Appendix A. Omitted Proofs

A.1. Proof of Theorem 1

Proof Before we start our proof, we give the definition of LR_S oracle and $\text{STAT}_{\mathcal{P}}(\tau)$ oracle to prepare the readers for the proof. LR_S oracle is based on the local randomizer which is defined as follows:

Definition 23 *(ϵ-local randomizer) An ϵ-local randomizer $\mathcal{R} : Z \to W$ is a randomized algorithm that $\forall z_1, z_2 \in Z$ and $\forall w \in W$, it satisfies:*

$$Pr[\mathcal{R}(z_1) = w] \leq e^{\epsilon}[\mathcal{R}(z_2) = w]$$

.

Definition 24 *(LR_S oracle Kasiviswanathan et al. (2011)) For a dataset $S \in Z^n$, an LR_S oracle takes an index i and a local randomizer $\mathcal{R}$ as inputs and outputs a random value w obtained by applying $\mathcal{R}(z_i)$.*

And we recall the definition of statistical queries.

Definition 25 *Let $\mathcal{P}$ be an distribution over a domain Z and $\tau > 0$. A statistical query oracle $STAT_{\mathcal{P}}(\tau)$ is an oracle that given any function $\phi : Z \to [-1, 1]$ as input, the statistical query oracle returns some value v such that $|v - \mathbb{E}_{z\sim\mathcal{P}}[\phi(z)]| \leq \tau$.*

Now we formally begin our proof. First, we prove that the algorithm given in Daniely and Feldman (2019) uses the same number of private data and public data. The core idea of the algorithm in Daniely and Feldman (2019) is that: when using the projected gradient descent to find a vector w that satisfies $\underset{(x,y)\sim\mathcal{P}}{Pr}[y \neq \text{sign}(\langle\hat{w}, x\rangle)] \leq \alpha$, the objective function can be decomposed as $F(w) = F_1(w) + F_2(w)$, where the (sub-)gradient of $F_1(w)$ (namely $\nabla F_1(w)$) is just a function of x while the gradient of $F_2(w)$ (namely $\nabla F_2(w)$) is independent of w. As a result, (sub)-gradient $\nabla F(w)$ can be computed non-interactively by calculating $\nabla F_1(w)$ with only public unlabeled data and calculating $\nabla F_2(w)$ with non-interactive statistic queries because $\nabla F_2(w)$ doesn't depend on w. So to make this algorithm achieve the PAC learning error α, the sample complexity of the private data and the public data should be the same. For more details, please refer to the proof of Lemma 4.3 in Daniely and Feldman (2019). So, to prove our theorem, we only have to prove that the sample complexity of the private data is $\widetilde{O}(\frac{d^{10}\log(1/\beta)}{\epsilon^2\cdot\gamma^{12}\alpha^6})$.

In the following, we give the private sample complexity of the algorithm in Daniely and Feldman (2019), which can be directly derived from the following two lemmas.

The first Lemma states that a statistic query oracle $\text{STAT}_{\mathcal{P}}(\tau)$ can be simulated with success probability $1 - \beta$ by ϵ-LDP algorithm using LR_S oracle.

Lemma 26 *Kasiviswanathan et al. (2011) Let $\mathcal{A}_{SQ}$ be an algorithm that makes at most t queries to $STAT_{\mathcal{P}}(\tau)$ oracle. Then for any $\epsilon > 0$ and $\beta > 0$, there is an ϵ-LDP algorithm $\mathcal{A}_{priv}$ that uses LR_s oracle for S containing $n = O(\frac{t\log(\frac{t}{\beta})}{(\epsilon\tau)^2})$ i.i.d. samples from $\mathcal{P}$ and produces the same output as $\mathcal{A}_{SQ}$ with probability at least $1 - \beta$. Further, if $\mathcal{A}_{SQ}$ is non-interactive then $\mathcal{A}_{priv}$ is non-interactive.*

The next lemma claims the existence a NLDP algorithm $\mathcal{A}_{SQ}$ that achieves PAC learning error α for any arbitrary $\alpha \in (0, 1)$.

Lemma 27 (Lemma 4.3 in Daniely and Feldman (2019)) *Let $\mathcal{P}$ be a distribution on $\mathcal{B}_2^d \times \{\pm 1\}$ such that there is a vector $w^* \in \mathcal{B}_2^d$ satisfying $Pr_{(x,y)\sim\mathcal{P}}[y\langle w^*, x\rangle \geq \gamma] = 1$. Then there is a non-interactive algorithm $\mathcal{A}_{SQ}$ that for every $\alpha \in (0,1)$, it uses $O(\frac{d^4}{\gamma^4\alpha^2})$ queries to STAT$_{\mathcal{P}}(\Omega(\frac{\gamma^4\alpha^2}{d^3}))$ and finds a vector $\hat{w}$ such that $\underset{(x,y)\sim\mathcal{P}}{Pr}[y \neq sign(\langle \hat{w}, x\rangle)] \leq \alpha$.*

Lemma 27 indicates that if we can find a non-interactive algorithm $\mathcal{A}_{SQ}$ that makes at most t queries to STAT$_{\mathcal{P}}(\tau)$ oracle, then with probability $1-\beta$, the existence of an ϵ-NLDP algorithm $\mathcal{A}_{priv}$ is guaranteed using $n = O(\frac{t\log(\frac{t}{\beta})}{(\epsilon\tau)^2})$ private data. So, by substituting $t = O(\frac{d^4}{\gamma^4\alpha^2})$ and $\tau = \Omega(\frac{\gamma^4\alpha^2}{d^3})$ in Lemma 26, the sample complexity of public data is straight forward.

■

A.2. Proof of Lemma 13

Proof To proof Lemma 13, we first study the excess empirical risk with the hinge loss $\ell(w,(x,y)) = \max\{0, 1 - y\langle w, x\rangle\}$ of the output w_t of the algorithm $\mathcal{H}_{priv}(\frac{1}{32R}, \epsilon, \delta, \widetilde{S}_t)$. First, we recall the following result of $\mathcal{H}_{priv}(\alpha, \epsilon, \delta, S)$ if each $\|x_i\|_2 \leq 1, |y_i| \leq 1$.

Lemma 28 (Theorem 30 in Wang et al. (2020)) *For any $0 < \epsilon, \delta < 1$, if each $\|x_i\|_2 \leq 1, |y_i| \leq 1$ for all $i \in [n]$, $\mathcal{H}_{priv}(\alpha, \epsilon, \delta, S)$ is (ϵ, δ)-NLDP. Moreover, for any error $\alpha \in (0,1)$, if the size of dataset n is sufficiently large such that $n \geq \widetilde{\Omega}(\frac{C^p p^{6p} d}{\epsilon^{4p+4}\alpha})$ with $p = O(\frac{1}{\alpha^3})$. Then the output w_n satisfies*

$$\mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}\max\{0, \frac{1}{R} - y\langle w, x\rangle\}] - \min_{||w||_2\leq 1}\frac{1}{n}\sum_{i=1}^{n}\max\{0, \frac{1}{R} - y\langle w, x\rangle\} \leq \alpha, \tag{1}$$

where $C > 0$ is a constant[3] and the expectation is taken over the internal randomness of the algorithm.

Note that in we need to assume $\|x_i\|_2 \leq 1$ in Lemma 28 while in our setting $\|x_i\|_2 \leq R$. Thus, we need to normalize the data to $\widetilde{S}_t$ first and revoke $\mathcal{H}_{priv}(\frac{1}{32R}, \epsilon, \delta, \widetilde{S}_t)$. By Lemma 28 we have when $\frac{n}{k} \geq \widetilde{\Omega}(d\text{Poly}(\frac{1}{\epsilon}, \log\frac{1}{\delta}))$

$$\mathbb{E}[\hat{L}(w_t, \widetilde{S}_t)] - \min_{||w||_2\leq 1}\hat{L}(w, \widetilde{S}_t) \leq \frac{1}{32R}, \tag{2}$$

where $\hat{L}(w_t, \widetilde{S}_t) = \frac{1}{|\widetilde{S}_t|}\sum_{(x_i,y_i)\in \widetilde{S}_t}\max\{0, \frac{1}{R} - y_i\langle w, \frac{x_i}{R}\rangle\}$. Thus, we have the following result via multiplying R in both side of (2).

Lemma 29 *When $n \geq \widetilde{\Omega}(dk\text{Poly}(\frac{1}{\epsilon}, \log\frac{1}{\delta}))$, each $w_t = \mathcal{H}_{priv}(\frac{1}{32R}, \epsilon, \delta, \widetilde{S}_t)$ for $t \in [k]$ satisfies*

$$\mathbb{E}[\hat{L}(w_t, S_t)] - \min_{||w||_2\leq 1}\hat{L}(w_t, S_t) \leq \frac{1}{32}, \tag{3}$$

where $\hat{L}(w_t, S_t)$ is the empirical risk of $\ell(w,(x,y)) = \max\{0, 1 - y\langle w, x\rangle\}$, and the expectation is taken over the internal randomness of the algorithm.

3. Note that Wang et al. (2020) only showed the case where $R = 2$. However, it is obvious to extend to the general R with the same proof.

The following lemma transforms the excess empirical risk in Lemma 29 to classification error.

Lemma 30 *Under the assumptions in Theorem 11, then for any $t \in [k]$, $\beta \in (0,1)$, with probability at least $1-\frac{\beta}{2}$, the following holds when $n \geq \widetilde{\Omega}(dk\mathit{Poly}(\frac{1}{\epsilon}, \log\frac{1}{\delta}))$ with $k = O(\log\frac{1}{\beta})$.*

$$\mathbb{E}[err_P(h_{w_t})] \leq \frac{1}{8}$$

where the expectation is taken over the random choice of the data in D and the internal randomness of $\mathcal{H}_{priv}$.

Proof [**Proof of Lemma 30**] We need the following lemma for our proof.

Lemma 31 (Anthony and Bartlett (2009)) *Let $\mathcal{H}$ be the set of $\{\pm 1\}$-valued functions defined on a set $\mathcal{X}$ and $\mathcal{P}$ is a probability distribution on $Z = \mathcal{X} \times \{\pm 1\}$. For $\eta \in (0,1)$, $\zeta > 0$, $Pr_{z \sim \mathcal{P}^n}[\exists h \in \mathcal{H} : err_{\mathcal{P}}(h) > (1+\zeta)err_z(h) + \eta] \leq 4\tau_{\mathcal{H}}(2n)e^{-\frac{\eta\zeta n}{4(\zeta+1)}}$, where $err_{\mathcal{P}}(h)$ is the population error, $err_z(h)$ is the empirical error on sample set z and $\tau_{\mathcal{H}}(\cdot)$ is the growth function of $\mathcal{H}$. If $\mathcal{H}$ is the hypothesis set of learning halfspaces, then $\tau_{\mathcal{H}}(2n) \leq (2n)^{d+1} + 1$ with d being the dimension of set $\mathcal{X}$.*

The following proof applies for any $t \in [k]$:

Based on our assumption, the halfspace is separable, so we know that $\min_{||w||_2 \leq 1} \hat{L}(w, S_t) = 0$. Since hinge loss is a convex surrogate for $0-1$ loss, we can get that

$\mathbb{E}[err_{S_t}(h_{w_t})] \leq \mathbb{E}[\hat{L}(w_t, S_t)] \leq \min_{||w||_2 \leq 1} \hat{L}(w, D) + \frac{1}{32} = \frac{1}{32}$, where the second inequality comes from (3).

Setting $\eta = \frac{1}{16}$ and $\zeta = 1$, for any $t \in [k]$, denoting $n_t = |S_t|$, then according to Lemma 31, we can get

$$\underset{S_t \sim \mathcal{P}^{n_t}}{Pr}\{\exists h_{w_t} \in \mathcal{H} : \mathbb{E}[err_{\mathcal{P}}(h_{w_t})] > 2 \cdot \frac{1}{32} + \frac{1}{16}\} \leq 4\tau_{\mathcal{H}}(2n_t)e^{-\frac{n_t}{128}}. \tag{4}$$

When $n_t = \widetilde{\Omega}(d\log\frac{1}{\beta}\text{Poly}(\log\frac{1}{\delta}, \frac{1}{\epsilon}))$, we have $4\tau_{\mathcal{H}}(2n_t)e^{-\frac{n_t}{128}} \leq \frac{\beta}{2k}$. Then (4) will become

$$\underset{S_t \sim \mathcal{P}^{n_t}}{Pr}\{\exists h_{w_t} \in \mathcal{H} : \mathbb{E}[err_{\mathcal{P}}(h_{w_t})] > \frac{1}{8}\} \leq \frac{\beta}{2k}.$$

Thus, take the union bound, we have with probability at least $1-\frac{\beta}{2}$ for any $t \in [k]$,

$$\mathbb{E}[err_{\mathcal{P}}(h_{w_t})] \leq \frac{1}{8}.$$

■

According to Lemma 30, for any $t \in [k]$, with probability at least $1-\frac{\beta}{2}$, we have

$$\mathbb{E}_{D,\mathcal{H}_{priv}}[err_{\mathcal{P}}(h_{w_t})] = \mathbb{E}_{D,\mathcal{H}_{priv}}\{\underset{(x,y)\sim\mathcal{P}}{Pr}[h_{w_t}(x) \neq y]\} \leq \frac{1}{8}.$$

Applying Hoeffding inequality, we have

$$Pr\{\underset{(x,y)\sim\mathcal{P}}{Pr}[\hat{f}(x)\neq y]-\frac{1}{8}>\frac{1}{4}\}\leq Pr\{\frac{1}{k}\sum_{t=1}^{k}\underset{(x,y)\sim\mathcal{P}}{Pr}[h_{w_t}(x)\neq y]-\frac{1}{8}>\frac{1}{16}\}$$
$$\leq Pr\{|\frac{1}{k}\sum_{t=1}^{k}\underset{(x,y)\sim\mathcal{P}}{Pr}[h_{w_t}(x)\neq y]-\frac{1}{8}|>\frac{1}{16}\}\leq 2e^{-\frac{k}{32}}$$

For the first inequality, denote the event $E_1=\{\underset{(x,y)\sim\mathcal{P}}{Pr}[\hat{f}(x)\neq y]-\frac{1}{8}>\frac{1}{4}\}$ and event $E_2=\{\frac{1}{k}\sum_{t=1}^{k}\underset{(x,y)\sim\mathcal{P}}{Pr}[h_{w_t}(x)\neq y]-\frac{1}{8}>\frac{1}{16}\}$. Thus, the first inequality holds if $E_1\subseteq E_2$. E_1 claims that with probability at least $\frac{3}{8}$ the classifier $\hat{f}$ will gives wrong prediction. That is more than half of $\{w_t\}_{t=1}^{k}$ give wrong predictions. Thus, $\frac{1}{k}\sum_{t=1}^{k}\underset{(x,y)\sim\mathcal{P}}{Pr}[h_{w_t}(x)\neq y]\geq\frac{\frac{k}{2}\times\frac{3}{8}}{k}=\frac{3}{16}$. The second inequality is due to $\mathbb{E}\{\frac{1}{k}\sum_{t=1}^{k}\underset{(x,y)\sim\mathcal{P}}{Pr}[h_{w_t}(x)\neq y]\}\leq\frac{1}{8}$.

When $k=O(\log(\frac{1}{\beta}))$, we have

$$Pr\{\underset{(x,y)\sim\mathcal{P}}{Pr}[\hat{f}(x)\neq y]>\frac{3}{16}\}\leq\frac{\beta}{2}$$

Therefore, with probability at least $1-\frac{\beta}{2}-\frac{\beta}{2}=1-\beta$, we have

$$\underset{(x,y)\sim\mathcal{P}}{Pr}[\hat{f}(x)\neq y]\leq\frac{3}{16}$$

■

A.3. Proof of Theorem 21

The proof of this theorem can be induced directly by the following two lemmas. The first lemma claims that Logistic Loss-NLDP outputs a classifier w^{priv} which is NLP and achieves a constant classification error C_{err} using $O(d\text{Poly}(\frac{1}{\epsilon}))$ private samples.

Lemma 32 *Algorithm 3 is (ϵ,δ)-NLDP and w^{priv} satisfies the following when $n=O(dPoly(\frac{1}{\epsilon}))$*

$$err_{\mathcal{P}}(h_{w^{priv}})\leq\frac{r^2}{144U}.$$

The second lemma claims that STWN (Algorithm 4) transforms a weak learner that achieves a constant classification error to a strong learner that achieves a classification error arbitrarily close to the Bayes-optimal error using only unlabeled samples.

Lemma 33 *Frei et al. (2021) If $(x,y)\sim\mathcal{P}$ is a mixture distribution with mean $\boldsymbol{\mu}$ satisfying $||\boldsymbol{\mu}||_2=\Theta(1)$ and $K,U,r>0$, assume $\widetilde{\ell}$ is well behaved for some $C_{\widetilde{\ell}}\geq 1$ and the temperature satisfies*

$\sigma \geq R \vee ||\boldsymbol{\mu}||_2$. Assume access to a pseudo labeler w_{pl} which achieves classification error less than $\frac{R^2}{72C_{\widetilde{\ell}}U}$, i.e., $err_{\mathcal{P}}(h_{w_{pl}}) \leq \frac{R^2}{72C_{\widetilde{\ell}}U}$. Let $\alpha, \beta \in (0,1)$, $B = \Omega\left(\frac{\log(\frac{1}{\beta})}{\alpha}\right)$, $T = \widetilde{\Omega}\left(\frac{d^2(\log(\frac{1}{\beta}))}{\alpha}\right)$ and step size $\eta = \widetilde{\Theta}\left(\frac{\alpha}{d(\log(\frac{1}{\beta}))^2}\right)$, running STWN (Algorithm 4) with $T \times B$ unlabeled samples, then with probability at least $1-\beta$, there exists $t^* < T$ such that $err_{\mathcal{P}}(h_{w^{(t^*)}}) \leq err_{\mathcal{P}}(h_{\boldsymbol{\mu}}) + \alpha$ where $err_{\mathcal{P}}(h_{\boldsymbol{\mu}})$ is the error of Bayes-optimal classifier.

In particular, let $B = O\left(\frac{\log(\frac{1}{\beta})}{\alpha}\right)$, $T = \widetilde{O}\left(\frac{d(\log(\frac{1}{\beta}))^2}{\alpha}\right)$, above conclusion holds using $T \times B = \widetilde{O}\left(\frac{d(\log(\frac{1}{\beta}))^3}{\alpha}\right)$ unlabeled data samples.

Proof of Theorem 21: Since in Algorithm 3 we use the logistic function as the well behaved loss, we have $C_{\widetilde{\ell}} = 2$. Moreover, under our assumption, the Bayes-optimal classifier is just w^* and thus $err_{\mathcal{P}}(h_{\boldsymbol{\mu}}) = 0$. Combing with Lemma 32 and Lemma 33 we finish the proof.

Proof [**Proof of Lemma 32**] To prove the lemma, we need the following lemma claiming the excess population loss of the output of Logistic Loss-NLDP: $\mathcal{T}_{priv}(\alpha, R, \epsilon, \delta, D)$

Lemma 34 (Theorem 6 in Zheng et al. (2017)) *For any $0 < \epsilon, \delta \leq 1$, if each $\|x_i\|_2 \leq R$ and $y \in \{-1, 1\}$ for all $i \in [n]$, and $\mathcal{W} = \{w : \|w\|_2 \leq \rho\}$, $\mathcal{T}_{priv}(\alpha, R, \rho, \epsilon, \delta, D)$ is (ϵ, δ)-NLDP. Moreover, for any given error $\alpha \in (0,1)$, if the size of dataset n is sufficiently large such that*

$$n \geq \widetilde{\Omega}\left(\left(\frac{8R\rho}{\alpha}\right)^{4R\rho \ln\ln \frac{8R\rho}{\alpha}} \left(\frac{4R\rho}{\epsilon}\right)^{2cR\rho \ln \frac{8R\rho}{\alpha}+2} \frac{1}{\alpha^2\epsilon^2}\right).$$

Then the output w_n satisfies $\mathbb{E}[L(w^{priv})] - \min_{w \in \mathcal{W}} L(w) \leq \alpha$, where $L(w^{priv})$ is the population risk of the logistic loss, i.e., $L(w) = \mathbb{E}_{(x,y)\sim\mathcal{P}}[\ell(w; x, y)]$, where $\ell(w; x, y) = \log(1 + e^{-y\langle x, w\rangle})$.

Apply the above Lemma 34 with $\alpha = \frac{C_{err}\log 2}{2} = \frac{\log 2r^2}{144U}$ and $\rho = \|\boldsymbol{\mu}\|_2$. Then using $n = O\left(d\text{Poly}(\frac{1}{\epsilon})\right)$ private samples, w^{priv} achieves the excess population loss no more than $\frac{\log 2r^2}{144U}$, i.e., $\mathbb{E}[L(w^{priv})] - \min_{\|w\|_2 \leq \|\boldsymbol{\mu}\|_2} \frac{\mathbb{E}[L(w)] \leq C_{err}\log 2}{2}$. Since $\|\boldsymbol{\mu}\|_2 \in \mathcal{W}$, thus,

$$\mathbb{E}[L(w^{priv})] \leq \mathbb{E}[L(\boldsymbol{\mu})] + \frac{C_{err}\log 2}{2}.$$

For the term of $\mathbb{E}[L(\boldsymbol{\mu})]$, recall the following lemma.

Lemma 35 (Lemma B.3 in Frei et al. (2021)) *Consider the logistic function $\ell(z) = \log(1 + e^{-z})$. Let $(x, y) \sim \mathcal{P}$ be a mixture distribution with mean $\boldsymbol{\mu}$ and parameters $K, U, R = \Theta(1) > 0$. Then if $\|\boldsymbol{\mu}\|_2 \geq 64K^2$ we have*

$$\mathbb{E}_{(x,y)\sim\mathcal{P}}[\ell(y\langle w, x\rangle)] \leq \exp(-\frac{\|\boldsymbol{\mu}\|_2}{3K}). \tag{5}$$

By using the previous lemma, we have

$$\mathbb{E}[L(w^{priv})] \leq \exp(-\frac{\|\boldsymbol{\mu}\|_2}{3K}) + C_{err}\log 2/2 \leq C_{err}\log 2,$$

where the last inequality is due to the assumption of $\|\boldsymbol{\mu}\|_2 \geq 3K\log(8/C_{err})$. Thus we have

$$
\begin{aligned}
&Pr[y \neq \text{sign}(\langle w^{priv}, x\rangle)] = Pr[y \cdot \langle w^{priv}, x\rangle < 0] = Pr[\ell(y \cdot \langle w^{priv}, x\rangle) > \ell(0)] \\
&\leq \frac{\mathbb{E}[\ell(y \cdot \langle w^{priv}, x\rangle)]}{\ell(0)} = \frac{\mathbb{E}[L(w^{priv})]}{\ell(0)} \leq \frac{r^2}{144U}
\end{aligned}
$$

where we use the monotonicity of the loss function and Markov's inequality. ■

Appendix B. Details of Hinge Loss-LDP and Logistic Loss-NLDP

Algorithm 5 Hinge Loss-NLDP: $\mathcal{H}_{priv}(\alpha, \epsilon, \delta, S)$

Input: Private data $S = \{(x_i, y_i)\}_{i=1}^n \in \mathbb{R}^d \times \{\pm 1\}$, where $||x_i||_2 \leq 1, ||y_i||_2 \leq 1$; Privacy parameters ϵ, δ; Error α.

Denote $P_p(x) = \sum_{j=0}^p c_i \binom{p}{j} x^j (1-x)^{p-j}$ as the p-th order Bernstein polynomial for the function f'_β, where $c_i = f'_\beta\left(\frac{i}{p}\right)$ and $f_\beta(x) = \frac{\frac{1}{R} - x + \sqrt{(\frac{1}{R} - x)^2 + \beta^2}}{2}$ with $\beta = \frac{\alpha}{4}$ and $p = \frac{2}{\beta^2 \alpha}$.
\\\\ The local user side:
for $i \in [n]$ **do**
  Set $\sigma_{i,0} \sim \mathcal{N}\left(0, \frac{32\log(1.25/\delta)}{\epsilon^2}\boldsymbol{I}_d\right)$ and $z_{i,0} \sim \mathcal{N}\left(0, \frac{32\log(1.25/\delta)}{\epsilon^2}\right)$
  Set $x_{i,0} = x_i + \sigma_{i,0}$ and $y_{i,0} = y_i + z_{i,0}$
  for $j \in [p(p+1)]$ **do**
    $x_{i,j} = x_i + \sigma_{i,j}$, where $\sigma_{i,j} \sim \mathcal{N}\left(0, \frac{8\log(1.25/\delta)p^2(p+1)^2}{\epsilon^2}\boldsymbol{I}_d\right)$
    $y_{i,j} = y_i + z_{i,j}$, where $z_{i,j} \sim \mathcal{N}\left(0, \frac{8\log(1.25/\delta)p^2(p+1)^2}{\epsilon^2}\right)$
  end for
  Send $\{x_{i,j}\}_{j=0}^{p(p+1)}$ and $\{y_{i,j}\}_{j=0}^{p(p+1)}$ to the server.
end for
\\\\ The server side:
for $t \in [n]$ **do**
  Randomly sample $i \in [n]$ uniformly and set $t_{i,0} = 1$
  for j=$\{0\} \cup [p]$ **do**
    $t_{i,j} = \Pi_{k=jp+1}^{jp+j} y_{i,k}\langle w_t, x_{i,k}\rangle$ and $t_{i,0} = 1$
    $s_{i,j} = \Pi_{k=jp+j+1}^{jp+p}(1 - y_{i,k}\langle w_t, x_{i,k}\rangle)$ and $s_{i,p} = 1$
  end for
  Denote $G(w_t, i) = (\sum_{j=0}^p c_j \binom{p}{j} t_{i,j} s_{i,j}) y_{i,0} x_{i,0}^T$
  Update SIGM (Algorithm 7) by $G(w_t, i)$
end for
Return w_n

Algorithm 6 Logistic Loss-NLDP: $\mathcal{T}_{priv}(\alpha, R, \rho, \epsilon, \delta, D)$

Input: Private data $S = \{(x_i, y_i)\}_{i=1}^n \in \mathbb{R}^d \times \{\pm 1\}$, where $||x_i||_2 \leq R, ||y_i||_2 \leq 1$; Privacy parameters ϵ, δ; Error α; Constraint set $\mathcal{W} = \{w : \|w\|_2 \leq \rho\}$.

Denote the logistic loss with scale $R\rho$: $\ell(w, x, y, R) = \log(1 + e^{-R\rho y\langle w, x\rangle}) = -yh_1(R\rho w^T x) + h_2(R\rho w^T x)$, where $h_1(z) = \frac{z}{2}$ and $h_2(z) = \frac{z}{2} + \log(1 + e^{-z})$. For the function $h_1'(R\rho\cdot) : [-1, 1] \mapsto \mathbb{R}$ and $h_2'(R\rho\cdot) : [-1, 1] \mapsto \mathbb{R}$, denote the Chebyshev polynomial with degree p for function $h_1'(R\rho\cdot)$ and $h_2'(R\rho\cdot)$ as $\sum_{i=1}^n c_{1k} x^k$ and $\sum_{i=1}^n c_{2k} x^k$ respectively, where the degree $p = O(R \ln \frac{R\rho}{\alpha})$.

\\ The local user side:

for $i \in [n]$ **do**
  Normalize the data $x_i' = \frac{x_i}{R}$.
  Set $\sigma_{i,0} \sim \mathcal{N}\left(0, \frac{32\log(1.25/\delta)}{\epsilon^2}\boldsymbol{I}_d\right)$ and $z_{i,0} \sim \mathcal{N}\left(0, \frac{32\log(1.25/\delta)}{\epsilon^2}\right)$
  Set $x_{i,0} = x_i' + \sigma_{i,0}$ and $y_{i,0} = y_i + z_{i,0}$
  for $j \in [p(p+1)]$ **do**
    $x_{i,j} = x_i' + \sigma_{i,j}$, where $\sigma_{i,j} \sim \mathcal{N}\left(0, \frac{8\log(1.25/\delta)p^2(p+1)^2}{\epsilon^2}\boldsymbol{I}_d\right)$
  end for
  for $j = p$ **do**
    $y_{i,j} = y_i + z_{i,j}$, where $z_{i,j} \sim \mathcal{N}\left(0, \frac{8\log(1.25/\delta)p^2}{\epsilon^2}\right)$
  end for
  Send $\{x_{i,j}\}_{j=0}^{p(p+1)}$ and $\{y_{i,j}\}_{j=0}^{p}$ to the server.
end for

\\ The server side:

for $t \in [n]$ **do**
  Randomly sample $i \in [n]$ uniformly and set $t_{i,0} = 1$
  for j=$\{0\} \cup [p]$ **do**
    $t_j = \Pi_{k=\frac{j(j-1)}{2}+1}^{\frac{j(j+1)}{2}}(w_t^T x_{i,k})$
  end for
  $\widetilde{G}(w_t; i) = \left(\sum_{k=0}^{p}(c_{2k} - c_{1k}y_{i,j})t_k(R\rho)^{k+1}\right)z_0.$
  Update SIGM (Algorithm 7) by $\widetilde{G}(w_t; i)$ to obtain w_{t+1}.
end for
Return w_{n+1}

Algorithm 7 Stochastic Intermediate Gradient Method (SIGM)

Input: The sequences $\{\alpha_i\}_{i\geq 0}$, $\{\beta_i\}_{i\geq 0}$, $\{B_i\}_{i\geq 0}$ functions $d(x) = \frac{||x||^2}{2}$, Bregman distance $V(x,z) = d(X) - d(Z) - \langle \nabla d(z), x - z\rangle$.

1: Compute $x_0 = \arg\min_{x\in\mathcal{C}}\{d(x)\}$.
2: Let ξ_0 be a realization of the random variable ξ.
3: Compute $y_0 = \arg\min_{x\in\mathcal{C}}\{\beta_0 d(x) + \alpha_0\langle G_{\gamma,\beta,\sigma}(x_0;\xi_0), x - x_0\rangle\}$
4: **for** $k \in \{0\} \cup [T-1]$ **do**
5: Compute $z_k = \arg\min_{x\in\mathcal{C}}\{\beta_k d(x) + \sum_{i=0}^{k} \alpha_i\langle G_{\gamma,\beta,\sigma}(x_i;\xi_i), x - x_i\rangle\}$
6: Let $x_{k+1} = \eta_k z_k + (1-\eta_k) y_k$
7: Let ξ_{k+1} be a realization of the random variable ξ
8: Compute $\hat{x}_{k+1} = \arg\min_{x\in\mathcal{C}}\{\beta_k V(x, z_k) + \alpha_{k+1}\langle G_{\gamma,\beta,\sigma}(x_{k+1};\xi_{k+1}), x - z_k\rangle\}$
9: Let $w_{k+1} = \eta\hat{x}_{k+1} + (1-\eta_k) y_k$
10: $y_{k+1} = \frac{A_{k+1} - B_{k+1}}{A_{k+1}} y_k + \frac{B_{k+1}}{A_{k+1}} w_{k+1}$
11: **end for**
12: Return y_T