Supplementary Material

A. Non-Contractivity of Mirror Descent

In this section, we provide counter examples that show that Mirror Descent is not a contraction in general. To this end, we consider the standard mirror descent algorithm with KL-regularization over the simplex $\Delta_{d-1} = \{x \in \mathbb{R}^d_+ : \|x\|_1 = 1\}$, that is, the following update with $h(x) = \sum_{j=1}^d x_j \log x_j$,

$$x^{k+1} = \operatorname*{argmin}_{x \in \Delta} \left\{ \langle \nabla f(x_k), x \rangle + \frac{1}{\eta_k} D_{\mathrm{h}}(x, x_k) \right\},$$

which yields the update

$$x^{k+1} = \frac{x^k \cdot e^{-\eta \nabla f(x^k)}}{\left\| x^k \cdot e^{-\eta \nabla f(x^k)} \right\|_1}. \tag{1}$$

We let $x_{k+1} = \mathsf{MD}_{\eta,f}(x_k)$ denote the above mirror descent update. The following lemma shows that mirror descent is not contractive even for linear functions.

Lemma A.1. *There exists a linear function $f : \Delta_2 \to \mathbb{R}$ such that for every $0 < \eta \le 1$, there are $x_0, y_0 \in \Delta_2$ such that the mirror descent update $x_1 = \mathsf{MD}_{\eta,f}(x_0)$ and $y_1 = \mathsf{MD}_{\eta,f}(y_0)$ have*

$$\|x_1 - y_1\|_1 \ge (1 + \eta/4) \|x_0 - y_0\|_1, \quad D_{\mathrm{h}}(x_1, y_1) \ge (1 + \eta/4) D_{\mathrm{h}}(x_0, y_0).$$

Proof. We consider a linear function $f(x_1, x_2, x_3) = -x_2 - x_3$, and two starting iterates for $n > 0$ to be chosen presently

$$x_0 = \left(1 - \frac{3}{n}, \frac{1}{n}, \frac{2}{n}\right), \quad y_0 = \left(1 - \frac{3}{n}, \frac{2}{n}, \frac{1}{n}\right).$$

First, notice that for this setting of parameters, we have that:

$$\|x_0 - y_0\|_1 = \frac{2}{n}, \quad D_{\mathrm{h}}(x_0, y_0) = D_{\mathrm{kl}}(x_0, y_0) = \frac{\log 2}{n}.$$

Using mirror descent update (1), we have

$$x_1 = \frac{1}{c}\left(x_{0,1}, x_{0,2}\, e^{\eta}, x_{0,3}\, e^{\eta}\right), \quad y_1 = \frac{1}{c}\left(y_{0,1}, y_{0,2}\, e^{\eta}, y_{0,3}\, e^{\eta}\right),$$

where $c = 1 + \frac{3}{n}(e^\eta - 1)$. Setting $n \ge 100(e^\eta - 1)/\eta$, we get that $c \le 1 + \eta/20$. Since $x_{0,1} = y_{0,1}$, we get that

$$\begin{aligned}
\|x_1 - y_1\|_1 &= \frac{e^\eta}{c} \|x_0 - y_0\|_1 \\
&\ge \frac{1 + \eta}{1 + \eta/20} \|x_0 - y_0\|_1 \\
&\ge \|x_0 - y_0\|_1 + \frac{\eta}{4} \|x_0 - y_0\|_1 .
\end{aligned}$$

Moreover, for KL-divergence we have

$$\begin{aligned}
D_{\mathrm{kl}}(x_1, y_1) &= \frac{e^\eta}{c} D_{\mathrm{kl}}(x_0, y_0) \\
&\ge (1 + \eta/4) D_{\mathrm{kl}}(x_0, y_0).
\end{aligned}$$

□

Although Lemma A.1 says that mirror descent update is not contractive even for linear functions, it does not preclude the possibility that mirror descent is stable. Indeed, the following lemma shows that mirror descent enjoys similar stability guarantees to SGD for linear functions. Extending this stability result to general convex functions is an interesting open question.

Lemma A.2. *Let $\mathcal{S} = (z_1, \ldots, z_n)$ and $\mathcal{S}' = (z_1', \ldots, z_n)$ be neighboring datasets where $x_i \in \mathbb{R}^d$ and $\|z_i\|_\infty \le L$. Consider the functions $f(x; z) = \langle z, x \rangle$. Let $\{x_k\}_{k=0}^T$ be the iterates of Algorithm 4 on $\mathcal{S}$ with $x_0 = \frac{1}{d} \cdot 1$ for R rounds and $\eta > 0$. Similarly, Let $\{y_k\}_{k=0}^T$ be the iterates of Algorithm 4 on $\mathcal{S}'$ with $y_0 = \frac{1}{d} \cdot 1$ for R rounds and $\eta > 0$. Then after R rounds ($T = Rn$ iterates),*

$$\|x_T - y_T\|_1^2 \le D_{\mathrm{kl}}(x_T, y_T) + D_{\mathrm{kl}}(y_T, x_T) \le 4\eta^2 L^2 R^2.$$

Proof. First, note that

$$\log \frac{x_k}{y_k} = \eta \sum_{i=1}^{k-1} (g_i' - g_i) + C,$$

where C is a constant vector, g_i and g_i' are the (sub)-gradients for $\mathcal{S}$ and $\mathcal{S}'$, respectively. Thus we have that

$$\begin{aligned} D_{\mathrm{kl}}(x_T, y_T) + D_{\mathrm{kl}}(y_T, x_T) &= \langle x_k - y_k, \log \frac{x_k}{y_k} \rangle \\ &= \eta \langle x_k - y_k, \sum_{i=1}^{k-1} (g_i' - g_i) \rangle \\ &\le \eta \sqrt{D_{\mathrm{kl}}(x_T, y_T) + D_{\mathrm{kl}}(y_T, x_T)} \sum_{i=1}^{k-1} \|g_i' - g_i\|_\infty \\ &\le 2\eta L R \sqrt{D_{\mathrm{kl}}(x_T, y_T) + D_{\mathrm{kl}}(y_T, x_T)}\ , \end{aligned}$$

where the first inequality follows from holder's inequality and the strong convexity of KL-divergence with respect to $\|\cdot\|_1$ (this is Pinsker's inequality; see e.g., Duchi, 2019) and the second inequality follows since the first sample z_1 (or z_1') appears R times. The claim follows. □

Algorithm 4 Stochastic Mirror Descent

```
Require: Dataset $\mathcal{S} = (z_1, \ldots, z_n) \in \mathcal{Z}^n$, step sizes $\eta$, initial point $x_0$, number of rounds $R$;
$k \leftarrow 0$
for $r = 1$ to $R$ do
    Sample a random permutation $\pi : [n] \to [n]$
    for $i = 1$ to $n$ do
        Set $g_k = \nabla f(x_k; z_{\pi(i)})$
        Find $x_{k+1} := \mathrm{argmin}_{x \in \Delta_{d-1}} \{\langle g_k, x - x_k \rangle + \frac{1}{\eta} D_{\mathrm{h}}(x, x_k)\}$ where $h(x) = \sum_{j=1}^d x_j \log x_j$
        $k \leftarrow k + 1$
    end for
end for
return $\bar{x}_T = \frac{1}{T} \sum_{k=1}^T x_k$
```

B. Rates for General ℓ_p-Geometry

In this section, we extend our algorithms to work for general ℓ_p-geometries for $p > 1$. Here, the optimization is over the domain $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\|_p \le 1\}$ and we consider functions $f : \mathcal{X} \to \mathbb{R}$ that are L-Lipschitz with respect to $\|\cdot\|_p$, that is, $\|g\|_q \le L$ for all x and sub-gradient $g \in \partial f(x)$ where $1/p + 1/q = 1$.

B.1. Algorithms for ERM for $1 \le p \le 2$

To extend Algorithm 1 to work for general geometries, we need to bound the sensitivity of the gradients. Consider $1 \le p \le 2$ then $q > 2$ which implies that $\|g\|_2 \le d^{1/2 - 1/q} \|g\|_q$, that is, the function is $d^{1/2-1/q} L$ with respect to $\|\cdot\|_2$.

Theorem 10. *Let $1 < p \le 2$, $h : \mathcal{X} \to \mathbb{R}$ be 1-strongly convex with respect to $\|\cdot\|_p$, $x^\star = \mathrm{argmin}_{x \in \mathcal{X}} \hat{F}(x; S)$, and assume $D_{\mathrm{h}}(x^\star, x_0) \le D^2$. Let $f(x; z)$ be convex and L-Lipschitz with respect to $\|\cdot\|_p$ for all $z \in \mathcal{Z}$. Setting $1 \le b$, $T = \frac{n^2}{b^2}$ and*

Algorithm 5 Noisy Mirror Descent for General Geometries

```
Require: Dataset $\mathcal{S} = (z_1, \dots, z_n) \in \mathcal{Z}^n$, $1 < p$ and convex set $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\|_p \le 1\}$, convex function $h : \mathcal{X} \to \mathbb{R}$,
    step sizes $\{\eta_k\}_{k=1}^T$, batch size $b$, initial point $x_0$, number of iterations $T$;
Find $q \ge 1$ such that $1/q + 1/p = 1$
for $k = 1$ to $T$ do
    Sample $S_1, \dots, S_b \sim \mathrm{Unif}(\mathcal{S})$
    Set $\hat{g}_k = \frac{1}{b} \sum_{i=1}^b \nabla f(x_k; S_i) + \zeta_i$ where $\zeta_i \sim \mathsf{N}(0, \sigma^2 I_d)$ with $\sigma = 100L\sqrt{d^{1-2/q}\log(1/\delta)}/b\varepsilon$
    Find $x_{k+1} := \mathrm{argmin}_{x \in \mathcal{X}} \{\langle \hat{g}_k, x - x_k \rangle + \frac{1}{\eta_k} D_{\mathrm{h}}(x, x_k)\}$
end for
return $\bar{x}_T = \frac{1}{T} \sum_{k=1}^T x_k$ (convex)
return $\hat{x}_T = \frac{2}{T(T+1)} \sum_{k=1}^T k x_k$ (strongly convex)
```

$\eta_k = \frac{D}{\sqrt{T}} \frac{1}{\sqrt{L^2 + 4d^{2/q}\sigma^2 \log d}}$, *Algorithm 5 is (ε, δ)-DP and*

$$\mathbb{E}[\hat{F}(\bar{x}_T; S) - \hat{F}(x^\star; S)] \le LD \cdot O\left(\frac{b}{n} + \frac{\sqrt{d \log \frac{1}{\delta}(1 + \log d \cdot 1\{p < 2\})}}{n\varepsilon}\right).$$

Moreover, if $f(x; z)$ is λ-strongly convex relative to $h(x)$, then setting $\eta_k = \frac{2}{\lambda(k+1)}$

$$\mathbb{E}[\hat{F}(\hat{x}_T; S) - \hat{F}(x^\star; S)] \le O\left(\frac{L^2 b^2}{\lambda n^2} + \frac{L^2 d \log \frac{1}{\delta}(1 + \log d \cdot 1\{p < 2\})}{\lambda n^2 \varepsilon^2}\right).$$

Proof. Following the proof of Theorem 3, privacy follows from similar arguments, and for utility we need to upper bound $\mathbb{E}[\|\tilde{g}_k\|_q]$. Note that for $p = q = 2$ we have $\mathbb{E}[\|\tilde{g}_k\|_q^2] \le d$. Otherwise we have

$$\mathbb{E}[\|\tilde{g}_k\|_q^2] \le 2L^2 + 2\mathbb{E}[\|\zeta_k\|_q^2] \le 2L^2 + 2d^{2/q}\mathbb{E}[\|\zeta_k\|_\infty^2] \le 2L^2 + 2d^{2/q}\mathbb{E}[\|\zeta_k\|_\infty^2] \le 2L^2 + 8d^{2/q}\sigma^2 \log d.$$

Now we complete the proof for $p < 2$. The same proof works for $p = 2$. The previous bound implies

$$\begin{aligned}
\mathbb{E}[\hat{F}(\bar{x}_T; S) - \hat{F}(x^\star; S)] &\le \frac{D^2}{T\eta} + \eta L^2 + 4\eta d^{2/q}\sigma^2 \log d \\
&\le 2D\sqrt{(L^2 + 4d^{2/q}\sigma^2 \log d)/T} \\
&\le LD \cdot O\left(\frac{b}{n} + \frac{\sqrt{d \log d \log \frac{1}{\delta}}}{n\varepsilon}\right),
\end{aligned}$$

where the second inequality follows from the choice of η. For the second part, Lemma 3.2 implies that

$$\mathbb{E}[\hat{F}(\hat{x}_T; S) - \hat{F}(x^\star; S)] \le \frac{L^2}{\lambda} O\left(\frac{b^2}{n^2} + \frac{d \log d \log \frac{1}{\delta}}{n^2 \varepsilon^2}\right). \qquad \square$$

B.2. Algorithms for SCO

We extend Algorithm 2 to work for general ℓ_p-geometries by using the general noisy mirror descent (Algorithm 5) to solve the optimization problem at each phase. The following theorem proves our main result for ℓ_p-geometry, that is, Theorem 5.

Theorem 11. *Let $1 < p \le 2$. Assume $\mathsf{diam}_p(\mathcal{X}) \le D$ and $f(x; z)$ is convex and L-Lipschitz with respect to $\|\cdot\|_p$ for all $z \in \mathcal{Z}$. If we set*

$$\eta = \frac{D}{L} \min\left\{1/\sqrt{(p-1)n}, \varepsilon/\sqrt{d \log \frac{1}{\delta}(1 + \log d \cdot 1\{p < 2\})}\right\},$$

then Algorithm 6 requires $O(\log n \cdot \min(n^{3/2}\sqrt{\log d}, n^2\varepsilon/\sqrt{d}))$ gradients and its output has

$$\mathbb{E}[F(x_k) - F(x^\star)] = LD \cdot O\left(\frac{1}{\sqrt{(p-1)n}} + \frac{\sqrt{d \log \frac{1}{\delta}(1 + \log d \cdot 1\{p < 2\})}}{(p-1)n\varepsilon} \right).$$

Proof. The proof follows from identical argument to the proof of Theorem 4 using the fact that $h_i(x) = \frac{1}{2(p-1)} \|x - x_{i-1}\|_p^2$ is 1-strongly convex with respect to $\|\cdot\|_p$. □

Algorithm 6 Localized Noisy Mirror Descent

```
Require: Dataset $\mathcal{S} = (z_1, \ldots, z_n) \in \mathcal{Z}^n$, $1 \le p$, constraint set $\mathcal{X}$, step size $\eta$, initial point $x_0$;
Set $k = \lceil \log n \rceil$
for $i = 1$ to $k$ do
    Set $n_i = 2^{-i} n$, $\eta_i = 2^{-4i}\eta$
    Apply Algorithm 5 with $(\varepsilon, \delta)$-DP, batch size $b_i = \max(\sqrt{n_i/\log d}, \sqrt{d/\varepsilon})$, $T = n_i^2/b_i^2$ and $h_i(x) = \frac{1}{2(p-1)} \|x - x_{i-1}\|_p^2$ for solving the ERM over $\mathcal{X}_i = \{x \in \mathcal{X} : \|x - x_{i-1}\|_p \le 2L\eta_i n_i (p-1)\}$:
        $$F_i(x) = \frac{1}{n_i} \sum_{j=1}^{n_i} f(x; z_j) + \frac{1}{\eta_i n_i (p-1)} \|x - x_{i-1}\|_p^2$$
    Let $x_i$ be the output of the private algorithm
end for
return the final iterate $x_k$
```

C. Implications for Strongly Convex Functions

When the function is strongly convex, we use standard reductions to the convex case to achieve better rates (Feldman et al., 2020a). Given a private algorithm $\mathcal{A}$ for the convex case, we use the following algorithm for the strongly convex case (see (Feldman et al., 2020a)): run $\mathcal{A}$ for $k = \lceil \log \log n \rceil$ iterates, each initialized at the output of the previous iterate and run for $n_i = 2^{i-2} n / \log n$. Using this reduction with our algorithms for convex functions, we have the following theorems for non-smooth and smooth functions.

Theorem 12. *Assume $\mathsf{diam}_1(\mathcal{X}) \le D$ and $f(x; z)$ is convex, L-Lipschitz, and λ-strongly convex with respect to $\|\cdot\|_1$ for all $z \in \mathcal{Z}$. Then using Algorithm 2 in the above algorithm results in an algorithm that uses $O(\log n \log \log n \cdot \min(n^{3/2}\sqrt{\log d}, n^2\varepsilon/\sqrt{d}))$ gradients and outputs $\hat{x}$ such that*

$$\mathbb{E}[F(\hat{x}) - F(x^\star)] = LD \cdot O\left(\frac{\log d}{n} + \frac{d \log^3 d \log \frac{1}{\delta}}{n^2 \varepsilon^2} \right).$$

Theorem 13. *Let $\delta \le 1/n$ and assume that $\mathsf{diam}_1(\mathcal{X}) \le D$, $m \le O(d)$ and that $f(x; z)$ is convex, L-Lipschitz, λ-strongly convex and β-smooth with respect to $\|\cdot\|_1$ where $\beta = O(L/D)$. Then using Algorithm 3 in the above algorithm results in an algorithm that uses $O(n)$ gradients and outputs $\hat{x}$ such that*

$$\mathbb{E}[F(\hat{x}) - F(x^\star)] \le LD \cdot O\left(\frac{\log d \log^2 n}{n} \right) + LD \cdot O\left(\frac{\log(1/\delta) \log m \log^2 n}{n\varepsilon} \right)^{4/3}.$$

The proof follows directly from the proof of Theorem 5.1 in (Feldman et al., 2020a), together with the bounds of Section 3 and Section 4.

D. Proofs of Section 3

D.1. Proof of Theorem 3

Proof. The privacy proof follows directly using moments accountant, that is, Theorem 1 in (Abadi et al., 2016), by noting the the ℓ_2-norm of the gradients is bounded by $\|\nabla f(x; z_i)\|_2 \le \|\nabla f(x; z_i)\|_\infty \sqrt{d} \le L\sqrt{d}$ for all $x \in \mathcal{X}$ and $z \in \mathcal{Z}$. Now we analyze the utility of the algorithm. To this end, we have that $\mathbb{E}[\|\hat{g}_k\|_\infty^2] \le 2L^2 + 2\mathbb{E}[\|\zeta_k\|_\infty^2] \le 2L^2 + 4\sigma^2 \log d$. Lemma 3.1 now implies that

$$\begin{aligned}\mathbb{E}[\hat{F}(\bar{x}_T; S) - \hat{F}(x^\star; S)] &\le \frac{D^2}{T\eta} + \eta L^2 + 2\eta\sigma^2 \log d \\ &\le 2\frac{D\sqrt{L^2 + 2\sigma^2 \log d}}{\sqrt{T}} \\ &\le O\left(LD\left(\frac{b}{n} + \frac{\sqrt{d \log d \log(1/\delta)}}{n\varepsilon}\right)\right),\end{aligned}$$

where the second inequality follows from the choice of η. Now we prove the claim for strongly convex functions. Lemma 3.2 implies that

$$\mathbb{E}[\hat{F}(\hat{x}_T; S) - \hat{F}(x^\star; S)] \le O\left(\frac{L^2 b^2}{\lambda n^2} + \frac{L^2 d \log d \log(1/\delta)}{\lambda n^2 \varepsilon^2}\right). \qquad \square$$

D.2. Proof of Lemma 3.2

Proof. First, by strong convexity we have

$$\begin{aligned}f(x_k) - f(x^\star) &\le \langle \nabla f(x_k), x_k - x^\star\rangle - \lambda D_{\mathrm{h}}(x^\star, x_k) \\ &= \langle g_k, x_k - x^\star\rangle + \langle \nabla f(x_k) - g_k, x_k - x^\star\rangle - \lambda D_{\mathrm{h}}(x^\star, x_k). \qquad (2)\end{aligned}$$

Let us now focus on the term $\langle g_k, x_k - x^\star\rangle$. The definition of x_{k+1} implies that for all $y \in \mathcal{X}$

$$\langle g_k + \frac{1}{\eta_k}(\nabla h(x_{k+1}) - \nabla h(x_k)), y - x_{k+1}\rangle \ge 0.$$

Substituting $y = x^\star$, we have

$$\begin{aligned}\langle g_k, x_k - x^\star\rangle &= \langle g_k, x_k - x_{k+1}\rangle + \langle g_k, x_{k+1} - x^\star\rangle \\ &\le \langle g_k, x_k - x_{k+1}\rangle + \frac{1}{\eta_k}\langle \nabla h(x_{k+1}) - \nabla h(x_k), x^\star - x_{k+1}\rangle \\ &\overset{(i)}{=} \langle g_k, x_k - x_{k+1}\rangle + \frac{1}{\eta_k}\left(D_{\mathrm{h}}(x^\star, x_k) - D_{\mathrm{h}}(x^\star, x_{k+1}) - D_{\mathrm{h}}(x_{k+1}, x_k)\right) \\ &\overset{(ii)}{\le} \frac{\eta_k}{2}\|g_k\|_\infty^2 + \frac{1}{2\eta_k}\|x_k - x_{k+1}\|_1^2 + \frac{1}{\eta_k}\left(D_{\mathrm{h}}(x^\star, x_k) - D_{\mathrm{h}}(x^\star, x_{k+1}) - D_{\mathrm{h}}(x_{k+1}, x_k)\right) \\ &\overset{(iii)}{\le} \frac{\eta_k}{2}\|g_k\|_\infty^2 + \frac{1}{\eta_k}\left(D_{\mathrm{h}}(x^\star, x_k) - D_{\mathrm{h}}(x^\star, x_{k+1})\right),\end{aligned}$$

where (i) follows from the definition of bregman divergence, (ii) follows from Fenchel-Young inequality, and (iii) follows since $h(x)$ is 1-strongly convex with respect to $\|\cdot\|_1$. Substituting into (2),

$$f(x_k) - f(x^\star) \le \frac{\eta_k}{2}\|g_k\|_\infty^2 + \langle \nabla f(x_k) - g_k, x_k - x^\star\rangle + \frac{1}{\eta_k}\left(D_{\mathrm{h}}(x^\star, x_k) - D_{\mathrm{h}}(x^\star, x_{k+1})\right) - \lambda D_{\mathrm{h}}(x^\star, x_k).$$

Multiplying by k and summing from $k = 1$ to T, we get

$$\begin{aligned}\sum_{k=1}^T k(f(x_k) - f(x^\star)) &\le \frac{1}{2\lambda}\sum_{k=1}^T \|g_k\|_\infty^2 + \langle \nabla f(x_k) - g_k, x_k - x^\star\rangle \\ &\quad + \frac{\lambda}{2}\left(k(k-1)D_{\mathrm{h}}(x^\star, x_k) - k(k+1)D_{\mathrm{h}}(x^\star, x_{k+1})\right) \\ &\le \frac{1}{2\lambda}\sum_{k=1}^T \|g_k\|_\infty^2 + \langle \nabla f(x_k) - g_k, x_k - x^\star\rangle.\end{aligned}$$

The claim now follows by taking expectations and using Jensen's inequality. □

D.3. Proof of Lemma 3.3

First, we prove that $\hat{x}_i \in \mathcal{X}_i$. The definition of $\hat{x}_i$ implies that

$$\frac{1}{n_i}\sum_{j=1}^{n_i} f(\hat{x}_i; z_j) + \frac{1}{\eta_i n_i (p-1)} \left\| \hat{x}_i - x_{i-1} \right\|_p^2 \le \frac{1}{n_i}\sum_{j=1}^{n_i} f(x_{i-1}; z_j).$$

Since $f(x;z)$ is L-Lipschitz, we get

$$\frac{1}{\eta_i n_i (p-1)} \left\| \hat{x}_i - x_{i-1} \right\|_p^2 \le L \left\| \hat{x}_i - x_{i-1} \right\|_1 \le 2L \left\| \hat{x}_i - x_{i-1} \right\|_p$$

where the last inequality follows from the choice of p (since $\|z\|_1 \le d^{1-1/p} \|z\|_p \le 2\|z\|_p$ for all $z \in \mathbb{R}^d$), hence we get $\|\hat{x}_i - x_{i-1}\|_p \le 2L\eta_i n_i (p-1)$. Thus, we have that $\hat{x}_i \in \mathcal{X}_i = \{x : \|x - x_{i-1}\|_p \le 2L\eta_i n_i (p-1)\}$.

Now, note that the function $F_i(x)$ is λ_i-strongly convex relative to $h_i(x) = \frac{1}{p-1}\|x - x_{i-1}\|_p^2$ where $\lambda_i = \frac{1}{\eta_i n_i}$. Moreover, the function $r_i(x) = \frac{1}{\eta_i n_i (p-1)}\|x - x_{i-1}\|_p^2$ is $4L$-Lipschitz with respect to $\|\cdot\|_1$ for $x \in \mathcal{X}_i$. Therefore using the bounds of Theorem 3 for noisy mirror descent and observing that $F_i(x)$ is λ_i-strongly convex with respect to $\|\cdot\|_p$,

$$\frac{\lambda_i}{2}\mathbb{E}[\|x_i - \hat{x}_i\|_p^2] \le \mathbb{E}[F_i(x_i) - F_i(\hat{x}_i)] \le O\left(\frac{L^2 d \log d \log(1/\delta)}{n_i^2 \varepsilon^2 \lambda_i}\right),$$

implying that $\mathbb{E}[\|x_i - \hat{x}_i\|_p^2] \le O\left(\frac{L^2 \eta_i^2 d \log d \log(1/\delta)}{\varepsilon^2}\right)$.

D.4. Proof of Lemma 3.4

The proof follows from Theorems 6 and 7 in (Shalev-Shwartz et al., 2009) by noting that the function $r(x; z_j) = f(x; z_j) + \frac{1}{\eta_i n_i (p-1)}\|x - x_{i-1}\|_p^2$ is $\frac{1}{\eta_i n_i}$-strongly convex and $O(L)$-Lipschitz with respect to $\|\cdot\|_1$ over $\mathcal{X}_i$.

E. Proofs of Section 4

To simplify notation, in this section we use the notion of (ε, δ)-indistinguishability; we say that two random variables X and Y are (ε, δ)-indistinguishable, denoted $X \approx_{(\varepsilon,\delta)} Y$, if for every $\mathcal{O}$, $\Pr(X \in \mathcal{O}) \le e^{\varepsilon}\Pr[Y \in \mathcal{O}] + \delta$ and $\Pr(Y \in \mathcal{O}) \le e^{\varepsilon}\Pr[X \in \mathcal{O}] + \delta$.

E.1. Proof of Lemma 4.1

The main idea for the privacy proof is that each sample in the set $S_{t,s}$ is used in the calculation of $v_{t,s}$ at most $N_{t,s} = 2^{t-|s|}$ times, hence setting the noise large enough so that each iterate is $\frac{\varepsilon}{N_{t,s}}$-DP, we get that the final mechanism is ε-DP using basic composition. Let us now provide a more formal argument. Let $\mathcal{S} = (z_1, \dots, z_{n-1}, z_n), \mathcal{S}' = (z_1, \dots, z_{n-1}, z_n')$ be two neighboring datasets with iterates $x = (x_1, \dots, x_K)$ and $x' = (x_1', \dots, x_K')$, respectively. We prove that x and x' are ε-indistinguishable, i.e., $x \approx_{(\varepsilon,0)} x'$. Let $S_{t,s}$ be the set (vertex) that contains the last sample (i.e., z_n or z_n') and let $j = |s|$ denote the depth of this vertex. We will prove privacy given that the n'th sample is in $S_{t,s}$, which will imply our general privacy guarantee as this holds for every choice of t and s.

Note that $|S_{t,s}| = 2^{-j}b$ and that this set is used in the calculation of v_k for at most 2^{t-j} (consecutive) iterates, namely these are leafs that are descendants of the vertex $u_{t,s}$. Let k_0 and k_1 be the first and last iterate such that the set $S_{t,s}$ is used for the calculation of v_k, hence $k_1 - k_0 + 1 \le 2^{t-j}$. The iterates $(x_1, \dots, x_{k_0 - 1})$ and $(x_1', \dots, x_{k_0-1}')$ do not depend on the last sample and therefore has the same distribution (hence 0-indistinguishable). Moreover, given that $(x_{k_0}, \dots, x_{k_1}) \approx_{(\varepsilon,0)} (x_{k_0}', \dots, x_{k_1}')$, it is clear that the remaining iterates $(x_{k_1+1}, \dots, x_K) \approx_{(\varepsilon,0)} (x_{k_1+1}', \dots, x_K')$ by post-processing as they do depend on the last sample only through the previous iterates. It is therefore enough to prove that $(x_{k_0}, \dots, x_{k_1}) \approx_{(\varepsilon,0)} (x_{k_0}', \dots, x_{k_1}')$. To this end, we prove that for each such iterate, $w_k \approx_{(\varepsilon/2^{t-j},0)} w_k'$, hence using post-processing and basic composition (Lemma 2.1) the iterates are ε-indistinguishable as $k_1 - k_0 + 1 \le 2^{t-j}$. Note that for every $k_0 \le k \le k_1$ the sensitivity $|\langle c_i, v_k - v_k' \rangle| \le \frac{DL}{2^{-j}b}$. Hence, using privacy guarantees of report noisy max [Dwork & Roth, 2014, claim 3.9], we have that $w_k \approx_{(\varepsilon/2^{t-j},0)} w_k'$ since $\lambda_{t,s} = \frac{2LD2^t}{b\varepsilon}$.

Now we prove the second part of the claim. Standard results for the expectation of the maximum of m Laplace random variables imply that $\mathbb{E}[\langle v_{t,s}, w_{t,s}\rangle] \le \min_{1 \le i \le m} \langle v_{t,s}, c_i \rangle + O(\frac{LD2^t}{b\varepsilon} \log m)$. Since $\mathcal{X} = \mathsf{conv}\{c_1, \dots, c_m\}$, we know that for any $v \in \mathbb{R}^d$, $\operatorname{argmin}_{w \in \mathcal{X}} \langle w, v \rangle \cap \{c_1, \dots, c_m\} \neq \emptyset$ [Talwar et al., 2015, fact 2.3] which proves the claim.

E.2. Proof of Lemma 4.2

The claim follows directly from the following lemma.

Lemma E.1. *Let (s,t) be a vertex and $\sigma^2 = (L^2 + \beta^2 D^2)/b$. For every index $1 \le i \le d$,*

$$\mathbb{E}\left[e^{\lambda(v_{t,s,i} - \nabla F_i(x_{t,s}))}\right] \le e^{O(1)\lambda^2\sigma^2}.$$

Lemma E.1 says that $v_{k,i} - \nabla F_i(x_k)$ is $O(\sigma^2)$-sub-Gaussian for every $1 \le i \le d$, hence standard results imply that the maximum of d sub-Gaussian random variables is $\mathbb{E}\left\|v_{t,s} - \nabla F(x_{t,s})\right\|_\infty \le O(\sigma)\sqrt{\log d}$. The claim follows.

Lemma E.1. Let us fix i for simplicity and let $B_{t,s} = v_{t,s,i} - \nabla F_i(x_{t,s})$. We prove the claim by induction on the depth of the vertex, i.e., $j = |s|$. If $j = 0$ then $s = \emptyset$ which implies that $v_{t,\emptyset} = \nabla f(x_{t,\emptyset}; S_{t,\emptyset})$ where $S_{t,\emptyset}$ is a sample of size b. Thus we have

$$\begin{aligned}
\mathbb{E}[e^{\lambda B_{t,\emptyset}}] &= \mathbb{E}\left[e^{\lambda(v_{t,\emptyset,i} - \nabla F_i(x_{t,\emptyset}))}\right] \\
&= \mathbb{E}\left[e^{\lambda(\frac{1}{b}\sum_{s \in S_{t,\emptyset}} \nabla f_i(x_{t,\emptyset}; s) - \nabla F_i(x_{t,\emptyset}))}\right] \\
&= \prod_{s \in S_{t,\emptyset}} \mathbb{E}[e^{\frac{\lambda}{b}(\nabla f_i(x_{t,\emptyset}; s) - \nabla F_i(x_{t,\emptyset}))}] \\
&\le e^{\lambda^2 L^2/2b},
\end{aligned}$$

where the last inequality follows since for a random variable $X \in [-L, L]$ and $\mathbb{E}[X] = 0$, we have $\mathbb{E}[e^{\lambda X}] \le e^{\lambda^2 L^2/2}$ [Duchi, 2019, example 3.6]. Assume now we have s with $|s| = j > 0$ and let $s = s'c$ where $c \in \{0, 1\}$. If $c = 0$ the claim clearly holds so we assume $c = 1$. Recall that in this case $v_{t,s} = v_{t,s'} + \nabla f(x_{t,s}; S_{t,s}) - \nabla f(x_{t,s'}; S_{t,s})$, hence $B_{t,s} = v_{t,s,i} - \nabla F_i(x_{t,s}) = B_{t,s'} + \nabla f_i(x_{t,s}; S_{t,s}) - \nabla f_i(x_{t,s'}; S_{t,s}) - \nabla F_i(x_{t,s}) + \nabla F_i(x_{t,s'})$ Letting $S_{<t,s} = \cup_{(t_1,s_1)<(t,s)} S_{t_1,s_1}$ be the set of all samples used up to vertex t, s, the law of iterated expectation implies

$$\begin{aligned}
\mathbb{E}[e^{\lambda B_{t,s}}] &= \mathbb{E}[e^{\lambda(B_{t,s'} + \nabla f_i(x_{t,s}; S_{t,s}) - \nabla f_i(x_{t,s'}; S_{t,s}) - \nabla F_i(x_{t,s}) + \nabla F_i(x_{t,s'}))}] \\
&= \mathbb{E}\left[\mathbb{E}[e^{\lambda(B_{t,s'} + \nabla f_i(x_{t,s}; S_{t,s}) - \nabla f_i(x_{t,s'}; S_{t,s}) - \nabla F_i(x_{t,s}) + \nabla F_i(x_{t,s'}))}] \mid S_{<(t,s)}\right] \\
&= \mathbb{E}\left[\mathbb{E}[e^{\lambda B_{t,s'})} \mid S_{<(t,s)}] \cdot \mathbb{E}[e^{\lambda(\nabla f_i(x_{t,s}; S_{t,s}) - \nabla f_i(x_{t,s'}; S_{t,s}) - \nabla F_i(x_{t,s}) + \nabla F_i(x_{t,s'}))} \mid S_{<t,s}]\right] \\
&= \mathbb{E}[e^{\lambda B_{t,s'})}] \cdot \mathbb{E}[e^{\lambda(\nabla f_i(x_{t,s}; S_{t,s}) - \nabla f_i(x_{t,s'}; S_{t,s}) - \nabla F_i(x_{t,s}) + \nabla F_i(x_{t,s'}))} \mid S_{<t,s}].
\end{aligned}$$

Since $f(\cdot; s)$ is β-smooth with respect to $\|\cdot\|_1$, we have that $|\nabla f_i(x_{t,s}; S_{t,s}) - \nabla f_i(x_{t,s'}; S_{t,s})| \le \beta \left\|x_{t,s} - x_{t,s'}\right\|_1$. Moreover, as $u_{t,s}$ is the right son of $u_{t,s'}$, the number of updates between $x_{t,s}$ and $x_{t,s'}$ is at most the number of leafs visited between these two vertices which is 2^{t-j}. Hence we get that

$$\left\|x_{t,s} - x_{t,s'}\right\|_1 \le D\eta_{t,s'} 2^{t-j} \le D2^{-j+2},$$

which implies that $|\nabla f_i(x_{t,s}; S_{t,s}) - \nabla f_i(x_{t,s'}; S_{t,s})| \le \beta D 2^{-j+2}$. Since $\mathbb{E}[\nabla f_i(x_{t,s}; S_{t,s}) - \nabla f_i(x_{t,s'}; S_{t,s}) \mid S_{<t,s}] = \nabla F_i(x_{t,s}) - \nabla F_i(x_{t,s'})$, by repeating similar arguments to the case $\ell = 0$, we get that

$$\begin{aligned}
\mathbb{E}[e^{\lambda(\nabla f_i(x_{t,s}; S_{t,s}) - \nabla f_i(x_{t,s'}; S_{t,s}) - \nabla F_i(x_{t,s}) + \nabla F_i(x_{t,s'}))} \mid S_{<t,s}] &\le e^{O(1)\lambda^2\beta^2 D^2 2^{-2j}/|S_{t,s}|} \\
&\le e^{O(1)\lambda^2\beta^2 D^2 2^{-j}/b}.
\end{aligned}$$

Overall we have that $\mathbb{E}[e^{\lambda B_{t,s}}] \le \mathbb{E}[e^{\lambda B_{t,s'}}] \cdot e^{O(1)\lambda^2\beta^2 D^2 2^{-j}/b}$. Applying this inductively, we get that for every (t, s)

$$\mathbb{E}[e^{\lambda B_{t,s}}] \le e^{O(1)\lambda^2(L^2 + \beta^2 D^2)/b}. \qquad \square$$

E.3. Proof of Lemma 4.3

For this proof, we use the following privacy amplification by shuffling.

Lemma E.2 (Feldman et al., 2020b, Theorem 3.8)**.** *Let $\mathcal{A}_i : \mathcal{T}^{i-1} \times \mathcal{Z} \to \mathcal{T}$ for $i \in [n]$ be a sequence of algorithm such that $\mathcal{A}_i(w_{1:i-1}, \cdot)$ is $(\varepsilon_0, \delta_0)$-DP for all values of $w_{1:i-1} \in \mathcal{T}^{i-1}$ with $\varepsilon_0 \le O(1)$. Let $\mathcal{A}_S : \mathcal{Z}^n \to \mathcal{T}^n$ be an algorithm that given $z_{1:n} \in \mathcal{Z}^n$, first samples a random permutation π, then sequentially computes $w_i = \mathcal{A}_i(w_{1:i-1}, z_{\pi(i)})$ for $i \in [n]$ and outputs $w_{1:n}$. Then for any δ such that $\varepsilon_0 \le \log(\frac{n}{16\log(2/\delta)})$, the algorithm $\mathcal{A}_s$ is $(\varepsilon, \delta + 20n\delta_0)$ where $\varepsilon \le O(\varepsilon_0\sqrt{\log(1/\delta)/n})$.*

We use the same notation as Lemma 4.1 where $\mathcal{S} = (z_1, \dots, z_{n-1}, z_n), \mathcal{S}' = (z_1, \dots, z_{n-1}, z'_n)$ denote two neighboring datasets with iterates $x = (x_1, \dots, x_K)$ and $x' = (x'_1, \dots, x'_K)$. Here, we prove privacy after conditioning on the event that the n'th sample is sampled at phase t and depth j. We need to show that the iterates are (ε, δ)-indistinguishable. We only need to prove privacy for the iterates at phase t as the iterates before phase t do not depend on the n'th sample and the iterates after phase t are (ε, δ)-indistinguishable by post-processing.

Let us now focus on the iterates at phase t. Let $u_1, \dots, u_p$ denote the vertices at level j that has samples $S_1, \dots, S_p$ each of size $|S_i| = 2^{-j}b$. We will have two steps in the proof. First, we use advanced composition to show that the iterates that are descendant of a vertex u_i are $(\varepsilon_0, \delta_0)$-DP where roughly $\varepsilon_0 = 2^{j/2}\varepsilon$. As we have $p = 2^j$ vertices at depth j, we then use the amplification by shuffling result to argue that the final privacy guarantee is (ε, δ)-DP (see Fig. 2 for a demonstration of the shuffling in our algorithm).

Let $\mathcal{A}_i$ be the algorithm that outputs the iterates corresponding to the leafs that are descendants of u_i; we denote this output by O_i. Note that the inputs of $\mathcal{A}_i$ are the samples at u_i, which we denote as S_i, and the previous outputs $O_1, \dots, O_{i-1}$. In this notation, we have that $O_i = \mathcal{A}_i(O_1, \dots, O_{i-1}, S_i)$. We let $\mathcal{A}_i$, S_i and O_i denote the above quantities when the input dataset is $\mathcal{S}_i$ and similarly $\mathcal{A}'_i$, S'_i and O'_i for $\mathcal{S}'$. To prove privacy, we need to show that $(O_1, \dots, O_p) \approx_{(\varepsilon,\delta)} (O'_1, \dots, O'_p)$, that is $(O_1, \dots, O_p)$ and $(O'_1, \dots, O'_p)$ are (ε, δ)-indistinguishable

To this end, we first describe an equivalent sampling procedure for the sets $S_1, \dots, S_p$. Given r samples, the algorithm basically constructs the sets $S_1, \dots, S_p$ by sampling uniformly at random p sets of size r/p without repetition. Instead, we consider the following sampling procedure. First, we randomly choose a set of size $p(r-1)$ samples that does not include the n'th sample and using this set we randomly choose $r/p - 1$ samples for each set S_i. Then, we shuffle the remaining p samples and add each sample to the corresponding set. It is clear that this sampling procedure is equivalent. We prove privacy conditional on the output of the first stage of the randomization procedure which will imply privacy unconditionally.

Assuming without loss of generality that the samples which remained in the second stage are $z_{n-p+1}, \dots, z_n$, and letting $\pi : [p] \to \{n-p+1, \dots, n\}$ denote the random permutation of the second stage, the algorithms $\mathcal{A}_i$ and $\mathcal{A}'_i$ can be written as a function of the previous outputs and the sample $z_{\pi(i)}$. This is true since the $\mathcal{S}$ and $\mathcal{S}'$ differ in one sample and therefore the first $r/p - 1$ samples in the sets S_i and S'_i are identical. Thus, we can write $O_i = \mathcal{A}_i(O_1, \dots, O_{i-1}, z_{\pi(i)})$.

Using the above notation, we are now ready to prove privacy. First, we show privacy for each i using advanced composition. Similarly to Lemma 4.1, as each iterate k which is a leaf of u_i has sensitivity $|\langle c_i, v_k - v'_k\rangle| \le \frac{DL}{2^{-j}b}$, we have that x_k and x'_k are $\frac{\varepsilon}{2^{T/2-j}\log(n/\delta)}$-indistinguishable since $\lambda_{t,s} = \frac{LD2^{T/2}\log(n/\delta)}{b\varepsilon}$. Since there are 2^{t-j} leafs of u_i, advanced composition (Lemma 2.2) implies that $O_i \approx_{(\varepsilon_0,\delta_0)} O'_i$ where $\varepsilon_0 = \frac{\varepsilon}{2^{T/2-j}\log(n/\delta)}\sqrt{2^{t-j}\log(1/\delta_0)} \le \frac{O(\varepsilon)}{\sqrt{\log(1/\delta)}2^{-j/2}}$ by setting $\delta_0 = \delta/n$.

Finally, we can use the amplification by shuffling result to finish the proof. First, note that we need $\varepsilon_0 \le \log(\frac{2^j}{16\log(2/\delta)})$ to be able to use Lemma E.2. If $2^j \le O(\log(1/\delta))$ then we do not need the amplification by shuffling result as $\varepsilon_0 \le O(\varepsilon 2^{j/2}/\sqrt{\log(1/\delta)}) \le O(\varepsilon)$. Otherwise 2^j is large enough so that we can use Lemma E.2. Since each $\mathcal{A}_i$ and $\mathcal{A}'_i$ are $(\varepsilon_0, \delta_0)$-DP and since the second stage shuffles the inputs to each algorithm, Lemma E.2 now implies that the outputs of the algorithms $\mathcal{A}_i$ and $\mathcal{A}'_i$ are $(\varepsilon_f, \delta + 20n\delta_0)$-DP where $\varepsilon_f \le \frac{\varepsilon_0\sqrt{\log(1/\delta)}}{2^{j/2}} \le O(\varepsilon)$ which proves the claim.

E.4. Proof of Theorem 7

The assumptions on β ensure that $2^T \le b$ and the assumptions on ε ensure $\varepsilon \le 2^{-T/2}\log(n/\delta)$ hence the privacy follows from Lemma 4.3. The utility analysis is similar to the proof of Theorem 6. Repeating the same arguments in the proof of Theorem 6 while using the new value of $\lambda_{t,s}$, we get

$$\mathbb{E}[F(x_K) - F(x^\star)] \le O\left(D(L+\beta D)\frac{\sqrt{\log d}}{\sqrt{b}} + \frac{\beta D^2}{2^T} + DL\frac{2^{T/2}\log(n/\delta)\log m}{b\varepsilon} \right).$$

As the number of samples is upper bounded by $T^2 \cdot b$, we set $T = \frac{2}{3}\log\left(\frac{b\varepsilon\beta D}{L\log(n/\delta)\log m}\right)$ and $b = n/\log^2 n$ to get the first part of the theorem. Note that the condition on β ensure the term inside the log is greater than 1.

F. Proofs for Section 5

F.1. Proofs for Lemma 5.1

Without loss of generality, we assume that $D = 1$. Moreover, similarly to the proof of Theorem 9, we prove lower bounds on the sample complexity to achieve a certain error which will imply our lower bound on the utility. For an algorithm $\mathcal{A}$ and data $\mathcal{S} \in \mathcal{Z}^n$, define the error of $\mathcal{A}$:

$$\mathsf{Err}(\mathcal{A}, \mathcal{S}) = \mathbb{E}\left[\sum_{j=1}^{d} |\bar{z}_j| 1\{\text{sign}(\mathcal{A}(\mathcal{S})_j) \ne \text{sign}(\bar{z}_j)\} \right].$$

The error of a $\mathcal{A}$ for datasets of size n is $\mathsf{Err}(\mathcal{A}, n) = \sup_{\mathcal{S}\in\mathcal{Z}^n} \mathsf{Err}(\mathcal{A}, \mathcal{S})$.

We let $n^\star(\alpha, \varepsilon)$ denote the minimal n such that there is an (ε, δ)-DP (with $\delta = n^{-\omega(1)}$) mechanism $\mathcal{A}$ such that $\mathsf{Err}(\mathcal{A}, n^\star(\alpha, \varepsilon)) \le \alpha$. We prove the following lower bound on the sample complexity which implies Lemma 5.1.

Proposition 1. *Let $z_i \in \{-1/d, 1/d\}^d$, $\alpha \le 1$, and $\varepsilon \le 1$. Then*

$$n^\star(\alpha, \varepsilon) \ge \Omega(1) \cdot \frac{\sqrt{d}}{\alpha\varepsilon\log d}.$$

The proof follows directly from the following two lemmas.

Lemma F.1 (Talwar et al. (2015), Theorem 3.2)**.** *Let the assumptions of Proposition 1 hold. Then*

$$n^\star(\alpha = 1/4, \varepsilon = 0.1) \ge \Omega(1) \cdot \frac{\sqrt{d}}{\log d}.$$

The following lemma shows how to extend the above lower bound to arbitrary accuracy and privacy parameters.

Lemma F.2. *Let $\varepsilon_0 \le 0.1$. For $\alpha \le \alpha_0/2$ and $\varepsilon \le \varepsilon_0/2$,*

$$n^\star(\alpha, \varepsilon) \ge \frac{\alpha_0\varepsilon_0}{\alpha\varepsilon} n^\star(\alpha_0, \varepsilon_0).$$

Proof. The proof follows the same arguments as in the proof of Lemma F.5. □

F.2. Proof of Theorem 9

In this section, we prove Theorem 9. We begin by recalling the lower bound of Talwar et al. (2015) and showing how it implies Lemma F.3.

Talwar et al. (2015) consider the family of quadratic functions where $f(x; a_i, b_i) = (a_i^T x - b_i)^2$ where $a_i \in \mathbb{R}^d$ and $b_i \in \mathbb{R}$. We assume $\mathcal{X} = \{x : \|x\|_1 \le D\}$, $\|a_i\|_\infty \le C$, and $|b_i| \le CD$. Note that the function f is L-Lipschitz and β-smooth with $L \le O(C^2 D)$ and $\beta \le O(C^2)$ and there is a choice of a_i, b_i that attains these. Theorem 3.1 in (Talwar et al., 2015) gives a lower bound of $1/n^{2/3}$ when $C = 1$, $D = 1$, and $d \ge \widetilde{\Omega}(n^{2/3})$. For general values of C and D, noticing that the function value is multiplied by $C^2 D^2$, the following lower bound follows as $LD = C^2 D^2$.

Lemma F.3. *Let $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\|_1 \le D\}$ and $d \ge \widetilde{\Omega}(n^{2/3})$. There is family of convex functions $f : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ that is L-Lipschitz and β-smooth with $\beta \le L/D$ such that any $(0.1, \delta)$-DP algorithm $\mathcal{A}$ with $\delta = o(1/n^2)$ has*

$$\sup_{\mathcal{S}\in\mathcal{Z}^n} \mathbb{E}\left[\hat{F}(\mathcal{A}(\mathcal{S}); \mathcal{S}) - \min_{x\in\mathcal{X}} \hat{F}(x; \mathcal{S}) \right] \ge \widetilde{\Omega}\left(\frac{LD}{n^{2/3}} \right).$$

Now we proceed to prove Theorem 9 and we assume without loss of generality that $L = 1$ and $D = 1$. We use techniques from (Steinke & Ullman, 2017) to extend the lower bound of Lemma F.3 to hold for arbitrary d and ε. To this end, instead of lower bounding the excess loss, it will be convenient to prove lower bounds on the sample size to achieve a certain excess loss α. More precisely, given a dataset $\mathcal{S} \in \mathcal{Z}^n$ and algorithm $\mathcal{A}$, we define its empirical excess loss on $\mathcal{S}$

$$\mathcal{E}(\mathcal{A}, \mathcal{S}) = \mathbb{E}\left[\hat{F}(\mathcal{A}(\mathcal{S}); \mathcal{S}) - \min_{x \in \mathcal{X}} \hat{F}(x; \mathcal{S})\right].$$

We also define its worst-case excess loss over all datasets of size n

$$\mathcal{E}(\mathcal{A}, n) = \sup_{\mathcal{S} \in \mathcal{Z}^n} \mathcal{E}(\mathcal{A}, \mathcal{S}).$$

We let $n^\star(\alpha, \varepsilon)$ be the minimal sample size that is required to achieve excess loss $\mathcal{E}(\mathcal{A}, n^\star(\alpha, \varepsilon)) \le \alpha$ using an (ε, δ)-DP algorithm $\mathcal{A}$ with $\delta = n^{-\omega(1)}$. We prove the following lemma which implies Theorem 9.

Lemma F.4. *Let the assumptions of Theorem 9 hold. Then*

$$n^\star(\alpha, \varepsilon) \ge \begin{cases} \widetilde{\Omega}\left(\frac{1}{\alpha^{3/2}\varepsilon}\right) & \text{if } \alpha = 1/d \\ \widetilde{\Omega}\left(\frac{\sqrt{d}}{\alpha\varepsilon}\right) & \text{if } \alpha \le 1/d \end{cases}$$

The proof of Lemma F.4 basically follows from the following two Lemmas.

Lemma F.5. *For* $0 < \alpha \le \alpha_0$ *and* $0 < \varepsilon \le \varepsilon_0 \le 0.1$,

$$n^\star(\alpha, \varepsilon) \ge \Omega\left(\frac{\alpha_0 \varepsilon_0}{\alpha\varepsilon} n^\star(\alpha_0, \varepsilon_0)\right).$$

Lemma F.6. *We have that*

$$n^\star(\alpha = 1/d, \varepsilon = 0.1) \ge \widetilde{\Omega}\left(d^{3/2}\right).$$

Before proving Lemmas F.5 and F.6, let us finish the proof of Lemma F.4. First, consider the case $\alpha = 1/d$. Lemma F.6 implies that

$$n^\star(\alpha = 1/d, \varepsilon) \ge \Omega\left(\frac{n^\star(\alpha = 1/d, \varepsilon = 0.1)}{\varepsilon}\right) \ge \widetilde{\Omega}\left(d^{3/2}/\varepsilon\right) = \widetilde{\Omega}\left(\frac{1}{\alpha^{3/2}\varepsilon}\right).$$

If $\alpha \le 1/d$, then similarly we have

$$n^\star(\alpha, \varepsilon) \ge \Omega\left(\frac{1}{d\alpha\varepsilon}\right) n^\star(\alpha = 1/d, \varepsilon = 0.1) \ge \widetilde{\Omega}\left(\frac{\sqrt{d}}{\alpha\varepsilon}\right).$$

Hence Lemma F.4 follows. Finally, we provide proofs for the remaining lemmas.

Lemma F.6. This lemma follows directly from Lemma F.3. Indeed, Lemma F.3 implies that if $d \ge \widetilde{\Omega}(n^{2/3})$ and $\varepsilon = 0.1$, the excess loss is lower bounded by $\mathcal{E}(\mathcal{A}, n) \ge \widetilde{\Omega}(1/n^{2/3})$. Stated differently, if $n \le \widetilde{O}(d^{3/2})$ then $\mathcal{E}(\mathcal{A}, n) \ge \widetilde{\Omega}(1/n^{2/3}) \ge \widetilde{\Omega}(1/d)$ which proves the claim. □

Lemma F.5. Given an (ε, δ)-DP algorithm $\mathcal{A}$ with $\mathcal{E}(\mathcal{A}, n) \le \alpha$, we show how to construct $\mathcal{A}'$ that is $(\varepsilon_0, 4\delta\varepsilon_0/\varepsilon)$-DP algorithm that works on datasets of size $n' = \Theta(\frac{\alpha\varepsilon}{\alpha_0\varepsilon_0} n)$ such that $\mathcal{E}(\mathcal{A}', n') \le \alpha_0$. This will prove the claim as we know that $n' \ge n(\alpha_0, \varepsilon_0)$. We now describe the construction of $\mathcal{A}'$. Given $\mathcal{S}' \in \mathcal{Z}^{n'}$ and $k > 0$ to be chosen presently, we define a new dataset $\mathcal{S}$ as follows: the first kn' samples are k copies of $\mathcal{S}'$ and the remaining $n - kn'$ are new samples $z \in \mathcal{Z}$ that have the loss function $f(x; z) = 0$ for all $x \in \mathcal{X}$. Clearly, these functions are convex, 0-Lipschitz, and 0-smooth. We then define $\mathcal{A}'(\mathcal{S}') = \mathcal{A}(\mathcal{S})$. Note that for all x we have that $\hat{F}(x; \mathcal{S}) = \frac{kn'}{n} \hat{F}(x; \mathcal{S}')$, which implies that

$$\begin{aligned}
\mathcal{E}(\mathcal{A}', \mathcal{S}') &= \mathbb{E}[\hat{F}(\mathcal{A}(\mathcal{S}); \mathcal{S}') - \min_{x \in \mathcal{X}} \hat{F}(x; \mathcal{S}')] \\
&= \frac{n}{kn'} \mathbb{E}[\hat{F}(\mathcal{A}(\mathcal{S}); \mathcal{S}) - \min_{x \in \mathcal{X}} \hat{F}(x; \mathcal{S})] \\
&= \frac{n}{kn'} \mathcal{E}(\mathcal{A}, \mathcal{S}) \le \frac{n\alpha}{kn'}.
\end{aligned}$$

Therefore if $n' \geq n\alpha/k\alpha_0$ we get $\mathcal{E}(\mathcal{A}', \mathcal{S}') \leq \alpha_0$. Hence it remains to argue for privacy. Using the group privacy property of private algorithms (Steinke & Ullman, 2017)(Fact 2.2), the algorithm $\mathcal{A}'$ is $(k\varepsilon, \frac{e^{k\varepsilon}-1}{e^{\varepsilon}-1}\delta)$-DP. Setting $k = \lfloor \log(1+\varepsilon_0)/\varepsilon \rfloor$ implies the claim as $e^{k\varepsilon} - 1 \leq \varepsilon_0$ and $k\varepsilon \leq \varepsilon_0$. □