|\ \mathcal{G}_t] \\
&= \int_{[0,1]} f'(x_{t+1} + \lambda(x_t - x_{t+1}); x_t - x_{t+1})d\lambda \\
&= f(x_t) - f(x_{t+1}),
\end{aligned}$$

where the second line we use the property of the oracle given in Assumption 1(b). Thus by taking the expectation, we have

$$\sum_{t=1}^{T}\mathbb{E}[\langle g(y_{t+1}), \eta_t m_t\rangle] = \mathbb{E}[f(x_1) - f(x_{T+1})] \le \Delta$$

Term iii: we would like to develop a telescopic sum for the third term, however this is non-trivial since the stepsize η_t is adaptive. Extensive algebraic manipulation is involved.

$$\begin{aligned}
&\sum_{t=1}^{T}\eta_t(-\|m_{t+1}\|^2 + \beta^2\|m_t\|^2) \\
=&\sum_{t=1}^{T}\frac{-\|m_{t+1}\|^2}{p\|m_t\| + q} + \beta^2\sum_{t=1}^{T}\frac{\|m_t\|^2}{p\|m_t\| + q} \\
=&\sum_{t=1}^{T}\left(\frac{-\|m_{t+1}\|^2}{p\|m_t\| + q} + \frac{\|m_{t+1}\|^2}{p\|m_{t+1}\| + q}\right) - \sum_{t=1}^{T}\frac{\|m_{t+1}\|^2}{p\|m_{t+1}\| + q} + \beta^2\sum_{t=1}^{T}\frac{\|m_t\|^2}{p\|m_t\| + q} \\
=&\sum_{t=1}^{T}\frac{p\|m_{t+1}\|^2(\|m_t\| - \|m_{t+1}\|)}{(p\|m_t\| + q)(p\|m_{t+1}\| + q)} + \beta^2\frac{\|m_1\|^2}{p\|m_1\| + q} + (\beta^2 - 1)\sum_{t=2}^{T+1}\frac{\|m_t\|^2}{p\|m_t\| + q}
\end{aligned} \tag{12}$$

The first equality follows by $\eta_t = \frac{1}{p\|m_t\|+q}$. The second equality subtract and add the same terms $\frac{\|m_{t+1}\|^2}{p\|m_{t+1}\|+q}$. The last equality regroups the terms.

We now prove the first term in (12) admits the following upper bound:

$$\frac{p\|m_{t+1}\|^2(\|m_t\| - \|m_{t+1}\|)}{(p\|m_t\|+q)(p\|m_{t+1}\|+q)} \leq (1-\beta)\frac{\|m_{t+1}\|^2}{p\|m_{t+1}\|+q} + \frac{(1-\beta)p\|g(y_{t+1})\|}{q}\frac{\|m_t\|^2}{p\|m_t\|+q} \tag{13}$$

Note that if $\|m_{t+1}\| \geq \|m_t\|$ then the inequality trivially holds. Thus, we only need to consider the case when $\|m_{t+1}\| \leq \|m_t\|$. By triangle inequality,

$$\begin{aligned}\|m_t\| - \|m_{t+1}\| \leq \|m_t - m_{t+1}\| &= (1-\beta)\|m_t - g(y_{t+1})\| \\ &\leq (1-\beta)(\|m_t\| + \|g(y_{t+1})\|).\end{aligned}$$

Therefore, substitue the above inequality into lefthand side of (13) and regroup the fractions,

$$\begin{aligned}\frac{p\|m_{t+1}\|^2(\|m_t\| - \|m_{t+1}\|)}{(p\|m_t\|+q)(p\|m_{t+1}\|+q)} &\leq \frac{p\|m_{t+1}\|^2(1-\beta)(\|m_t\| + \|g(y_{t+1})\|)}{(p\|m_t\|+q)(p\|m_{t+1}\|+q)} \\ &= (1-\beta)\frac{\|m_{t+1}\|^2}{p\|m_{t+1}\|+q}\frac{p\|m_t\|}{p\|m_t\|+q} + \frac{(1-\beta)p\|g(y_{t+1})\|}{p\|m_t\|+q}\frac{\|m_{t+1}\|^2}{p\|m_{t+1}\|+q} \\ &\leq (1-\beta)\frac{\|m_{t+1}\|^2}{p\|m_{t+1}\|+q} + \frac{(1-\beta)p\|g(y_{t+1})\|}{q}\frac{\|m_t\|^2}{p\|m_t\|+q},\end{aligned}$$

where the last step we use the fact that $\|m_{t+1}\| \leq \|m_t\|$ and the function $x \to x^2/(px+q)$ is increasing on $\mathbb{R}_+$. Now, taking expectation on both sides of (13) yields

$$\begin{aligned}\mathbb{E}\left[\frac{p\|m_{t+1}\|^2(\|m_t\| - \|m_{t+1}\|)}{(p\|m_t\|+q)(p\|m_{t+1}\|+q)}\right] &\leq (1-\beta)\mathbb{E}\left[\frac{\|m_{t+1}\|^2}{p\|m_{t+1}\|+q}\right] + \frac{p(1-\beta)}{q}\mathbb{E}\left[\|g(y_{t+1})\|\frac{\|m_t\|^2}{p\|m_t\|+q}\right] \\ &= (1-\beta)\mathbb{E}\left[\frac{\|m_{t+1}\|^2}{p\|m_{t+1}\|+q}\right] + \frac{p(1-\beta)}{q}\mathbb{E}\left[\mathbb{E}\left[\|g(y_{t+1})\||\mathcal{G}_t\right]\frac{\|m_t\|^2}{p\|m_t\|+q}\right] \\ &\leq (1-\beta)\mathbb{E}\left[\frac{\|m_{t+1}\|^2}{p\|m_{t+1}\|+q}\right] + \frac{p(1-\beta)G}{q}\mathbb{E}\left[\frac{\|m_t\|^2}{p\|m_t\|+q}\right] \\ &\leq (1-\beta)\mathbb{E}\left[\frac{\|m_{t+1}\|^2}{p\|m_{t+1}\|+q}\right] + \frac{\beta(1-\beta)}{2}\mathbb{E}\left[\frac{\|m_t\|^2}{p\|m_t\|+q}\right]\end{aligned}$$

where the third inequality follows by the fact that $\mathbb{E}[\|g(y_{t+1}\||\mathcal{G}_t] \leq \sqrt{L^2+\sigma^2}$ and the last inequality follows from our choice of parameters ensuring $pG/q \leq \beta/2$.

Now we are ready to proceed the telescopic summing. Summing up over t and yields

$$\begin{aligned}&\sum_{t=1}^{T}\mathbb{E}\left[\eta_t(-\|m_{t+1}\|^2 + \beta^2\|m_t\|^2)\right] \\ &\leq (1-\beta)\sum_{t=1}^{T}\mathbb{E}\left[\frac{\|m_{t+1}\|^2}{p\|m_{t+1}\|+q}\right] + \frac{\beta-\beta^2}{2}\sum_{t=1}^{T}\mathbb{E}\left[\frac{\|m_t\|^2}{p\|m_t\|+q}\right] + \beta^2\mathbb{E}\left[\frac{\|m_1\|^2}{p\|m_1\|+q}\right] + (\beta^2-1)\sum_{t=2}^{T+1}\mathbb{E}\left[\frac{\|m_t\|^2}{p\|m_t\|+q}\right] \\ &= \frac{\beta^2+\beta}{2}\mathbb{E}\left[\frac{\|m_1\|^2}{p\|m_1\|+q}\right] + \frac{\beta^2-\beta}{2}\sum_{t=2}^{T+1}\mathbb{E}\left[\frac{\|m_t\|^2}{p\|m_t\|+q}\right] \\ &= \beta^2\mathbb{E}\left[\frac{\|m_1\|^2}{p\|m_1\|+q}\right] + \frac{\beta^2-\beta}{2}\sum_{t=1}^{T+1}\mathbb{E}\left[\frac{\|m_t\|^2}{p\|m_t\|+q}\right] \\ &\leq \frac{\beta^2G^2}{q} + \frac{\beta^2-\beta}{2}\sum_{t=1}^{T+1}\mathbb{E}\left[\frac{\|m_t\|^2}{p\|m_t\|+q}\right]\end{aligned}$$

The first inequality uses (13). The third line and the foruth line regroup the terms. The last line follows by $p\|m_1\| + q \geq q$ and $\mathbb{E}[\|m_1\|^2] \leq G^2$.

Combine all term i, ii and iii in (10) yields

$$\frac{\beta - \beta^2}{2} \sum_{t=1}^{T+1} \mathbb{E}\left[\frac{\|m_t\|^2}{p\|m_t\| + q}\right] \leq 2\beta(1-\beta)\mathbb{E}[f(x_1) - f(x_{T+1})] + \frac{\beta^2 G^2}{q} + T(1-\beta)^2 \frac{G^2}{q}.$$

Multiply both side by $\frac{2q}{T(\beta-\beta^2)}$ we get

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\frac{q\|m_t\|^2}{p\|m_t\| + q}\right] \leq \frac{4q\Delta}{T} + \frac{2\beta G^2}{T(1-\beta)} + \frac{2(1-\beta)G^2}{\beta} \tag{14}$$

We may assume $\epsilon \leq G$, otherwise any x_t is a (δ, ϵ)-stationary point. Then by choosing $\beta = 1 - \frac{\epsilon^2}{64G^2}$, $p = \frac{64G^2 \ln(16G/\epsilon)}{\delta\epsilon^2}$, $q = \frac{256G^3 \ln(16G/\epsilon)}{\delta\epsilon^2}$, $T = \frac{2^{16}G^3\Delta \ln(16G/\epsilon)}{\epsilon^4\delta} \max\{1, \frac{G\delta}{8\Delta}\}$, have

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\frac{4G\|m_t\|^2}{\|m_t\| + 4G}\right] \leq \frac{\epsilon^2}{17} \tag{15}$$

Note that the function $x \to x^2/(x + 4G)$ is convex, thus by Jensen's inequality, for any t, we have

$$\frac{4G\mathbb{E}\left[\|m_t\|\right]^2}{\mathbb{E}[\|m_t\|] + 4G} \leq \mathbb{E}\left[\frac{4G\|m_t\|^2}{\|m_t\| + 4G}\right] \tag{16}$$

Let's denote

$$m_{avg} = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\|m_t\|\right],$$

then again by Jensen's inequality,

$$\frac{4Gm_{avg}^2}{m_{avg} + 4G} \leq \frac{1}{T}\sum_{t=1}^{T} \frac{4G\mathbb{E}\left[\|m_t\|\right]^2}{\mathbb{E}[\|m_t\|] + 4G} \leq \frac{\epsilon^2}{17}$$

Solving the quadratic inequality with respect to m_{avg} and using $\epsilon \leq G$, we have

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\|m_t\|\right] \leq \frac{\epsilon}{4}.$$

In contrast to the smooth case, we cannot directly conclude from this inequality since m_t is not the gradient at x_t. Indeed, it is the convex combination of all previous stochastic gradients. Therefore, we still need to find a reference point such that m_t is approximately in the δ-subdifferential of the reference point. Note that

$$m_t = \sum_{i=t-K+1}^{t} \alpha_i g(y_i) + \beta^K m_{t-K}$$

Intuitively, when K is sufficiently large, the contribution of the last term in m_t is negligible. In which case, we could deduce m_t is approximately in $\partial f(x_{t-K} + \delta B)$. More precisely, with $\beta = 1 - \frac{\epsilon^2}{64G^2}$, as long as $K \geq \frac{64G^2}{\epsilon^2} \ln(\frac{16G}{\epsilon})$, we have

$$\beta^K \leq \frac{\epsilon}{16G}.$$

This is a simple analysis result using the fact that $ln(1-x) \leq -x$. Then by Assumption on the oracle, we know that $\mathbb{E}[g(y_i)|\mathcal{Y}_i] \in \partial f(y_i)$ and $\|y_i - x_{t-K}\| \leq \frac{K}{p} \leq \delta$ for any $i \in [t-K+1, t]$. Thus,

$$\mathbb{E}[g(y_i)|x_{t-K}] \in \partial f(x_{t-K} + \delta B).$$

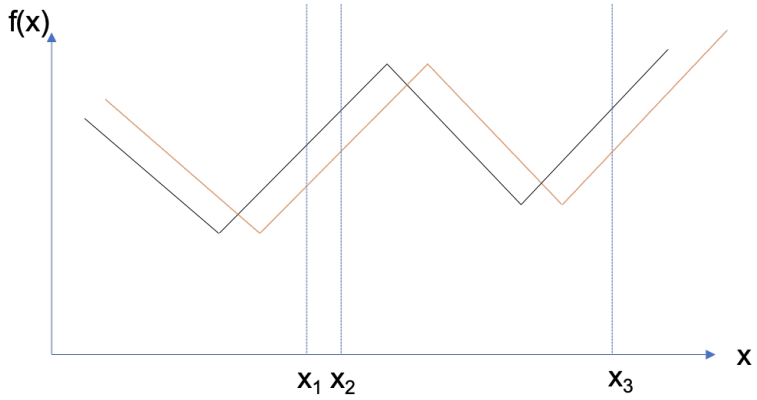

Figure 2.

Consequently, the convex combination

$$\frac{1}{\sum \alpha_i} \sum_{i=t-K+1}^{t} \alpha_i \mathbb{E}[g(y_i)|x_{t-K}] \in \partial f(x_{t-K} + \delta B).$$

Note that $\sum \alpha_i = 1 - \beta^K$, the above inclusion could be rewritten as

$$\frac{1}{1-\beta^K}(\mathbb{E}[m_t|x_{t-K}] - \beta^K m_{t-K}) \in \partial f(x_{t-K} + \delta B).$$

This implies that conditioned on x_{t-K}

$$d(0, \partial f(x_{t-K} + \delta B)) \leq \frac{1}{1-\beta^K} \left(\|\mathbb{E}[m_t \mid x_{t-K}]\| + \beta^K \|m_{t-K}\| \right) \leq \frac{1}{1-\beta^K} \left(\mathbb{E}[\|m_t\||x_{t-K}] + \beta^K \|m_{t-K}\| \right).$$

Therefore, by taking the expectation,

$$\mathbb{E}[d(0, \partial f(x_{t-K} + \delta B))] \leq \frac{1}{1-\beta^K} \left(\mathbb{E}[\|m_t\|] + \beta^K G \right) \leq \frac{1}{1-\frac{1}{16}} (\mathbb{E}[\|m_t\|] + \frac{\epsilon}{16}) = \frac{16}{15} \mathbb{E}[\|m_t\|] + \frac{\epsilon}{15}.$$

Finally, averaging over $t = 1$ to T yields,

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[d(0, \partial f(x_{t-K} + \delta B))] \leq \frac{16}{15T} \sum_{t=1}^{T} \mathbb{E}[\|m_t\|] + \frac{\epsilon}{15} \leq \frac{\epsilon}{3}$$

When $t < K$, $\partial f(x_{t-K} + \delta B)$ simply means $\partial f(x_1 + \delta B)$. As a result, if we randomly out put $x_{\max\{1,t-K\}}$ among $t \in [1, T]$, then with at least probability $2/3$, the δ-subdifferential set contains an element with norm smaller than ϵ. To achieve $1 - \gamma$ probability result for arbitrary γ, it suffices to repeat the algorithm $\log(1/\gamma)$ times.

□

E. Proof of Theorem 11

Proof. The proof idea is similar to Proof of Theorem 5. Since the algorithm does not have access to function value, our resisting strategy now always returns

$$\nabla f(x) = 1.$$

If we can prove that for any set of points $x_k, k \in [1, K], K \leq \frac{\Delta}{8\delta}$, there exists two one dimensional functions such that they satisfy the resisting strategy $\nabla f(x_k) = 1, k \in [1, K]$, and that the two functions do not have two stationary points that are

δ close to each other, then we know no randomized/deterministic can return an $(\delta, \epsilon)-$stationary points with probability more than $1/2$ for both functions simultaneously. In other word, no algorithm that query K points can distinguish these two functions. Hence we proved the theorem following the definition of complexity in (5).

From now on, let $x_k, k \in [1, K]$ be the sequence of points queried after sorting in ascending order. Below, we construct two functions such that $\nabla f(x_k) = 1, k \in [1, K]$, and that the two functions do not have two stationary points that are δ close to each other. Assume WLOG that x_k are ascending. First, we define $f : \mathbb{R} \to \mathbb{R}$ as follows:

$$
\begin{aligned}
&f(x_0) = 0,\\
&f'(x) = -1 \quad \text{if} \quad x \le x_1 - 2\delta,\\
&f'(x) = 1 \quad \text{if exists } i \in [K] \text{ such that} \quad |x - x_i| \le 2\delta,\\
&f'(x) = -1 \quad \text{if exists } i \in [K] \text{ such that} \quad x \in [x_i + 2\delta, \frac{x_i + x_{i+1}}{2}],\\
&f'(x) = 1 \quad \text{if exists } i \in [K] \text{ such that} \quad x \in [\frac{x_i + x_{i+1}}{2}, x_{i+1} - 2\delta],\\
&f'(x) = 1 \quad \text{if} \quad x \ge x_K + 2\delta
\end{aligned}
$$

A schematic picture is shown in Figure 2. It is clear that this function satisfies the resisting strategy. It also has stationary points that are at least 4δ apart. Therefore, simply by shifting the function by 1.5δ, we get the second function.

The only thing left to check is that $\sup_k f(x_k) - \inf_x f(x) \le \Delta$. By construction, we note that the value from x_i to x_{i+1} is non decreasing and increase by at most 4δ

$$\sup_k f(x_k) - f(x_0) \le 4\delta K \le \Delta/2. \tag{17}$$

We further notice that the global minimum of the function is achieved at $x_0 - 2\delta$, and $f(x_0 - 2\delta) = -2\delta \le 4\delta K \le \Delta/2$. Combined with (17), we get,

$$\sup_k f(x_k) - \inf_x f(x) \le \Delta. \tag{18}$$

□