# Supplementary Material for "Collect at Once, Use Effectively: Making Non-interactive Locally Private Learning Possible"

## 1. Omitted Proofs in Section 3

**Lemma 1** (Lemma 3 in Main Body)**.** *Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n \sim$ i.i.d.$\mathcal{D}$ with $\boldsymbol{\mu} = \mathbb{E}_{\mathcal{D}}[\boldsymbol{x}]$ and supp$(\mathcal{D}) \subseteq \mathcal{B}(0,1)$. Let $G$ and $\{\boldsymbol{y}_i\}_{i=1}^n$ defined in the above procedure. For each of group $S_j$ fixed, we have the following with probability $2/3$:*

$$\Big\| \frac{1}{|S_j|} \sum_{\boldsymbol{y}_i \in S_j} \boldsymbol{y}_i - G\boldsymbol{\mu} \Big\|_1 \le O\left( \frac{p \log(nd)}{\epsilon \sqrt{|S_j|}} \right) \tag{1}$$

*Proof.* Apparently $\frac{1}{|S_j|}\sum_{i\in S_j} \boldsymbol{r}_i \sim \mathcal{N}(0, \frac{2\log(1.25/\delta)}{|S_j|\epsilon^2} I_d)$. So we have $\|\frac{1}{|S_j|}\sum_{i\in S_j}\boldsymbol{r}_i\|_1 \le O\left(\frac{p\log n}{\epsilon\sqrt{|S_j|}}\right)$ with probability $\frac{1}{9}$. We then turn to bound the loss incurred by random sample of data.

$$\begin{aligned} &\mathbb{E}\Big\| \boldsymbol{\mu} - \frac{1}{|S_j|}\sum_{i\in S_j} \boldsymbol{x}_i \Big\|^2 = \frac{1}{|S_j|}\sum_{l=1}^d \mathrm{var}(x_{1l}) \\ \le & \frac{1}{|S_j|}\sum_{l=1}^d \mathbb{E}[x_{1l}^2] \le \frac{1}{|S_j|}. \end{aligned} \tag{2}$$

According to Markov Inequality, we have

$$\mathcal{P}\left\{ \Big\|\boldsymbol{\mu} - \frac{1}{|S_j|}\sum_{i\in S_j}\boldsymbol{x}_i\Big\|^2 \ge \frac{9}{|S_j|} \right\} \le \frac{1}{9}$$

Given $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n$ fixed under this event, we can easily derive upper bounds on entries of $G(\boldsymbol{\mu} - \frac{1}{|S_j|}\sum_{i\in S_j}\boldsymbol{x}_i)$: for $\boldsymbol{g} \sim \mathcal{N}(0, I_d)$ and $\boldsymbol{q} = \boldsymbol{\mu} - \frac{1}{|S_j|}\sum_{i\in S_j}\boldsymbol{x}_i$, we have $|\boldsymbol{g}^T\boldsymbol{q}| \le 12\sqrt{\frac{\log d}{|S_j|}}$ with probability $1 - \frac{1}{9d}$. By union bound we have the following with probability $\frac{2}{9}$:

$$\Big\| G(\boldsymbol{\mu} - \frac{1}{|S_j|}\sum_{i\in S_j}\boldsymbol{x}_i)\Big\|_1 \le O\left(\sqrt{\frac{p\log d}{|S_j|}}\right).$$

Putting the two inequalities together using union bound, we get the result. □

**Lemma 2** (Lemma 6 in Main Body)**.** *Under the assumptions made in Section 3.2, given projection matrix $\Phi$, with high probability over the randomness of private mechanism, we have*

$$\bar{L}(\boldsymbol{w}^{priv}; \bar{X}, \boldsymbol{y}) - \bar{L}(\hat{\boldsymbol{w}}^*; \bar{X}, \boldsymbol{y}) \leqslant \tilde{O}\left(\sqrt{\frac{m}{n\epsilon^2}}\right) \tag{3}$$

*Proof.* Note, once we prove the uniform convergence of $|\hat{L}(\boldsymbol{w}; Z, \boldsymbol{v}) - \bar{L}(\boldsymbol{w}; \bar{X}, \boldsymbol{y})| \leqslant O\left(\sqrt{\frac{m}{n\epsilon^2}}\right)$ for any $\boldsymbol{w} \in \mathcal{C}$, then the conclusion holds directly. Now, we will prove the uniform convergence. Note $Z = \bar{X} + E$, where $E \in \mathbb{R}^{n\times m}$, and each entry $e_{ij} \sim \mathcal{N}(0, \sigma^2)$, $\boldsymbol{v} = \boldsymbol{y} + \boldsymbol{r}$, where $\boldsymbol{r} \sim \mathcal{N}(0, \sigma^2 I_n)$. Denote $\bar{\boldsymbol{w}} = \Phi^T\boldsymbol{w}$.

$$\begin{aligned} &\left|\hat{L}(\boldsymbol{w}; Z, \boldsymbol{v}) - \bar{L}(\boldsymbol{w}; \bar{X}, \boldsymbol{y})\right| \\ =&\left| \frac{1}{2n}\bar{\boldsymbol{w}}^T(Q - \bar{X}^T\bar{X})\bar{\boldsymbol{w}} - \frac{1}{n}\left(\boldsymbol{v}^T Z\bar{\boldsymbol{w}} - \boldsymbol{y}^T\bar{X}\bar{w}\right)\right| \\ \leqslant& \frac{1}{2n}\left\|Q - \bar{X}^T\bar{X}\right\|_2 \|\bar{\boldsymbol{w}}\|_2^2 + \frac{1}{n}\left|\boldsymbol{v}^T Z\bar{\boldsymbol{w}} - \boldsymbol{y}^T\bar{X}\bar{\boldsymbol{w}}\right| \\ \leqslant& \frac{1}{2n}\left\|Q - \bar{X}^T\bar{X}\right\|_F \|\bar{\boldsymbol{w}}\|_2^2 + \frac{1}{n}|\boldsymbol{v}^T Z\bar{\boldsymbol{w}} - \boldsymbol{y}^T\bar{X}\bar{\boldsymbol{w}}| \\ \leqslant& \tfrac{1}{2n}\left\|Z^TZ - n\sigma^2 I_m - \bar{X}^T\bar{X}\right\|_F \|\bar{\boldsymbol{w}}\|_2^2 + \tfrac{1}{n}|\boldsymbol{v}^T Z\bar{\boldsymbol{w}} - \boldsymbol{y}^T\bar{X}\bar{\boldsymbol{w}}| \\ \leqslant& \frac{1}{2n}\left\|E^TE - n\sigma^2 I_m\right\|_F \|\bar{\boldsymbol{w}}\|_2^2 + \frac{1}{n}\left\|\bar{X}^TE\right\|_F \|\bar{\boldsymbol{w}}\|_2^2 + \\ & \frac{1}{n}\left(\left\|E^T\boldsymbol{y}\right\|_2 + \left\|\bar{X}^T\boldsymbol{r}\right\|_2 + \left\|E^T\boldsymbol{r}\right\|_2\right)\|\bar{\boldsymbol{w}}\|_2 \end{aligned}$$

From the property of random projection, we know $\|\bar{\boldsymbol{w}}\|_2 \leqslant 1$ with high probability. Besides, as each entry in $E$ is i.i.d. Gaussian, and $\mathbb{E}[E^TE] = n\sigma^2 I_m$, thus we have $\frac{1}{2n}\left\|E^TE - n\sigma^2 I_m\right\|_2 \leqslant O\left(\sigma\sqrt{\frac{\log m}{n}}\right)$ with high probability according to lemma 3, hence $\frac{1}{2n}\left\|E^TE - n\sigma^2 I_m\right\|_F \leqslant O(\sigma\sqrt{\frac{m\log m}{n}})$ with high probability.

As $\frac{1}{n^2}\left\|\bar{X}^TE\right\|_F^2 = \frac{1}{n^2}\sum_{j=1}^m(\sum_{i=1}^m(\boldsymbol{q}_j^T\boldsymbol{e}_i)^2)$, where $\boldsymbol{q}_j, \boldsymbol{e}_i$ are the $j$-th and $i$-th column of $\bar{X}$ and $E$ respectively. For each $j \in [m]$, $\frac{1}{n^2}\sum_{i=1}^m(\boldsymbol{q}_j^T\boldsymbol{e}_i)^2$ obeys Chi-square distribution (with some scaling), thus with high probability, $\frac{1}{n^2}\sum_{i=1}^m(\boldsymbol{q}_j^T\boldsymbol{e}_i)^2 \leqslant O\left(\frac{m\|\boldsymbol{q}_j\|^2\sigma^2}{n^2}\right)$. Therefore, by union bound, we have $\frac{1}{n^2}\sum_{j=1}^m(\sum_{i=1}^m(\boldsymbol{q}_j^T\boldsymbol{e}_i)^2) \leqslant O\left(\frac{m\sum_j\|\boldsymbol{q}_j\|^2\sigma^2}{n^2}\right) = O\left(\frac{m\sigma^2}{n}\right)$, as $\sum_j\|\boldsymbol{q}_j\|^2 =$

$\left\|\bar{X}\right\|_F^2 \leqslant n$. Hence, there is $\frac{1}{n}\left\|\bar{X}^T E\right\|_F \leqslant O\left(\sqrt{\frac{m\sigma^2}{n}}\right)$ with high probability. Using similar augument, we have $\frac{1}{n}\left\|\bar{E}^T \boldsymbol{y}\right\|_2 \leqslant O\left(\sqrt{\frac{m\sigma^2}{n}}\right)$, $\frac{1}{n}\left\|\bar{E}^T \boldsymbol{r}\right\|_2 \leqslant O\left(\sqrt{\frac{m\sigma^2}{n}}\right)$ with high probability. For $\frac{1}{n}\left\|\bar{X}^T r\right\|$, according to matrix concentration inequality (Theorem 4.1.1 in (Tropp et al., 2015)), we have $\frac{1}{n}\left\|\bar{X}^T \boldsymbol{r}\right\|_2 \leqslant O\left(\frac{1}{\sqrt{n}}\right)$.

Combine all these results together, we obtain the desired conclusion. □

**Lemma 3** ((Vershynin, 2009)). *Suppose $\boldsymbol{x} \in \mathbb{R}^d$ be a random vector satisfies $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T] = I_d$. Denote $\|\boldsymbol{x}\|_{\phi_1} = M$, where $\|\cdot\|_{\psi_1}$ represents Orlicz $\psi_1$-norm. Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be independent copies of $\boldsymbol{x}$, then for every $\epsilon \in (0,1)$, we have*

$$\boldsymbol{Pr}\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^T - I_d\right\|_2 > \epsilon\right) \leqslant de^{-n\epsilon^2/4M^2}$$

**Theorem 1** (Theorem 3 in Main Body). *Under the assumption in this section, set $m = \Theta\left(\sqrt{n\epsilon^2 \log d}\right)$ for $\beta > 0$, then with high probability , there is*

$$L(\boldsymbol{w}^{priv}) - L(\boldsymbol{w}^*) = \tilde{O}\left(\left(\frac{\log d}{n\epsilon^2}\right)^{1/4}\right)$$

*Proof.* On one hand,

$$\begin{aligned}
& L(\boldsymbol{w}^{priv}) - L(\boldsymbol{w}^*) \\
=& L(\boldsymbol{w}^{priv}) - \bar{L}(\boldsymbol{w}^{priv}) + \bar{L}(\boldsymbol{w}^{priv}) - \bar{L}(\hat{\boldsymbol{w}}^*) \\
& + \bar{L}(\hat{\boldsymbol{w}}^*) - \bar{L}(\boldsymbol{w}^*) + \bar{L}(\boldsymbol{w}^*) - L(\boldsymbol{w}^*) \\
\leqslant & \left[L(\boldsymbol{w}^{priv}) - \bar{L}(\boldsymbol{w}^{priv}) + \bar{L}(\boldsymbol{w}^*) - L(\boldsymbol{w}^*)\right] \\
& + \bar{L}(\boldsymbol{w}^{priv}) - \bar{L}(\hat{\boldsymbol{w}}^*) \\
\leqslant & G[\max_i\{|\langle \boldsymbol{w}^{priv}, \boldsymbol{x}_i\rangle - \langle \Phi^T\boldsymbol{w}^{priv}, \Phi^T\boldsymbol{x}_i\rangle|\} \\
& + \max_i\{|\langle \boldsymbol{w}^*, \boldsymbol{x}_i\rangle - \langle \Phi^T\boldsymbol{w}^*, \Phi^T\boldsymbol{x}_i\rangle|\}] \\
& + [\bar{L}(\boldsymbol{w}^{priv}) - \bar{L}(\hat{\boldsymbol{w}}^*)] \qquad (4) \\
& \text{(where } G \text{ is the Lipschitz constant)}
\end{aligned}$$

On the other hand, for $\forall \boldsymbol{w} \in \mathcal{C}, \forall \boldsymbol{x} \in D$, there is

$$\begin{aligned}
& |\langle \boldsymbol{w}, \boldsymbol{x}\rangle - \langle \Phi^T\boldsymbol{w}, \Phi^T\boldsymbol{x}\rangle| \\
=& \left|\frac{\left\|\Phi^T(\boldsymbol{w}+\boldsymbol{x})\right\|_2^2 - \left\|\Phi^T(\boldsymbol{w}-\boldsymbol{x})\right\|_2^2}{4} - \frac{\|\boldsymbol{w}+\boldsymbol{x}\|_2^2 - \|\boldsymbol{w}-\boldsymbol{x}\|_2^2}{4}\right| \\
\leqslant & \left|\frac{\left\|\Phi^T(\boldsymbol{w}+\boldsymbol{x})\right\|_2^2 - \|\boldsymbol{w}+\boldsymbol{x}\|_2^2}{4}\right| + \left|\frac{\left\|\Phi^T(\boldsymbol{w}-\boldsymbol{x})\right\|_2^2 - \|\boldsymbol{w}-\boldsymbol{x}\|_2^2}{4}\right|
\end{aligned}$$

According to the results of random projection w.r.t. additive error (Dirksen, 2016), we know with high probability, there is $|\langle \boldsymbol{w}, \boldsymbol{x}\rangle - \langle \Phi^T\boldsymbol{w}, \Phi^T\boldsymbol{x}\rangle| \leqslant O\left(\sqrt{\frac{\log d}{m}}\right)$, for $\forall \boldsymbol{w} \in \mathcal{C}, \forall \boldsymbol{x} \in D$. Therefore, the first term in equation (4) is less than $O\left(\sqrt{\frac{\log d}{m}}\right)$.

From lemma 2, we know $\bar{L}(\bar{\boldsymbol{w}}^{priv}) - \bar{L}(\bar{\boldsymbol{w}}^*) \leqslant \tilde{O}\left(\sqrt{\frac{m}{n\epsilon^2}}\right)$ holds with high probability. Combine these two inequalities, it is easy to determine the optimal $m$, then obtain the conclusion. □

**Corollary 1** (Corollary 2 in Main Body). *Algorithm LDP kernel mechanism satisfies $(\epsilon, \delta)$-LDP, and with high probability, there is*

$$L_{\hat{H}}(\hat{\boldsymbol{w}}^{priv}) - L_H(f^*) \leqslant \tilde{O}\left(\left(\frac{d}{n\epsilon^2}\right)^{1/4}\right)$$

$$\sup_{\boldsymbol{x}\in\mathcal{X}} |\Phi(\boldsymbol{x})^T f^* - (\hat{\Phi}(\boldsymbol{x}))^T\hat{\boldsymbol{w}}^{priv}| \leqslant \tilde{O}\left(\left(\frac{d}{n\epsilon^2}\right)^{1/8}\right)$$

*Proof.* Algorithm satisfies local privacy is obvious. For excess risk, as $L_{\hat{H}}(\hat{\boldsymbol{w}}^{priv}) - L_H(f^*) = L_{\hat{H}}(\hat{\boldsymbol{w}}^{priv}) - L_{\hat{H}}(g^*) + L_{\hat{H}}(g^*) - L_H(f^*)$, follow nearly the same proof of lemma 5 of sparse linear regression, we have $L_{\hat{H}}(\hat{\boldsymbol{w}}^{priv}) - L_{\hat{H}}(g^*) \leqslant \tilde{O}\left(\sqrt{\frac{d_p}{n\epsilon^2}}\right)$. On the other hand, nearly borrow the proof of Lemma 17 in (Rubinstein et al., 2012) and property of RRF , we have

$$L_{\hat{H}}(g^*) - L_H(f^*) \leqslant \tilde{O}\left(\sqrt{\frac{d}{d_p}}\right)$$

Combine above two inequalities, and choose optimal $d_p$ as $\tilde{O}\left(\sqrt{dn\epsilon^2}\right)$, we obtain the first inequality of the conclusion. Then combine lemma 7 in this paper, it is easy to obtaint the second inequality. □

## 2. Omitted contents and proofs in Section 4

### 2.1. Relations between smooth generalized linear losses (SGLL) and generalized linear models (GLM)

Note that a model is called GLM, if for $\boldsymbol{x}, \boldsymbol{w}^* \in \mathbb{R}^d$, label $y$ with respect to $\boldsymbol{x}$ is given by a distribution which belongs to the exponential family:

$$p(y|\boldsymbol{x}, \boldsymbol{w}^*) = \exp\left(\frac{y\theta - b(\theta)}{\Phi} + c(y, \Phi)\right) \qquad (5)$$

where $\theta, \Phi$ are parameters, and $b(\theta), c(y, \Phi)$ are known functions. Besides, there is an one-to-one continuous differentiable transformation $g(\cdot)$ such that $g(b'(\theta)) = \boldsymbol{x}^T\boldsymbol{w}^*$.

According to the key equality $g(b'(\theta)) = \boldsymbol{x}^T\boldsymbol{w}^*$, usually we can obtain smooth function $\theta = h_1(\boldsymbol{x}^T\boldsymbol{w}^*), b(\theta) = h_2(\boldsymbol{x}^T\boldsymbol{w}^*)$, and what's more, univariate function

$h_i(x)(i = 1, 2)$ satisfies the absolutely smooth property.

For such GLM, if we consider optimizing the expected negative logarithmic probability $-\mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \log p(\boldsymbol{x}, y; \boldsymbol{w})$, once discarding unrelated terms to $\boldsymbol{w}$, we obtain the new population loss, $L(\boldsymbol{w}) := \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \ell(\boldsymbol{w}; \boldsymbol{x}, y)$, where $\ell(\boldsymbol{w}; \boldsymbol{x}, y) = -yh_1(\boldsymbol{x}^T\boldsymbol{w}) + h_2(\boldsymbol{x}^T\boldsymbol{w})$, exactly the form of smooth generalized linear loss defined in section 4. Hence our SGLL is a natural loss defined by GLM with additional smoothness assumptions.

### 2.2. Omitted proofs

**Lemma 4** (Lemma 8 in Main Body)**.** *Given any $\alpha > 0$, by setting $k = c \ln \frac{1}{\alpha}, p = \lceil k + e\mu_2(k; r)\rceil$, where $c$ is a constant, we have $\left\|\hat{f}_p(x) - f(x)\right\|_\infty \leqslant \alpha$.*

*Proof.* As $f, f', \cdots, f^{(k-1)}$ are absolutely continuous over $[-1, 1]$, and $\left\|f^{(k)}\right\|_T \leqslant \mu_1(k; r)\mu_2(k; r)^k$, according to the results in (Trefethen, 2008), we have

$$\begin{aligned} \left\|\hat{f}_p(x) - f(x)\right\|_\infty \leqslant & \frac{2\left\|f^{(k)}\right\|_T}{\pi k(p-k)^k} \\ \leqslant & \frac{2\mu_1(k; r)}{\pi k e^k} \end{aligned} \tag{6}$$

It is easy to see there exists $c > 0$, such that the term (6) is less than $\alpha$ with chosen $k$, hence the conclusion holds. □

**Lemma 5** (Lemma 9 in Main Body)**.** *For any $\gamma > 0$, setting $k = c \ln \frac{4r}{\gamma}, p = \lceil k + 2\mu_2(k; r)\rceil$, then algorithm 7 outputs a $(\gamma, \beta, \sigma)$ stochastic oracle, where $\sigma = \tilde{O}\left(\sigma_0 + \gamma + \frac{p^{2p+1}(4r)^{p+1}}{\epsilon^{p+2}}\right)$.*

*Proof.* According to lemma 4, we know the approximation error, $|\hat{m}(\boldsymbol{w}; \boldsymbol{x}, y) - m(\boldsymbol{w}; \boldsymbol{x}, y)| \leqslant \frac{\gamma}{2r}$. For any fixed $(\boldsymbol{x}, y)$, from the construction of stochastic inexact gradient oracle, there is $\mathbb{E}[\tilde{G}(\boldsymbol{w}; b)|\boldsymbol{x}, y] = \hat{G}(\boldsymbol{w}; \boldsymbol{x}, y)$. Denote $\hat{g}(\boldsymbol{w}) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[\hat{G}(\boldsymbol{w}; \boldsymbol{x}, y)]$, thus we have

$$\begin{aligned} \mathbb{E}\left[\left\|\tilde{G}(\boldsymbol{w}; b) - \hat{g}(\boldsymbol{w})\right\|^2\right] = & \mathbb{E}\left[\left\|\tilde{G}(\boldsymbol{w}; b) - \hat{G}(\boldsymbol{w}; \boldsymbol{x}, y)\right\|^2\right] \\ & + \mathbb{E}\left[\left\|\hat{G}(\boldsymbol{w}; \boldsymbol{x}, y) - \hat{g}(\boldsymbol{w})\right\|^2\right] \end{aligned}$$

For above two terms, combined with results given in lemma 6, we we obtain

$$\mathbb{E}\left[\left\|\tilde{G}(\boldsymbol{w}; b) - g(\boldsymbol{w})\right\|^2\right] \leqslant \tilde{O}\left(\left(\frac{r(2rp)^{p+1}}{\epsilon^{p+2}} + \gamma + \sigma_0\right)^2\right)$$

.

As $L(\boldsymbol{v}) - L(\boldsymbol{w}) - \hat{g}(\boldsymbol{w})^T(\boldsymbol{v} - \boldsymbol{w}) = L(\boldsymbol{v}) - L(\boldsymbol{w}) - g(\boldsymbol{w})^T(\boldsymbol{v} - \boldsymbol{w}) + (g(\boldsymbol{w}) - \hat{g}(\boldsymbol{w}))^T(\boldsymbol{v} - \boldsymbol{w})$, and from the approximation error, we know $|(g(\boldsymbol{w}) - \hat{g}(\boldsymbol{w}))^T(\boldsymbol{v} - \boldsymbol{w})| \leqslant \frac{\gamma}{2}$. What's more, as $L(\boldsymbol{w})$ is convex and $\beta$-smooth, that is $0 \leqslant L(\boldsymbol{v}) - L(\boldsymbol{w}) - g(\boldsymbol{w})^T(\boldsymbol{v} - \boldsymbol{w}) \leqslant \frac{\beta}{2}\|\boldsymbol{v} - \boldsymbol{w}\|^2$. Combined these inequalities, we obtain

$$\begin{aligned} & -\tfrac{\gamma}{2} \leqslant L(\boldsymbol{v}) - L(\boldsymbol{w}) - \hat{g}(\boldsymbol{w})^T(\boldsymbol{v} - \boldsymbol{w}) \leqslant \tfrac{\beta}{2}\|\boldsymbol{v} - \boldsymbol{w}\|^2 + \tfrac{\gamma}{2} \\ \Longleftrightarrow & 0 \leqslant L(\boldsymbol{v}) - (L(\boldsymbol{w}) - \tfrac{\gamma}{2}) - \hat{g}(\boldsymbol{w})^T(\boldsymbol{v} - \boldsymbol{w}) \leqslant \tfrac{\beta}{2}\|\boldsymbol{v} - \boldsymbol{w}\|^2 + \gamma \end{aligned}$$

Note the function value oracles in the stochastic oracle definition (either $F_{\gamma,\beta,\sigma}(\cdot)$ or $f_{\gamma,\beta,\sigma}(\cdot)$) do not play any role in the optimization algorithm, hence we can set it as $L(\boldsymbol{w}) - \frac{\gamma}{2}$, though we do not know how to calculate. □

**Lemma 6.** *Based on above statements, we have*

$$\mathbb{E}\left[\left\|\tilde{G}(\boldsymbol{w}; b) - \hat{G}(\boldsymbol{w}; \boldsymbol{x}, y)\right\|^2\right] \leqslant \tilde{O}\left(\frac{p^{4p+2}(4r)^{2p+2}}{\epsilon^{2p+4}}\right)$$
$$\mathbb{E}\left[\left\|\hat{G}(\boldsymbol{w}; \boldsymbol{x}, y) - \hat{g}(\boldsymbol{w})\right\|^2\right] \leqslant (\gamma + \sigma_0)^2$$

*Proof.* First, we calculate the variance of each $t_k$, $\text{var}(t_j) \leqslant \prod_{i=j(j-1)/2+1}^{j(j+1)/2} (\text{var}(\boldsymbol{w}^T\boldsymbol{z}_i) + (\mathbb{E}[\boldsymbol{w}^T\boldsymbol{z}_i])^2) \leqslant \tilde{O}\left((\frac{p(p+1)}{\epsilon})^{2j}\right)$.

Next, we upper bound the coefficient $c_k$ (as it is the same for $c_{1k}$ and $c_{2k}$, hence we use $c_k$ for short). Note $c_k = \sum_{m=k}^p a_m b_{mk}$, where $a_m$ is the coefficient of original function represented by Chebyshev basis, $b_{mk}$ is the coefficient of order $k$ monomial in Chebyshev basis $T_m(x)$, where $0 \leqslant k \leqslant m$. According to the formula of $T_m(x)$ given in (Qazi & Rahman, 2007) and well-known Stirling's approximation, after some translation, we have

$$\begin{aligned} |b_{mk}| \leqslant & \max_{\theta\in(0,\frac{1}{2})} O\left(\sqrt{m} \cdot \left[\frac{(1-\theta)^{1-\theta}}{\theta^\theta(1-2\theta)^{1-2\theta}}\right]^m\right) \\ \leqslant & O\left(\sqrt{m}2^m\right) \end{aligned}$$

Besides, from the absolutely smooth property of $h_i'(x)(i \in \{1, 2\})$ and the convergence results in (Trefethen, 2008), we have $a_m \leqslant O\left(\frac{1}{m^2}\right)$, thus $c_k = \sum_{m=k}^p a_m b_{mk} \leqslant O\left(2^p\right)$. Hence, there is

$$\begin{aligned} \text{var}\left[(c_{2k} - c_{1k}z_y)t_k r^{k+1}\right] \leqslant & r^{2k+2}\mathbb{E}\left[((c_{2k} - c_{1k}z_y)t_k)^2\right] \\ \leqslant & O\left(\frac{p^{4k+2}(4r)^{2p+2}}{\epsilon^{2k+2}}\right) \end{aligned}$$

As each $(c_{2k} - c_{1k}z_y)t_k r^{k+1}$ is independent with each other (for different $k$), which leads to

$$\text{var}\left[\sum_{k=0}^p (c_{2k} - c_{1k}z_y)t_k r^{k+1}\right] \leqslant O\left(\frac{p^{4p+2}(4r)^{2p+2}}{\epsilon^{2p+2}}\right)$$

Moreover, $\mathrm{var}(\boldsymbol{z}_0) \leqslant O\left(\frac{1}{\epsilon^2}\right)$. Therefore,

$$\mathbb{E}\left[\left\|\tilde{G}(\boldsymbol{w};b) - \hat{G}(\boldsymbol{w};\boldsymbol{x},y)\right\|^2\right] \leqslant \tilde{O}\left(\frac{p^{4p+2}(4r)^{2p+2}}{\epsilon^{2p+4}}\right)$$

For second inequality in the conclusion, there is

$$\begin{aligned}
&\mathbb{E}\left[\left\|\hat{G}(\boldsymbol{w};\boldsymbol{x},y) - \hat{g}(\boldsymbol{w})\right\|^2\right] \\
\leqslant &\mathbb{E}\left[\left\|\hat{G}(\boldsymbol{w};\boldsymbol{x},y) - G(\boldsymbol{w};\boldsymbol{x},y) + G(\boldsymbol{w};\boldsymbol{x},y) - g(\boldsymbol{w}) + g(\boldsymbol{w}) - \hat{g}(\boldsymbol{w})\right\|^2\right] \\
\leqslant &\gamma^2 + \sigma_0^2 + 2\sigma_0\gamma = (\gamma + \sigma_0)^2
\end{aligned}$$

□

**Proposition 1.** $f(x) = \ln(1 + e^{-x})$ *is absolutely smooth with* $\mu_1(k;r) = r\sqrt{4k\pi^3}, \mu_2(k;r) = \frac{rk}{e}$

*Proof.* For any $r, k > 0$, the absolutely continuous of $f^{(k)}(rx)$ is obvious, now consider $\left\|f^{(k+1)}(rx)\right\|_T$:

$$\begin{aligned}
\left\|f^{(k+1)}\right\|_T = &\int_{-1}^{1} \frac{|f^{(k+2)}(rx)|}{\sqrt{1-x^2}} \mathrm{d}x \\
\leqslant &\pi \left\|f^{(k+2)}(rx)\right\|_\infty \\
\leqslant &\pi r^{k+2} \left\|\textstyle\sum_{j=1}^{k+1} (-1)^{k+j} A_{k+1,j-1} f^j (1-f)^{k+2-j}\right\|_\infty \\
\leqslant &\pi r^{k+2} \sum_{j=1}^{k+1} A_{k+1,j-1} \\
\leqslant &\pi (k+1)! r^{k+2} \\
\leqslant &\sqrt{4\pi^3} r^{k+2} (k+1)^{k+3/2} e^{-k-1} \\
= &r\sqrt{4\pi^3(k+1)} \left(\frac{r(k+1)}{e}\right)^{k+1}
\end{aligned}$$

□

**Theorem 2** (Theorem 6 in Main Body)**.** *For any* $\alpha > 0$, *set* $\gamma = \frac{\alpha}{2}, k = c\ln\frac{4r}{\gamma}, p = \lceil k + 2\mu_2(k;r) \rceil$, *if* $n > O\left((\frac{8r}{\alpha})^{4r\ln\ln(8r/\alpha)} \left(\frac{4r}{\epsilon}\right)^{2cr\ln(8r/\alpha)+2} \left(\frac{1}{\alpha^2\epsilon^2}\right)\right)$, *using algorithms 6,7,8, then we have* $L(\boldsymbol{w}^{priv}) - L(\boldsymbol{w}^*) \leqslant \alpha$.

*Proof.* According to lemma 10 in main body, with a $(\gamma, \beta, \sigma)$ stochastic oracle, SIGM algorithm converges with rate $O\left(\frac{\sigma}{\sqrt{n}} + \gamma\right)$. In order to have $O\left(\frac{\sigma}{\sqrt{n}} + \gamma\right) \leqslant \alpha$, it suffices if $n > O\left(\frac{p^{4p+2}(4r)^{2p+2}}{\alpha^2\epsilon^{2p+4}}\right) = O\left((\frac{8r}{\alpha})^{4r\ln\ln(8r/\alpha)} \left(\frac{4r}{\epsilon}\right)^{2cr\ln(8r/\alpha)+2} \left(\frac{1}{\alpha^2\epsilon^2}\right)\right)$, as $\sigma = O\left(\frac{p^{2p+1}(4r)^{p+1}}{\epsilon^{p+2}}\right)$ according to lemma 5 (ignoring negligible $\sigma_0, \gamma$). □

## References


Dirksen, Sjoerd. Dimensionality reduction with subgaussian matrices: a unified theory. *Foundations of Computational Mathematics*, 16(5):1367–1396, 2016.

Qazi, MA and Rahman, QI. Some coefficient estimates for polynomials on the unit interval. *Serdica Mathematical Journal*, 33(4):449p–474p, 2007.

Rubinstein, B., Bartlett, P. L., Huang, L., and Taft, N. Learning in a large function space: Privacy-preserving mechanisms for svm learning. *Journal of Privacy and Confidentiality*, 4(1):4, 2012.

Trefethen, Lloyd N. Is gauss quadrature better than clenshaw–curtis? *SIAM review*, 50(1):67–87, 2008.

Tropp, Joel A et al. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

Vershynin, Roman. A note on sums of independent random matrices after ahlswede-winter. *Lecture notes*, 2009.