1 Convergence Guarantees

Assumption 1. The functions ℓ, and g are continuously differentiable and both have L-Lipschitz continuous gradients. Additionally, ℓ is lower-bounded; or equivalently, we can assume that $\ell(x) \to \infty$ when $|x| \to \infty$.

Assumption 2. The Mangasarian-Fromovitz constraint qualification (MFCQ) is satisfied for all $x \in \mathbb{R}^n$, i.e. $\forall x \in \mathbb{R}^n$, $\exists w \in \mathbb{R}^d$ s.t. $\nabla g_i(x) w > 0$ for all $i \in I(x)$, where $I(x) = \{i \in \mathbb{Z} \mid g_i(x) \leq 0\}$.

Note. MFCQ is automatically satisfied for all $x \in \mathbb{R}^n$ with $\psi^i(x^i) \neq 0$ for all $i = 1, 2, \ldots, n$.

Assumption 3. The step sizes $(\gamma_k)_{k\geq 0}$ is given by $\gamma_k = \gamma_0/(k+1)$, where γ_0 is a positive constant.

Additionally, we state the following assumptions for the property of the mini-batch of data samples drawn from the dataset, ensuring an unbiased stochastic gradient with bounded variance.

Assumption 4. Suppose that $\pi(\cdot)$ is the data distribution and ξ represents samples of data drawn in a mini-batch. We also denote $\pi(\xi)$ as the probability density function of ξ defined on the probability space $\mathtt{Z}$. We have

$$\ell(x) = \int_{\mathtt{Z}} \ell(x;\xi)\pi(\xi)\mathrm{d}\xi.$$

We are now ready to state the following assumptions.

(a) The stochastic gradient is unbiased, i.e.

$$\mathbb{E}_{\xi\sim\pi}[\nabla \ell(x;\xi)] = \nabla \ell(x), \forall x \in \mathbb{R}^n.$$

(b) The gradient and the stochastic gradient are bounded, i.e.

$$\mathbb{E}_{\xi\sim\pi}\left[\|\nabla \ell(x;\xi)\|\right] \leq M_\ell^2, \forall x \in \mathbb{R}^n.$$

(c) The stochastic gradient has a bounded variance, i.e.

$$\mathbb{E}_{\xi\sim\pi}\left[\|\nabla \ell(x;\xi) - \nabla \ell(x)\|^2\right] \leq \sigma^2, \forall x \in \mathbb{R}^n.$$

Theorem 1. Let $\gamma_k = 1/(k+1)$. Under Assumption 1, 3, 4, and $0 < \varepsilon \leq \inf_{1\leq i\leq d}\inf_{1\leq j\leq K^i}\left|c_j^i - c_{j+1}^i\right|^4/16$, where $\{c_j^i\}$ are the quantization levels, $\ell(\hat{w}_k)$ converges and $\lim_{k\to\infty} d(\hat{w}_k, \mathcal{Z}_\varepsilon) = 0$ almost surely. Here, $\hat{w}_k = w_{t+N_0}$ with probability $1/(H_k(t+N_0+1))$, where $H_k = \sum_{t=0}^{k-1} 1/(t+N_0+1)$ and $N_0 > 0$ is a sufficiently large integer.

Lemma 1. Under Assumption 1, 2, 4, it holds that $\limsup_{k\to\infty} \mathrm{d}(w_k, C_\varepsilon) = 0$ almost surely.

Lemma 2. Denote $[c_-, c_+]$ the set $C_\varepsilon^i \cap [(c_j^i + c_{j-1}^i)/2, (c_j^i + c_{j+1}^i)/2)$, where C_ε^i is the projection of C_ε on the i-th coordinate. Let $0 < \varepsilon \leq \inf_{1\leq i\leq d}\inf_{1\leq j\leq K^i}\left|c_j^i - c_{j+1}^i\right|^2/4$ and $(c_j^i + c_{j-1}^i)/2 - c_- < \varepsilon_1 < 0, 0 < \varepsilon_2 < (c_j^i + c_{j+1}^i)/2 - c_+$. The following statements on the small perturbations $\varepsilon_1, \varepsilon_2$ are true:

(a) $\left|-\alpha\psi_\varepsilon^i(c_- + \varepsilon_1)/(\psi_\varepsilon'^i(c_- + \varepsilon_1))\right| = O(\varepsilon_1)$;

(b) $\left|-\alpha\psi_\varepsilon^i(c_+ + \varepsilon_2)/(\psi_\varepsilon'^i(c_+ + \varepsilon_2))\right| = O(\varepsilon_2)$;

Proof. We will prove the case of c_- and it is easy to see (by symmetry) that the statement holds for the case of c_+.

Notice that $\psi_\varepsilon^i(c_-) = \varepsilon + (c_- - c_j^i)(c_- - c_{j-1}^i) = 0$.

$\psi_\varepsilon^i(c_- + \varepsilon_1) = \varepsilon + (c_- + \varepsilon_1 - c_j^i)(c_- + \varepsilon_1 - c_{j-1}^i) = \varepsilon_1(2c_- - c_{j-1}^i - c_j^i) + \varepsilon_1^2 < 0$.

Also, $\psi_\varepsilon'^i(c_- + \varepsilon_1) = 2(c_- + \varepsilon_1) - c_{j-1}^i - c_j^i > 0$.

Thus,

$$
\begin{aligned}
\frac{-\alpha\psi_\varepsilon^i(c_- + \varepsilon_1)}{\psi_\varepsilon'^i(c_- + \varepsilon_1)} &= \frac{-\alpha(\varepsilon_1(2c_- - c_{j-1}^i - c_j^i) + \varepsilon_1^2)}{2(c_- + \varepsilon_1) - c_{j-1}^i - c_j^i} \\
&= \frac{-\alpha(1 + \varepsilon_1/(2c_- - c_{j-1}^i - c_j^i))}{1/\varepsilon_1 + 2/(2c_- - c_{j-1}^i - c_j^i)} \\
&\leq \frac{-\alpha\varepsilon_1/(2c_- - c_{j-1}^i - c_j^i)}{2/(2c_- - c_{j-1}^i - c_j^i)} \\
&\leq -\alpha\varepsilon_1/2.
\end{aligned}
$$

□

Lemma 3. (Descent Lemma for `ASkewSGD`) There exists $\varepsilon_1 > 0, K \geq 0$, such that $\forall k \geq K$, $1 - \gamma_k L/2 > 1/2$ and

$$
\mathbb{E}\left[\ell(w_{k+1})|w_k\right] \leq \ell(w_k) - \frac{\gamma_k}{2}\sum_{i\notin I_{k,\varepsilon_1}}\left\|[\nabla\ell(w_k)]^i\right\|^2 + \frac{\gamma_k^2\sigma^2 L}{2} + \gamma_k\sum_{i\in I_{k,\varepsilon_1}} MO(\varepsilon_1) + \gamma_k^2\sum_{i\in I_{k,\varepsilon_1}}\frac{L}{2}O(\varepsilon_1^2),
$$

where $i \notin I_{k,\varepsilon}$ if $w_k^i \in (c_-^i + \varepsilon_1, c_+^i - \varepsilon_1)$ or $v_k^i = -[\widehat{\nabla}\ell(w_k)]^i$.

Proof. By Lemma 1, there exists $\varepsilon_1 > 0$ and K_1 such that $\mathrm{d}(w_k, C_\varepsilon) < \varepsilon_1$ for all $k \geq K_1$.

By Assumption 3, γ_k can reach an arbitrarily small positive value so that $1 - \gamma_k L/2 > 1/2$. Denote the smallest possible k as K_2.

Set $K = \min\{K_1, K_2\}$.

Consider the update rule for the model parameter $w_{k+1} = w_k + \gamma_k v_k$.

By the smoothness of ℓ, we obtain

$$
\ell(w_{k+1}) \leq \ell(w_k) + \gamma_k v_k^\top \nabla\ell(w_k) + \frac{\gamma_k^2 L}{2}\|v_k\|^2.
$$

Taking the conditional expectation $\mathbb{E}\left[\cdot|w_k\right]$, we obtain

$$
\begin{aligned}
\mathbb{E}\left[\ell(w_{k+1})|w_k\right] \leq\ & \ell(w_k) + \gamma_k\mathbb{E}\left[v_k^\top\nabla\ell(w_k)|w_k\right] + \frac{\gamma_k^2 L}{2}\mathbb{E}\left[\|v_k\|^2|w_k\right] \\
\leq\ & \ell(w_k) - \sum_{i\notin I_{k,\varepsilon_1}}\gamma_k\mathbb{E}\left[v_k^i[\nabla\ell(w_k)]^i|w_k\right] + \sum_{i\in I_{k,\varepsilon_1}}\gamma_k\mathbb{E}\left[v_k^i[\nabla\ell(w_k)]^i|w_k\right] \\
& + \sum_{i\notin I_{k,\varepsilon_1}}\frac{\gamma_k^2 L}{2}\mathbb{E}\left[\left\|v_k^i\right\|^2\middle|w_k\right] + \sum_{i\in I_{k,\varepsilon_1}}\frac{\gamma_k^2 L}{2}\mathbb{E}\left[\left\|v_k^i\right\|^2\middle|w_k\right] \\
\leq\ & \ell(w_k) - \sum_{i\notin I_{k,\varepsilon_1}}\gamma_k\mathbb{E}\left[[\widehat{\nabla}\ell(w_k)]^i[\nabla\ell(w_k)]^i\middle|w_k\right] + \sum_{i\in I_{k,\varepsilon_1}}\gamma_k\mathbb{E}\left[v_k^i[\nabla\ell(w_k)]^i\middle|w_k\right] \\
& + \sum_{i\notin I_{k,\varepsilon_1}}\frac{\gamma_k^2 L}{2}\mathbb{E}\left[\left\|[\widehat{\nabla}\ell(w_k)]^i\right\|^2\middle|w_k\right] + \sum_{i\in I_{k,\varepsilon_1}}\frac{\gamma_k^2 L}{2}\mathbb{E}\left[\left\|v_k^i\right\|^2\middle|w_k\right] \\
\leq\ & \ell(w_k) - \sum_{i\notin I_{k,\varepsilon_1}}\gamma_k\left\|[\nabla\ell(w_k)]^i\right\|^2 + \sum_{i\in I_{k,\varepsilon_1}}\gamma_k\mathbb{E}\left[v_k^i[\nabla\ell(w_k)]^i\middle|w_k\right] \\
& + \sum_{i\notin I_{k,\varepsilon_1}}\frac{\gamma_k^2 L}{2}\mathbb{E}\left[\left\|[\widehat{\nabla}\ell(w_k)]^i\right\|^2\middle|w_k\right] + \sum_{i\in I_{k,\varepsilon_1}}\frac{\gamma_k^2 L}{2}\mathbb{E}\left[\left\|v_k^i\right\|^2\middle|w_k\right].
\end{aligned}
$$

Note that

$$\begin{aligned}\mathbb{E}\left[\left\|[\widehat{\nabla}\ell(w_k)]^i\right\|^2\middle|w_k\right] &= \mathbb{E}\left[\left\|[\widehat{\nabla}\ell(w_k)]^i-[\nabla\ell(w_k)]^i+[\nabla\ell(w_k)]^i\right\|^2\middle|w_k\right]\\ &= \mathbb{E}\left[\left\|[\widehat{\nabla}\ell(w_k)]^i-[\nabla\ell(w_k)]^i\right\|^2\middle|w_k\right]\\ &\quad+2[\nabla\ell(w_k)]^i\mathbb{E}\left[[\widehat{\nabla}\ell(w_k)]^i-[\nabla\ell(w_k)]^i\middle|w_k\right]\\ &\quad+\left\|[\nabla\ell(w_k)]^i\right\|^2\\ &\le \sigma^2+\left\|[\nabla\ell(w_k)]^i\right\|^2.\end{aligned}$$

We notice that $i\notin I_{k,\varepsilon_1}$ implies that v_k^i takes $-\alpha\psi_\varepsilon^i(w_k^i)/(\psi_\varepsilon'^i(w_k^i))$, and following this hereby, we simply use the notation v_k^i to denote the pushing force. Thus,

$$\begin{aligned}\mathbb{E}\left[\ell(w_{k+1})|w_k\right] \le\;& \ell(w_k)-(\gamma_k-\frac{\gamma_k^2L}{2})\sum_{i\notin I_{k,\varepsilon_1}}\left\|[\nabla\ell(w_k)]^i\right\|^2+\sum_{i\in I_{k,\varepsilon_1}}\gamma_k\mathbb{E}\left[v_k^i[\nabla\ell(w_k)]^i\middle|w_k\right]\\ &+\sum_{i\in I_{k,\varepsilon_1}}\frac{\gamma_k^2L}{2}\mathbb{E}\left[\left\|v_k^i\right\|^2\middle|w_k\right]+\frac{\gamma_k^2\sigma^2L}{2}\\ \le\;& \ell(w_k)-\frac{\gamma_k}{2}\sum_{i\notin I_{k,\varepsilon_1}}\left\|[\nabla\ell(w_k)]^i\right\|^2+\sum_{i\in I_{k,\varepsilon_1}}\gamma_k\mathbb{E}\left[\left|v_k^i\right|\left|[\nabla\ell(w_k)]^i\right|\middle|w_k\right]\\ &+\sum_{i\in I_{k,\varepsilon_1}}\frac{\gamma_k^2L}{2}\mathbb{E}\left[\left\|v_k^i\right\|^2\middle|w_k\right]+\frac{\gamma_k^2\sigma^2L}{2}\\ \le\;& \ell(w_k)-\frac{\gamma_k}{2}\sum_{i\notin I_{k,\varepsilon_1}}\left\|[\nabla\ell(w_k)]^i\right\|^2+\sum_{i\in I_{k,\varepsilon_1}}\gamma_kM_\ell O(\varepsilon_1)\\ &+\sum_{i\in I_{k,\varepsilon_1}}\frac{\gamma_k^2L}{2}O(\varepsilon_1^2)+\frac{\gamma_k^2\sigma^2L}{2}.\end{aligned}$$

□

Lemma 4. (Telescoping Sum Argument) Let $T>0$. There exists $\varepsilon_1>0, K\ge 0$, such that $\forall k\ge K$,

$$\begin{aligned}&\sum_{k=0}^{T-1}\frac{\gamma_0}{2(k+K-1)}\left(\sum_{i\notin I_{k+K,\varepsilon}}\mathbb{E}\left[\left\|[\nabla\ell(w_{k+K})]^i\right\|^2\right]+\sum_{i\in I_{k+K,\varepsilon}}\mathbb{E}\left[\left\|v_{k+K}\right\|^2\right]\right)\\ \le\;& \mathbb{E}[\ell(w_K)-\ell(w_{K+T})]-\sum_{k=0}^{T-1}\frac{\gamma_0^2\sigma^2L}{2(k+K+1)^2}+\sum_{k=0}^{T-1}\frac{\gamma_0}{k+K+1}O(\varepsilon_1).\end{aligned}$$

Proof. Continuing from Lemma 3, we now have

$$\mathbb{E}\left[\ell(w_{k+1})|w_k\right] \le \ell(w_k)-\frac{\gamma_k}{2}\sum_{i\notin I_{k,\varepsilon_1}}\left\|[\nabla\ell(w_k)]^i\right\|^2+\gamma_kO(\varepsilon_1)+\frac{\gamma_k^2\sigma^2L}{2}.$$

Taking the full expectation of the last inequality, we have

$$\mathbb{E}\left[\ell(w_{k+1})\right] \le \mathbb{E}\left[\ell(w_k)\right]-\frac{\gamma_k}{2}\mathbb{E}\left[\sum_{i\notin I_{k,\varepsilon_1}}\left\|[\nabla\ell(w_k)]^i\right\|^2\right]+\gamma_kO(\varepsilon_1)+\frac{\gamma_k^2\sigma^2L}{2},$$

$$\frac{\gamma_k}{2}\mathbb{E}\left[\sum_{i\notin I_{k,\varepsilon_1}} \left\|[\nabla \ell(w_k)]^i\right\|^2\right] \le \mathbb{E}\left[\ell(w_k)\right] - \mathbb{E}\left[\ell(w_{k+1})\right] + \gamma_k O(\varepsilon_1) + \frac{\gamma_k^2\sigma^2 L}{2}$$

$$\frac{\gamma_k}{2}\mathbb{E}\left[\sum_{i\notin I_{k,\varepsilon_1}} \left\|[\nabla \ell(w_k)]^i\right\|^2\right] + \frac{\gamma_k}{2}\mathbb{E}\left[\sum_{i\in I_{k,\varepsilon_1}} \left\|v_k^i\right\|^2\right] \le \mathbb{E}\left[\ell(w_k)\right] - \mathbb{E}\left[\ell(w_{k+1})\right] + \gamma_k O(\varepsilon_1) + \frac{\gamma_k^2\sigma^2 L}{2}.$$

Summing over T epochs and rearranging the terms, we now have

$$\begin{aligned}
&\sum_{k=0}^{T-1}\frac{\gamma_{k+K}}{2}\left(\sum_{i\notin I_{k+K,\varepsilon}}\mathbb{E}\left[\left\|[\nabla \ell(w_{k+K})]^i\right\|^2\right] + \sum_{i\in I_{k+K,\varepsilon}}\mathbb{E}\left[\left\|v_{k+K}\right\|^2\right]\right)\\
\le\;& \mathbb{E}[\ell(w_K) - \ell(w_{K+T})] - \sum_{k=0}^{T-1}\frac{\gamma_{k+K}^2\sigma^2 L}{2} + \sum_{k=0}^{T-1}\gamma_k O(\varepsilon_1).
\end{aligned}$$

Substituting $\gamma_{k+K} = \gamma_0/(k+K-1)$ concludes the proof. □

Lemma 5. (Heavy Tail Substitution) There exists $\varepsilon_1 > 0, K \ge 0$, such that when $T \to \infty$, $\sum_{i\notin \bar{I}_\varepsilon}\mathbb{E}\left[\left\|[\nabla \ell(\hat{w}_T)]^i\right\|^2\right] + \sum_{i\in \bar{I}_\varepsilon}\mathbb{E}\left[\left\|\hat{v}_T\right\|^2\right] \to O(\varepsilon_1)$, where $\hat{w}_T = w_{t+K}, \hat{v}_T = v_{t+K}$ with probability $1/(H_T(t+K+1))$, $\bar{I}_\varepsilon$ is dependent on $\hat{w}_T, \hat{v}_T$, and $H_T = \sum_{i=0}^{T-1} 1/(i+K+1)$.

Proof.

$$\mathbb{E}\left[\left\|[\nabla \ell(\hat{w}_T)]^i\right\|^2\right] = \sum_{t=0}^{T-1}\mathbb{P}(\hat{w}_T = w_{t+K})\mathbb{E}\left[\left\|[\nabla \ell(w_{t+K})]^i\right\|^2\right] = \sum_{t=0}^{T-1}\frac{\mathbb{E}\left[\left\|[\nabla \ell(w_{t+K})]^i\right\|^2\right]}{H_T(t+K+1)}$$

Similarly,

$$\mathbb{E}\left[\left\|\hat{v}_T^i\right\|^2\right] = \sum_{t=0}^{T-1}\mathbb{P}(\hat{v}_T = v_{t+K})\mathbb{E}\left[\left\|v_T^i\right\|^2\right] = \sum_{t=0}^{T-1}\frac{\mathbb{E}\left[\left\|v_T^i\right\|^2\right]}{H_T(t+K+1)}$$

Therefore,

$$\begin{aligned}
&\sum_{i\notin \bar{I}_\varepsilon}\mathbb{E}\left[\left\|[\nabla \ell(\hat{w}_T)]^i\right\|^2\right] + \sum_{i\in \bar{I}_\varepsilon}\mathbb{E}\left[\left\|\hat{v}_T\right\|^2\right]\\
\le\;& \frac{2}{\gamma_0 H_T}\left(\mathbb{E}[\ell(w_K) - \ell(w_{K+T})] - \sum_{k=0}^{T-1}\frac{\gamma_0^2\sigma^2 L}{2(k+K+1)^2} + \sum_{k=0}^{T-1}\frac{\gamma_0}{k+K+1}O(\varepsilon_1)\right).
\end{aligned}$$

The above inequality implies that for some constant $C > 0$,

$$\sum_{i\notin \bar{I}_\varepsilon}\mathbb{E}\left[\left\|[\nabla \ell(\hat{w}_T)]^i\right\|^2\right] + \sum_{i\in \bar{I}_\varepsilon}\mathbb{E}\left[\left\|\hat{v}_T\right\|^2\right] \le \frac{C}{\log T} + O(\varepsilon_1).$$

Thus, when $T \to \infty$,

$$\sum_{i\notin \bar{I}_\varepsilon}\mathbb{E}\left[\left\|[\nabla \ell(\hat{w}_T)]^i\right\|^2\right] + \sum_{i\in \bar{I}_\varepsilon}\mathbb{E}\left[\left\|\hat{v}_T\right\|^2\right] \to O(\varepsilon_1)$$

and this concludes the theorem.

Note that the update step also goes to $O(\varepsilon_1)$ and this implies the convergence of the cost function

$\ell(\hat{w}_k)$. $\square$