Principled Approaches for Private Adaptation from a Public Source

Raef Bassily
The Ohio State University
& Google Research NY
bassily.1@osu.edu

Mehryar Mohri
Google Research & Courant Institute
of Mathematical Sciences, NY
mohri@google.com

Ananda Theertha Suresh
Google Research, NY
theertha@google.com

Abstract

A key problem in a variety of applications is that of domain adaptation from a public source domain, for which a relatively large amount of labeled data with no privacy constraints is at one's disposal, to a private target domain, for which a private sample is available with very few or no labeled data. In regression problems, where there are no privacy constraints on the source or target data, a *discrepancy minimization* approach was shown to outperform a number of other adaptation algorithm baselines. Building on that approach, we initiate a principled study of differentially private adaptation from a source domain with public labeled data to a target domain with unlabeled private data. We design differentially private discrepancy-based adaptation algorithms for this problem. The design and analysis of our private algorithms critically hinge upon several key properties we prove for a smooth approximation of the weighted discrepancy, such as its smoothness with respect to the ℓ_1-norm and the sensitivity of its gradient. We formally show that our adaptation algorithms benefit from strong generalization and privacy guarantees.

1 INTRODUCTION

In a variety of applications in practice, the amount of labeled data available from the domain of interest is too modest to train an accurate model. Instead, the learner must resort to using labeled samples from an alternative source domain, whose distribution is expected to be close to that of the target domain. Additionally, typically a large amount of *unlabeled data* from the target domain is also at one's disposal.

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

The problem of generalizing from that distinct source domain to a target domain for which few or no labeled data is available is a fundamental challenge in learning theory and algorithmic design known as the *domain adaptation problem*. We study a privacy-constrained and thus even more demanding scenario of domain adaptation, motivated by the critical data restrictions in modern applications: in practice, often the labeled data available from the source domain is public with no privacy constraints, but the unlabeled data from the target domain is subject to privacy constraints.

Differential privacy has become the gold standard of privacy-preserving data analysis as it offers formal and quantitative privacy guarantees and enjoys many attractive properties from an algorithmic design perspective [DR14]. Despite the remarkable progress in the field of differentially private machine learning, the problem of differentially private domain adaptation is still not well-understood. Here, we present new differentially private adaptation algorithms for the scenario described above and provide formal guarantees on their expected accuracy. Note that there has been a sequence of publications that provide formal differentially private learning guarantees assuming access to public data [CH11, BNS13, BTT18, ABM19, NB20, BCM+20]. However, their results are not applicable to the adaptation problem we study since they rely on the assumption that the source and target domains coincide. Recently, [JCYS21] proposed a differentially private correlation alignment approach for domain adaptation when both source and target data are private. [WLZ+20] proposed algorithms for deep domain adaptation for classification, however they do not provide theoretical guarantees on the proposed algorithms.

The design of our algorithms and their guarantees benefit from the theoretical analysis of domain adaptation by a series of prior publications. [MMR09] and [CM14] introduced the notion of *discrepancy*, which they used to give a general analysis of single-source adaptation for arbitrary loss functions. The notion of discrepancy is a divergence measure tailored to domain adaptation (see also [MM12, KM15, KM20]). Unlike other divergence measures between distributions such as the ℓ_1-distance, discrepancy takes into account the loss function and the hypothesis

set and, crucially, can be estimated from finite samples. The authors presented Rademacher complexity learning bounds in terms of the discrepancy for arbitrary hypothesis sets and loss functions, as well as pointwise learning bounds for kernel-based hypothesis sets. In the special case of the zero-one loss, the notion of discrepancy coincides with the d_A-*distance* between distributions introduced by [KBG04] and used in [BBCP06]. These authors used this notion to derive learning bounds for the zero-one loss, (see also the follow-up publications [BCK+08, BDBC+10]) in terms of a quantity denoted by $\lambda_{\mathcal{H}}$ that depends on the hypothesis set $\mathcal{H}$ and the distribution, but that cannot be estimated from observations.

For regression problems with no privacy constraints on the source or target data, [CM14] gave a *discrepancy minimization algorithm* based on a reweighting of the losses of the sample points. They further demonstrated that their algorithm outperformed all other baselines in a variety of tasks. Building on that approach, we design new differentially private discrepancy-based algorithms for adaptation from a source domain with public labeled data to a target domain with unlabeled private data. In Section 3, we briefly present some background material on the discrepancy analysis of adaptation motivating that approach.

The design and analysis of our private algorithms crucially hinge upon several key properties we prove for a smooth approximation of the weighted discrepancy, such as its smoothness with respect to the ℓ_1-norm and the sensitivity of its gradient. In Section 4, we describe that smooth approximation and give a detailed analysis of its crucial properties, which enable the construction of our private algorithms.

In Section 5, we present a new, computationally efficient, differentially private adaptation algorithm seeking to directly minimize the sum of the weighted empirical loss and the discrepancy. With respect to the work of [MMR09] and [CM14], our novel contribution is this one-stage algorithm (and a differentially private counterpart) directly seeking to miminize the learning bound, unlike that of previous work that consisted of a two-stage method, first seeking a sample reweighting that minimizes the empirical discrepancy, next fixing the weights thereby obtained and solving a weighted regression.

Since attaining the minimum in this case is generally intractable, due to the non-convexity of the objective, instead, our algorithm finds an approximate stationary point of this objective. Our algorithm is comprised of a sequence of Frank-Wolfe updates, where each update consists of a differentially private update of the weights and a non-private update of the predictor. In fact, our algorithm can be used in much more general settings of private non-convex optimization over a product of domains with different geometries. We formally prove the privacy and convergence guarantees of our algorithm in a general problem setting, and then derive its generalization guarantees in the context of adaptation. Our result in this section offers two main contributions to the growing body of work on convergence to stationary points in non-convex optimization [FLLZ18, MWCC18, CDHS17, NP06, GL13, ACD+19]. First, to the best of our knowledge, our work is the first to provide an algorithm with a strong convergence guarantee when the non-convex objective is defined over a product of domains with different geometries. Second, we achieve this under the constraint of differential privacy, which requires a different design and analysis paradigm than existing non-private non-convex optimization algorithms.

In Section 6, we present new two-stage private adaptation algorithms that can be viewed as private counterparts of the discrepancy minimization algorithm of [CM14]. As with that algorithm, the first stage aims at finding a reweighting of the source sample that minimizes the discrepancy, and the second stage aims at minimizing a regularized weighted empirical loss based on the reweighting found in the first stage. Since the second stage does not involve private data, only the first stage requires a private solution. Our solutions are based on private variants of Frank-Wolfe and Mirror-Descent algorithms, and they are computationally efficient. We describe these solutions in detail and give privacy and learning bounds for both algorithms. We further compare the benefits of these algorithms as a function of the sample sizes.

Finally, in Section 7, we conduct a set of proof-of-concept experiments showing that our algorithms yield good accuracy while attaining reasonable privacy guarantees.

We start with preliminary concepts and definitions relevant to our analysis.

2 PRELIMINARIES

Let $\mathcal{X} \subset \mathbb{R}^d$ denote the input space and $\mathcal{Y}$ the output space, which we assume to be a measurable subset of $\mathbb{R}$. We assume that $\mathcal{X}$ is included in the ℓ_2 ball of radius r, $\mathbb{B}_2^d(r)$. We will also assume that $\mathcal{Y}$ is included in a bounded interval of diameter $Y > 0$. Let $\mathcal{H}$ be a family of hypotheses mapping from $\mathcal{X}$ to $\mathcal{Y}$. We focus on the family of *linear hypotheses* $\mathcal{H} = \{x \mapsto w \cdot x : \|w\| \leq \Lambda\}$. We will be mainly interested in the regression setting, though some of our results can be extended to other contexts. For any $h \in \mathcal{H}$, we denote by $\ell(h(x), y) = (h(x) - y)^2$ the familiar squared loss of h for the labeled point $(x, y) \in \mathcal{X} \times \mathcal{Y}$. We denote by $M > 0$ an upper bound on the loss: $\ell(h(x), y) \leq M$, for all (x, y).

Learning scenario. We identify a domain with a distribution over $\mathcal{X} \times \mathcal{Y}$ and refer to the source domain as the one corresponding to a distribution $\mathcal{Q}$ and the target domain, the one corresponding to a distribution $\mathcal{P}$. We assume that the learner receives a sample $S = ((x_1, y_1), \ldots, (x_m, y_m))$ of m labeled points drawn i.i.d. from a distribution $\mathcal{Q}$

over $\mathcal{X} \times \mathcal{Y}$ and that it also has access to a large sample $T = (\widetilde{x}_1, \ldots, \widetilde{x}_n)$ of n unlabeled points drawn i.i.d. from $\mathcal{P}_\mathcal{X}$, the input marginal distribution associated to $\mathcal{P}$. We view the data from $\mathcal{Q}$, that is sample S, as *public data*, and the data from $\mathcal{P}$, sample T, as *private data*.

The objective of the learner is to use the samples S and T to select a hypothesis $h \in \mathcal{H}$ with small expected loss with respect to the target domain: $\mathcal{L}(\mathcal{P}, h) = \mathbb{E}_{(x,y)\sim\mathcal{P}}[\ell(h(x), y)]$. In the absence of any constraints, this coincides with the standard problem of single-source domain adaptation, studied in a very broad recent literature, starting with the theoretical studies of [BBCP06, MMR09, CM14].

Discrepancy notions. Clearly, the success of adaptation depends on the closeness of the distributions $\mathcal{P}$ and $\mathcal{Q}$, which can be measured according to various divergences. The notion of *discrepancy* has been shown to be appropriate measure of divergence between distributions in the context of domain adaptation. We will distinguish the so-called $\mathcal{Y}$-*discrepancy* $\mathrm{dis}_\mathcal{Y}(\mathcal{P}, \mathcal{Q})$, which can only be estimated when sufficient labeled data is available from both distributions, and the standard discrepancy $\mathrm{dis}(\mathcal{P}, \mathcal{Q})$, which can be estimated from finite unlabeled samples from both distributions:

$$\mathrm{Dis}_\mathcal{Y}(\mathcal{P}, \mathcal{Q}) = \max_{h\in\mathcal{H}}\Big\{\mathcal{L}(\mathcal{P}, h) - \mathcal{L}(\mathcal{Q}, h)\Big\}$$

$$\mathrm{Dis}(\mathcal{P}, \mathcal{Q}) = \max_{h,h'\in\mathcal{H}}\Big\{\mathop{\mathbb{E}}_{x\sim\mathcal{P}_\mathcal{X}}[\ell(h(x), h'(x))] - \mathop{\mathbb{E}}_{x\sim\mathcal{Q}_\mathcal{X}}[\ell(h(x), h'(x))]\Big\}.$$

We will be using the two-sided versions of these expressions, for example $\mathrm{dis}(\mathcal{P}, \mathcal{Q}) = \max\{\mathrm{Dis}(\mathcal{P}, \mathcal{Q}), \mathrm{Dis}(\mathcal{Q}, \mathcal{P})\}$, though part of our analysis holds with one-sided definitions too.

Matrix definitions. We will adopt the following matrix definitions and notation. We denote by $\mathbb{M}_d$ the set of real-valued $d \times d$ matrices and by $\mathbb{S}_d$ the subset of $\mathbb{M}_d$ formed by symmetric matrices. We will denote by $\langle \cdot, \cdot \rangle$ the Frobenius product defined for all $\mathbf{M}, \mathbf{M}' \in \mathbb{M}_d$ by $\langle \mathbf{M}, \mathbf{M}' \rangle = \sum_{i,j=1}^d \mathbf{M}_{ij}\mathbf{M}'_{ij} = \mathrm{Tr}(\mathbf{M}^\top \mathbf{M}')$. For any matrix $\mathbf{M} \in \mathbb{S}_d$, we denote by $\lambda_i(\mathbf{M})$ the ith eigenvalue of $\mathbf{M}$ in decreasing order and will also denote by $\lambda_{\max}(\mathbf{M}) = \lambda_1(\mathbf{M})$ its largest eigenvalue, and by $\lambda_{\min}(\mathbf{M}) = \lambda_d(\mathbf{M})$ its smallest eigenvalue. We also denote by $\lambda(\mathbf{M})$ the vector of eigenvalues of $\mathbf{M}$. For any $p \in [1, +\infty]$, we will denote by $\|\mathbf{M}\|_{(p)}$ the p-Schatten norm of $\mathbf{M}$ defined by $\|\mathbf{M}\|_{(p)} = \|\lambda(\mathbf{M})\|_p = \left[\sum_{i=1}^d |\lambda_i(\mathbf{M})|^p\right]^{\frac{1}{p}}$. Note that $p = +\infty$ corresponds to the spectral norm: $\|\mathbf{M}\|_{(\infty)} = \|\lambda(\mathbf{M})\|_\infty$, which we also denote by $\|\mathbf{M}\|_2$.

Smoothness. We will say that a continuously differentiable function f defined over a vector space $\mathcal{E}$ is γ-smooth for norm $\|\cdot\|$ if $\forall x, x' \in \mathcal{E}, \|\nabla f(x) - \nabla f(x')\|_* \leq \gamma\|x - x'\|$, where $\|\cdot\|_*$ is the dual norm associated to $\|\cdot\|$. When f is twice differentiable, it is known that the condition on the Hessian $\forall x, z \in \mathcal{E}, \left|z^\top \nabla^2 f(x) z\right| \leq \gamma\|z\|^2$, implies that f is $\|\cdot\|$-γ-smooth [Sid19][Chapter 5; lemma 8].

Differential Privacy [DMNS06, DKM+06]. Fix $\varepsilon, \delta > 0$. A (randomized) algorithm $\mathcal{A}\colon \mathcal{Z}^n \to \mathcal{R}$ is said to be (ε, δ)-differentially private if for all pairs of datasets $T, T' \in \mathcal{Z}$ that differ in exactly one entry, and every measurable $\mathcal{O} \subseteq \mathcal{R}$, we have: $\mathbb{P}(\mathcal{A}(T) \in \mathcal{O}) \leq e^\varepsilon \cdot \mathbb{P}(\mathcal{A}(T') \in \mathcal{O}) + \delta$. We consider differentially private algorithms that have access to an auxiliary public sample S in addition to their input private dataset T. In that case, we view the public sample S as being "hardwired" to the algorithm, and the constraint of differential privacy is imposed only with respect to the private dataset.

3 DISCREPANCY-BASED GENERALIZATION BOUNDS

In this section, we briefly present some background material on discrepancy-based generalization guarantees. A more detailed discussion is presented in Appendix A. Let the *output label-discrepancy* $\eta_\mathcal{H}(S, \widetilde{T})$ be defined as follows:

$$\eta_\mathcal{H}(S, \widetilde{T}) = \min_{h_0\in\mathcal{H}}\left\{\sup_{(x,y)\in S}|y - h_0(x)| + \sup_{(x,y)\in\widetilde{T}}|y - h_0(x)|\right\},$$

where $\widetilde{T}$ is the labeled version of T (i.e., $\widetilde{T}$ is T associated with its true, hidden labels). Note that $\mathrm{dis}(\mathcal{P}_\mathcal{X}, \mathsf{q})$ measures the difference of the distributions on the input domain. In contrast, $\eta_\mathcal{H}(S, \widetilde{T})$ accounts for the difference of the output labels in S and T. Note that under the covariate-shift and separability assumption, we have $\eta_\mathcal{H}(S, \widetilde{T}) = 0$. In general, adaptation is not possible when $\eta_\mathcal{H}(S, \widetilde{T})$ is large since the labels received on the training sample would then be very different from the target ones. Thus, we will assume, as in previous work, that we have $\eta_\mathcal{H}(S, \widetilde{T}) \ll 1$. Then, the following learning bound, expressed in terms of the empirical unlabeled discrepancy $\mathrm{dis}\big(\widehat{\mathcal{P}}_\mathcal{X}, \mathsf{q}\big)$, $\eta_\mathcal{H}(S, \widetilde{T})$, and the Rademacher complexity of $\mathcal{H}$, holds with probability at least $1 - \beta$ for all $h \in \mathcal{H}$ and all distributions q over S [CM14, CMMM19]:

$$\mathcal{L}(\mathcal{P}, h) \leq \sum_{i=1}^m \mathsf{q}_i \ell(h(x_i), y_i) + \mathrm{dis}\big(\widehat{\mathcal{P}}_\mathcal{X}, \mathsf{q}\big) + \eta_\mathcal{H}(S, \widetilde{T}) + 2M\mathfrak{R}_n(\mathcal{H}) + M\sqrt{\frac{\log\frac{1}{\beta}}{2n}}. \quad (1)$$

When $\mathcal{H}$ is the class of linear predictors and the support of $\mathcal{P}_\mathcal{X}$ is included in the ℓ_2-ball of radius r, the Rademacher complexity can be explicitly upper-bounded as follows: $\mathfrak{R}_n(\mathcal{H}) \leq \sqrt{\frac{r^2\Lambda^2}{n}}$ [MRT18]. [CM14] proposed an adaptation algorithm motivated by these learning bounds and other pointwise guarantees expressed in terms of discrepancy. Their algorithm can be viewed as a two-stage method seeking to minimize the first two terms of this learning

bound. It consists of first finding a minimizer q of the weighted discrepancy (second term) and then minimizing a regularized q-weighted empirical loss (first term) with respect to h for that value of q. In Section 6, we design private adaptation algorithms for a similar two-stage approach. We first give a new, direct approach for private adaptation based on a new differentially private non-convex optimization algorithm, which we discuss in Section 5. This algorithm can be viewed as a direct, single-stage approach for private domain adaptation that seeks to find h and q directly minimizing the first two terms of the bound.

The privacy and accuracy guarantees of our algorithms crucially rely on a careful analysis of a smooth approximation of the discrepancy term, which we present in the following section.

4 DISCREPANCY ANALYSIS AND SMOOTH APPROXIMATION

4.1 Analysis

For the squared loss and $\mathcal{H} = \{x \mapsto w \cdot x : \|w\| \leq \Lambda\}$, the weighted discrepancy term of the learning bound (1) can be expressed in terms of the spectral norm of a matrix that is an affine function of q.

Lemma 1 ([MMR09, CM14]). *For any distribution q over $S_{\mathcal{X}}$, the following equality holds:*

$$\begin{aligned}\mathrm{dis}(\widehat{P}, \mathsf{q}) &= 4\Lambda^2 \|\mathbf{M}(\mathsf{q})\|_2 \\ &= 4\Lambda^2 \max\{\lambda_{\max}(\mathbf{M}(\mathsf{q})), \lambda_{\max}(-\mathbf{M}(\mathsf{q}))\},\end{aligned}$$

where $\mathbf{M}(\mathsf{q}) = \mathbf{M}_0 - \sum_{i=1}^m q_i \mathbf{M}_i$ *and where* $\mathbf{M}_0 = \sum_{x \in \mathcal{X}} \widehat{\mathcal{P}}_{\mathcal{X}}(x) x x^\top$, *and* $\mathbf{M}_i = x_i x_i^\top$, $i \in [m]$.

For completeness, a short proof of this result is given in Appendix B. In view of that, the learning bound (1) suggests seeking $h \in \mathcal{H}$ and $\mathsf{q} \in \Delta_m$ to minimize the first two terms:

$$\min_{\substack{h \in \mathcal{H} \\ \mathsf{q} \in \Delta_m}} \sum_{i=1}^m \mathsf{q}_i \ell(h(x_i), y_i) + 4\Lambda^2 \|\mathbf{M}(\mathsf{q})\|_2. \tag{2}$$

Note that the second term of the bound is sub-differentiable but it is not differentiable both because of the underlying maximum operator and because the maximum eigenvalue is not differentiable at points where its multiplicity is more than one. Furthermore, the first term of the objective function is convex with respect to h and convex with respect to q, but it is not jointly convex in both.

Our private algorithms require bounded sensitivity of the gradients as well as smoothness of the objective, which would not hold given the first issue mentioned. Thus, instead, we will define a uniform α-smooth approximation of the second term, for which we analyze the smoothness and gradient sensitivity in detail.

4.2 Softmax smooth approximation

A natural approximation of $\lambda_{\max}(\mathsf{q})$ is based on the softmax approximation: $F(\mathsf{q}) = \frac{1}{\mu} \log\big[\sum_{i=1}^d e^{\mu \lambda_i(\mathbf{M}(\mathsf{q}))}\big]$. Eigenvalues are not differentiable everywhere. However, F is infinitely differentiable since it can be expressed as a composition of the $\log$, the trace and the matrix exponential, which are all infinitely differentiable and since $\mathbf{M}(\mathsf{q})$ is an affine function of q: $F(\mathsf{q}) = \frac{1}{\mu} \log\Big[\mathrm{Tr}\big[\exp(\mu \mathbf{M}(\mathsf{q}))\big]\Big]$. The matrix exponential can be computed in $O(d^3)$, using an SVD of matrix $\mathbf{M}(\mathsf{q})$. The following inequalities follow directly the properties of the softmax:

$$\lambda_{\max}(\mathbf{M}(\mathsf{q})) \leq F(\mathsf{q}) \leq \lambda_{\max}(\mathbf{M}(\mathsf{q})) + \frac{\log(\mathrm{rank}(\mathbf{M}(\mathsf{q})))}{\mu}. \tag{3}$$

Note that we have $\mathrm{rank}(\mathbf{M}(\mathsf{q})) \leq \min(m+n, d)$. Thus, for $\mu = \frac{\log(m+n)}{\alpha}$, $F(\mathsf{q})$ gives a uniform α-approximation of $\lambda_{\max}(\mathbf{M}(\mathsf{q}))$. The components of the gradient of F are given by

$$[\nabla F(\mathsf{q})]_j = -\frac{\langle \exp(\mu \mathbf{M}(\mathsf{q})), \mathbf{M}_j \rangle}{\mathrm{Tr}(\exp(\mu \mathbf{M}(\mathsf{q})))}, \qquad j \in [m] \tag{4}$$

where $\langle \cdot, \cdot \rangle$ denotes the Frobenius inner product. Both the smoothness and sensitivity of ∇F will be needed for the derivation of our algorithm. We now analyze these properties of function F, using function f which is defined for any symmetric matrix $\mathbf{M} \in \mathbb{S}_d$ as follows:

$$f(\mathbf{M}) = \frac{1}{\mu} \log\left[\sum_{k=0}^{+\infty} \frac{\mu^k}{k!} \langle \mathbf{M}^k, \mathbf{I} \rangle\right].$$

We have $F(\mathsf{q}) = f(\mathbf{M}(\mathsf{q}))$. The following result provides the desired smoothness result needed for F, which we prove by using the μ-smoothness of f.

Theorem 1. *The softmax approximation F is $\mu\big(\max_{i \in [m]} \|x_i\|_2^4\big)$-smooth for $\|\cdot\|_1$.*

The proof is given in Appendix C.1. Next, we analyze the *sensitivity of* ∇F, that is the maximum variation in ℓ_∞-norm of $\nabla F(\mathsf{q})$ when a single point x in the sample of size n drawn from $\widehat{\mathcal{P}}_{\mathcal{X}}$ is changed to another one x'.

Theorem 2. *The gradient of the softmax approximation F is $\frac{2\mu r^2}{n} \max_{i \in [m]} \|x_i\|_2^2$-sensitive.*

Proof. For $\mathbf{M}(\mathsf{q}))$ and $\mathbf{M}'(\mathsf{q}))$ differing only by point x and x' in $\widehat{\mathcal{P}}_{\mathcal{X}}$, we have:

$$\|\mathbf{M}(\mathsf{q}) - \mathbf{M}'(\mathsf{q})\|_2 = \left\| \frac{1}{n} \big[x x^\top - x' x'^\top\big] \right\|_2 \leq \frac{2r^2}{n}.$$

Note that we have $F(\mathsf{q}) = f(\mathbf{M}(\mathsf{q}))$. Thus, the gradient of F can be expressed as follows:

$$\nabla F(\mathsf{q}) = -[\langle \nabla f(\mathbf{M}(\mathsf{q})), \mathbf{M}_i \rangle]_{i \in [m]}.$$

Thus, the sensitivity of the gradient of F can be bounded as follows:

$$\begin{aligned}
&\max_{i\in[m]} |\langle \nabla f(\mathbf{M}(\mathsf{q})) - \nabla f(\mathbf{M}'(\mathsf{q})), \mathbf{M}_i\rangle| \\
&\leq \max_{i\in[m]} \|\nabla f(\mathbf{M}(\mathsf{q})) - \nabla f(\mathbf{M}'(\mathsf{q}))\|_{(1)} \|\mathbf{M}_i\|_{(\infty)} \\
&\leq \mu \max_{i\in[m]} \|\mathbf{M}(\mathsf{q}) - \mathbf{M}'(\mathsf{q})\|_{(\infty)} \|\mathbf{M}_i\|_{(\infty)} \\
&= \mu \max_{i\in[m]} \|\mathbf{M}(\mathsf{q}) - \mathbf{M}'(\mathsf{q})\|_2 \|\mathbf{M}_i\|_2 \\
&\leq \frac{2\mu r^2}{n} \max_{i\in[m]} \|x_i\|_2^2,
\end{aligned}$$

where the first inequality holds by Hölder's inequality, the second by the μ-smoothness of f, and the third by the definition of $\|\cdot\|_{(\infty)}$. This completes the proof. □

Note that the softmax function f is known to be convex [BV14]. Since $\mathbf{M}(\mathsf{q})$ is an affine function of q and that composition with affine functions preserves convexity, this shows that F is also a convex function. The following further shows that F is $\max_{i\in[m]}\|x_i\|_2^2$-Lipschitz.

Theorem 3. *For any $\mathsf{q} \in \Delta_m$, the gradient of F is bounded as follows: $\|\nabla F(\mathsf{q})\|_\infty \leq \max_{i\in[m]}\|x_i\|_2^2$.*

The proof is given in Appendix C.1. In view of the expression of the weighted discrepancy $\mathrm{dis}(\widehat{\mathcal{P}}, \mathsf{q}) = \max\{\lambda_{\max}(\mathbf{M}(\mathsf{q})), \lambda_{\max}(-\mathbf{M}(\mathsf{q}))\}$, the smooth approximation $F(\mathsf{q})$ of the maximum eigenvalue of $\mathbf{M}(\mathsf{q})$ leads immediately to a smooth approximation $\widetilde{F}(\mathsf{q}) = f(\widetilde{\mathbf{M}}(\mathsf{q}))$ of $\mathrm{dis}(\widehat{\mathcal{P}}, \mathsf{q})$, with

$$\widetilde{\mathbf{M}}(\mathsf{q}) = \begin{bmatrix} \mathbf{M}(\mathsf{q}) & \mathbf{0} \\ \mathbf{0} & -\mathbf{M}(\mathsf{q}) \end{bmatrix}.$$

Thus, $\widetilde{F}$ inherits the key properties of F gathered in the following corollary.

Corollary 1. *The following properties holds for $\widetilde{F}$:*

1. *$\widetilde{F}$ is convex and is a uniform $\frac{\log(2\min\{m+n,d\})}{\mu}$-approximation of $\mathsf{q} \mapsto \mathrm{dis}(\widehat{P}, \mathsf{q})$.*
2. *$\widetilde{F}$ is $\mu\big(\max_{i\in[m]}\|x_i\|_2^4\big)$-smooth for $\|\cdot\|_1$.*
3. *$\|\nabla\widetilde{F}\|_\infty$ is $\frac{2\mu r^2}{n}\max_{i\in[m]}\|x_i\|_2^2$-sensitive.*
4. *for any $\mathsf{q} \in \Delta_m$, $\|\nabla\widetilde{F}(\mathsf{q})\|_\infty \leq \max_{i\in[m]}\|x_i\|_2^2$.*

The proof is given in Appendix C.2. In Appendix C.3, we also present and analyze a p-norm smooth approximation of the discrepancy. This approximation can be used to design private adaptation algorithms with a relative deviation guarantee that can be more favorable in some contexts.

5 PRIVATE ADAPTATION ALGORITHM VIA NON-CONVEX OPTIMIZATION

Here, we describe a novel private domain adaptation algorithm for regression problems. Our algorithm is based on a single-stage approach that consists of directly optimizing an objective function based on the learning bound (1).

Let $\widetilde{F}_T(\mathsf{q})$, $\mathsf{q} \in \Delta_m$, denote the smooth approximation discussed in Section 4.2, where the additional subscript T is used to emphasize the dependence on the private dataset T. Then, by Lemma 1 and the results of Section 4.2, the following objective function is a smooth approximation of the first two terms of the learning bound (1):

$$L_T(\mathsf{q}, w) \triangleq \sum_{i=1}^m \mathsf{q}_i(\langle w, x_i\rangle - y_i)^2 + 4\Lambda^2\widetilde{F}_T(\mathsf{q}).$$

This is a non-convex function of (q, w) and no tractable method is known for finding a global minimizer. Instead, our algorithm returns an approximate stationary point, which is the most reasonable alternative.

Note that, as shown in Section 4.2, $L_T(\mathsf{q}, w)$ is smooth in q with respect to $\|\cdot\|_1$. By definition of the squared loss, it is also smooth in w with respect to $\|\cdot\|_2$. These smoothness properties enable us to design our private solution. Given the approximation guarantee (3), the data-dependent terms in the learning bound (1) can thus be approximated by $L_T(\mathsf{q}, w)$. Hence, our strategy here is to find an approximate stationary point $(\widehat{\mathsf{q}}, \widehat{w})$ of L_T via our private algorithm, and then derive a learning bound in terms of $L_T(\widehat{\mathsf{q}}, \widehat{w})$. The following gives a formal definition of an approximate stationary point.

Definition 1 (α-approximate stationary point)**.** *Let $f\colon \mathcal{C} \to \mathbb{R}$ be a differentiable function over a convex and compact subset $\mathcal{C}$ of a normed vector space. Let $\alpha \geq 0$. We say that $u \in \mathcal{C}$ is an α-approximate stationary point of f if the stationarity gap of f at u, defined as $\mathrm{Gap}_f(u) \triangleq \max_{v\in\mathcal{C}}\langle -\nabla f(u), v - u\rangle$ is at most α.*

We first give a generic differentially-private algorithm for approximating a stationary point of a smooth non-convex objective $f_T\colon \mathcal{Q}\times\mathcal{W} \to \mathbb{R}$, that is defined by a private dataset T and satisfies certain smoothness and Lipschitz-continuity conditions. Let $p_1, p_2 \geq 1$. We assume that $\mathcal{Q}$ is a convex set whose $\|\cdot\|_{p_1}$-diameter is bounded by D_{q}. We refer to $\mathcal{Q}$ as a $(D_{\mathsf{q}}, \|\cdot\|_{p_1})$-bounded set in that case. Similarly, we will assume that $\mathcal{W}$ is a convex $(D_w, \|\cdot\|_{p_2})$-bounded set. We now give formal definitions of the smoothness and Lipschitz-continuity conditions we will be assuming.

Definition 2 ($((\gamma_{\mathsf{q}}, \|\cdot\|_{p_1}), (\gamma_w, \|\cdot\|_{p_2}))$-Lipschitz function)**.** *Let $\gamma_{\mathsf{q}}, \gamma_w \geq 0$. We say that $f\colon \mathcal{Q}\times\mathcal{W} \to \mathbb{R}$ is $((\gamma_{\mathsf{q}}, \|\cdot\|_{p_1}), (\gamma_w, \|\cdot\|_{p_2}))$-Lipschitz if for any $w \in \mathcal{W}$, $f(\cdot, w)$ is γ_{q}-Lipschitz with respect to $\|\cdot\|_{p_1}$ over $\mathcal{Q}$, and for every $\mathsf{q} \in \mathcal{Q}$, $f(\mathsf{q},)$ is γ_w-Lipschitz with respect to $\|\cdot\|_{p_2}$ over $\mathcal{W}$.*

Definition 3 ($((\mu_{\mathsf{q}}, \|\cdot\|_{p_1}), (\mu_w, \|\cdot\|_{p_2}))$-smooth function)**.** *This notion is defined analogously. We say that f is $((\mu_{\mathsf{q}}, \|\cdot\|_{p_1}), (\mu_w, \|\cdot\|_{p_2}))$-smooth if for any $w \in \mathcal{W}$, $f(\cdot, w)$ is μ_{q}-smooth with respect to $\|\cdot\|_{p_1}$ over $\mathcal{Q}$, and for every $\mathsf{q} \in \mathcal{Q}$, $f(\mathsf{q},)$ is μ_w-smooth with respect to $\|\cdot\|_{p_2}$ over $\mathcal{W}$.*

Our private algorithm (Algorithm 1) takes as input an objective $f_T: \mathcal{Q} \times \mathcal{W} \to \mathbb{R}$, where $\mathcal{Q}$ is a convex polyhedral set with bounded $\|\cdot\|_1$-diameter and $\mathcal{W}$ is a convex set with bounded $\|\cdot\|_2$-diameter. Thus, this covers our objective function L_T as a special case.

The algorithm is comprised of a number rounds, where in each round, two private Frank-Wolfe update steps are performed; one for q and another for w. The privacy mechanism for each is different due to the different geometries of $\mathcal{Q}$ and $\mathcal{W}$. We note that in the special case where $f_T = L_T$, there is no need to privatize the Frank-Wolfe step for w due to the fact that such update step depends only on the q-weighted empirical loss over the public data and the fact that differential privacy is closed under post-processing (the previous update step for q is carried out in a differentially private manner).

When f_T satisfies the Lipschitzness and smoothness properties defined above with respect to $\|\cdot\|_1$ and $\|\cdot\|_2$, we give formal convergence guarantees to a stationary point in terms of a high-probability bound on the *stationarity gap* of the output (see Definition 1). Despite the different geometries of $\mathcal{Q}$ and $\mathcal{W}$, our final bound is roughly the sum of the bounds we would obtain if we ran two separate Frank-Wolfe algorithms (one over $\mathcal{Q}$ and the other over $\mathcal{W}$). This is mainly due to the hybrid Lipschitzness and smoothness conditions (with respect to $\|\cdot\|_1$ for q and with respect to $\|\cdot\|_2$ for w), which enable us to decompose the bound on the convergence rate effectively into two terms: one for q and one for w.

Novelty of our algorithm: Convergence to stationary points in non-private, non-convex optimization has received significant attention recently, see, e.g., [FLLZ18, MWCC18, CDHS17, NP06, GL13, ACD+19]. Our result in this section offers two major contributions to this body of work. First, to the best of our knowledge, our work is the first to provide a strong convergence guarantee in scenarios where the non-convex objective is defined over a product of domains with different ℓ_p geometries. Second, we do this while guaranteeing *differentially privacy*.

Theorem 4. *Algorithm 1 is (ε, δ)-differentially private. Suppose $f_T: \mathcal{Q} \times \mathcal{W} \to \mathbb{R}$ is $((\gamma_{\mathsf{q}}, \|\cdot\|_1), (\gamma_w, \|\cdot\|_2))$-Lipschitz and $((\mu_{\mathsf{q}}, \|\cdot\|_1), (\mu_w, \|\cdot\|_2))$-smooth. Assume further that for all $\mathsf{q} \in \mathcal{Q}$, and $w, w' \in \mathcal{W}$, $\|\nabla_{\mathsf{q}} f_T(\mathsf{q}, w) - \nabla_{\mathsf{q}} f_T(\mathsf{q}, w')\|_\infty \le \gamma_{\mathsf{q},w} \|w - w'\|_2$. Then, for any $\beta \in (0,1)$, there exists a choice of K and μ such that, with probability at least $1-\beta$, the stationarity gap of the output $\widehat{w}$, $\mathrm{Gap}_{f_T}(\widehat{\mathsf{q}}, \widehat{w})$, is upper bounded by*

$$5\sqrt{\bar{D}\left(\sigma_{\mathsf{q}}^0 \log\left(\frac{\bar{D}J}{\sigma_{\mathsf{q}}^0 \beta}\right) + D_w \sigma_w^0 \sqrt{d \log\left(\frac{\bar{D}}{D_w \sigma_w^0 \beta}\right)}\right)},$$

where

$$\bar{D} = \sqrt{(D_{\mathsf{q}}\gamma_{\mathsf{q}} + D_w \gamma_w)(D_{\mathsf{q}}^2 \mu_{\mathsf{q}} + D_w^2 \mu_w + 2\gamma_{\mathsf{q},w} D_{\mathsf{q}} D_w)},$$

Algorithm 1 Private Frank-Wolfe for approximating stationary points of $f_T : \mathcal{Q} \times \mathcal{W} \to \mathbb{R}$

Require: Private dataset: $T = (z_1, \ldots, z_n) \in \mathcal{Z}^n$, privacy parameters (ε, δ), a convex $(D_{\mathsf{q}}, \|\cdot\|_1)$-bounded polyhedral set: $\mathcal{Q} \subset \mathbb{R}^m$ with J vertices $\mathcal{V} = (v_1, \ldots, v_J)$, a convex $(D_w, \|\cdot\|_2)$-bounded set $\mathcal{W} \subset \mathbb{R}^d$, a function $f_T(\mathsf{q}, w)$, $\mathsf{q} \in \mathcal{Q}, w \in \mathcal{W}$ (defined via the dataset T), bound on the global $\|\cdot\|_\infty$-sensitivity of $\nabla_{\mathsf{q}} f_T(\mathsf{q}, w)$: $\tau_{\mathsf{q}} > 0$, bound on the global $\|\cdot\|_2$-sensitivity of $\nabla_w f_T(\mathsf{q}, w)$: $\tau_w \ge 0$, step size: η, number of iterations: K.

1: Set $\sigma_{\mathsf{q}} := \frac{4\tau_{\mathsf{q}}\sqrt{2K\log(\frac{1}{\delta})}}{\varepsilon}$.
2: Set $\sigma_w := \frac{4\tau_w\sqrt{2K\log(\frac{1}{\delta})}}{\varepsilon}$.
3: Choose arbitrarily $(\mathsf{q}^0, w^0) \in \mathcal{Q} \times \mathcal{W}$.
4: **for** $k = 0$ to $K-1$ **do**
5: $\quad \nabla_{\mathsf{q}}^k := \nabla_{\mathsf{q}} f_T(\mathsf{q}^k, w^k)$.
6: $\quad$ Draw $\left(b_v^k : v \in \mathcal{V}\right)$ independently $\sim \mathsf{Lap}(\sigma_{\mathsf{q}})$.
7: $\quad v_{\mathsf{q}}^k := \operatorname{argmin}_{v \in \mathcal{V}} \left\{\langle \nabla_{\mathsf{q}}^k, v\rangle + b_v^k\right\}$.
8: $\quad G_{\mathsf{q}}^k := -\left(\langle \nabla_{\mathsf{q}}^k, v_{\mathsf{q}}^k - \mathsf{q}^k\rangle + b_v^k\right)$.
9: $\quad \mathsf{q}^{k+1} := (1-\eta)\mathsf{q}^k + \eta v_{\mathsf{q}}^k$.
10: $\quad \nabla_w^k := \nabla_w f_T(\mathsf{q}^k, w^k)$.
11: $\quad \widehat{\nabla}_w^k := \nabla_w^k + \mathbf{g}^k$, where $\mathbf{g}^k \sim \mathcal{N}(\mathbf{0}, \sigma_w^2 \mathbb{I}_d)$.
12: $\quad u_w^k := \operatorname{argmin}_{u \in \mathcal{W}} \langle \widehat{\nabla}_w^k, u\rangle$.
13: $\quad G_w^k := -\langle \widehat{\nabla}_w^k, u_w^k - w^k\rangle$.
14: $\quad w^{k+1} := (1-\eta)w^k + \eta u_w^k$.
15: **end for**
16: **return** $(\widehat{\mathsf{q}}, \widehat{w}) = (\mathsf{q}^{k^*}, w^{k^*})$, where $k^* = \operatorname{argmin}_{k \in [K]} (G_{\mathsf{q}}^k + G_w^k)$.

$\sigma_{\mathsf{q}}^0 = \frac{\sigma_{\mathsf{q}}}{\sqrt{K}}$, and $\sigma_w^0 = \frac{\sigma_w}{\sqrt{K}}$ (where $\sigma_{\mathsf{q}}, \sigma_w$ are as given in steps 1 and 2 of Algorithm 1).

The proof of Theorem 4 is given in Appendix D.1. We note that our adaptation objective $L_T(\mathsf{q}, w)$ satisfies all the conditions in Theorem 4.

Instantiating Algorithm 1 with $L_T(\mathsf{q}, w)$: Our objective function $L_T(\mathsf{q}, w) = \sum_{i=1}^m \mathsf{q}_i \ell(h_w(x_i), y_i) + 4\Lambda^2 \widetilde{F}_T(\mathsf{q})$ (where $\ell(h_w(x_i), y_i) = (\langle w, x_i\rangle - y_i)^2$ is the squared loss) is an instance of f_T in Theorem 4. Recall that we assume $\mathcal{W} \subseteq \mathbb{B}_2^d(\Lambda)$, $\mathcal{X} \subset \mathbb{B}_2^d(r)$, and $\mathcal{Y} \subseteq [-Y, +Y]$ for some $Y > 0$. Also, recall that we denote the maximum norm of the feature vectors in the public dataset, $\max_{i \in [m]} \|x_i\|_2$, by $\hat{r}$.

First, note that $\mathcal{Q}$ in Algorithm 1 is instantiated with the simplex Δ_m, thus $\mathcal{V}$ is $\{\mathbf{e}_1, \ldots, \mathbf{e}_m\}$. Second, since the private dataset T only appears in $\widetilde{F}_T$, note that $\nabla_w L_T(\mathsf{q}, w)$ does not involve T. Thus, σ_w in Algorithm 1 can be set to zero. That is, we do not need to privatize the Frank-Wolfe steps over w. Third, note that the global $\|\cdot\|_\infty$-sensitivity of $\nabla_{\mathsf{q}} L_T(\mathsf{q}, w)$, τ_{q}, is the same as that of

$4\Lambda^2\nabla_{\mathsf{q}}\widetilde{F}$, which follows from Corollary 1 (Part 3), namely, $\tau_{\mathsf{q}} = \frac{8\Lambda^2\mu r^2\hat{r}^2}{n}$, where μ is the approximation parameter of the soft-max and $\hat{r} = \max_{i\in[m]}\|x_i\|_2$. Fourth, note that $L_T(\cdot,\cdot)$ is $\Big((\gamma_{\mathsf{q}},\|\cdot\|_1),(\gamma_w,\|\cdot\|_2)\Big)$-Lipschitz, where $\gamma_{\mathsf{q}} = (\Lambda\hat{r}+Y)^2+4\Lambda^2\hat{r}^2$, which follows from $\|\nabla_{\mathsf{q}}L_T(\mathsf{q},w)\|_\infty \le \max_{i\in[m]}\ell(h_w(x_i),y_i)+4\Lambda^2\|\nabla_{\mathsf{q}}\widetilde{F}(\mathsf{q})\|_\infty$ together with Corollary 1 (Part 4), and $\gamma_w = 2(\Lambda\hat{r}+Y)\hat{r}$, which follows directly from the $\|\cdot\|_2$-bound on the gradient of the squared loss over $\mathbb{B}_2^d(\Lambda)$. Moreover, $L_T(\cdot,\cdot)$ is $\Big((\mu_{\mathsf{q}},\|\cdot\|_1),(\mu_w,\|\cdot\|_2)\Big)$-smooth, where $\mu_{\mathsf{q}} = 4\Lambda^2\mu\hat{r}^4$, which follows from the fact that the smoothness of L_T with respect to q is given by the smoothness of $4\Lambda^2\widetilde{F}_T$, which follows from Corollary 1 (Part 2), and $\mu_w = \hat{r}^2$, which follows from the fact that the squared loss $\ell(h_w(x),y)$ is $\|x\|_2^2$-smooth with respect to $\|\cdot\|_2$. Additionally, the condition on $\|\nabla_{\mathsf{q}}L_T(\mathsf{q},w)-\nabla_{\mathsf{q}}L_T(\mathsf{q},w')\|_\infty$ in Theorem 4 is satisfied in our case with $\gamma_{\mathsf{q},w} = 2\hat{r}(\Lambda\hat{r}+Y)$, which follows from the fact that $\|\nabla_{\mathsf{q}}L_T(\mathsf{q},w)-\nabla_{\mathsf{q}}L_T(\mathsf{q},w')\|_\infty = \max_{i\in[m]}|\ell(h_w(x_i),y_i)-\ell(h_{w'}(x_i),y_i)|$ together with the Lipschitzness property of the squared loss over $\mathbb{B}_2^d(\Lambda)$. Finally, note that $D_{\mathsf{q}} = 2$ and $D_w = 2\Lambda$.

As a result, we immediately obtain the following corollary.

Corollary 2. *Let $L_T(\mathsf{q},w) = \sum_{i=1}^m \mathsf{q}_i(\langle w,x_i\rangle - y_i)^2 + 4\Lambda^2\widetilde{F}_T(\mathsf{q})$ be the input objective to Algorithm 1. Let $\beta\in(0,1)$. Then, there exists a choice of K and η such that, with probability at least $1-\beta$, the output of the algorithm is an approximate stationary point of L_T with stationarity gap upper bounded as follows:*

$$\mathrm{Gap}_{L_T}(\widehat{\mathsf{q}},\widehat{w}) \le \widetilde{O}\left(\frac{\mu^{3/4}}{\sqrt{\varepsilon n}}\right).$$

Here, $\widetilde{O}(\cdot)$ hides poly-logarithmic factors in m.

In view of that, the learning bound (1) implies that with probability $\ge 1-2\beta$ over the choice of the public and private datasets and the algorithm's internal randomness, the expected loss of the predictor $h_{\widehat{w}}$ (defined by the output $\widehat{w}$) with respect to the target domain is bounded as

$$\mathcal{L}(\mathcal{P},h_{\widehat{w}}) \le L_T(\widehat{\mathsf{q}},\widehat{w}) + \widetilde{O}\left(\frac{1}{\mu}+\frac{1}{\sqrt{n}}\right) + \eta_{\mathcal{H}}(S,\widetilde{T}).$$

Note that $(\widehat{\mathsf{q}},\widehat{w})$ is an approximate stationary point of L_T. In practice, $(\widehat{\mathsf{q}},\widehat{w})$ can be an approximation of a good local minimum of L_T as demonstrated by our experiments. In such situations, the bound above implies a good prediction accuracy for the output predictor. Note also that this bound is given in terms of the softmax approximation parameter μ. In general, this parameter should be treated as a hyperparameter and tuned appropriately to minimize the above bound. One reasonable choice of μ can be obtained by balancing the bound on the stationarity gap with the error term $\log(m+n)/\mu$ due to the softmax approximation. The corresponding value of the parameter is then given by $\mu = \widetilde{O}((\varepsilon n)^{2/7})$.

6 TWO-STAGE PRIVATE ADAPTATION ALGORITHMS

Here, we discuss a two-stage approach to private adaptation that consists of first privately obtaining q that (approximately) minimizes the empirical discrepancy and next fixing q to that value and minimizing the empirical q-weighted loss over $h\in\mathcal{H}$. In the absence of privacy constraints, this coincides with the algorithm of [CM14], which has been shown to both benefit from the theoretical guarantees and to outperform all other baselines.

We give here two private algorithms based on that two-stage paradigm. More specifically, the first stage consists of privately finding an approximate minimizer $\mathsf{q}\in\Delta_m$ for an ℓ_2-regularized version of the discrepancy:

$$\min_{\mathsf{q}\in\Delta_m}\|\mathbf{M}(\mathsf{q})\|_2 + \frac{\lambda}{2}\|\mathsf{q}\|_2^2 \tag{5}$$

The second stage simply consists of fixing the solution q obtained in the first stage and seeking $h\in\mathcal{H}$ that minimizes the q-weighted empirical loss:

$$\min_{w\in\mathbb{B}_2^d(\Lambda)}\sum_{i=1}^m \mathsf{q}_i\ell(\langle w,x_i\rangle,y_i), \tag{6}$$

where $\mathbb{B}_2^d(\Lambda)$ is the Euclidean ball in $\mathbb{R}$ of radius Λ. Equivalently, we can define an ℓ_2-regularized version of the weighted empirical loss and minimize it over $\mathbb{R}^d$; namely, solve $\min_{w\in\mathbb{R}^d}\sum_{i=1}^m \mathsf{q}_i\ell(\langle w,x_i\rangle,y_i)+\widetilde{\lambda}\|w\|_2^2$ where $\widetilde{\lambda}>0$ is a hyperparameter. Regularization in the first stage is done to ensure that the resulting weights q are not too sparse since sparse solutions can lead to poor output model in the second stage of the adaptation algorithm.

In the second stage, no private data is involved. Thus, in this section, we focus on private algorithms for the first stage. We give two private algorithms for that discrepancy minimization stage. Our private algorithms seek to minimize an ℓ_2-regularized version of the smooth approximation of the discrepancy, $\widetilde{F}$, discussed in Section 4.2. To emphasize its dependence on the private unlabeled dataset T, we will use the notation $\widetilde{F}_T$. Thus, our algorithms seek to privately minimize the following ℓ_2-regularized version of $\widetilde{F}_T$:

$$\widetilde{F}_T^{\lambda} \triangleq \widetilde{F}_T(\mathsf{q}) + \frac{\lambda}{2}\|\mathsf{q}\|_2^2.$$

As mentioned earlier, the regularization term is used to avoid sparse solutions q that may impact the accuracy of the output model in the second stage of the adaptation algorithm. Our algorithms are based on private variants of the Frank-Wolfe algorithm and the Mirror Descent algorithm. The general structure of these algorithms follow known private

constructions devised in the context of differentially private empirical risk minimization [TGTZ15, BGN21, AFKT21]. However, we note that the guarantees of both algorithms crucially rely on the smoothness and sensitivity properties of the approximation proved in Section 4. Solving the optimization with respect to the smooth approximation of the discrepancy enables us to bound the sensitivity of the gradients (see Theorem 2), which helps us devise private algorithms for this problem.

We start with a formal description of our noisy Frank-Wolfe algorithm (Algorithm 2), followed by a formal statement of its guarantees.

Algorithm 2 Noisy Frank-Wolfe for minimizing (regularized) smoothed discrepancy

Require: Private unlabeled dataset $T = (\widetilde{x}_1, \ldots, \widetilde{x}_n) \in \mathcal{X}^n$, public unlabeled dataset $S_{\mathcal{X}} = (x_1, \ldots, x_m) \in \mathcal{X}^m$, privacy parameters (ε, δ), smooth-approximation parameter μ, regularization parameter λ, # of iterations K.

Let $r = \max_{x \in \mathcal{X}} \|x\|_2$.
Let $\hat{r} = \max_{i \in [m]} \|x_i\|_2$.
Let Δ_m be the $(m-1)$-dimensional probability simplex.
Define $\widetilde{F}_T^{\lambda}(\mathsf{q}) \triangleq \widetilde{F}_T(\mathsf{q}) + \frac{\lambda}{2}\|\mathsf{q}\|_2^2$, $\mathsf{q} \in \Delta_m$.
Choose an arbitrary point $\mathsf{q}_1 \in \Delta_m$.
Set $\sigma = \frac{4\mu r^2 \hat{r}^2 \sqrt{2K \log(\frac{1}{\delta})}}{n\varepsilon}$.
for $k = 1$ to K **do**
  Compute $\nabla \widetilde{F}_T^{\lambda}(\mathsf{q}_k) = \nabla \widetilde{F}_T(\mathsf{q}_k) + \lambda \mathsf{q}_k$, where $\nabla \widetilde{F}_T(\mathsf{q}_k)$ is computed as described in Section 4.2.
  Draw $\{b_{i,k}\}_{i \in [m]}$ i.i.d. from $\mathsf{Lap}(\sigma)$.
  Find $j_k = \operatorname{argmin}_{i \in [m]} \{\langle \mathbf{e}_i, \nabla \widetilde{F}_T^{\lambda}(\mathsf{q}_k)\rangle + b_{i,k}\}$, where $\{\mathbf{e}_i\}_{i \in [m]}$ are the standard unit vectors in $\mathbb{R}^m$.
  Update $\mathsf{q}_{k+1} = (1 - \eta_k)\mathsf{q}_k + \eta_k \mathbf{e}_{j_k}$, where $\eta_k = \frac{3}{k+2}$.
end for
return $\widehat{\mathsf{q}} = \mathsf{q}_K$.

Theorem 5. *The Noisy Frank-Wolfe algorithm (Algorithm 2) is (ε, δ)-differentially private. Let $\mathsf{q}^* \in \operatorname{argmin}_{\mathsf{q} \in \Delta_m} \mathrm{dis}(\widehat{P}, \mathsf{q})$. Then, there exists a choice of the parameters of Algorithm 2 such that, with high probability over the algorithm's internal randomness, the output $\widehat{\mathsf{q}}$ satisfies*

$$\mathrm{dis}(\widehat{P}, \widehat{\mathsf{q}}) \leq \mathrm{dis}(\widehat{P}, \mathsf{q}^*) + \frac{\lambda}{2}\|\mathsf{q}^*\|_2^2 + \widetilde{O}\left(\frac{1}{(\varepsilon n)^{1/3}}\right),$$

where $\widetilde{O}(\cdot)$ is hiding a poly-logarithmic factor in m.

As shown in Theorem 5, the smoothness we created in $\widetilde{F}_T$ enables us to use a private variant of the Frank-Wolfe algorithm, whose optimization error scales only logarithmically with m.

Next, we give a formal description of our noisy mirror descent algorithm (Algorithm 3) followed by a formal statement of its guarantees.

Algorithm 3 Noisy Mirror-Descent for minimizing $\widetilde{F}_T^{\lambda}$

Require: Private unlabeled dataset $T = (\widetilde{x}_1, \ldots, \widetilde{x}_n) \in \mathcal{X}^n$, public unlabeled dataset $S_{\mathcal{X}} = (x_1, \ldots, x_m) \in \mathcal{X}^m$, privacy parameters (ε, δ), smooth-approximation parameter μ, number of iterations K.

Let $r = \max_{x \in \mathcal{X}} \|x\|_2$.
Let $\hat{r} = \max_{i \in [m]} \|x_i\|_2$.
Let Δ_m be the $(m-1)$-dimensional probability simplex.
Let $p = 1 + \frac{1}{\log(m)}$.
Set $\sigma = \frac{4\mu r^2 \hat{r}^2 \sqrt{2Km \log(\frac{1}{\delta})}}{n\varepsilon}$
Set $\eta = \frac{2}{(\hat{r}^2 + \lambda)} \sqrt{\frac{\log(m)}{K}}$.
Choose an arbitrary point $\mathsf{q}_1 \in \Delta_m$.
for $k = 1$ to K **do**
  Compute $\widehat{\nabla}_k = \nabla \widetilde{F}_T(\mathsf{q}_k) + \lambda \mathsf{q}_k + Z_k$, where $Z_k \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbb{I}_m)$.
  Update $\mathsf{q}_{k+1} = \arg\min_{\mathsf{q} \in \Delta_m} \left\{ \langle \widehat{\nabla}_k, \mathsf{q} - \mathsf{q}_k \rangle + \frac{\|\mathsf{q} - \mathsf{q}_k\|_p^2}{\eta(p-1)} \right\}$.
end for
return $\widehat{\mathsf{q}} = \frac{1}{K} \sum_{k=1}^{K} \mathsf{q}_k$

Theorem 6. *The Noisy Mirror Descent algorithm (Algorithm 3) is (ε, δ)-differentially private. Let $\mathsf{q}^* \in \operatorname{argmin}_{\mathsf{q} \in \Delta_m} \mathrm{dis}(\widehat{P}, \mathsf{q})$. There exists a choice of the parameters of Algorithm 3 such that with high probability over the algorithm's randomness, the output $\widehat{\mathsf{q}}$ satisfies*

$$\mathrm{dis}(\widehat{P}, \widehat{\mathsf{q}}) \leq \mathrm{dis}(\widehat{P}, \mathsf{q}^*) + \frac{\lambda}{2}\|\mathsf{q}^*\|_2^2 + \widetilde{O}\left(\frac{m^{1/4}}{\sqrt{\varepsilon n}}\right).$$

Note that, compared to the guarantees of the private Frank-Wolfe algorithm in Theorem 5, the optimization error of the Noisy Mirror Descent algorithm (Theorem 6) exhibits a better dependence on n at the expense of worse dependence on m. In Appendix E, we give full proofs of these theorems.

Implication on the learning guarantee: Note that by standard stability arguments, the minimum weighted empirical loss of the second stage when training with q^* is close to the minimum weighted empirical loss when training with $\widehat{\mathsf{q}}$ when the discrepancy between $\widehat{\mathsf{q}}$ and q^* is small [MMR09]. Theorems 5 and 6 precisely supply guarantees for that closeness in discrepancy via the inequality $\mathrm{dis}(\widehat{\mathsf{q}}, \mathsf{q}^*) \leq \mathrm{dis}(\widehat{\mathcal{P}}, \widehat{\mathsf{q}}) - \mathrm{dis}(\widehat{\mathcal{P}}, \mathsf{q}^*)$, thereby guaranteeing the closeness of the loss of our private predictor (output of the second stage) to the minimum q^*-weighted empirical loss. This, together with the learning bound (1), immediately provide a bound on the expected loss of our private predictor.

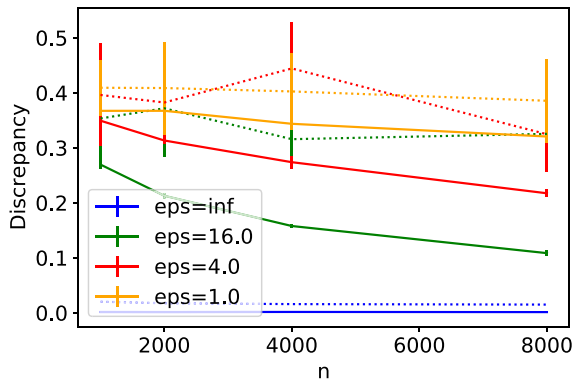

Figure 1: Value of the spectral norm $\|\mathbf{M}(\mathsf{q})\|_2$ for the output of noisy Frank-Wolfe (solid lines) and noisy Mirror descent (dotted lines) discrepancy minimization as a function of the number of samples from the private dataset n.

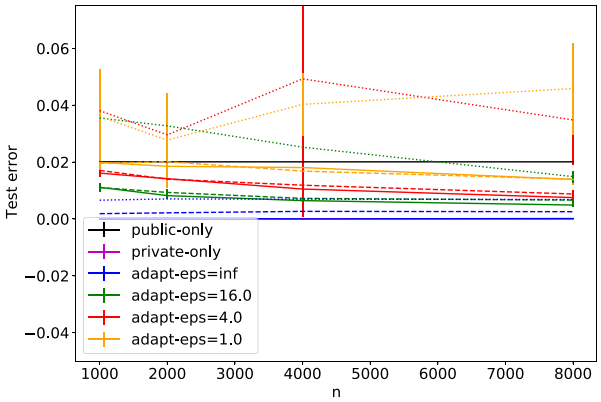

Figure 2: Test error as a function of the number of samples from the private dataset n. The solid lines correspond to the single-stage algorithm, the dotted lines to the two-stage mirror decent algorithm, and dashed lines to the two-stage Frank Wolfe algorithm.

7 EMPIRICAL RESULTS

The objective of this section is to provide proof-of-concept experiments to demonstrate that reasonable privacy guarantees can be achieved, when using our private domain adaptation algorithms. We use a setting similar to that of [CM14, Section 7.1] and demonstrate that the utility of private adaptation degrades gracefully with increased privacy guarantees and that the single-stage Frank-Wolfe algorithm performs best most scenarios.

We carried out experiments with the following synthetic dataset. Let $d = 10$ and $\sigma^2 = 1/(9d)$. We chose $P_{\mathcal{X}}$ to be a spherical Gaussian centered around $(-1/\sqrt{2d}, 1/\sqrt{2d}, \ldots, -1/\sqrt{2d}, 1/\sqrt{2d})$ and with variance σ^2 in all directions. Let $Q_{\mathcal{X}}$ be a Gaussian distribution with mean $(1/\sqrt{2d}, \ldots, 1/\sqrt{2d})$ and with variance σ^2 in all directions. We defined the labeling function via $f(x) = x \cdot \bar{1}$ if $\bar{1} \cdot x > 0$, $\left(\frac{1}{2} x \cdot \bar{1}\right)$ otherwise, where $\bar{1} = (1/\sqrt{d}, \ldots, 1/\sqrt{d})$. We chose the target distribution to be $P_{\mathcal{X}}$ and the source distribution as a mixture of $P_{\mathcal{X}}$ and $Q_{\mathcal{X}}$ with the weight of $P_{\mathcal{X}}$ set to 25%. We fixed the number of source samples to be $1{,}000$ and varied the number of unlabeled target samples from $1{,}000$ to $8{,}000$. All experiments were repeated ten times for statistical consistency. We set $K = 1{,}000$, $\lambda = 0.001$, the privacy parameter $\delta = 1/8{,}000$, and varied ϵ in experiments. The standard deviations were calculated over 10 runs in experiments.

In this setup, we first ran differentially private discrepancy minimization using Algorithms 2 and 3. We plotted $\|\mathbf{M}(\mathsf{q})\|_2$ for different values of ϵ in Figure 1. The performance of the noisy Frank-Wolfe algorithm degrades smoothly with ϵ and improves with n. However the performance of the noisy mirror decent algorithm is much worse. This is in line with the theoretical guarantees as $m = \Omega(n^{2/3})$ in these experiments and noisy Frank-Wolfe algorithm has a better convergence guarantee in this regime. We expect mirror descent to perform better with much larger values of n. Furthermore, observe that the noisy mirror descent has a high standard deviation compared to Frank-Wolfe algorithm as the noise added in mirror descent scales polynomially in m, whereas it scales only logarithmically in m for the Frank-Wolfe algorithm.

We next compared our single-stage (Algorithm 1) and the two-stage differentially private algorithms with the model trained only with the public dataset (Figure 2). As an oracle baseline, we also plotted the model trained with the labeled private dataset. Note that this model uses extra information that is not available during training and is plotted for illustration purposes only. The single-stage Frank-Wolfe algorithm without privacy admits the same performance as the model trained on the labeled private dataset. It performs better than the two-stage Frank-Wolfe algorithm, however the gap decreases as the privacy guarantee ϵ improves. The performance of the mirror descent algorithm without differential privacy is similar to that of Frank-Wolfe algorithm, however as theory indicates, the performance degrades quickly with the privacy parameter. Similar to Figure 1, the performance of the noisy mirror descent algorithm is much worse and has a high standard deviation.

8 CONCLUSION

We presented new differentially private adaptation algorithms with formal theoretical guarantees. Our analysis can form the basis for the study of privacy for other related adaptation scenarios, including scenarios where a small amount of (private) labeled data is also available from the target domain and those with multiple sources. Our single-stage private algorithm is further likely to be of independent interest for private optimization of other similar objective functions. The solutions we presented are for regression problems, as with the non-private algorithm of [CM14]. We leave it to future work to leverage similar ideas and techniques to derive principled private adaptation algorithms from a public source for classification problems.

Acknowledgements

RB's research is supported by NSF CAREER Award 2144532, NSF Award AF-1908281, and Google Faculty Research Award.

References

[ABM19] Noga Alon, Raef Bassily, and Shay Moran. Limits of private learning with access to public data. *NeuRIPS 2019, also available at arXiv:1910.11519 [cs.LG]*, 2019.

[ACD+19] Yossi Arjevani, Yair Carmon, John C. Duchi, Dylan J. Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization, 2019.

[ACG+16] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[AFKT21] Hilal Asi, Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: Optimal rates in L1 geometry. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 393–403. PMLR, 2021.

[BBCP06] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In Bernhard Schölkopf, John C. Platt, and Thomas Hofmann, editors, *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006*, pages 137–144. MIT Press, 2006.

[BCK+08] John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman. Learning bounds for domain adaptation. In *Advances in neural information processing systems*, pages 129–136, 2008.

[BCM+20] Raef Bassily, Albert Cheu, Shay Moran, Aleksandar Nikolov, Jonathan Ullman, and Steven Wu. Private query release assisted by public data. In *International Conference on Machine Learning*, pages 695–703. PMLR, 2020.

[BDBC+10] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.

[BGM21] Raef Bassily, Cristóbal Guzmán, and Michael Menart. Differentially private stochastic optimization: New results in convex and non-convex settings. *Advances in Neural Information Processing Systems*, 34, 2021.

[BGN21] Raef Bassily, Cristóbal Guzmán, and Anupama Nandi. Non-euclidean differentially private stochastic convex optimization. *arXiv preprint arXiv:2103.01278*, 2021.

[BLST10] Raghav Bhaskar, Srivatsan Laxman, Adam Smith, and Abhradeep Thakurta. Discovering frequent patterns in sensitive data. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 503–512, 2010.

[BNS13] Amos Beimel, Kobbi Nissim, and Uri Stemmer. Private learning and sanitization: Pure vs. approximate differential privacy. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 363–378. Springer, 2013.

[BTT18] Raef Bassily, Abhradeep Thakurta, and Om Thakkar. Model-agnostic private learning. In *Advances in Neural Information Processing Systems 31*, pages 7102–7112. Curran Associates, Inc., 2018.

[BV14] Stephen P. Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2014.

[CDHS17] Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. "convex until proven guilty": Dimension-free acceleration of gradient descent on non-convex functions. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 654–663. JMLR.org, 2017.

[CH11] Kamalika Chaudhuri and Daniel Hsu. Sample complexity bounds for differentially private learning. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 155–186, 2011.

[CM14] Corinna Cortes and Mehryar Mohri. Domain adaptation and sample bias correction theory and algorithm for regression. *Theor. Comput. Sci.*, 519:103–126, 2014.

[CMMM19] Corinna Cortes, Mehryar Mohri, and Andrés Muñoz Medina. Adaptation based on generalized discrepancy. *J. Mach. Learn. Res.*, 20:1:1–1:30, 2019.

[DKM+06] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 486–503. Springer, 2006.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[DR14] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[FLLZ18] Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[GL13] Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[JCYS21] Kaizhong Jin, Xiang Cheng, Jiaxi Yang, and Kaiyuan Shen. Differentially private correlation alignment for domain adaptation. In *IJCAI*, volume 21, pages 3649–3655, 2021.

[JN08] Anatoli Juditsky and Arkadi Nemirovski. Large deviations of vector-valued martingales in 2-smooth normed spaces. Rapport de recherche hal-00318071, HAL, 2008.

[KBG04] Daniel Kifer, Shai Ben-David, and Johannes Gehrke. Detecting change in data streams. In Mario A. Nascimento, M. Tamer Özsu, Donald Kossmann, Renée J. Miller, José A. Blakeley, and K. Bernhard Schiefer, editors, *(e)Proceedings of the Thirtieth International Conference on Very Large Data Bases, VLDB 2004, Toronto, Canada, August 31 - September 3 2004*, pages 180–191. Morgan Kaufmann, 2004.

[KM15] Vitaly Kuznetsov and Mehryar Mohri. Learning theory and algorithms for forecasting non-stationary time series. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *Proceedings of NIPS*, pages 541–549, 2015.

[KM20] Vitaly Kuznetsov and Mehryar Mohri. Discrepancy-based theory and algorithms for forecasting non-stationary time series. *Ann. Math. Artif. Intell.*, 88(4):367–399, 2020.

[MM12] Mehryar Mohri and Andres Muñoz Medina. New analysis and algorithm for learning with drifting distributions. In Nader H. Bshouty, Gilles Stoltz, Nicolas Vayatis, and Thomas Zeugmann, editors, *Proceesings of ALT*, volume 7568 of *Lecture Notes in Computer Science*, pages 124–138. Springer, 2012.

[MMR09] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009*, 2009.

[MRT18] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. Adaptive computation and machine learning. MIT Press, 2018. Second edition.

[MWCC18] Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix completion. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3345–3354. PMLR, 10–15 Jul 2018.

[NB20] Anupama Nandi and Raef Bassily. Privately answering classification queries in the agnostic pac model. In *Algorithmic Learning Theory*, pages 687–703, 2020.

[Nes07] Yurii Nesterov. Smoothing technique and its applications in semidefinite optimization. *Math. Program.*, 110:245–259, 2007.

[NJLS09] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

[NP06] Yurii Nesterov and Boris Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108:177–205, 2006.

[NY83] A.S. Nemirovsky and D.B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. A Wiley-Interscience publication. Wiley, 1983.

[Sid19] Aaron Sidford. Introduction to optimization theory - MS&E213 / CS269O.

Stanford Course Notes, 2019. https://web.stanford.edu/~sidford/courses/19fa_opt_theory/.

[TGTZ15] Kunal Talwar, Abhradeep Guha Thakurta, and Li Zhang. Nearly optimal private lasso. *Advances in Neural Information Processing Systems*, 28:3025–3033, 2015.

[WLZ+20] Qian Wang, Zixi Li, Qin Zou, Lingchen Zhao, and Song Wang. Deep domain adaptation with differential privacy. *IEEE Transactions on Information Forensics and Security*, 15:3093–3106, 2020.

Contents of Appendix

A Background on discrepancy-based generalization bounds

In this section, we briefly present some background material on discrepancy-based generalization guarantees.

The following learning bound was given by [CMMM19]: for any $\beta > 0$, with probably at least $1 - \beta$ over the draw of a sample $S \sim \mathcal{Q}^m$, for any distribution q over $S_{\mathcal{X}}$, for all $h \in \mathcal{H}$, the following inequality holds:

$$\mathcal{L}(\mathcal{P}, h) \leq \sum_{i=1}^{m} \mathsf{q}_i \ell(h(x_i), y_i) + \mathrm{dis}_{\mathcal{Y}}\big(\widehat{\mathcal{P}}, \mathsf{q}\big) + 2\mathfrak{R}_n(\ell \circ \mathcal{H}) + M\sqrt{\frac{\log \frac{1}{\beta}}{2n}}. \tag{7}$$

This bound is tight in the sense that for the hypothesis reaching the maximum in the definition of the $\mathcal{Y}$-discrepancy, the bound coincides with the standard Rademacher complexity bound on $\widehat{\mathcal{P}}$ [CMMM19]. The bound suggests choosing $h \in \mathcal{H}$ and the distribution q to minimize the right-hand side. The first term of the bound is not jointly convex with respect to h and q. Instead, the algorithm suggested by [CM14] (see also [MMR09]) consists of a two-stage procedure: first choose q to minimize the q-weighted empirical discrepancy, next fix q and choose h to minimize the q-weighted empirical loss $\sum_{i=1}^{m} \mathsf{q}_i \ell(h(x_i), y_i)$.

In practice, we do not have labeled data from $\mathcal{P}$ or too few to be able to accurately minimize the $\mathcal{Y}$-discrepancy, since otherwise adaptation would not be even necessary and we could directly use labeled data from $\mathcal{P}$ for training. Instead, we upper bound the $\mathcal{Y}$-discrepancy in terms of the discrepancy $\mathrm{dis}(\mathcal{P}_{\mathcal{X}}, \mathsf{q})$ and the *output label-discrepancy* $\eta_{\mathcal{H}}(S, \widetilde{T})$ defined as follows:

$$\eta_{\mathcal{H}}(S, \widetilde{T}) = \min_{h_0 \in \mathcal{H}} \left\{ \sup_{(x,y) \in S} |y - h_0(x)| + \sup_{(x,y) \in \widetilde{T}} |y - h_0(x)| \right\},$$

where $\widetilde{T}$ is the labeled version of T (i.e., $\widetilde{T}$ is T associated with its true, hidden labels). Note that $\mathrm{dis}(\mathcal{P}_{\mathcal{X}}, \mathsf{q})$ measures the difference of the distributions on the input domain. In contrast, $\eta_{\mathcal{H}}(S, \widetilde{T})$ accounts for the difference of the output labels in S and T. We will assume that $\eta_{\mathcal{H}}(S, \widetilde{T}) \ll 1$. Note that under the covariate shift assumption and separable case, we have $\eta_{\mathcal{H}}(S, \widetilde{T}) = 0$. In general, adaptation is not possible when $\eta_{\mathcal{H}}(S, \widetilde{T})$ can be large since the labels received on the training sample can be different from the target ones.

We will say that a loss function ℓ is γ-*admissible* if $|\ell(h(x), y) - \ell(h'(x), y)| \leq \gamma |h(x) - h'(x)|$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and $h \in \mathcal{H}$ [CMMM19]. Note that this is a slightly weaker condition than that of γ-Lipschitzness of the loss with respect to its first argument.

Theorem 7. *Let ℓ be a γ-admissible loss. Then, the following upper bound holds:*

$$\mathrm{dis}_{\mathcal{Y}}(\mathcal{P}, \mathcal{Q}) \leq \mathrm{dis}(\mathcal{P}_{\mathcal{X}}, \mathsf{q}) + \gamma\, \eta_{\mathcal{H}}(\mathrm{supp}(\mathcal{P}), \mathrm{supp}(\mathcal{Q})).$$

The proof is given in Appendix B. Note that the squared loss is $2M$-admissible: since the function $x \mapsto x^2$ is 2-Lipschitz on $[0, 1]$, we have $|\ell(h(x), y) - \ell(h'(x), y)| = M\left|\frac{\ell(h(x),y)}{M} - \frac{\ell(h'(x),y)}{M}\right| \leq 2M|h(x) - h'(x)|$. Thus, the learning bound (7) can be expressed in terms of the discrepancy and the Rademacher complexity of $\mathcal{H}$ as follows, using the fact $\mathfrak{R}_n(\ell \circ \mathcal{H}) \leq 2M\mathfrak{R}_n(\mathcal{H})$ [MRT18][Prop. 11.2]:

$$\mathcal{L}(\mathcal{P}, h) \leq \sum_{i=1}^{m} \mathsf{q}_i \ell(h(x_i), y_i) + \mathrm{dis}\big(\widehat{\mathcal{P}}_{\mathcal{X}}, \mathsf{q}\big) + \eta_{\mathcal{H}}(S, S') + 2M\mathfrak{R}_n(\mathcal{H}) + M\sqrt{\frac{\log \frac{1}{\beta}}{2n}}.$$

In this work, we focus on the family of linear hypotheses $\mathcal{H} = \{x \mapsto w \cdot x : \|w\| \leq \Lambda\}$ and we assume that the support of $\mathcal{P}_{\mathcal{X}}$ is included in the ℓ_2 ball of radius r. Thus, the Rademacher complexity can be explicitly upper bounded as: $\mathfrak{R}_n(\mathcal{H}) \leq \sqrt{\frac{r^2\Lambda^2}{n}}$ [MRT18].

B Discrepancy analysis and bounds

Theorem 7. *Let ℓ be a γ-admissible loss. Then, the following upper bound holds:*

$$\mathrm{dis}_{\mathcal{Y}}(\mathcal{P}, \mathcal{Q}) \leq \mathrm{dis}(\mathcal{P}_{\mathcal{X}}, \mathsf{q}) + \gamma\, \eta_{\mathcal{H}}(\mathrm{supp}(\mathcal{P}), \mathrm{supp}(\mathcal{Q})).$$

Proof. For any hypothesis h_0 in $\mathcal{H}$, we can write

$$\begin{aligned}
\mathrm{dis}_{\mathcal{Y}}(\widehat{\mathcal{P}}, \mathsf{q}) &= \sup_{h \in \mathcal{H}} \left| \mathop{\mathbb{E}}_{(x,y)\sim\widehat{\mathcal{P}}}[\ell(h(x), y)] - \sum_{i=1}^{m} \mathsf{q}_i \ell(h(x_i), y_i) \right| \\
&\leq \sup_{h \in \mathcal{H}} \left| \mathop{\mathbb{E}}_{x\sim\widehat{\mathcal{P}}}[\ell(h(x), h_0(x))] - \sum_{i=1}^{m} \mathsf{q}_i \ell(h(x_i), h_0(x_i)) \right| \\
&\quad + \sup_{h \in \mathcal{H}} \left| \mathop{\mathbb{E}}_{(x,y)\sim\widehat{\mathcal{P}}}[\ell(h(x), y)] - \mathop{\mathbb{E}}_{x\sim\widehat{\mathcal{P}}}[\ell(h(x), h_0(x))] \right| \\
&\quad + \sup_{h \in \mathcal{H}} \left| \sum_{i=1}^{m} \mathsf{q}_i \ell(h(x_i), h_0(x_i)) - \sum_{i=1}^{m} \mathsf{q}_i \ell(h(x_i), y_i) \right| \\
&\leq \mathrm{dis}(\widehat{\mathcal{P}}_{\mathcal{X}}, \mathsf{q}) + \gamma \mathop{\mathbb{E}}_{(x,y)\sim\widehat{\mathcal{P}}}[|y - h_0(x)|] + \gamma \sum_{i=1}^{m} \mathsf{q}_i |y_i - h_0(x_i)| \\
&\leq \mathrm{dis}(\widehat{\mathcal{P}}_{\mathcal{X}}, \mathsf{q}) + \gamma \left\{ \sup_{(x,y)\in\mathrm{supp}(\widehat{\mathcal{P}})} |y - h_0(x)| + \sup_{(x,y)\in\mathrm{supp}(\widehat{\mathcal{Q}})} |y - h_0(x)| \right\} \\
&= \mathrm{dis}(\widehat{\mathcal{P}}_{\mathcal{X}}, \mathsf{q}) + \gamma \eta_{\mathcal{H}}(\mathrm{supp}(\mathcal{P}), \mathrm{supp}(\mathcal{Q})),
\end{aligned}$$

which completes the proof. □

Lemma 1 ([MMR09, CM14])**.** *For any distribution q over $S_{\mathcal{X}}$, the following equality holds:*

$$\begin{aligned}
\mathrm{dis}(\widehat{P}, \mathsf{q}) &= 4\Lambda^2 \|\mathbf{M}(\mathsf{q})\|_2 \\
&= 4\Lambda^2 \max\{\lambda_{\max}(\mathbf{M}(\mathsf{q})), \lambda_{\max}(-\mathbf{M}(\mathsf{q}))\},
\end{aligned}$$

where $\mathbf{M}(\mathsf{q}) = \mathbf{M}_0 - \sum_{i=1}^{m} q_i \mathbf{M}_i$ and where $\mathbf{M}_0 = \sum_{x\in\mathcal{X}} \widehat{\mathcal{P}}_{\mathcal{X}}(x) x x^\top$, and $\mathbf{M}_i = x_i x_i^\top$, $i \in [m]$.

Proof.

$$\begin{aligned}
\mathrm{dis}(\widehat{P}, \mathsf{q}) &= \max_{\|w\|, \|w'\| \leq \Lambda} \left| \mathop{\mathbb{E}}_{x\sim\mathsf{q}}\left[[(w - w') \cdot x]^2\right] - \mathop{\mathbb{E}}_{x\sim\widehat{\mathcal{P}}_X}\left[[(w - w') \cdot x]^2\right] \right| \\
&= \max_{\|w\|, \|w'\| \leq \Lambda} \left| \sum_{x\in\mathcal{X}} \left[\widehat{P}(x) - \mathsf{q}(x)\right] [(w - w') \cdot x]^2 \right| \\
&= \max_{\|u\| \leq 2\Lambda} \left| \sum_{x\in\mathcal{X}} \left[\widehat{P}(x) - \mathsf{q}(x)\right] [u \cdot x]^2 \right| \\
&= \max_{\|u\| \leq 2\Lambda} \left| u^\top \left[\sum_{x\in\mathcal{X}} \left(\widehat{P}(x) - \mathsf{q}(x)\right) x x^\top \right] u \right| \\
&= 4\Lambda^2 \max_{\|u\| \leq 1} \left| u^\top \left[\mathbf{M}_0 - \sum_{i=1}^{m} \mathsf{q}_i \mathbf{M}_i \right] u \right| \\
&= 4\Lambda^2 \max_{\|u\| \leq 1} \left| u^\top \mathbf{M}(\mathsf{q}) u \right| \\
&= 4\Lambda^2 \max_{\|u\| = 1} \left| u^\top \mathbf{M}(\mathsf{q}) u \right| \\
&= 4\Lambda^2 \max\{\lambda_{\max}(\mathbf{M}(\mathsf{q})), \lambda_{\max}(-\mathbf{M}(\mathsf{q}))\}.
\end{aligned}$$

This completes the proof. □

C Smooth approximations

C.1 Softmax approximation

Proposition 1. *Assume that f is γ-smooth with respect to $\|\cdot\|_2$, then F is $\gamma\big(\max_{i\in[m]}\|x_i\|_2^4\big)$-smooth with respect to $\|\cdot\|_1$.*

Proof. For any $\mathsf{q}, \mathsf{q}' \in \Delta(m)$, the following upper bound on the spectral norm of $\mathbf{M}(\mathsf{q}) - \mathbf{M}(\mathsf{q}')$ holds:

$$\begin{aligned}
\|\mathbf{M}(\mathsf{q}) - \mathbf{M}(\mathsf{q}')\|_2 &= \left\|\sum_{i=1}^m (\mathsf{q}_i - \mathsf{q}'_i) x_i x_i^\top\right\|_2 && (8)\\
&\le \sum_{i=1}^m |\mathsf{q}_i - \mathsf{q}'_i| \left\|x_i x_i^\top\right\|_2 \\
&\le \|\mathsf{q} - \mathsf{q}'\|_1 \max_{i\in[m]} \left\|x_i x_i^\top\right\|_2 && \text{(Hölder's ineq.)}\\
&= \|\mathsf{q} - \mathsf{q}'\|_1 \max_{i\in[m]} \|x_i\|_2^2. && (x_i x_i^\top \text{ admits a single non-zero eigenvalue, } \|x_i\|_2^2)
\end{aligned}$$

We have $F(\mathsf{q}) = f(\mathbf{M}(\mathsf{q}))$, thus the gradient of F can be expressed as follows:

$$\nabla F(\mathsf{q}) = -\big[\langle \nabla f(\mathbf{M}(\mathsf{q})), \mathbf{M}_i\rangle\big]_{i\in[m]}.$$

Thus, for any $\mathsf{q}, \mathsf{q}' \in \Delta(m)$, we have:

$$\begin{aligned}
\|\nabla F(\mathsf{q}) - \nabla F(\mathsf{q}')\|_\infty &= \max_{i\in[m]} |\langle \nabla f(\mathbf{M}(\mathsf{q})) - \nabla f(\mathbf{M}(\mathsf{q}')), \mathbf{M}_i\rangle| \\
&\le \max_{i\in[m]} \|\nabla f(\mathbf{M}(\mathsf{q})) - \nabla f(\mathbf{M}(\mathsf{q}'))\|_{(1)} \|\mathbf{M}_i\|_{(\infty)} && \text{(Hölder's ineq.)}\\
&\le \gamma \max_{i\in[m]} \|\mathbf{M}(\mathsf{q}) - \mathbf{M}(\mathsf{q}')\|_{(\infty)} \|\mathbf{M}_i\|_{(\infty)} && (\gamma\text{-smoothness of } f)\\
&= \gamma \max_{i\in[m]} \|\mathbf{M}(\mathsf{q}) - \mathbf{M}(\mathsf{q}')\|_2 \|\mathbf{M}_i\|_2 && \text{(definition of } \|\cdot\|_{(\infty)})\\
&\le \gamma \max_{i\in[m]} \left\{\|\mathsf{q} - \mathsf{q}'\|_1 \max_{i\in[m]} \|x_i\|_2^2\right\} \|x_i\|_2^2 && \text{(inequality (8))}\\
&= \gamma \left(\max_{i\in[m]} \|x_i\|_2^4\right) \|\mathsf{q} - \mathsf{q}'\|_1.
\end{aligned}$$

This completes the proof. □

We will use the following bound for the Hessian of f.

Lemma 2 ([Nes07])**.** *The following upper bound holds for the Hessian of f for any two symmetric matrices $\mathbf{M}, \mathbf{U} \in \mathbb{S}_d$:*

$$\left\langle \nabla^2 f(\mathbf{M})\, \mathbf{U}, \mathbf{U}\right\rangle \le \mu \|\mathbf{U}\|_2^2,$$

where $\|\mathbf{U}\|_2 = \|\lambda(\mathbf{U})\|_\infty$ denotes the spectral norm of $\mathbf{U}$.

Theorem 1. *The softmax approximation F is $\mu\big(\max_{i\in[m]}\|x_i\|_2^4\big)$-smooth for $\|\cdot\|_1$.*

Proof. In view of Lemma 2, f is $\|\cdot\|_2$-μ-smooth. The result thus follows by Proposition 1. □

Theorem 2. *The gradient of the softmax approximation F is $\frac{2\mu r^2}{n} \max_{i\in[m]}\|x_i\|_2^2$-sensitive.*

Proof. For $\mathbf{M}(\mathsf{q}))$ and $\mathbf{M}'(\mathsf{q}))$ differing only by point x and x' in $\widehat{\mathcal{P}}_{\mathcal{X}}$, we have:

$$\|\mathbf{M}(\mathsf{q}) - \mathbf{M}'(\mathsf{q})\|_2 = \left\|\frac{1}{n}\big[xx^\top - x'x'^\top\big]\right\|_2 \le \frac{2r^2}{n}. \tag{9}$$

Thus, following the proof of Proposition 1, the sensitivity is bounded by

$$\begin{aligned}
\max_{i\in[m]} |\langle \nabla f(\mathbf{M}(\mathsf{q})) - \nabla f(\mathbf{M}'(\mathsf{q})), \mathbf{M}_i\rangle| &\le \max_{i\in[m]} \|\nabla f(\mathbf{M}(\mathsf{q})) - \nabla f(\mathbf{M}'(\mathsf{q}))\|_{(1)} \|\mathbf{M}_i\|_{(\infty)} && \text{(Hölder's ineq.)}\\
&\le \mu \max_{i\in[m]} \|\mathbf{M}(\mathsf{q}) - \mathbf{M}(\mathsf{q}')\|_{(\infty)} \|\mathbf{M}_i\|_{(\infty)} && (\mu\text{-smoothness of } f)\\
&= \mu \max_{i\in[m]} \|\mathbf{M}(\mathsf{q}) - \mathbf{M}(\mathsf{q}')\|_2 \|\mathbf{M}_i\|_2 && \text{(definition of } \|\cdot\|_{(\infty)})\\
&\le \frac{2\mu r^2}{n} \max_{i\in[m]} \|x_i\|_2^2.
\end{aligned}$$

This completes the proof. □

Proposition 2. *The following inequality holds for the spectral norm of the Hessian of F:*

$$\|\nabla^2 F\|_2 \le \mu \left\| \sum_{i=1}^m x_i x_i^\top \right\|_2 \le \mu \left[\sum_{i=1}^m \|x_i\|_2^2 \right].$$

Proof. The second-partial derivatives of $F(\mathsf{q})$ can be expressed as follows:

$$\begin{aligned}
\frac{\partial^2 S}{\partial \mathsf{q}_i \partial \mathsf{q}_j} &= -\left\langle \frac{\partial}{\partial \mathsf{q}_j} \nabla f(\mathbf{M}(\mathsf{q})), \mathbf{M}_i \right\rangle\\
&= +\langle \nabla^2 f(\mathbf{M}(\mathsf{q}))\, \mathbf{M}_j, \mathbf{M}_i\rangle.
\end{aligned}$$

Thus, using the shorthand $\overline{\mathbf{M}} = \sum_{i=1}^m \mathbf{X}_i \mathbf{M}_i$, for any $\mathbf{X} \in \mathbb{R}^m$, we can write:

$$\begin{aligned}
\mathbf{X}^\top \nabla^2 F \mathbf{X} &= \sum_{i,j=1}^d \mathbf{X}_i \mathbf{X}_j \left\langle \nabla^2 f(\mathbf{M}(\mathsf{q}))\, \mathbf{M}_j, \mathbf{M}_i \right\rangle\\
&= \left\langle \nabla^2 f(\mathbf{M}(\mathsf{q})) \left(\sum_{j=1}^d \mathbf{X}_j \mathbf{M}_j \right), \left(\sum_{i=1}^d \mathbf{X}_i \mathbf{M}_i \right) \right\rangle\\
&= \left\langle \nabla^2 f(\mathbf{M}(\mathsf{q})) \left(\overline{\mathbf{M}} \right), \left(\overline{\mathbf{M}} \right) \right\rangle\\
&\le \mu \|\overline{\mathbf{M}}\|_2^2 && \text{(Lemma 2)}\\
&= \mu \left(\left\| \sum_{i=1}^m \mathbf{X}_i x_i x_i^\top \right\|_2 \right)^2\\
&= \mu \left(\max_{\|u\|\le 1} \left| \sum_{i=1}^m \mathbf{X}_i u^\top x_i x_i^\top u \right| \right)^2 && \text{(def. of spectral norm)}\\
&= \mu \left(\max_{\|u\|\le 1} \left| \sum_{i=1}^m \mathbf{X}_i (u^\top x_i)^2 \right| \right)^2\\
&\le \mu \left(\max_{\|u\|\le 1} \|\mathbf{X}\| \sqrt{\sum_{i=1}^m (u^\top x_i)^2} \right)^2 && \text{(Cauchy-Schwarz ineq.)}\\
&= \mu \left(\|\mathbf{X}\| \sqrt{\max_{\|u\|\le 1} \sum_{i=1}^m (u^\top x_i)^2} \right)^2\\
&= \mu \left\| \sum_{i=1}^m x_i x_i^\top \right\|_2 \|\mathbf{X}\|^2.
\end{aligned}$$

This completes the proof. □

Theorem 3. *For any $\mathsf{q} \in \Delta_m$, the gradient of F is bounded as follows: $\|\nabla F(\mathsf{q})\|_\infty \le \max_{i\in[m]} \|x_i\|_2^2$.*

Proof. By inequality (4), for any $i \in [m]$, we have

$$\begin{aligned}
|[\nabla F(\mathsf{q})]_i| &= \left| \frac{\langle \exp(\mu \mathbf{M}(\mathsf{q})), \mathbf{M}_i \rangle}{\operatorname{Tr}(\exp(\mu \mathbf{M}(\mathsf{q})))} \right| \\
&= \frac{x_i^\top \exp(\mu \mathbf{M}(\mathsf{q})) x_i}{\operatorname{Tr}(\exp(\mu \mathbf{M}(\mathsf{q})))} \\
&\leq \|x_i\|_2^2 \frac{\max_{\|u\|_2=1} u^\top \exp(\mu \mathbf{M}(\mathsf{q})) u}{\operatorname{Tr}(\exp(\mu \mathbf{M}(\mathsf{q})))} \\
&= \|x_i\|_2^2 \frac{\lambda_{\max}(\exp(\mu \mathbf{M}(\mathsf{q})))}{\operatorname{Tr}(\exp(\mu \mathbf{M}(\mathsf{q})))} \leq \|x_i\|_2^2.
\end{aligned}$$

This completes the proof. □

C.2 Properties of $\widetilde{F}$

Corollary 1. *The following properties holds for $\widetilde{F}$:*

1. *$\widetilde{F}$ is convex and is a uniform $\frac{\log(2\min\{m+n,d\})}{\mu}$-approximation of $\mathsf{q} \mapsto \operatorname{dis}(\widehat{P}, \mathsf{q})$.*
2. *$\widetilde{F}$ is $\mu\big(\max_{i\in[m]} \|x_i\|_2^4\big)$-smooth for $\|\cdot\|_1$.*
3. *$\|\nabla \widetilde{F}\|_\infty$ is $\frac{2\mu r^2}{n} \max_{i\in[m]} \|x_i\|_2^2$-sensitive.*
4. *for any $\mathsf{q} \in \Delta_m$, $\|\nabla \widetilde{F}(\mathsf{q})\|_\infty \leq \max_{i\in[m]} \|x_i\|_2^2$.*

Proof. The results follow directly the definition of $\widetilde{F}$ and Theorems 1, 2, 3 and the discussion above. In particular, since $\widetilde{F}(\mathsf{q}) = f(\widetilde{\mathbf{M}}(\mathsf{q}))$, the gradient of $\widetilde{F}$ can be expressed as follows in terms of f:

$$\nabla \widetilde{F}(\mathsf{q}) = -\big\langle \nabla f(\widetilde{\mathbf{M}}(\mathsf{q})), \operatorname{diag}(\mathbf{M}_i, -\mathbf{M}_i) \big\rangle.$$

Thus, for any $i \in [m]$, we have:

$$\big[\nabla \widetilde{F}(\mathsf{q})\big]_i = -\frac{\big\langle \exp\big(\mu \widetilde{\mathbf{M}}(\mathsf{q})\big), \operatorname{diag}(\mathbf{M}_i, -\mathbf{M}_i) \big\rangle}{\operatorname{Tr}\big(\exp\big(\mu \widetilde{\mathbf{M}}(\mathsf{q})\big)\big)}.$$

In particular, we can write:

$$\begin{aligned}
&\big|\big[\nabla \widetilde{F}(\mathsf{q})\big]_i\big| \\
&= -\frac{x_i^\top [\exp(\mu \mathbf{M}(\mathsf{q})) - \exp(-\mu \mathbf{M}(\mathsf{q}))] x_i}{\operatorname{Tr}(\exp(\mu \mathbf{M}(\mathsf{q}))) + \operatorname{Tr}(\exp(-\mu \mathbf{M}(\mathsf{q})))} \\
&\leq \|x_i\|_2^2 \max_{\|u\|_2^2=1} \left| \frac{u^\top [\exp(\mu \mathbf{M}(\mathsf{q})) - \exp(-\mu \mathbf{M}(\mathsf{q}))] u}{\operatorname{Tr}(\exp(\mu \mathbf{M}(\mathsf{q}))) + \operatorname{Tr}(\exp(-\mu \mathbf{M}(\mathsf{q})))} \right| \\
&\leq \|x_i\|_2^2 \frac{\lambda_{\max}(\exp(\mu \mathbf{M}(\mathsf{q}))) + \lambda_{\max}(\exp(-\mu \mathbf{M}(\mathsf{q})))}{\operatorname{Tr}(\exp(\mu \mathbf{M}(\mathsf{q}))) + \operatorname{Tr}(\exp(-\mu \mathbf{M}(\mathsf{q})))} \\
&\leq \|x_i\|_2^2.
\end{aligned}$$

This completes the proof. □

In the following, we further give explicit proofs of some of these statements.

Proposition 3. *Assume that f is γ-smooth with respect to $\|\cdot\|_2$, then $\widetilde{F}$ is $\gamma\big(\max_{i\in[m]} \|x_i\|_2^4\big)$-smooth with respect to $\|\cdot\|_1$.*

Proof. For any $\mathsf{q}, \mathsf{q}' \in \Delta(m)$, the following upper bound on the spectral norm of $\mathbf{M}(\mathsf{q}) - \mathbf{M}(\mathsf{q}')$ holds:

$$\begin{aligned}
\|\widetilde{\mathbf{M}}(\mathsf{q}) - \widetilde{\mathbf{M}}(\mathsf{q}')\|_2 &= \|\mathrm{diag}(\mathbf{M}(\mathsf{q}) - \mathbf{M}(\mathsf{q}'), -[\mathrm{diag}(\mathbf{M}(\mathsf{q}) - \mathbf{M}(\mathsf{q}'))]\|_2 && (10)\\
&= \|\mathbf{M}(\mathsf{q}) - \mathbf{M}(\mathsf{q}')\|_2 && (11)\\
&= \left\|\sum_{i=1}^m (\mathsf{q}_i - \mathsf{q}'_i) x_i x_i^\top\right\|_2 && (12)\\
&\leq \sum_{i=1}^m |\mathsf{q}_i - \mathsf{q}'_i| \left\|x_i x_i^\top\right\|_2 \\
&\leq \|\mathsf{q} - \mathsf{q}'\|_1 \max_{i\in[m]} \left\|x_i x_i^\top\right\|_2 && \text{(Hölder's ineq.)}\\
&= \|\mathsf{q} - \mathsf{q}'\|_1 \max_{i\in[m]} \|x_i\|_2^2. && (x_i x_i^\top \text{ admits a single non-zero eigenvalue, } \|x_i\|_2^2)
\end{aligned}$$

We have $F(\mathsf{q}) = f(\mathbf{M}(\mathsf{q}))$, thus the gradient of F can be expressed as follows:

$$\nabla F(\mathsf{q}) = -[\langle \nabla f(\mathbf{M}(\mathsf{q})), \mathbf{M}_i\rangle]_{i\in[m]}.$$

Thus, for any $\mathsf{q}, \mathsf{q}' \in \Delta(m)$, we have:

$$\begin{aligned}
\|\nabla \widetilde{F}(\mathsf{q}) - \nabla \widetilde{F}(\mathsf{q}')\|_\infty &= \max_{i\in[m]} \left|\langle \nabla f(\widetilde{\mathbf{M}}(\mathsf{q})) - \nabla f(\widetilde{\mathbf{M}}(\mathsf{q}')), \mathrm{diag}(\mathbf{M}_i, -\mathbf{M}_i)\rangle\right| \\
&\leq \max_{i\in[m]} \left\|\nabla f(\widetilde{\mathbf{M}}(\mathsf{q})) - \nabla f(\widetilde{\mathbf{M}}(\mathsf{q}'))\right\|_{(1)} \|\mathrm{diag}(\mathbf{M}_i, -\mathbf{M}_i)\|_{(\infty)} && \text{(Hölder's ineq.)}\\
&\leq \gamma \max_{i\in[m]} \left\|\widetilde{\mathbf{M}}(\mathsf{q}) - \widetilde{\mathbf{M}}(\mathsf{q}')\right\|_{(\infty)} \|\mathbf{M}_i\|_{(\infty)} && (\gamma\text{-smoothness of } f)\\
&= \gamma \max_{i\in[m]} \left\|\widetilde{\mathbf{M}}(\mathsf{q}) - \widetilde{\mathbf{M}}(\mathsf{q}')\right\|_2 \|\mathbf{M}_i\|_2 && (\text{definition of } \|\cdot\|_{(\infty)})\\
&\leq \gamma \max_{i\in[m]} \left\{\|\mathsf{q} - \mathsf{q}'\|_1 \max_{i\in[m]} \|x_i\|_2^2\right\} \|x_i\|_2^2 && (\text{inequality (10)})\\
&= \gamma \left(\max_{i\in[m]} \|x_i\|_2^4\right) \|\mathsf{q} - \mathsf{q}'\|_1.
\end{aligned}$$

This completes the proof. □

Proof. For $\mathbf{M}(\mathsf{q}))$ and $\mathbf{M}'(\mathsf{q}))$ differing only by point x and x' in $\widehat{\mathcal{P}}_X$, we have:

$$\begin{aligned}
\|\widetilde{\mathbf{M}}(\mathsf{q}) - \widetilde{\mathbf{M}}'(\mathsf{q})\|_2 &= \|\mathbf{M}(\mathsf{q}) - \mathbf{M}'(\mathsf{q})\|_2 && (13)\\
&= \left\|\frac{1}{n}\left[xx^\top - x'x'^\top\right]\right\|_2 \leq \frac{2r^2}{n}. && (14)
\end{aligned}$$

Thus, following the proof of Proposition 3, the sensitivity is bounded by

$$\begin{aligned}
&\max_{i\in[m]} \left|\langle \nabla f(\widetilde{\mathbf{M}}(\mathsf{q})) - \nabla f(\widetilde{\mathbf{M}}'(\mathsf{q})), \mathrm{diag}(\mathbf{M}_i, -\mathbf{M}_i)\rangle\right| \\
&\leq \max_{i\in[m]} \left\|\nabla f(\widetilde{\mathbf{M}}(\mathsf{q})) - \nabla f(\widetilde{\mathbf{M}}'(\mathsf{q}))\right\|_{(1)} \|\mathrm{diag}(\mathbf{M}_i, -\mathbf{M}_i)\|_{(\infty)} && \text{(Hölder's ineq.)}\\
&\leq \mu \max_{i\in[m]} \left\|\widetilde{\mathbf{M}}(\mathsf{q}) - \widetilde{\mathbf{M}}(\mathsf{q}')\right\|_{(\infty)} \|\mathbf{M}_i\|_{(\infty)} && (\mu\text{-smoothness of } f)\\
&= \mu \max_{i\in[m]} \|\mathbf{M}(\mathsf{q}) - \mathbf{M}(\mathsf{q}')\|_2 \|\mathbf{M}_i\|_2 && (\text{definition of } \|\cdot\|_{(\infty)})\\
&\leq \frac{2\mu r^2}{n} \max_{i\in[m]} \|x_i\|_2^2.
\end{aligned}$$

This completes the proof. □

C.3 p-norm approximation

Here, we described an alternative approximation for which we also prove smoothness and gradient sensitivity guarantees. This approximation can be used to design private adaptation algorithms with a relative deviation guarantee that can be more favorable in some contexts, since, as we shall see, the approximation guarantee is modulo a multiplicative term. Unless the softmax approximation, however, here we approximate the square of the norm-2 of the matrix.

A smooth approximation of $\|\mathbf{M}(\mathsf{q})\|_2^2 = \big\|\lambda\big(\mathbf{M}(\mathsf{q})\big)\big\|_\infty^2$ can be defined as follows:

$$G(\mathsf{q}) = \mathrm{Tr}\big[\mathbf{M}(\mathsf{q})^{2p}\big]^{\frac{1}{p}} = \left[\sum_{i=1}^d \lambda_i(\mathbf{M}(\mathsf{q}))^{2p}\right]^{\frac{1}{p}},$$

for p sufficiently large. The following inequalities hold for this approximation:

$$\big\|\lambda\big(\mathbf{M}(\mathsf{q})\big)\big\|_\infty^2 \le G(\mathsf{q}) \le [\mathrm{rank}(\mathbf{M}(\mathsf{q}))]^{\frac{1}{p}} \big\|\lambda\big(\mathbf{M}(\mathsf{q})\big)\big\|_\infty^2.$$

The gradient of the smooth approximation is given for all $i \in [1, m]$ by:

$$[\nabla G(\mathbf{M}(\mathsf{q}))]_i = -2\big\langle \mathbf{M}^{2p-1}(\mathsf{q}), \mathbf{M}_i\big\rangle \,\mathrm{Tr}[\mathbf{M}^{2p}(\mathsf{q})]^{\frac{1}{p}-1}. \tag{15}$$

We can write $G(\mathsf{q}) = g(\mathbf{M}(\mathsf{q}))$ where g is defined for all $\mathbf{M} \in \mathbb{S}_d$ by

$$g(\mathbf{M}) = \mathrm{Tr}\big[\mathbf{M}^{2p}\big]^{\frac{1}{p}} = \big\langle \mathbf{M}^{2p}, \mathbf{I}\big\rangle^{\frac{1}{p}}.$$

The following result provides the desired smoothness result needed for G, which we prove by using the smoothness property of g.

Theorem 8. *The p-norm approximation function G is $(2p-1)\big(\max_{i\in[m]}\|x_i\|_2^4\big)$-smooth for $\|\cdot\|_1$.*

The proof is given in Appendix C.3. Next, we present a sensitivity bounds for G.

Theorem 9. *Assume that the support of $\mathcal{P}$ is included in the ℓ_2 ball of radius r. Then, the gradient of the p-norm approximation G is $\frac{2(2p-1)r^2}{n}\max_{i\in[m]}\|x_i\|_2^2$-sensitive.*

The proof is given in Appendix C.3.

Proposition 4. *Assume that g is γ-smooth with respect to the norm $\|\cdot\|_{(2p)}$:*

$$\forall \mathbf{M}, \mathbf{M}' \in \mathbb{S}_d, \quad \|\nabla g(\mathbf{M}) - \nabla g(\mathbf{M}')\|_{(r)} \le \gamma \|\nabla g(\mathbf{M}) - \nabla g(\mathbf{M}')\|_{(2p)},$$

with $\frac{1}{r} + \frac{1}{2p} = 1$. Then, G is $\gamma\big(\max_{i\in[m]}\|x_i\|_2^4\big)$-smooth:

$$\forall \mathsf{q}, \mathsf{q}' \in \mathbb{R}^d, \quad \|\nabla G(\mathsf{q}) - \nabla G(\mathsf{q}')\|_\infty \le \gamma\left(\max_{i\in[m]}\|x_i\|_2^4\right)\|\mathsf{q} - \mathsf{q}'\|_1.$$

Proof. The proof is similar to that of Proposition 1. For any $\mathsf{q}, \mathsf{q}' \in \Delta(m)$, the following upper bound on the norm-$(2p)$ of $\mathbf{M}(\mathsf{q}) - \mathbf{M}(\mathsf{q}')$ holds:

$$\begin{aligned}
\|\mathbf{M}(\mathsf{q}) - \mathbf{M}(\mathsf{q}')\|_{(2p)} &= \left\|\sum_{i=1}^m (\mathsf{q}_i - \mathsf{q}'_i) x_i x_i^\top\right\|_{(2p)} && (16)\\
&\le \sum_{i=1}^m |\mathsf{q}_i - \mathsf{q}'_i| \left\|x_i x_i^\top\right\|_{(2p)} \\
&\le \|\mathsf{q} - \mathsf{q}'\|_1 \max_{i\in[m]} \left\|x_i x_i^\top\right\|_{(2p)} && \text{(Hölder's ineq.)}\\
&= \|\mathsf{q} - \mathsf{q}'\|_1 \max_{i\in[m]} \|x_i\|_2^2. && (x_i x_i^\top \text{ admits a single non-zero eigenvalue, } \|x_i\|_2^2)
\end{aligned}$$

We have $G(\mathsf{q}) = g(\mathbf{M}(\mathsf{q}))$, thus the gradient of G can be expressed as follows:

$$\nabla G(\mathsf{q}) = -[\langle \nabla g(\mathbf{M}(\mathsf{q})), \mathbf{M}_i\rangle]_{i\in[m]}.$$

Thus, for any $\mathsf{q}, \mathsf{q}' \in \Delta(m)$, we have:

$$\begin{aligned}
\|\nabla G(\mathbf{M}(\mathsf{q})) - \nabla G(\mathbf{M}(\mathsf{q}'))\|_\infty &= \max_{i\in[m]} |\langle \nabla g(\mathbf{M}(\mathsf{q})) - \nabla g(\mathbf{M}(\mathsf{q}')), \mathbf{M}_i\rangle| \\
&\le \max_{i\in[m]} \|\nabla g(\mathbf{M}(\mathsf{q})) - \nabla g(\mathbf{M}(\mathsf{q}'))\|_{(r)} \|\mathbf{M}_i\|_{(2p)} && \text{(Hölder's ineq.)} \\
&\le \gamma \max_{i\in[m]} \|\mathbf{M}(\mathsf{q}) - \mathbf{M}(\mathsf{q}')\|_{(2p)} \|\mathbf{M}_i\|_{(2p)} && (\gamma\text{-smoothness of } f) \\
&\le \gamma \max_{i\in[m]} \left\{ \|\mathsf{q} - \mathsf{q}'\|_1 \max_{i\in[m]} \|x_i\|_2^2 \right\} \|x_i\|_2^2 && \text{(inequality (16))} \\
&= \gamma \left(\max_{i\in[m]} \|x_i\|_2^4 \right) \|\mathsf{q} - \mathsf{q}'\|_1 .
\end{aligned}$$

This completes the proof. □

We will use the following bound for the Hessian of g.

Lemma 3 ([Nes07])**.** *The following upper bound holds for the Hessian of f for any two symmetric matrices $\mathbf{M}, \mathbf{U} \in \mathbb{S}_d$:*

$$\left\langle \nabla^2 g(\mathbf{M})\, \mathbf{U}, \mathbf{U} \right\rangle \le (2p-1) \|\lambda(\mathbf{U})\|_{2p}^2,$$

where $\|\lambda(\mathbf{U})\|_{2p}^2 = \left(\mathrm{Tr}\left[\mathbf{U}^{2p}\right]\right)^{\frac{1}{p}}$.

Theorem 8. *The p-norm approximation function G is $(2p-1)\left(\max_{i\in[m]} \|x_i\|_2^4\right)$-smooth for $\|\cdot\|_1$.*

Proof. In view of Lemma 3, g is $\|\cdot\|_{(2p)}$-$(2p-1)$-smooth. The result thus follows by Proposition 4. □

Theorem 9. *Assume that the support of $\mathcal{P}$ is included in the ℓ_2 ball of radius r. Then, the gradient of the p-norm approximation G is $\frac{2(2p-1)r^2}{n} \max_{i\in[m]} \|x_i\|_2^2$-sensitive.*

Proof. For $\mathbf{M}(\mathsf{q}))$ and $\mathbf{M}'(\mathsf{q}))$ differing only by point x and x' in $\widehat{\mathcal{P}}_{\mathcal{X}}$, we have:

$$\|\mathbf{M}(\mathsf{q}) - \mathbf{M}'(\mathsf{q})\|_2 = \left\| \frac{1}{n} \left[x x^\top - x' x'^\top \right] \right\|_2 \le \frac{2r^2}{n}. \tag{17}$$

Thus, following the proof of Proposition 4, the sensitivity is bounded by

$$\begin{aligned}
\max_{i\in[m]} |\langle \nabla g(\mathbf{M}(\mathsf{q})) - \nabla g(\mathbf{M}'(\mathsf{q})), \mathbf{M}_i\rangle| &\le \max_{i\in[m]} \|\nabla g(\mathbf{M}(\mathsf{q})) - \nabla g(\mathbf{M}'(\mathsf{q}))\|_{(r)} \|\mathbf{M}_i\|_{(2p)} && \text{(Hölder's ineq.)} \\
&\le (2p-1) \max_{i\in[m]} \|\mathbf{M}(\mathsf{q}) - \mathbf{M}(\mathsf{q}')\|_{(2p)} \|\mathbf{M}_i\|_{(2p)} && (\gamma\text{-smoothness of } f) \\
&\le \frac{2(2p-1)r^2}{n} \max_{i\in[m]} \|x_i\|_2^2. && \text{(inequality (17))}
\end{aligned}$$

This completes the proof. □

Proposition 5. *The following inequality holds for the spectral norm of the Hessian of F:*

$$\|\nabla^2 G\|_2 \le (2p-1) \left[\sum_{i=1}^m \|x_i\|_2^2 \right].$$

Proof. As in the proof of Proposition 2, we have:

$$\frac{\partial^2 G}{\partial \mathsf{q}_i \partial \mathsf{q}_j} = -\left\langle \frac{\partial}{\partial \mathsf{q}_j} \nabla f(\mathbf{M}(\mathsf{q})), \mathbf{M}_i \right\rangle = +\langle \nabla^2 g(\mathbf{M}(\mathsf{q}))\, \mathbf{M}_j, \mathbf{M}_i \rangle.$$

Thus, using the shorthand $\overline{\mathbf{M}} = \sum_{i=1}^{m} \mathbf{X}_i \mathbf{M}_i$, for any $\mathbf{X} \in \mathbb{R}^m$, we can write:

$$
\begin{aligned}
\mathbf{X}^\top \nabla^2 G \mathbf{X} &= \sum_{i,j=1}^{d} \mathbf{X}_i \mathbf{X}_j \left\langle \nabla^2 g(\mathbf{M}(\mathsf{q}))\, \mathbf{M}_j, \mathbf{M}_i \right\rangle \\
&= \left\langle \nabla^2 f(\mathbf{M}(\mathsf{q})) \left(\sum_{j=1}^{d} \mathbf{X}_j \mathbf{M}_j \right), \left(\sum_{i=1}^{d} \mathbf{X}_i \mathbf{M}_i \right) \right\rangle \\
&= \left\langle \nabla^2 f(\mathbf{M}(\mathsf{q})) \left(\overline{\mathbf{M}} \right), \left(\overline{\mathbf{M}} \right) \right\rangle \\
&\leq (2p-1) \|\overline{\mathbf{M}}\|_{(2p)}^2 && \text{(Lemma 3)} \\
&= (2p-1) \left(\left\| \sum_{i=1}^{m} \mathbf{X}_i x_i x_i^\top \right\|_{(2p)} \right)^2 \\
&= (2p-1) \left(\mathrm{Tr}\left(\left[\sum_{i=1}^{m} \mathbf{X}_i x_i x_i^\top \right]^{2p} \right) \right)^{\frac{1}{p}} \\
&\leq (2p-1) \left(\sum_{i=1}^{m} |\mathbf{X}_i| \left\| x_i x_i^\top \right\|_{(2p)} \right)^2 \\
&\leq (2p-1) \|\mathbf{X}\|_2^2 \sum_{i=1}^{m} \left\| x_i x_i^\top \right\|_{(2p)}^2 \\
&= (2p-1) \|\mathbf{X}\|_2^2 \sum_{i=1}^{m} \|x_i\|_2^2 .
\end{aligned}
$$

This completes the proof.

□

D Proofs of Section 5

D.1 Proof of Theorem 4

Theorem 4. *Algorithm 1 is (ε, δ)-differentially private. Suppose $f_T : \mathcal{Q} \times \mathcal{W} \to \mathbb{R}$ is $((\gamma_{\mathsf{q}}, \|\cdot\|_1), (\gamma_w, \|\cdot\|_2))$-Lipschitz and $((\mu_{\mathsf{q}}, \|\cdot\|_1), (\mu_w, \|\cdot\|_2))$-smooth. Assume further that for all $\mathsf{q} \in \mathcal{Q}$, and $w, w' \in \mathcal{W}$, $\|\nabla_{\mathsf{q}} f_T(\mathsf{q}, w) - \nabla_{\mathsf{q}} f_T(\mathsf{q}, w')\|_\infty \le \gamma_{\mathsf{q},w} \|w - w'\|_2$. Then, for any $\beta \in (0,1)$, there exists a choice of K and μ such that, with probability at least $1-\beta$, the stationarity gap of the output $\widehat{w}$, $\mathrm{Gap}_{f_T}(\widehat{\mathsf{q}}, \widehat{w})$, is upper bounded by*

$$5\sqrt{\bar{D}\left(\sigma_{\mathsf{q}}^0 \log\left(\frac{\bar{D}J}{\sigma_{\mathsf{q}}^0 \beta}\right) + D_w \sigma_w^0 \sqrt{d \log\left(\frac{\bar{D}}{D_w \sigma_w^0 \beta}\right)}\right)},$$

where

$$\bar{D} = \sqrt{(D_{\mathsf{q}}\gamma_{\mathsf{q}} + D_w\gamma_w)(D_{\mathsf{q}}^2\mu_{\mathsf{q}} + D_w^2\mu_w + 2\gamma_{\mathsf{q},w} D_{\mathsf{q}} D_w)},$$

$\sigma_{\mathsf{q}}^0 = \frac{\sigma_{\mathsf{q}}}{\sqrt{K}}$, *and* $\sigma_w^0 = \frac{\sigma_w}{\sqrt{K}}$ *(where $\sigma_{\mathsf{q}}, \sigma_w$ are as given in steps 1 and 2 of Algorithm 1).*

The statement holds with the following choice of K and μ:

$$K = \frac{\sqrt{2}\bar{D}}{\sigma_{\mathsf{q}}^0 \log\left(\frac{\bar{D}J}{\sigma_{\mathsf{q}}^0\beta}\right) + D_w\sigma_w^0\sqrt{d\log\left(\frac{\bar{D}}{D_w\sigma_w^0\beta}\right)}} \qquad \eta = \sqrt{\frac{2(D_{\mathsf{q}}\gamma_{\mathsf{q}} + D_w\gamma_w)}{(D_{\mathsf{q}}^2\mu_{\mathsf{q}} + D_w^2\mu_w + 2\gamma_{\mathsf{q},w}D_{\mathsf{q}}D_w)K}},$$

where $\bar{D} = \sqrt{(D_{\mathsf{q}}\gamma_{\mathsf{q}} + D_w\gamma_w)(D_{\mathsf{q}}^2\mu_{\mathsf{q}} + D_w^2\mu_w + 2\gamma_{\mathsf{q},w}D_{\mathsf{q}}D_w)}$, $\sigma_{\mathsf{q}}^0 = \frac{\sigma_{\mathsf{q}}}{\sqrt{K}}$, and $\sigma_w^0 = \frac{\sigma_w}{\sqrt{K}}$ (where $\sigma_{\mathsf{q}}, \sigma_w$ are as given in steps 1 and 2).

Proof. The privacy proof follows by combining the guarantees of the Report-Noisy-Min mechanism (steps 1, 7, and 8) and the Gaussian mechanism (steps 2 and 11) together with the application of the advanced composition theorem of differential privacy over the K rounds of the algorithm.

We now prove the convergence (stationarity gap) guarantee. By the smoothness of f_T, we have

$$\begin{aligned}
&f_T(\mathsf{q}^{k+1}, w^{k+1}) \\
&\le f_T(\mathsf{q}^k, w^{k+1}) + \langle \nabla_{\mathsf{q}}^k, \mathsf{q}^{k+1} - \mathsf{q}^k \rangle + \langle \nabla_{\mathsf{q}} f_T(\mathsf{q}^k, w^{k+1}) - \nabla_{\mathsf{q}} f_T(\mathsf{q}^k, w^k), \mathsf{q}^{k+1} - \mathsf{q}^k \rangle + \frac{\mu_{\mathsf{q}}}{2}\|\mathsf{q}^{k+1} - \mathsf{q}^k\|_1^2 \\
&\le f_T(\mathsf{q}^k, w^{k+1}) + \langle \nabla_{\mathsf{q}}^k, \mathsf{q}^{k+1} - \mathsf{q}^k \rangle + \gamma_{\mathsf{q},w}\|w^{k+1} - w^k\|_2 \|\mathsf{q}^{k+1} - \mathsf{q}^k\|_1 + \frac{\mu_{\mathsf{q}}}{2}\|\mathsf{q}^{k+1} - \mathsf{q}^k\|_1^2 \\
&\le f_T(\mathsf{q}^k, w^k) + \langle \nabla_{\mathsf{q}}^k, \mathsf{q}^{k+1} - \mathsf{q}^k \rangle + \langle \nabla_w^k, w^{k+1} - w^k \rangle + \gamma_{\mathsf{q},w}\|w^{k+1} - w^k\|_2 \|\mathsf{q}^{k+1} - \mathsf{q}^k\|_1 + \frac{\mu_{\mathsf{q}}}{2}\|\mathsf{q}^{k+1} - \mathsf{q}^k\|_1^2 \\
&\quad + \frac{\mu_w}{2}\|w^{k+1} - w^k\|_2^2 \\
&\le f_T(\mathsf{q}^k, w^k) + \eta\langle \nabla_{\mathsf{q}}^k, v_{\mathsf{q}}^k - \mathsf{q}^k \rangle + \eta\langle \nabla_w^k, u_w^k - w^k \rangle + \gamma_{\mathsf{q},w}\eta^2 D_{\mathsf{q}} D_w + \frac{\eta^2\mu_{\mathsf{q}}D_{\mathsf{q}}^2}{2} + \frac{\eta^2\mu_w D_w^2}{2} \\
&\le f_T(\mathsf{q}^k, w^k) + \eta\langle \nabla_{\mathsf{q}}^k, v_{\mathsf{q}}^k - \mathsf{q}^k \rangle + \eta\langle \widehat{\nabla}_w^k, u_w^k - w^k \rangle + \eta\langle \nabla_w^k - \widehat{\nabla}_w^k, u_w^k - w^k \rangle + \gamma_{\mathsf{q},w}\eta^2 D_{\mathsf{q}} D_w + \frac{\eta^2\mu_{\mathsf{q}}D_{\mathsf{q}}^2}{2} + \frac{\eta^2\mu_w D_w^2}{2}
\end{aligned}$$

Define $v_*^k \triangleq \operatorname{argmin}_{v \in \mathcal{V}} \langle \nabla_{\mathsf{q}}^k, v \rangle$ and $\alpha^k \triangleq \langle \nabla_{\mathsf{q}}^k, v_{\mathsf{q}}^k - v_*^k \rangle$. Also, define $u_*^k \triangleq \operatorname{argmin}_{u \in \mathcal{W}} \langle \nabla_w^k, u \rangle$. Hence, noting that $\langle \widehat{\nabla}_w^k, u_w^k - w^k \rangle \le \langle \widehat{\nabla}_w^k, u_*^k - w^k \rangle$ (which follows from the definition of u_w^k in Step 12 in Algorithm 1), the bound on $f_T(\mathsf{q}^{k+1}, w^{k+1})$ above

can be further upper bounded as

$$\begin{aligned}
&f_T(\mathsf{q}^k, w^k) + \eta\langle \nabla^k_{\mathsf{q}}, v^k_{\mathsf{q}} - \mathsf{q}^k\rangle + \eta\Big[\langle \widehat{\nabla}^k_w, u^k_* - w^k\rangle + \langle \nabla^k_w - \widehat{\nabla}^k_w, u^k_w - w^k\rangle\Big] + \gamma_{\mathsf{q},w}\eta^2 D_{\mathsf{q}} D_w \\
&\quad + \frac{\eta^2\mu_{\mathsf{q}} D^2_{\mathsf{q}}}{2} + \frac{\eta^2\mu_w D^2_w}{2} \\
\leq &f_T(\mathsf{q}^k, w^k) + \eta\langle \nabla^k_{\mathsf{q}}, v^k_{\mathsf{q}} - \mathsf{q}^k\rangle + \eta\Big[\langle \nabla^k_w, u^k_* - w^k\rangle + \langle \widehat{\nabla}^k_w - \nabla^k_w, u^k_* - w^k\rangle + \langle \nabla^k_w - \widehat{\nabla}^k_w, u^k_w - w^k\rangle\Big] \\
&\quad + \gamma_{\mathsf{q},w}\eta^2 D_{\mathsf{q}} D_w + \frac{\eta^2\mu_{\mathsf{q}} D^2_{\mathsf{q}}}{2} + \frac{\eta^2\mu_w D^2_w}{2} \\
\leq &f_T(\mathsf{q}^k, w^k) + \eta\langle \nabla^k_{\mathsf{q}}, v^k_{\mathsf{q}} - \mathsf{q}^k\rangle + + \eta\Big[\langle \nabla^k_w, u^k_* - w^k\rangle + \langle \widehat{\nabla}^k_w - \nabla^k_w, u^k_* - u^k_w\rangle\Big] + \gamma_{\mathsf{q},w}\eta^2 D_{\mathsf{q}} D_w \\
&\quad + \frac{\eta^2\mu_{\mathsf{q}} D^2_{\mathsf{q}}}{2} + \frac{\eta^2\mu_w D^2_w}{2} \\
\leq &f_T(\mathsf{q}^k, w^k) + \eta\Big[\langle \nabla^k_{\mathsf{q}}, v^k_* - \mathsf{q}^k\rangle + \alpha^k\Big] + \eta\Big[\langle \nabla^k_w, u^k_* - w^k\rangle + \langle \widehat{\nabla}^k_w - \nabla^k_w, u^k_* - u^k_w\rangle\Big] + \gamma_{\mathsf{q},w}\eta^2 D_{\mathsf{q}} D_w \\
&\quad + \frac{\eta^2\mu_{\mathsf{q}} D^2_{\mathsf{q}}}{2} + \frac{\eta^2\mu_w D^2_w}{2} \\
\leq &f_T(\mathsf{q}^k, w^k) + \eta\Big[\langle \nabla^k_{\mathsf{q}}, v^k_* - \mathsf{q}^k\rangle + \langle \nabla^k_w, u^k_* - w^k\rangle\Big] + \eta\alpha^k + \eta D_w \|\widehat{\nabla}^k_w - \nabla^k_w\|_2 + \gamma_{\mathsf{q},w}\eta^2 D_{\mathsf{q}} D_w \\
&\quad + \frac{\eta^2\mu_{\mathsf{q}} D^2_{\mathsf{q}}}{2} + \frac{\eta^2\mu_w D^2_w}{2}
\end{aligned}$$

Note $\langle \nabla^k_{\mathsf{q}}, v^k_* - \mathsf{q}^k\rangle + \langle \nabla^k_w, u^k_* - w^k\rangle = -\operatorname{Gap}_{f_T}(\mathsf{q}^k, w^k)$. Moreover, with standard bounds on then tail of Laplacian and Gaussian random variables, with probability at least $1-\beta$, for all $k \in [K]$, $\alpha^k \leq \sigma_{\mathsf{q}} \log(2JK/\beta)$ and $\|\widehat{\nabla}^k_w - \nabla^k_w\|_2 \leq \sigma_w\sqrt{d\log(2K/\beta)}$. We will condition on this event for the rest of the proof. Hence, the bound becomes:

$$f_T(\mathsf{q}^k, w^k) - \eta \operatorname{Gap}_{f_T}(\mathsf{q}^k, w^k) + \eta\sigma_{\mathsf{q}} \log\left(\frac{2JK}{\beta}\right) + \eta D_w \sigma_w \sqrt{d\log\left(\frac{2K}{\beta}\right)} + \gamma_{\mathsf{q},w}\eta^2 D_{\mathsf{q}} D_w + \frac{\eta^2\mu_{\mathsf{q}} D^2_{\mathsf{q}}}{2} + \frac{\eta^2\mu_w D^2_w}{2}$$

Rearranging terms, and then averaging over $k \in [K]$, we get

$$\begin{aligned}
\frac{1}{K}\sum_{k=1}^K \operatorname{Gap}_{f_T}(\mathsf{q}^k, w^k) &\leq \frac{f_T(\mathsf{q}^0, w^0) - f_T(\mathsf{q}^{K+1}, w^{K+1})}{\eta K} + \eta\left[\gamma_{\mathsf{q},w} D_{\mathsf{q}} D_w + \frac{\mu_{\mathsf{q}} D^2_{\mathsf{q}}}{2} + \frac{\mu_w D^2_w}{2}\right] \qquad (18)\\
&\quad + \sigma_{\mathsf{q}} \log\left(\frac{2JK}{\beta}\right) + D_w\sigma_w\sqrt{d\log\left(\frac{2K}{\beta}\right)} \\
&\leq \frac{D_{\mathsf{q}}\gamma_{\mathsf{q}} + D_w\gamma_w}{\eta K} + \eta\left[\gamma_{\mathsf{q},w} D_{\mathsf{q}} D_w + \frac{\mu_{\mathsf{q}} D^2_{\mathsf{q}} + \mu_w D^2_w}{2}\right] \qquad (19)\\
&\quad + \sqrt{K}\left[\sigma^0_{\mathsf{q}} \log\left(\frac{2JK}{\beta}\right) + D_w\sigma^0_w\sqrt{d\log\left(\frac{2K}{\beta}\right)}\right]. \qquad (20)
\end{aligned}$$

Optimizing this bound in η and K results in the settings of K and η in the theorem statement. Substituting with these settings and simplifying, we get that the average gap is upper bounded by $\mathsf{A}/2$, where A is the bound in the theorem; namely, $\mathsf{A} = 5\sqrt{\bar{D}\left(\sigma^0_{\mathsf{q}} \log\left(\frac{\bar{D}J}{\sigma^0_{\mathsf{q}}\beta}\right) + D_w\sigma^0_w\sqrt{d\log\left(\frac{\bar{D}}{D_w\sigma^0_w\beta}\right)}\right)}$.

Now, to conclude the proof, we show that $\operatorname{Gap}_{f_T}(\widehat{\mathsf{q}}, \widehat{w}) \leq \frac{1}{K}\sum_{k=1}^K \operatorname{Gap}_{f_T}(\mathsf{q}^k, w^k) + \mathsf{A}/2$. By the definition of $\widehat{\mathsf{q}}, \widehat{w}$ and using a similar analysis as above (and using the tail bounds on the Gaussian and Laplace r.v.s as before), observe that $\operatorname{Gap}_{f_T}(\widehat{\mathsf{q}}, \widehat{w})$ can be upper bounded as

$$\begin{aligned}
\operatorname{Gap}_{f_T}(\widehat{\mathsf{q}}, \widehat{w}) &= \operatorname{Gap}_{f_T}(\mathsf{q}^{k^*}, w^{k^*}) \\
&\leq \min_{k\in[K]} \operatorname{Gap}_{f_T}(\mathsf{q}^k, w^k) + \sigma_{\mathsf{q}} \log\left(\frac{2JK}{\beta}\right) + D_w\sigma_w\sqrt{d\log\left(\frac{2K}{\beta}\right)} \\
&\leq \frac{1}{K}\sum_{k=1}^K \operatorname{Gap}_{f_T}(\mathsf{q}^k, w^k) + \sqrt{K}\left[\sigma^0_{\mathsf{q}} \log\left(\frac{2JK}{\beta}\right) + D_w\sigma^0_w\sqrt{d\log\left(\frac{2K}{\beta}\right)}\right].
\end{aligned}$$

Observe that the term $\sqrt{K}\left[\sigma_{\mathsf{q}}^0 \log(\frac{2JK}{\beta}) + D_w \sigma_w^0 \sqrt{d \log(\frac{2K}{\beta})}\right]$ above is the last term in (20). Hence, by substituting with the values of K and η, we can show that this term is upper bounded by $\mathsf{A}/2$. This leads to the following:

$$\mathrm{Gap}_{f_T}(\widehat{\mathsf{q}}, \widehat{w}) \leq \frac{1}{K} \sum_{k=1}^{K} \mathrm{Gap}_{f_T}(\mathsf{q}^k, w^k) + \mathsf{A}/2,$$

which completes the proof. □

D.2 Proof of Corollary 2

Corollary 2. *Let $L_T(\mathsf{q}, w) = \sum_{i=1}^m \mathsf{q}_i(\langle w, x_i\rangle - y_i)^2 + 4\Lambda^2 \widetilde{F}_T(\mathsf{q})$ be the input objective to Algorithm 1. Let $\beta \in (0,1)$. Then, there exists a choice of K and η such that, with probability at least $1-\beta$, the output of the algorithm is an approximate stationary point of L_T with stationarity gap upper bounded as follows:*

$$\mathrm{Gap}_{L_T}(\widehat{\mathsf{q}}, \widehat{w}) \leq \widetilde{O}\left(\frac{\mu^{3/4}}{\sqrt{\varepsilon n}}\right).$$

Here, $\widetilde{O}(\cdot)$ hides poly-logarithmic factors in m.

Given Theorem 4, the stationarity gap is more precisely bounded as

$$\mathrm{Gap}_{L_T}(\widehat{\mathsf{q}}, \widehat{w}) \leq \frac{32(1+2\mu\hat{r}^2)^{\frac{1}{4}}(\Lambda\hat{r})^{\frac{3}{2}} r \sqrt{(\Lambda\hat{r}+Y)\mu \log(\frac{mn}{\beta})} \log^{\frac{1}{4}}(1/\delta)}{\sqrt{\varepsilon n}},$$

when we choose K and η as follows:

$$K = \frac{\varepsilon n(\Lambda\hat{r}+Y)\sqrt{1+2\mu\hat{r}^2}}{4\Lambda\hat{r}r^2\mu \log\left(\frac{mn}{\beta}\right)\sqrt{\log(\frac{1}{\delta})}} \qquad \eta = \frac{\sqrt{2}(\Lambda\hat{r}+Y)}{\Lambda\hat{r}\sqrt{(1+2\mu\hat{r}^2)K}}.$$

Thus, the learning bound (1) implies that with probability $\geq 1-2\beta$ over the choice of the public and private datasets and the algorithm's internal randomness, the expected loss of the predictor $h_{\widehat{w}}$ (defined by the output $\widehat{w}$) with respect to the target domain is bounded as follows:

$$\mathcal{L}(\mathcal{P}, h_{\widehat{w}}) \leq L_T(\widehat{\mathsf{q}}, \widehat{w}) + \frac{2\log(m+n)}{\mu} + \frac{2\Lambda r(\Lambda r+Y)^2}{\sqrt{n}} + (\Lambda r+Y)^2\sqrt{\frac{\log\frac{1}{\beta}}{2n}} + \eta_{\mathcal{H}}(S, \widetilde{T}).$$

E Proofs of Section 6

E.1 Proof of Theorem 5

Theorem 5. *The Noisy Frank-Wolfe algorithm (Algorithm 2) is (ε, δ)-differentially private. Let $\mathsf{q}^* \in \operatorname{argmin}_{\mathsf{q}\in\Delta_m} \operatorname{dis}(\widehat{P}, \mathsf{q})$. Then, there exists a choice of the parameters of Algorithm 2 such that, with high probability over the algorithm's internal randomness, the output $\widehat{\mathsf{q}}$ satisfies*

$$\operatorname{dis}(\widehat{P}, \widehat{\mathsf{q}}) \leq \operatorname{dis}(\widehat{P}, \mathsf{q}^*) + \frac{\lambda}{2}\|\mathsf{q}^*\|_2^2 + \widetilde{O}\left(\frac{1}{(\varepsilon n)^{1/3}}\right),$$

where $\widetilde{O}(\cdot)$ is hiding a poly-logarithmic factor in m.

The above theorem follows as a corollary of the following theorem.

Theorem 10. *Algorithm 2 is (ε, δ)-differentially private. Let $\beta \in (0,1)$. With probability $1-\beta$ over the algorithm's randomness (the Laplace noise), the output $\widehat{\mathsf{q}}$ satisfies*

$$\widetilde{F}_T^{\lambda}(\widehat{\mathsf{q}}) \leq \min_{\mathsf{q}\in\Delta_m} \widetilde{F}_T^{\lambda}(\mathsf{q}) + \frac{2(\mu\hat{r}^4 + \lambda)}{K} + \frac{8\mu r^2\hat{r}^2\sqrt{2K\log(\frac{1}{\delta})}\log(K)\log(\frac{mK}{\beta})}{\varepsilon n}.$$

The proof relies on the smoothness property of $\widetilde{F}_T^{\lambda}$ and the sensitivity bound on $\nabla\widetilde{F}_T(\mathsf{q})$. Using the approximation guarantee of $\widetilde{F}_T$ given in Corollary 1 together with Theorem 10 above, we reach the result of Theorem 5. The discrepancy guarantee in Theorem 5 can be more precisely stated as the following corollary of Theorem 10.

Corollary 3. *Let $\mathsf{q}^* \in \operatorname*{argmin}_{\mathsf{q}\in\Delta_m} \operatorname{dis}(\widehat{P}, \mathsf{q})$. Let $\beta \in (0,1)$. There exists a choice of K and μ in Algorithm 2 for which the following holds: assuming w.l.o.g. that $\lambda \leq \mu\hat{r}^4$, with probability at least $1-\beta$, the output $\widehat{\mathsf{q}}$ satisfies*

$$\operatorname{dis}(\widehat{P}, \widehat{\mathsf{q}}) \leq \operatorname{dis}(\widehat{P}, \mathsf{q}^*) + \frac{\lambda}{2}\|\mathsf{q}^*\|_2^2 + \widetilde{O}\left(\frac{\Lambda^4\hat{r}^{4/3}r^{2/3}}{(\varepsilon n)^{1/3}}\right),$$

where Λ is the $\|\cdot\|_2$-bound on the predictors in $\mathcal{H}$.

Proof of Theorem 10 For the privacy guarantee of Algorithm 2, first note that the global $\|\cdot\|_\infty$-sensitivity of $\nabla\widetilde{F}_T^{\lambda}$ (with respect to replacing any data point in the private dataset) is the same as that of $\nabla\widetilde{F}_T$, which is bounded by $\frac{2\mu r^2\hat{r}^2}{n}$ as established in Corollary 1 (Part 3). Hence, by the setting of the scale of the Laplace noise and the privacy guarantee of the Report-Noisy-Max mechanism [DR14, BLST10], it follows that a single iteration of Algorithm 2 is $\left(\frac{\varepsilon}{\sqrt{8K\log(\frac{1}{\delta})}}, 0\right)$-differentially private. The advanced composition theorem of differential privacy [DR14] thus implies that the algorithm is (ε, δ)-differentially private.

We now prove the convergence guarantee. Let $\tilde{\mathsf{q}} \in \operatorname*{argmin}_{\mathsf{q}\in\Delta_m} S_\lambda(\mathsf{q})$. First, by Corollary 1 (Part 2), $\widetilde{F}_T$ is $\mu\hat{r}^4$-smooth with respect to $\|\cdot\|_1$. Note also that $\frac{\lambda}{2}\|\mathsf{q}\|_2^2$ is λ-smooth over $\mathsf{q} \in \Delta_m$ with respect to $\|\cdot\|_1$. This follows from the fact that for any $\mathsf{q}, \mathsf{q}' \in \Delta_m$,

$$\|\nabla\left(\frac{\lambda}{2}\|\mathsf{q}\|_2^2\right) - \nabla\left(\frac{\lambda}{2}\|\mathsf{q}'\|_2^2\right)\|_\infty = \lambda\|\mathsf{q} - \mathsf{q}'\|_\infty \leq \lambda\|\mathsf{q} - \mathsf{q}'\|_1.$$

Hence, we get that the objective $\widetilde{F}_T^{\lambda}$ is $(\mu\hat{r}^4 + \lambda)$-smooth with respect to $\|\cdot\|_1$ over Δ_m. Thus, by standard analysis of the Noisy Frank-Wolfe algorithm (see, e.g., [TGTZ15, BGM21]), we have

$$\widetilde{F}_T^{\lambda}(\widehat{\mathsf{q}}) - \widetilde{F}_T^{\lambda}(\tilde{\mathsf{q}}) \leq \frac{2(\mu\hat{r}^4 + \lambda)}{K} + \sum_{k=1}^{K} \eta_k\alpha_k,$$

where $\alpha_k \triangleq \langle\nabla\widetilde{F}_T^{\lambda}(\mathsf{q}_k), \mathbf{e}_{j_k}\rangle - \min_{i\in[m]}\langle\nabla\widetilde{F}_T^{\lambda}(\mathsf{q}_k), \mathbf{e}_i\rangle$. By the tail properties of the Laplace distribution together with the union bound, we get that with probability $\geq 1-\beta$, for all $k \in [K]$, $\alpha_k \leq \sigma\log(Km/\beta) = \frac{4\mu r^2\hat{r}^2\sqrt{2K\log(\frac{1}{\delta})}\log(Km/\beta)}{n\varepsilon}$. Hence,

given the setting of η_k, with probability $\geq 1-\beta$, the above bound simplifies to

$$\widetilde{F}_T^{\lambda}(\widehat{\mathsf{q}}) - \widetilde{F}_T^{\lambda}(\tilde{\mathsf{q}}) \leq \frac{2(\mu\hat{r}^4+\lambda)}{K} + \frac{8\mu r^2\hat{r}^2\sqrt{2K\log(\frac{1}{\delta})}\log(K)\log(Km/\beta)}{n\varepsilon},$$

which completes the proof.

Proof of Corollary 3. The result can be obtained with the following choices of K and μ:

$$K = \frac{\hat{r}^{4/3}(\varepsilon n)^{2/3}}{3r^{4/3}\log^{1/3}(\frac{1}{\delta})\log^{2/3}(n)\log^{2/3}(\frac{mn}{\beta})} \qquad \mu = \sqrt{\frac{K\log(m+n)}{8\hat{r}^4}}.$$

E.2 Proof of Theorem 6

Next, we give an alternative private algorithm for minimizing the regularized smooth approximation of the discrepancy, $\widetilde{F}_T^{\lambda}$. Compared to the guarantees of the private Frank-Wolfe algorithm, the optimization error of this algorithm exhibits a better dependence on n at the expense of worse dependence on m. In particular, the excess error with respect to the minimum discrepancy scales as $\widetilde{O}\Big(\frac{m^{1/4}}{\sqrt{n}}\Big)$ (see Corollary 4). When $m = \widetilde{O}\big(n^{2/3}\big)$, Algorithm 3 benefits from more favorable generalization error guarantees than Algorithm 2.

Theorem 6. *The Noisy Mirror Descent algorithm (Algorithm 3) is (ε,δ)-differentially private. Let $\mathsf{q}^* \in \operatorname{argmin}_{\mathsf{q}\in\Delta_m} \operatorname{dis}(\widehat{P},\mathsf{q})$. There exists a choice of the parameters of Algorithm 3 such that with high probability over the algorithm's randomness, the output $\widehat{\mathsf{q}}$ satisfies*

$$\operatorname{dis}(\widehat{P},\widehat{\mathsf{q}}) \leq \operatorname{dis}(\widehat{P},\mathsf{q}^*) + \frac{\lambda}{2}\|\mathsf{q}^*\|_2^2 + \widetilde{O}\Big(\frac{m^{1/4}}{\sqrt{\varepsilon n}}\Big).$$

The above theorem follows as a corollary of the following theorem.

Theorem 11. *Algorithm 3 is (ε,δ)-differentially private. Let $\beta \in (0,1)$. If we set*

$$K = \frac{(\hat{r}^2+\lambda)^2\varepsilon^2 n^2}{128\mu^2\hat{r}^4 r^4 m\log(\frac{2m}{\beta})\log(\frac{1}{\delta})}.$$

then with probability at least $1-\beta$ over the algorithm's randomness (Gaussian noise), the output $\widehat{\mathsf{q}}$ satisfies

$$\widetilde{F}_T^{\lambda}(\widehat{\mathsf{q}}) \leq \min_{\mathsf{q}\in\Delta_m} \widetilde{F}_T^{\lambda}(\mathsf{q}) + \frac{46(\lambda+\hat{r}^2)\mu r^2\hat{r}^2\log(\frac{2m}{\beta})\sqrt{m\log(\frac{1}{\delta})}}{\varepsilon n}.$$

Using the approximation guarantee of $\widetilde{F}_T$ given in Corollary 1 together with Theorem 11 above, we reach the result of Theorem 6. The discrepancy guarantee in Theorem 6 can be more precisely stated as the following corollary of Theorem 11.

Corollary 4. *Let $\mathsf{q}^* \in \arg\min_{\mathsf{q}\in\Delta_m} \operatorname{dis}(\widehat{P},\mathsf{q})$. Let $\beta \in (0,1)$. In Algorithm 3, set K as in Theorem 11. Then, there exists a choice of μ such that the following holds: assuming w.l.o.g. that $\lambda = O(\hat{r}^2)$, with probability at least $1-\beta$, the output $\widehat{\mathsf{q}}$ satisfies*

$$\operatorname{dis}(\widehat{P},\widehat{\mathsf{q}}) \leq \operatorname{dis}(\widehat{P},\mathsf{q}^*)) + \frac{\lambda}{2}\|\mathsf{q}^*\|_2^2 + \widetilde{O}\Big(\frac{\Lambda^2 r\hat{r}^2 m^{1/4}}{\sqrt{\varepsilon n}}\Big),$$

where Λ is the $\|\cdot\|_2$-bound on the predictors in $\mathcal{H}$.

Proof of Theorem 11 First, we show that Algorithm 3 is (ε,δ)-differentially private. Note that for any $\mathsf{q} \in \Delta_m$ the $\|\cdot\|_2$-sensitivity of $\nabla\widetilde{F}_T^{\lambda}$ can be upper bounded as $\|\nabla\widetilde{F}_T^{\lambda}(\mathsf{q}) - \nabla\widetilde{F}_{T'}^{\lambda}(\mathsf{q})\|_2 = \|\nabla\widetilde{F}_T(\mathsf{q}) - \nabla\widetilde{F}_{T'}(\mathsf{q})\|_2 \leq \sqrt{m}\|\nabla\widetilde{F}_T(\mathsf{q}) - \nabla\widetilde{F}_{T'}(\mathsf{q})\|_\infty \leq \frac{2\mu r^2\hat{r}^2\sqrt{m}}{n}$, where the last inequality follows from the sensitivity bound in Corollary 1. Thus, given the setting of the Gaussian noise in the algorithm, the privacy guarantee of the Gaussian mechanism [DKM+06, DR14] together with the Moments Accountant technique [ACG+16] show the claimed privacy guarantee.

Next, we prove the convergence guarantee. The analysis here is similar to the analysis of noisy mirror descent in [BGN21, AFKT21]. First, it is known that $\Phi(\mathsf{q}) \triangleq \frac{\|\mathsf{q}\|_p^2}{p-1}$, where $p = 1 + \frac{1}{\log(m)}$, is 1-strongly convex with respect to $\|\cdot\|_1$ (see, e.g., [NY83]). Moreover, $D_\Phi \triangleq \max_{\mathsf{q},\mathsf{q}'} |\Phi(\mathsf{q}) - \Phi(\mathsf{q}')| \le 2\log(m)$. Note also that $\widetilde{F}_T^\lambda$ is $\gamma \triangleq (\hat{r}^2 + \lambda)$-Lipschitz w.r.t $\|\cdot\|_1$, which follows from the Lipschitz property of $\widetilde{F}_T$ (Corollary 1) and the fact that $\frac{\lambda}{2}\|\mathsf{q}\|_2^2$ is λ-Lipschitz with respect to $\|\cdot\|_1$ over Δ_m. Hence, by standard analysis of (noisy) mirror descent [NY83, NJLS09], we have (letting $\tilde{\mathsf{q}} = \operatorname{argmin}_{\mathsf{q}\in\Delta_m} \widetilde{F}_T^\lambda(\mathsf{q})$)

$$\begin{aligned}\widetilde{F}_T^\lambda(\widehat{\mathsf{q}}) - \widetilde{F}_T^\lambda(\tilde{\mathsf{q}}) &\le \frac{D_\Phi}{2\eta K} + \frac{\eta\gamma^2}{2} + \frac{\eta}{2K}\sum_{k=1}^K \|Z_k\|_\infty^2 \\ &\le \frac{2\log(m)}{\eta K} + \frac{\eta(\lambda + \hat{r}^2)^2}{2} + \frac{\eta}{2K}\sum_{k=1}^K \|Z_k\|_\infty^2\end{aligned}$$

where $\{Z_k : k \in [K]\}$ are i.i.d. from $\mathcal{N}(\mathbf{0}, \sigma^2\mathbb{I}_m)$. By a concentration argument in non-Euclidean norms [JN08, Theorem 2.1], with probability $\ge 1-\beta$, we have $\frac{1}{K}\sum_{k=1}^K \|Z_k\|_\infty^2 \le 4\sigma^2\log(\frac{2m}{\beta})$. Hence, with probability $\ge 1-\beta$, we have

$$\widetilde{F}_T^\lambda(\widehat{\mathsf{q}}) - \widetilde{F}_T^\lambda(\tilde{\mathsf{q}}) \le \frac{2\log(m)}{\eta K} + \frac{\eta(\lambda + \hat{r}^2)^2}{2} + 2\eta\sigma^2\log(\frac{2m}{\beta})$$

Thus, given the setting of σ (Step 5 of Algorithm 3), optimizing the bound above in η and K yields $\eta = \frac{2}{(\hat{r}^2+\lambda)}\sqrt{\frac{\log(m)}{K}}$ and $K = \frac{(\hat{r}^2+\lambda)^2\varepsilon^2 n^2}{128\mu^2\hat{r}^4 r^4 m\log(\frac{2m}{\beta})\log(\frac{1}{\delta})}$. Plugging these values in the above bound yields the claimed bound.

Proof of Corollary 4. The following is the choice of μ yielding the statement of the corollary:

$$\mu = \frac{\sqrt{\varepsilon n}\log^{1/4}(m+n)}{4r\hat{r}\sqrt{(\lambda+\hat{r}^2)\log(\frac{2m}{\beta})\big[m\log(\frac{1}{\delta})\big]^{1/4}}}.$$