Private Rate-Constrained Optimization with Applications to Fair Learning

Mohammad Yaghini*[1], Tudor Cebere*[2], Michael Menart[1], Aurélien Bellet[2], Nicolas Papernot[1]

[1] *University of Toronto and Vector Institute, Toronto, Canada*
[2] *Inria, Université de Montpellier, INSERM, France*

June 13, 2025

Abstract

Many problems in trustworthy ML can be formulated as minimization of the model error under constraints on the prediction rates of the model for suitably-chosen marginals, including most group fairness constraints (demographic parity, equality of odds, etc.). In this work, we study such constrained minimization problems under differential privacy (DP). Standard DP optimization techniques like DP-SGD rely on the loss function's decomposability into per-sample contributions. However, rate constraints introduce inter-sample dependencies, violating the decomposability requirement. To address this, we develop RaCO-DP, a DP variant of the Stochastic Gradient Descent-Ascent (SGDA) algorithm which solves the Lagrangian formulation of rate constraint problems. We demonstrate that the additional privacy cost of incorporating these constraints reduces to privately estimating a histogram over the mini-batch at each optimization step. We prove the convergence of our algorithm through a novel analysis of SGDA that leverages the linear structure of the dual parameter. Finally, empirical results on learning under group fairness constraints demonstrate that our method Pareto-dominates existing private learning approaches in fairness-utility trade-offs.

1 Introduction

From fair learning [59, 27, 2, 16] to robust optimization [13, 15] and cost-sensitive learning [45, 15], many Machine Learning (ML) tasks can be formulated as constrained optimization problems. In such problems, the goal is to minimize the model's overall error subject to *rate constraints*, which enforce specific conditions on prediction rates across subsets of the training data. For instance, ensuring fairness in a resume screening system might require similar rates of positive outcomes (e.g., resumes selected for human review) across gender groups, a criterion known as equalized odds [32]. Similarly, in a medical setting with imbalanced data, a decision-support system may need to strictly limit its false negative rate, e.g., to reduce the risk of misdiagnosing cancerous tumours as benign.

These learning tasks often require training models on sensitive data, such as employee or patient records in our previous examples. Publicly releasing these models can expose data owners to privacy attacks [49] and result in the disclosure of personal data without consent [52]. In the absence of proper safeguards, these risks can harm individuals, undermine trust in AI systems, and discourage data sharing in critical applications like medical research. Despite recent advances, differentially private (DP) constrained optimization has focused almost exclusively on fairness constraints [35, 46, 55, 11, 43]. Methods tailored to fairness do not extend to the broader family of rate constraints that arise in practice. We bridge this gap with the first general DP framework for arbitrary rate-constrained problems. Our approach both expands DP's reach to previously incompatible applications and pushes the Pareto frontier of utility, privacy, and constraint satisfaction, including fairness, beyond the current state of the art.

*Equal contribution.

Differential Privacy (DP) is the standard framework for private data analysis that has been successfully applied to training and releasing *unconstrained* models with formal privacy guarantees. A key approach is the widely used DP-SGD algorithm [50, 9, 1], which ensures DP by clipping per-sample gradients and adding calibrated noise to the averaged gradient. This process bounds each data sample's influence. Incorporating rate constraints presents a challenge, however, because unlike typical training losses, these constraint functions (or their regularizer counterparts) *do not readily decompose into per-sample terms*. This fundamental incompatibility with per-sample processing makes it difficult to integrate them with standard DP-SGD and its variants. Our key contribution is the introduction of a DP optimization algorithm that bridges this gap, enabling private optimization subject to rate constraints.

To address the challenge of privately enforcing rate constraints, we propose *generalized rate constraints*. Generalized rate constraints allow us to express all rate constraints in a common form based on statistics (i.e., histograms) on *disjoint* subgroups within the dataset. This common structure is the key to our privacy solution: it allows us to efficiently gather all the necessary information to evaluate these constraints under DP. Beyond its advantages for privacy, our method offers greater flexibility compared to prior work [15] by extending to scenarios with multiple output classes.

With generalized rate constraints at hand, we introduce RaCO-DP, a framework for optimizing machine learning models under rate constraints with differential privacy (DP). RaCO-DP is a differentially private variant of the Stochastic Gradient Descent Ascent (SGDA) algorithm which leverages a Langrangian formulation and generalized constraints to overcoming the decomposability obstacle. Our core insight leverages the structure that the generalized rate constraints provide: we can efficiently compute differentially private statistics (e.g., histograms) for these subgroups. By privatizing these statistics at each step, we enable private evaluation of the constraint function, and its per-sample constraint gradients. We provide a formal convergence analysis of RaCO-DP, proving that even for non-convex optimization problems, our method converges to an approximate stationarity point (i.e. a local optimum). We introduce a novel approach to analyzing SGDA that accounts for bias in gradient estimates and exploits the linear structure of the dual update to enhance convergence speed.

As a concrete application, we showcase our algorithm private learning with group fairness constraints, specifically demographic parity [23], and false negative rate constraints, highlighting the versatility of our approach. We present new state-of-the-art (SOTA) results on 4 datasets showing RaCO-DP Pareto-dominates the previous SOTA method [43] in terms of accuracy and fairness trade-off curves and nearly closes the optimality gap between private and non-private models. Additionally, our method offers two advantages over existing approaches designed specifically for fairness constraints. First, it provides stronger privacy guarantees than prior approaches that only consider the privacy of the sensitive label [35, 55, 43] (akin to label privacy [26]). Second, it allows practitioners to *directly* specify the maximum allowed disparity, unlike previous penalization-based methods such as [43] that offer only indirect control and require hyperparameter tuning to achieve the desired fairness level.

2 Background

2.1 Differential Privacy

Differential Privacy (DP) [22] has become the de-facto standard in privacy-preserving ML thanks to the robustness of its guarantees, its desirable behaviour under post-processing and composition, and its extensive algorithmic framework. We recall the definition below and refer to [21] for more details. We denote by $\mathcal{D}$ the space of datasets of some fixed size.

Definition 2.1 (Differential Privacy)**.** *A randomized mechanism $\mathcal{M}$ is (ε, δ)-DP if for all datasets $D, D' \in \mathcal{D}$ differing in one datapoint and for all events $\mathcal{O}$: $P[\mathcal{M}(D) \in \mathcal{O}] \leq e^{\varepsilon} P[\mathcal{M}(D') \in \mathcal{O}] + \delta$.*

In the above definition, $\delta \in (0, 1)$ can be thought of as a very small failure probability, and $\varepsilon > 0$ is the privacy loss; smaller ϵ and δ correspond to stronger privacy guarantees.

Differentially Private Stochastic Gradient Descent (DP-SGD) [50, 9, 1] serves as the foundational algorithm in private ML. Given a training dataset D and model parameters $\theta \in \mathbb{R}^d$, DP-SGD aims to privately solve the

empirical risk minimization problem $\min_{\theta\in\mathbb{R}^d} \ell(\theta)$, where $\ell(\theta) = \frac{1}{|D|}\sum_{x\in D} \ell(\theta; x)$ and $\ell(\theta, \cdot)$ is a loss function differentiable in θ. DP-SGD follows the standard SGD update, but guarantees differential privacy by (i) capping each data point's influence on the gradient through gradient clipping, and (ii) injecting Gaussian noise into the clipped gradients.

Each iteration $t \in [T]$ of DP-SGD incurs a privacy loss $(\varepsilon_t, \delta_t)$. Privacy composition [22, 36] and privacy accounting [1, 30, 20] are techniques that aggregate these per-step privacy losses into a total privacy guarantee (ε, δ) that holds for the entire optimization process.

2.2 Constrained Optimization via Lagrangian

In this work, we aim to solve *constrained* empirical risk minimization problems of the form:

$$\min_{\theta\in\mathbb{R}^d} \left\{\ell(\theta) := \frac{1}{|D|}\sum_{x\in D} \ell(\theta; x)\right\} \quad \text{s.t.} \quad \forall j \in [J],\ \Gamma_j(\theta) \le \gamma_j. \tag{1}$$

where $\Gamma_j : \mathbb{R}^d \mapsto \mathbb{R}^+$ are the constraint functions and $\gamma \in (\mathbb{R}^+)^J$ are slack parameters. We focus on inequality constraints, as equality constraints are generally infeasible under differential privacy [17].

Due to the difficulty of solving (1) directly, we instead solve an equivalent min-max optimization problem with respect to the Lagrangian:

$$\min_{\theta\in\mathbb{R}^d} \max_{\lambda\in\Lambda} \{\mathcal{L}(\theta, \lambda) := \ell(\theta) + R(\theta, \lambda)\}, \quad \text{where: } R(\theta, \lambda) = \sum_{j=1}^{J} \lambda_j(\Gamma_j(\theta) - \gamma_j). \tag{2}$$

Here, $\lambda \in \Lambda \subseteq (\mathbb{R}^+)^J$ are the Lagrange multipliers, often referred to as the dual parameter, while θ is the primal parameter. One of the simplest algorithms to solve (2) is the generalization of (S)GD known as (Stochastic) Gradient Descent-Ascent, (S)GDA [47].

Definition 2.2 (GDA)**.** *At each iteration t:*

$$\theta^{(t+1)} \leftarrow \theta^{(t)} - \eta_\theta \nabla_\theta \mathcal{L}(\theta^{(t)}, \lambda^{(t)}), \qquad \text{and} \qquad \lambda^{(t+1)} \leftarrow \Pi_\Lambda\big(\lambda^{(t)} + \eta_\lambda \nabla_\lambda \mathcal{L}(\theta^{(t)}, \lambda^{(t)})\big). \tag{3}$$

where η_θ and η_λ are step sizes and Π_Λ performs orthogonal projection onto Λ. The stochastic version (SGDA) replaces exact gradients with stochastic estimates.

2.3 Rate Constraints

In this work, we focus on constraints that relate to prediction behavior over subsets of the dataset. Consider a model $h : \mathcal{X} \times \mathbb{R}^d \mapsto \mathbb{R}^K$ that maps inputs from feature space $\mathcal{X}$ to real-valued prediction scores over the label set $\mathcal{Y} = \{1, \ldots, K\}$ using parameters $\theta \in \mathbb{R}^d$. Formally, (hard) prediction rates count the number of points in a dataset D for which the model predicts a certain label $k \in \mathcal{Y}$: $P_k^{\text{hard}}(D; \theta) = \frac{1}{|D|}\sum_{x\in D} \mathbb{1}_{[\arg\max_{k'\in[K]}\{h(\theta;x)_{k'}\}=k]}$.

As the indicator function is non-differentiable and thus challenging to optimize, we will use differentiable versions of these constraints. We rely on the tempered softmax function $\sigma_\tau(z)_k = \frac{\exp(-\tau z_k)}{\sum_{l=1}^{K}\exp(-\tau z_l)}$, where the temperature parameter $\tau \in \mathbb{R}^+$ controls the sharpness of the probability distribution. This allows us to define soft prediction rates:

$$P_k(D; \theta, \tau) = \frac{1}{|D|}\sum_{x\in D} \sigma_\tau(h(\theta; x))_k, \quad \text{for } k \in \mathcal{Y}. \tag{4}$$

Observe that $\lim_{\tau\to\infty} P_k(D; \theta, \tau) = P_k^{\text{hard}}(D; \theta)$. For brevity, unless otherwise specified, we drop explicit mention of temperature and provide more discussion of this hyperparameter experimentally in Appendix E. Our convergence analysis in Section 5 holds for arbitrary τ.

Rate constraints, as defined by Goh et al. [28] and Cotter et al. [15] for the case of binary classification ($\mathcal{Y} = \{0, 1\}$), are linear combinations of a classifier's prediction rates across different data partitions:

$$\sum_{q\in[Q]} \alpha_q P_1(D_q, \theta) + \beta_q P_0(D_q, \theta) \le \gamma, \tag{5}$$

$Q > 0$, $\alpha_q, \beta_q \in \mathbb{R}$ are mixing coefficients, $D_1, ..., D_Q \subseteq D$ and γ is the slack parameter. Examples of rate constraints in this form include ensuring that the fraction of positive predictions across different demographic groups stays within a specified threshold, or requiring the model to achieve a minimum level of precision or recall for both classes [16].

Such constraints are limited to binary classification, and a multiclass generalization is not immediate. More critically, it is unclear how to evaluate rate constraints in (5) while efficiently preserving privacy.

3 Private Learning with Generalized Rate Constraints

Generalized rate constraints. We propose a generalized form of rate constraints that (i) is applicable to the multi-class setting, and (ii) exploits a structure shared across constraints that will allow accurate private estimation.

This shared structure is a partition $\{D_1, ..., D_Q\}$ of the dataset D for some $Q > 0$, which we refer to as the "global" partition. We then allow each rate constraint to incorporate prediction rates over any recombination of sub-datasets in the global partition. This structure is flexible, as it allows each rate constraint to have its own "local" datasets, provided that each of these datasets can be formed as a union of sets from the global partition. For example, in the context of fairness constraints, the global partition corresponds to the the sensitive groups (e.g., Hispanic, Black, Caucasian), and local datasets for one of the constraint is {Hispanic, Non-Hispanic} where Non-Hispanic = {Black, Caucasian}. See Figure 1.

Formally, given this global partition, we assume for each $j \in [J]$ there exists a family of subsets of $[Q]$, denoted $\mathcal{I}_j \subseteq 2^{[Q]}$, and a weight vector $\alpha_j \in \mathbb{R}^{|\mathcal{I}_j| \cdot K}$, such that the constraint Γ_j can be written in the following form:[1]

$$\Gamma_j(\theta) = \sum_{I \in \mathcal{I}_j} \sum_{k=1}^{K} \alpha_{j,I,k} P_k(\cup_{i \in I} D_i; \theta). \tag{6}$$

Assuming such a global partition is not restrictive, as the trivial partition $D_q = x_q$ (with $Q = |D|$) can always be used. However, as Section 4 will show, a smaller partition size Q enables a better privacy-utility trade-off through more effective noise use.

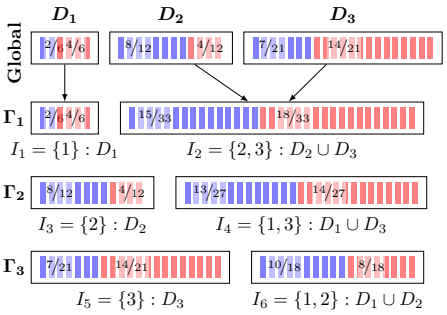


Figure 1: **Each rate constraint of the form (6) builds local datasets based on the global partition.** A class-1 (class-0) prediction is shown with a blue (red) square. Prediction rates P_0, P_1 are shown as fractions. As an example, let D_1, D_2, and D_3 be the set of Hispanic, Black, and Caucasian individuals in the dataset, respectively. Constraint Γ_1 builds its local datasets as $\{\{D_1\}, \{D_2 \cup D_3\}\}$, i.e. {Hispanic, Non-Hispanic}, from the global partition $\{D_1, D_2, D_3\}$ using the set of index subsets $\mathcal{I}_1 = \{I_1, I_2\}$.

Remark 1. *For common rate constraints, the best global partition will be readily apparent (see the application to fairness below). The smallest global partition can however be defined explicitly. Let $\bar{D}_1, ..., \bar{D}_{\bar{Q}}$ be the (possibly non-disjoint) subsets of D over which the functions $\Gamma_1, ..., \Gamma_J$ compute a prediction rate as per Eq. (6). Then the smallest global partition assigns any two data points x and x' in the same D_q if and only if $\{\forall \bar{q} \in [\bar{Q}], \{x, x'\} \cap \bar{D}_q \in \{\emptyset, \{x, x'\}\}\}$.*

Note that each rate constraint Γ_j is uniquely defined by the subset family $\mathcal{I}_j$ and vector α_j. Both parameters are public, specifying only the constraint's structure and containing no sensitive data.

Application to fair learning. Group fairness in machine learning aims to prevent models from making biased decisions across different sensitive groups. We show below our general form of rate constraints (4) allows to capture the popular group fairness notion of demographic parity [6]. More generally, all common group fairness measures can be formulated as rate constraints [15]. We provide details on formulating other fairness notions in Appendix A.

[1]With some abuse of notation, we denote by $\alpha_{j,I,k}$ the entry of α_j corresponding to subset I and label k.

Definition 3.1 (Demographic Parity)**.** *Assume each feature vector $x \in \mathcal{X}$ contains a sensitive attribute, denoted as Z, taking on values in $\mathcal{Z} \subset \mathbb{Z}$. A classifier $h(\theta; \cdot)$ satisfies demographic parity with respect to sensitive attribute Z if the probability of predicting any class k is independent of Z:*

$$\Pr[\hat{Y} = k \mid Z = z] = \Pr[\hat{Y} = k], \quad \forall z \in \mathcal{Z}, \forall k \in \mathcal{Y},$$

where $\hat{Y} = h(\theta; X)$ is the predicted label. In practice, we do not have access to the true probabilities, so it is common to estimate them by empirical prediction rates P_k. Using a slack parameter γ gives:

$$P_k(D[Z = z]; \theta) - P_k(D[Z \neq z]; \theta) \leq \gamma, \tag{7}$$

for each $z \in \mathcal{Z}$ and $k \in \mathcal{Y}$. Demographic parity thus leads to $J = |\mathcal{Z}| \cdot |\mathcal{Y}|$ rate constraints of the form (6). The global partition is of size $Q = |\mathcal{Z}|$ with elements $D_z = D[Z = z]$, $\forall z \in \mathcal{Z}$, and for the constraint corresponding to elements $z \in \mathcal{Z}$ and $y \in \mathcal{Y}$, we have $\mathcal{I} = \{\{z\}, [|\mathcal{Z}|] \setminus z\}$. The associated vector α has $\alpha_{\{z\},y} = 1$ and $\alpha_{\{[|\mathcal{Z}|]\setminus z\},y} = -1$, with the rest of the components set to 0.

Objective. With these definitions in place, we can now state our objective: solve problem (2) with generalized rate constraints of the form (6) under DP.

4 RaCO-DP: Private Rate-Constrained Optimization

We introduce RaCO-DP (Algorithm 1), an algorithm for rate-constrained optimization that extends SGDA (Definition 2.2) to satisfy DP. Each iteration t of the algorithm operates on a mini-batch $B^{(t)}$ and consists of three key components that we will describe in detail in this section:

1. **Private histogram computation:** For each class $k \in [K]$ and part $q \in [Q]$ of the partition, we privately estimate the sum of model predictions $\sigma(h(\theta_t; x))_k$ over the points x in the mini-batch $B^{(t)}$ that belong to D_q, storing these counts in a histogram $H^{(t)}$ (Section 4.1).
2. **Private primal updates:** We derive per-sample gradients for the primal update of SGDA based on post-processing the histogram $H^{(t)}$. We then clip, average and privatize these gradients using the Gaussian mechanism (Section 4.2).
3. **Private dual updates:** We compute all constraint values by again post-processing the histogram $H^{(t)}$, allowing us to perform the dual update at no additional privacy cost (Section 4.3).

Algorithm 1 RaCO-DP

Require: Dataset D, Parameter θ_0, learning rates η_θ, η_λ, Gaussian noise variance σ, Laplace parameter b, clipping norm C, sampling rate r, slack parameter vector γ, loss function $\ell(\theta; \cdot)$

Initialize $\lambda^{(0)} \leftarrow [0]$
for each $t \in \{0, \ldots, T-1\}$ **do**
  $B^{(t)} \leftarrow \text{PoissonSample}(D, r)$
  Private Histogram:
  $\hat{H}^{(t)}_{q,k} \leftarrow \sum_{x \in D_q \cap B^{(t)}} \sigma(h(\theta^{(t)}; x))_k + \text{Lap}(\frac{1}{b})$
  Primal Update:
  $Z^{(t)} \sim \text{Gaussian}(0, \mathbb{I}_d \sigma^2)$
  $\forall x \in B^{(t)}$: set $g^{(t)}_{x,\theta}$ as (see also Eq.(10))
  $\frac{1}{r|D|} \nabla_\theta \ell(\theta^{(t)}; x) + \nabla_\theta \hat{R}(\theta^{(t)}, \lambda^{(t)}; \hat{H}^{(t)}, x)$
  $g^{(t)}_\theta \leftarrow \left(\sum_{x \in B^{(t)}} \text{clip}(g^{(t)}_{x,\theta}, \frac{C}{r|D|}) \right) + Z^{(t)}$
  $\theta^{(t+1)} \leftarrow \theta^{(t)} - \eta_\theta g^{(t)}_\theta$
  Dual Update:
  $[g^{(t)}_\lambda]_j \leftarrow \Gamma^{\text{post}}_j(\hat{H}^{(t)}) - \gamma_j, \quad \forall j \in [J]$
  $\lambda^{(t+1)} \leftarrow \Pi_\Lambda(\lambda^{(t)} + \eta_\lambda g^{(t)}_\lambda)$ (see Eqn. (12)
end for
return $\theta^{(T)}$

4.1 Private Histogram Computation

Our algorithm's primal and dual updates compute prediction rates across dataset partition parts $D_1, \ldots, D_Q$ for each class. To track these prediction rates privately, we construct a histogram $H^{(t)} \in \mathbb{R}^{Q \times K}$ that counts (soft) model predictions for each combination of part q and class k. For a given model parameter θ, each sample x_i in the mini-batch $B^{(t)}$ belongs to exactly one part D_q of the partition but can influence the counts of all K classes through the softmax probabilities $\sigma(h(\theta; x))_k$. This non-private histogram is constructed by accumulating these softmax vectors:

$$H^{(t)}_{q,k} = \textstyle\sum_{D_q \cap B^{(t)}} \sigma(h(\theta; x_i))_k. \tag{8}$$

To make this histogram differentially private, we use the Laplace mechanism [22]. The ℓ_1 sensitivity of $H^{(t)}$ is 1 because each sample belongs to exactly one element of the global partition and its softmax predictions sum to 1 across classes. Therefore, we can achieve ε-DP by adding independent Laplace noise to each element:

$$\hat{H}^{(t)}_{q,k} = H^{(t)}_{q,k} + \mathrm{Lap}(1/\varepsilon), \quad \forall q \in [Q], k \in \mathcal{Y}. \tag{9}$$

Remark 2. *We focus on the Laplace mechanism for simplicity, but we note that our framework can readily accommodate other differentially private histogram mechanisms that may provide better utility in some regimes, e.g., when the histogram is high-dimensional and sparse [56].*

In the following sections, we will see that $\hat{H}^{(t)}$ contains all the necessary information to compute the quantities related to the rate constraints required for both the primal and dual updates, thereby avoiding additional privacy costs that would arise from composing multiple queries.

4.2 Privately Computing the Primal Gradient

A key requirement in DP-SGDA is that each sample must have a bounded contribution to the gradient updates. To satisfy this requirement in RaCO-DP, we need to decompose the Lagrangian into per-sample terms. While the loss term $\ell(\theta)$ naturally decomposes into per-sample terms as in standard DP-SGD, the regularizer R is more challenging.

Given a mini-batch B, define $B_{\cap I} = B \cap (\bigcup_{i \in I} D_i)$ and recall that $\mathcal{I}_j$ and $\alpha_j \in \mathbb{R}^{|\mathcal{I}_j| \times K}$ denote the family of subsets and weight vector associated with constraint Γ_j. The minibatch-level regularizer is,

$$R(\theta, \lambda; B) = \sum_{j=1}^{J} \lambda_j \Big(\big(\sum_{I \in \mathcal{I}_j} \sum_{k}^{K} \alpha_{j,I,k} P_k(B_{\cap I}; \theta) \big) - \gamma_j \Big) = \sum_{j=1}^{J} \lambda_j \Big(\big(\sum_{I \in \mathcal{I}_j} \sum_{k}^{K} \sum_{x \in B_{\cap I}} \frac{\alpha_{j,I,k}}{|B_{\cap I}|} \sigma(h(\theta; x)_k) \big) - \gamma_j \Big).$$

Note that this may be a *biased* estimate of $R(\theta, \lambda)$ due to the normalization term $|B_{\cap I}|$. We account for this bias in our convergence analysis (Section 5).

We would like a per-sample decomposition of $R(\theta, \lambda; B)$. The main obstacle in such a decomposition are the quantities $|B_{\cap I}|$, which depend on the entire mini-batch. We overcome this by first noting that $|B_{\cap I}| = \sum_{i \in I} \sum_k H^{(t)}_{i,k}$, leading to the following per-sample regularizer estimator at point $x \in D$:

$$\hat{R}(\theta, \lambda; H, x) = \sum_{j=1}^{J} \lambda_j \Big(\big(\sum_{I \in \mathcal{I}_j} \sum_{k_1 \in K} \frac{\alpha_{j,I,k_1} \, \mathbb{1}_{[x \in B_{\cap I}]}}{\sum_{i \in I} \sum_{k_2 \in K} H^{(t)}_{i,k_2}} \sigma(h(\theta; x))_{k_1} \big) - \gamma_j \Big),$$

and thus the overall per-sample gradient is given by,

$$\frac{\nabla_\theta \ell(\theta; x)}{r|D|} + \sum_{j=1}^{J} \sum_{I \in \mathcal{I}_j} \sum_{k_1 \in [K]} \frac{\lambda_j \; \alpha_{j,I,k_1} \; \mathbb{1}_{[x \in B_{\cap I}]}}{\sum_{i \in I} \sum_{k_2 \in [K]} H^{(t)}_{i,k_2}} \nabla_\theta [\sigma(h(\theta; x))_{k_1}]. \tag{10}$$

Then, since H depends on the mini-batch B, we use instead its differentially private version $\hat{H}$. Note that the normalizing term $\frac{1}{r|D|}$ is necessary to correctly implement clipping in Line 8 of Algorithm 1.

Remark 3. *For specific constraints, the estimation of $|B_{\cap I}|$ can be further refined. For example, when each $\mathcal{I}_j$ is itself a partition of $[Q]$, the sensitivity of $\sum_i \sum_{k_2} H^{(t)}_{i,k_2}$ is at most 1, and adding Laplace noise directly results in a tighter estimate of $|B_{\cap I}|$.*

With this per-sample decomposition, we can apply standard DP-SGD techniques: clipping per-sample gradients, averaging them over the mini-batch, and adding Gaussian noise to preserve privacy.

4.3 Privately Computing the Constraint Dual Gradient

For each constraint Γ_j and corresponding slack parameter γ_j, the gradient of the Lagrangian w.r.t λ is:

$$\nabla_\lambda \mathcal{L}(\theta^{(t)}, \lambda^{(t)})_j = \Gamma_j(\theta^{(t)}; B^{(t)}) - \gamma_j, \tag{11}$$

The dual update in RaCO-DP thus requires evaluating rate constraints on the current mini-batch $B^{(t)}$, incurring a privacy cost. To avoid this additional cost, we introduce a post-processing function $\Gamma_j^{\text{post}} : \mathbb{R}^{Q\times K} \mapsto \mathbb{R}$ that operates directly on the private histogram $\hat{H}^{(t)}$. This function replaces each sum of model predictions $\sum_{i\in D_q\cap B^{(t)}} \sigma(h(\theta; x_i))_k$ with the corresponding histogram count $\hat{H}^{(t)}_{q,k}$:

$$\Gamma_j^{\text{post}}(\hat{H}^{(t)}) = \sum_{I\in\mathcal{I}_j} \sum_{k_1\in[K]} \frac{\alpha_{j,I,k_1} \sum_{i\in I} \hat{H}^{(t)}_{i,k_1}}{\sum_{i\in I}\sum_{k_2\in[K]} \hat{H}^{(t)}_{i,k_2}}. \tag{12}$$

Since $\hat{H}^{(t)}$ is already differentially private, the post-processing property of DP ensures that this computation requires no additional privacy budget.

Remark 4. *As standard in private optimization, the mini-batches $B^{(t)}$ are constructed through Poisson sampling (each datapoint is independently included with probability r). This allows us to leverage privacy amplification by subsampling [37, 5] for both the Laplace mechanism used in histogram computation (Section 4.1) and the Gaussian mechanism used for the per-sample gradient (Section 4.2).*

RaCO-DP's efficiency relies on a private histogram $H^{(t)}$ which enables both per-sample gradient computation and private constraint evaluation, key for handling rate constraints with DP guarantees. One has the following privacy guarantees for Algorithm 1, which follow from composing the (subsampled) Laplace and Gaussian mechanisms over the T iterations.

Theorem 4.1. *Let $b \geq 2\max\left\{\frac{1}{\epsilon}, \frac{r\sqrt{T\log(T/\delta)}}{\epsilon}\right\}$ and $\sigma \geq 10\max\left\{\frac{C\log(T/\delta)}{r|D|\epsilon}, \frac{C\sqrt{T}\log(T/\delta)}{|D|\epsilon}\right\}$, then Algorithm 1 is (ϵ, δ)-DP.*

The proof is in Appendix B. We present this result primarily for to provide intuition about parameter scaling. In our experiments, we use a tighter privacy accountant that improves both constants and logarithmic factors. Additionally, in our convergence analysis, we offer a more detailed examination when the Lagrangian is Lipschitz and the algorithm is run without clipping.

5 Convergence and Utility Analysis

Consider the function defined as $\Phi(\theta) = \max_{\lambda\in\Lambda}\{\mathcal{L}(\theta, \lambda)\}$. Ideally, one would want to show that Algorithm 1 approximately minimizes Φ. However, due to the fact that Φ may be non-convex, finding an approximate minimizer is intractable in general. In fact, because the constraint functions, $\{\Gamma_j\}$, may be non-convex in θ, if $\Lambda = (\mathbb{R}^+)^J$ then even finding a point where Φ is finite may be intractable. As such, we must make two standard concessions. First, we will assume Λ is a compact convex set of bounded diameter. Intuitively, this bounds the penalty applied when the constraints are not satisfied. Further, instead of guaranteeing that Algorithm 1 finds an approximate minimizer of Φ, we will show the algorithm finds an approximate stationary point. We note that stationarity is a standard convergence measure in non-convex optimization, and provide more discussion in Appendix D.1. Our subsequent analysis in fact provides for a slightly stronger, but more technical, notion of stationarity than we provide here; see Appendix D.6 for more details.

Definition 5.1. *((α, ν)-stationary point) A point θ is an (α, ν)-stationary point if $\exists\theta'$ s.t. $\|\theta - \theta'\| \leq \nu$ and $\min_{v\in\partial\Phi(\theta')} \|v\| \leq \alpha$, with $\partial\Phi$ the subdifferential of Φ.*

When the loss is Lipschitz and smooth, SGDA converges to an approximate stationary point of Φ [41]. Unfortunately, Algorithm 1 may have *biased* gradients, and the scale of noise present in $\hat{g}_\theta$ and $\hat{g}_\lambda$ may vary dramatically depending on d and J. Thus, our main goal in this section is two-fold. First, we aim to formally show that despite using biased gradients, Algorithm 1 provably finds an approximate stationary

point. Specifically, when the error in the primal updates (w.r.t. $\|\cdot\|_2$) is at most τ_θ and the error in the dual updates is at most τ_λ (w.r.t. $\|\cdot\|_\infty$), we show SGDA on a nonconvex-linear loss finds a stationary point roughly with $\alpha = O(\frac{\sqrt{1+\tau_\theta}}{T^{1/4}} + \tau_\theta + \sqrt{\tau_\lambda} + \frac{1}{\sqrt{T}})$; see Appendix D.6. Second, we characterize the impact of the noise in $\hat{g}_\theta$ and $\hat{g}_\lambda$ added due to privacy. This involves correctly balancing the number of iterations T with the scale of noise needed to ensure privacy, which increases with T. This leads to the following result for Algorithm 1 run without clipping.

Theorem 5.2 (Informal). *Let $n = \min_{q\in[Q]}\{|D_q|\}$. Assume $h(\theta; x)$ and $\ell(\theta)$ are both Lipschitz and smooth. Then, under appropriate choices of parameters, Algorithm 1 is (ϵ, δ)-DP and with probability at least $1-\rho$ there exists $t \in [T]$ s.t. θ_t is an (α, α)-stationary point of Φ with,*

$$\alpha = O\Bigg(\Bigg(\frac{\sqrt{d\log(\frac{JKn}{\rho})}\log(\frac{n}{\delta})}{n\epsilon}\Bigg)^{\frac{1}{3}} + \frac{K^{\frac{1}{4}}\sqrt{\log(\frac{n}{\delta})}\log^{\frac{1}{4}}(\frac{JKn}{\rho})}{(n\epsilon)^{1/4}}\Bigg),$$

up to dependence on problem constants. We provide a complete statement and full proof in Appendix D.2. In Appendix D.7, we derive Lipschitz and smoothness constants for $\mathcal{L}$ from those of the classifier.

There are several key steps involved in achieving this result. First, we provide a general convergence proof for SGDA under the assumption that gradients have bounded error (Theorem D.5). Notably, in contrast to previous analyses, this result 1) allows for biased gradients; 2) depends on the ℓ_∞ error in the dual gradient estimate (rather than the ℓ_2 error); and 3) achieves faster convergence (in terms of T) by leveraging the linear structure of the dual. To point (3), our analysis shows SGDA in this setting can converge as fast as $\frac{1}{T^{1/4}}$ instead of $\frac{1}{T^{1/6}}$ (shown by [41]), essentially matching the rate observed in comparable minimization (rather than min-max) settings. See Theorem D.5 for this specific result. The next step in proving Theorem 5.2 is to control the error of the gradient estimates while balancing the noise necessary for privacy. We show that this error scales proportional to $\tilde{O}(\frac{\sqrt{dT}}{n\epsilon} + \frac{1}{\sqrt{n}})$ in the primal and $\tilde{O}(\frac{\sqrt{KT}\log J}{n\epsilon} + \sqrt{\frac{\log J}{n\epsilon}})$ in the dual. We defer the reader to Lemma D.4 in Appendix D for a more detailed accounting of this error.

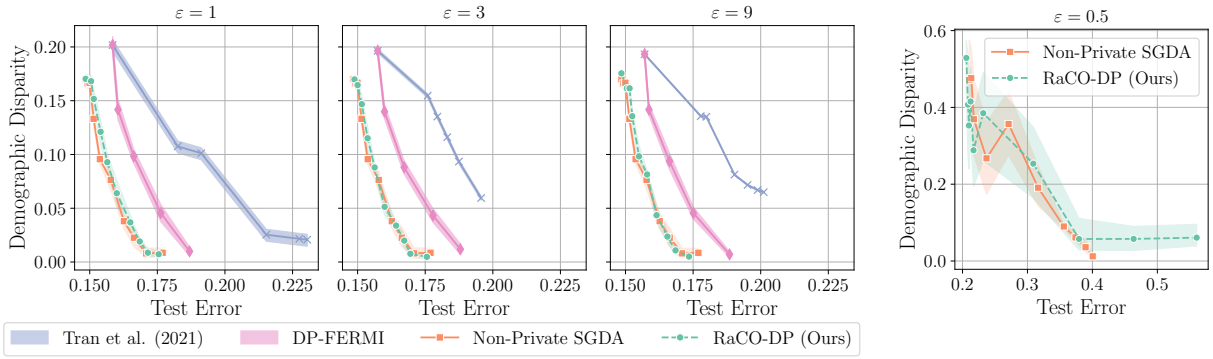


Figure 2: **(Left) Disparity-Error trade-off curves** of DP fair training algorithms on `Adult` under demographic parity constraints. RaCO-DP Pareto dominates the SOTA method (DP-FERMI), closing the optimality gap with non-private (SGDA). **(Right) RaCO-DP vs. Non-Private SGDA on** `ACSEmployment` with 18 constraints, showing that RaCO-DP adapts to multiple sensitive groups.

6 Experimental Results

To demonstrate RaCO-DP's versatility, we evaluate it on two constraint types: demographic parity (using four standard fairness benchmarks: `Adult` [10], `Credit-Card` [58], `Parkinsons` [42], and `Folkstables` [19]) and false negative rate (FNR) constraints (using `Adult` and `heart` [3]). Dataset descriptions and prediction task details are deferred to Appendix E.1.

Baselines. For demographic parity, RaCO-DP is benchmarked against three competing approaches: DP-FERMI [43], the framework of Tran, Dinh, and Fioretto [53], and the method of Jagielski et al. [35]. We report the results of these competing approaches as reported in Lowy, Gupta, and Razaviyayn [43] and adopt a similar setup for our algorithm to make the results comparable (see below). For FNR constraints (Appendix A.2), as we are unaware of prior work addressing these under DP, we employ non-private SGDA as a reference point to quantify DP's impact.

Experimental setup. We evaluate all algorithms for $\varepsilon \in \{1, 3, 9\}$, and $\delta = 10^{-5}$., using logistic regression models, consistent with prior work. For all datasets, train-test splits are $75\% - 25\%$. Unlike Lowy, Gupta, and Razaviyayn [43]'s setup, we reserve 15% of the training set for validation. As in prior work, we do not account for privacy loss from hyperparameter tuning [48]. Complete training setup details, including hyperparameter tuning, are in Appendix E.2.

We use the numerical accountant from Doroshenko et al. [20] to track privacy loss during training. Based on hyperparameters affecting privacy (σ, b, and B) and for each privacy budget (ε, δ), we determine maximum training steps T ensuring total privacy loss stays within the specified budget. Recent work by Lebeda et al. [40] and Chua et al. [14] shows that the common use of shuffled fixed-size mini-batches can violate guarantees from privacy accountants, which assume Poisson sampling. Therefore, we use Poisson sampling for the mini-batches.

Fairness results. The results in Figure 2 prove that RaCO-DP achieves superior performance compared to all other methods, providing SOTA results for the privacy-utility-fairness tradeoff, approaching the accuracy of non-private SGDA on the `Adult` dataset, with additional results in Appendix E.3.

Computational Performance. We benchmark the average wall-time per step for SGD, DP-SGD, RaCO-DP, and DP-FERMI using their public code [31]). On identical hardware, RaCO-DP trains *three orders of magnitude* faster than DP-FERMI on Adult, see Appendix E.5.

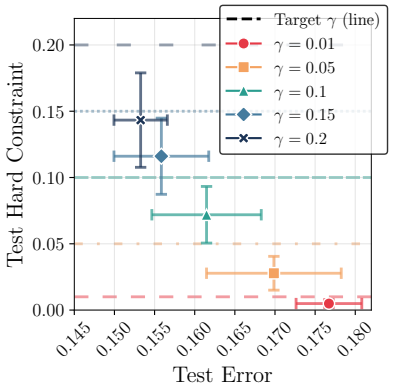


Figure 3: **Satisfiability on Adult.** Trade-off between test error and constraint violation for different target values γ (dashed lines), averaged over 20 runs. RaCO-DP achieves demographic parity constraint satisfaction.

Constraint satisfaction. The results in Figure 3 demonstrate that RaCO-DP consistently achieves the pre-specified constraint values γ as measured by the *hard* rate constraints. This marks a significant improvement over existing approaches, which typically rely on indirect hyperparameter tuning to influence constraint satisfaction. Our direct constraint optimization approach through the Lagrangian formulation is simpler for practitioners and provides reliable performance. Additionally, these results show that tempered-sigmoid temperature $\tau = 1$ is sufficient to enforce hard-constraints.

FNR results. As shown in Figure 5 of Appendix E, under FNR constraints, we nearly match non-private SGDA for the regime where $\gamma > 0.1$. The regularizer requires large gradients for $\gamma < 0.1$, but the clipping norm C imposes a limit that limits its effectiveness, as discussed below. In Appendix E we also showcase the performance of RaCO-DP on `heart.`, a medical classification dataset.

Limitations. Koloskova, Hendrikx, and Stich [38] show that gradient clipping biases SGD, persisting even with vanishing step sizes. In our setting, this bias can push convergence outside the feasible set, making the clipping norm C a critical hyperparameter.To demonstrate this is a general clipping issue rather than specific to RaCO-DP, we examine a strict FNR $= 0$ constraint in logistic regression on Adult, without DP noise ($\sigma = 0$, $b = \infty$). As shown in Figure 6 of Appendix E, using a clipping norm below 12.5 makes RaCO-DP fail to meet the constraint, though less restrictive constraints allow smaller norms. Another limitation is the usage of soft constraints for the dual update. Hard constraints can be used in practice for the update step, although this departs from the theoretical guarantees offered by RaCO-DP. However, as shown in Appendix E.4 this modification offers limited utility benefits.

7 Related Work

To the best of our knowledge, private learning under general rate constraints has not been explored in prior work, except in the specific case of group fairness constraints. Accordingly, we review related work at the intersection of differential privacy and fairness, a key application area where early research has identified fundamental trade-offs between these two objectives [17].

Existing work can be grouped into three main categories. A first line of research [4, 25, 51, 53, 39, 24, 54, 44] examines how privacy mechanisms can inadvertently harm fairness. For instance, Esipova et al. [24] characterizes the disparate impact of DP-SGD due to gradient misalignment caused by clipping. The second line of work focuses on protecting the privacy of sensitive attributes used to enforce fairness constraints [35, 46, 55]. Unlike standard DP, which protects the entire dataset, privacy for sensitive attributes requires injecting less noise, leading to improved model performance. However, this approach has significant limitations. Individuals can still be re-identified through their non-sensitive attributes, and if sensitive and non-sensitive features are correlated, an adversary may still be able to infer sensitive attributes. Due to these vulnerabilities, such works are not directly comparable to ours.

The third line of work, closest to ours, seeks to jointly enforce DP and group fairness [11, 43]. Berrada et al. [11] use DP-SGD without fairness mitigation, finding well-generalized models show no major privacy-fairness trade-off. However, their notion of fairness is error disparity, arguably measuring subpopulation generalization. Our work shows that while mitigation is needed for demographic parity, appropriate algorithm design can surmount the fairness-privacy trade-off. Lowy, Gupta, and Razaviyayn [43] propose a proxy objective for stochastic optimization of group fairness measures. In contrast, RaCO-DP supports a broad range of rate constraints that can be freely combined, without needing a task-specific objective. Additionally, Lowy, Gupta, and Razaviyayn [43] assumes Lipschitz continuity of model parameters—a strong assumption we relax, instead using clipping for bounded gradients. While its theoretical convergence rate is slower, RaCO-DP offers greater generality, complicating direct comparisons. Notably, Lowy, Gupta, and Razaviyayn [43] presents results for both sensitive attribute DP and standard DP (Definition 2.1), but their public code and evaluations focus on the weaker notion. Despite a stronger privacy guarantee, RaCO-DP achieves superior privacy-fairness trade-offs.

Fairness aside, SGDA has been extensively studied for minimax optimization [47, 33, 41], with Yang et al. [57] proposing its first DP analogue. Private minmax optimization has been heavily studied recently [12, 60, 7, 8, 29], although work on non-convex losses is limited to [43].

8 Conclusion

We introduced a DP algorithm using private histograms for training rate-constrained models. RaCO-DP demonstrates strong performance across various datasets and constraint types, often nearing non-private baselines while meeting privacy and constraint criteria. Our findings suggest that privacy-fairness trade-offs may be less significant than previously believed. Future work could explore private learning under individual fairness constraints, which cannot be formulated as rate constraints.

9 Acknowledgments

We would like to acknowledge our sponsors, who support our research with financial and in-kind contributions: Apple, CIFAR through the Canada CIFAR AI Chair, Meta, Microsoft, NSERC through the Discovery Grant and an Alliance Grant with ServiceNow and DRDC, the Ontario Early Researcher Award, the Schmidt Sciences foundation through the AI2050 Early Career Fellow program. Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute. Michael Menart is also supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), grant RGPIN-2021-03206. The work of Tudor Cebere and Aurélien Bellet is supported by grant ANR-20-CE23-0015 (Project PRIDE) and the ANR 22-PECY-0002 IPOP (Interdisciplinary Project on Privacy) project of the Cybersecurity PEPR. This work was performed using HPC resources from GENCI–IDRIS (Grant 2023-AD011014018R1).

References

[1] Martin Abadi et al. "Deep Learning with Differential Privacy". In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 2016.

[2] Alekh Agarwal et al. "A Reductions Approach to Fair Classification". In: *Proceedings of the 35th International Conference on Machine Learning*. Proceedings of Machine Learning Research. PMLR, 2018.

[3] Alex Teboul. *Heart Disease Health Indicators Dataset*. 2022. URL: `https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset` (visited on 05/16/2025).

[4] Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. "Differential Privacy Has Disparate Impact on Model Accuracy". In: *Advances in Neural Information Processing Systems*. 2019.

[5] Borja Balle, Gilles Barthe, and Marco Gaboardi. "Privacy amplification by subsampling: Tight analyses via couplings and divergences". In: *Advances in neural information processing systems* 31 (2018).

[6] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.

[7] Raef Bassily, Cristóbal Guzmán, and Michael Menart. "Differentially Private Algorithms for the Stochastic Saddle Point Problem with Optimal Rates for the Strong Gap". In: *Proceedings of Thirty Sixth Conference on Learning Theory*. Proceedings of Machine Learning Research. 2023.

[8] Raef Bassily, Cristóbal A Guzmán, and Michael Menart. "Private Algorithms for Stochastic Saddle Points and Variational Inequalities: Beyond Euclidean Geometry". In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. 2024.

[9] Raef Bassily, Adam D. Smith, and Abhradeep Thakurta. "Private Empirical Risk Minimization: Efficient Algorithms and Tight Error Bounds". In: *FOCS*. 2014.

[10] Barry Becker and Ronny Kohavi. *Adult*. UCI Machine Learning Repository. 1996.

[11] Leonard Berrada et al. *Unlocking Accuracy and Fairness in Differentially Private Image Classification*. 2023.

[12] Digvijay Boob and Cristóbal Guzmán. "Optimal algorithms for differentially private stochastic monotone variational inequalities and saddle-point problems". In: *Mathematical Programming* (2023).

[13] Robert S. Chen et al. "Robust Optimization for Non-Convex Objectives". In: *NIPS*. 2017.

[14] Lynn Chua et al. "Scalable DP-SGD: Shuffling vs. poisson subsampling". In: *arXiv preprint arXiv:2411.04205* (2024).

[15] Andrew Cotter et al. "Optimization with Non-Differentiable Constraints with Applications to Fairness, Recall, Churn, and Other Goals". In: *Journal of Machine Learning Research* (2019).

[16] Andrew Cotter et al. "Training Well-Generalizing Classifiers for Fairness Metrics and Other Data-Dependent Constraints". In: *Proceedings of the 36th International Conference on Machine Learning*. International Conference on Machine Learning. PMLR, 2019.

[17] Rachel Cummings et al. "On the compatibility of privacy and fairness". In: *Adjunct publication of the 27th conference on user modeling, adaptation and personalization*. 2019.

[18] Damek Davis and Dmitriy Drusvyatskiy. "Stochastic Model-Based Minimization of Weakly Convex Functions". In: *SIAM Journal on Optimization* 29 (2019).

[19] Frances Ding et al. "Retiring Adult: New Datasets for Fair Machine Learning". In: *Advances in Neural Information Processing Systems* 34 (2021).

[20] Vadym Doroshenko et al. "Connect the dots: Tighter discrete approximations of privacy loss distributions". In: *arXiv preprint arXiv:2207.04380* (2022).

[21] Cynthia Dwork and Aaron Roth. "The Algorithmic Foundations of Differential Privacy". In: *Foundations and Trends® in Theoretical Computer Science* (2014).

[22] Cynthia Dwork et al. "Calibrating Noise to Sensitivity in Private Data Analysis". In: *Theory of Cryptography*. 2006.

[23] Cynthia Dwork et al. "Fairness through awareness". In: *Proceedings of the 3rd innovations in theoretical computer science conference*. 2012.

[24] Maria S Esipova et al. "Disparate Impact in Differential Privacy from Gradient Misalignment". In: *The Eleventh International Conference on Learning Representations*. 2022.

[25] Tom Farrand et al. "Neither private nor fair: Impact of data imbalance on utility and fairness in differential privacy". In: *Proceedings of the 2020 Workshop on Privacy-Preserving Machine Learning in Practice*. 2020.

[26] Badih Ghazi et al. "Deep Learning with Label Differential Privacy". In: *Advances in Neural Information Processing Systems*. Vol. 34. 2021.

[27] Naman Goel, Mohammad Yaghini, and Boi Faltings. "Non-Discriminatory Machine Learning through Convex Fairness Criteria". In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 2018.

[28] Gabriel Goh et al. "Satisfying real-world goals with dataset constraints". In: *Advances in neural information processing systems* 29 (2016).

[29] Tomas Gonzalez, Cristobal Guzman, and Courtney Paquette. "Mirror Descent Algorithms with Nearly Dimension-Independent Rates for Differentially-Private Stochastic Saddle-Point Problems extended abstract". In: *Proceedings of Thirty Seventh Conference on Learning Theory*. Proceedings of Machine Learning Research. 2024.

[30] Sivakanth Gopi, Yin Tat Lee, and Lukas Wutschitz. "Numerical Composition of Differential Privacy". In: *Advances in Neural Information Processing Systems*. Vol. 34. 2021.

[31] Devansh Gupta. *Stochastic-Differentially-Private-and-Fair-Learning*. May 6, 2023. URL: `https://github.com/devanshgupta160/Stochastic-Differentially-Private-and-Fair-Learning` (visited on 09/27/2023).

[32] Moritz Hardt, Eric Price, and Nati Srebro. "Equality of opportunity in supervised learning". In: *Advances in neural information processing systems* 29 (2016).

[33] Martin Heusel et al. "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017.

[34] Hans Hofmann. *Statlog (German Credit Data)*. UCI Machine Learning Repository. 1994.

[35] Matthew Jagielski et al. "Differentially private fair learning". In: *International Conference on Machine Learning*. PMLR. 2019.

[36] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. "The composition theorem for differential privacy". In: *International conference on machine learning*. PMLR. 2015.

[37] Shiva Prasad Kasiviswanathan et al. "What can we learn privately?" In: *SIAM Journal on Computing* 40 (2011).

[38] Anastasia Koloskova, Hadrien Hendrikx, and Sebastian U Stich. "Revisiting Gradient Clipping: Stochastic bias and tight convergence guarantees". In: *International Conference on Machine Learning*. PMLR. 2023.

[39] Bogdan Kulynych et al. "What You See is What You Get: Principled Deep Learning via Distributional Generalization". In: *Advances in Neural Information Processing Systems*. 2022.

[40] Christian Janos Lebeda et al. "Avoiding Pitfalls for Privacy Accounting of Subsampled Mechanisms under Composition". In: *arXiv preprint arXiv:2405.20769* (2024).

[41] Tianyi Lin, Chi Jin, and Michael Jordan. "On Gradient Descent Ascent for Nonconvex-Concave Minimax Problems". In: *Proceedings of the 37th International Conference on Machine Learning*. International Conference on Machine Learning. PMLR, Nov. 21, 2020.

[42] Max Little. *Parkinsons*. UCI Machine Learning Repository. 2007.

[43] Andrew Lowy, Devansh Gupta, and Meisam Razaviyayn. "Stochastic Differentially Private and Fair Learning". In: *The Eleventh International Conference on Learning Representations*. 2023.

[44] Paul Mangold et al. “Differential Privacy has Bounded Impact on Fairness in Classification”. In: *ICML*. 2023.

[45] Ibomoiye Domor Mienye and Yanxia Sun. “Performance analysis of cost-sensitive learning methods with application to imbalanced medical data”. In: *Informatics in Medicine Unlocked* (2021).

[46] Hussein Mozannar, Mesrob Ohannessian, and Nathan Srebro. “Fair learning with private demographic data”. In: *International Conference on Machine Learning*. PMLR. 2020.

[47] A. Nemirovski et al. “Robust Stochastic Approximation Approach to Stochastic Programming”. In: (2009).

[48] Nicolas Papernot and Thomas Steinke. “Hyperparameter Tuning with Renyi Differential Privacy”. In: *International Conference on Learning Representations*. 2022.

[49] Reza Shokri et al. “Membership Inference Attacks against Machine Learning Models”. In: *2017 IEEE Symposium on Security and Privacy (SP)*. 2017.

[50] Shuang Song, Kamalika Chaudhuri, and Anand D. Sarwate. “Stochastic gradient descent with differentially private updates”. In: *2013 IEEE Global Conference on Signal and Information Processing*. 2013.

[51] Vinith M. Suriyakumar et al. “Chasing Your Long Tails: Differentially Private Prediction in Health Care Settings”. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 2021.

[52] Latanya Sweeney. “K-Anonymity: A Model for Protecting Privacy”. In: *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* (Oct. 1, 2002).

[53] Cuong Tran, My Dinh, and Ferdinando Fioretto. “Differentially Private Empirical Risk Minimization under the Fairness Lens”. In: *Advances in Neural Information Processing Systems*. Vol. 34. 2021.

[54] Cuong Tran and Ferdinando Fioretto. “On the Fairness Impacts of Private Ensembles Models”. In: *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*. 2023.

[55] Cuong Tran et al. “SF-PATE: Scalable, Fair, and Private Aggregation of Teacher Ensembles”. In: *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*. International Joint Conferences on Artificial Intelligence Organization, 2023.

[56] Arjun Wilkins et al. “Exact Privacy Analysis of the Gaussian Sparse Histogram Mechanism”. In: *Journal of Privacy and Confidentiality* 14.1 (2024).

[57] Zhenhuan Yang et al. “Differentially Private SGDA for Minimax Problems”. In: *The 38th Conference on Uncertainty in Artificial Intelligence*. 2022.

[58] I-Cheng Yeh. *Default of Credit Card Clients*. UCI Machine Learning Repository. 2009.

[59] Muhammad Bilal Zafar et al. “Fairness Constraints: Mechanisms for Fair Classification”. In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. 2017.

[60] Liang Zhang et al. “Bring Your Own Algorithm for Optimal Differentially Private Stochastic Minimax Optimization”. In: *Advances in Neural Information Processing Systems*. Vol. 35. Curran Associates, Inc., 2022.

A Application to Other Rate Constraints

A.1 Fairness Constraints

Fairness in machine learning aims to prevent models from making biased decisions based on sensitive attributes. We aim to train a classifier under fairness constraints by formulating a constrained optimization problem. We consider two popular group fairness [6] metrics: demographic parity and equality of odds. All group fairness measures can be formulated as rate-constraints [15], for individual fairness [23] it is easier to bound the per-sample contribution and privatize it with clipping and noising in the style of DP-SGD, thus a rate-constrained solution is not required, hence we focus on group fairness metrics.

Definition A.1 (Demographic Parity)**.** *A classifier $h(\theta;\cdot)$ satisfies demographic parity with respect to sensitive attribute $Z \in \mathcal{Z} = \{1, ..., |\mathcal{Z}|\}$ if the probability of predicting any class k is independent of Z:*

$$\Pr[\hat{Y} = k \mid Z = z] = \Pr[\hat{Y} = k], \quad \forall z \in \mathcal{Z}, \forall k \in \mathcal{Y},$$

where $\hat{Y} = h(\theta; x)$ is the predicted label.

In practice, we do not have access to the true probabilities, so it is common to estimate them by empirical prediction rates P_k. Using a slack parameter γ, this gives:

$$P_k(D[Z = z]; \theta) - P_k(D[Z \neq z]; \theta) \leq \gamma \quad \forall z \in \mathcal{Z}, \forall k \in \mathcal{Y} \tag{13}$$

Demographic parity thus leads to $J = |\mathcal{Z}| \cdot |\mathcal{Y}|$ rate constraints of the form specified in Eqn. (6). The global partition is of size $Q = |\mathcal{Z}|$ with elements $D_z = D[Z = z]$, $\forall z \in \mathcal{Z}$, and for the constraint corresponding to elements $z \in \mathcal{Z}$ and $y \in \mathcal{Y}$, we have $\mathcal{I} = \{\{z\}, [|\mathcal{Z}|] \setminus z\}$. The associated vector α has $\alpha_{\{z\},y} = 1$ and $\alpha_{\{[|\mathcal{Z}|]\setminus z\},y} = -1$, with the rest of the components set to 0.

Definition A.2 (Equality of Odds)**.** *A classifier $h(\theta;\cdot)$ satisfies equality of odds if the probability of predicting any class k is conditionally independent of the sensitive attribute Z given the ground truth:*

$$\Pr[\hat{Y} = k \mid Y = k', Z = z] = \Pr[Y = k', \hat{Y} = k, Z = z'], \quad \forall z', z \in \mathcal{Z}, \forall k, k' \in \mathcal{Y} \tag{14}$$

We note that the original notion of equalized odds is for binary sensitive attribute. For non-binary sensitive attributes, we can extend equalized odds to equalize rates between all subpopulations (as above), or we can consider a counter-factual definition of equalized odds:

$$\Pr[\hat{Y} = k \mid Y = k', Z = z] = \Pr[Y = k', \hat{Y} = k, Z \neq z], \quad \forall, z \in \mathcal{Z}, \forall k, k' \in \mathcal{Y} \tag{15}$$

In the above formulation, we seek to achieve equal odds for each subpopulation compared to other subpopulations combined (e.g. white vs. non-white, etc.). It is clear that in the binary sensitive attribute, the definitions are the same. Our framework can handle either variant by changing the adjusting the local partitioning (see Section 3) but we adopt the counter-factual definition.

We observe that the only difference between the equality of odds and demographic parity is the additional conditioning on the ground truth, which we will reflect as the additional predicate $Y = k'$ in our base rates to define the following constraint:

$$P_k(D[Y = k', Z = z]; \theta) - P_k(D[Y = k', Z = z']; \theta) \leq \gamma \quad \forall z \in \mathcal{Z}, \forall k, k' \in \mathcal{Y} \tag{16}$$

Equality of odds leads to $J = |\mathcal{Y}|^2 \times |\mathcal{Z}|$ number of constraints. With regards to implementing Eqn. (6), we can use a global partition with $|\mathcal{Y}| \times |\mathcal{Z}|$ where each element is the subset of D with some fixed ground truth label k and class z. The constraint for some $k \in [K]$ and $z \in \mathcal{Z}$ then has $\mathcal{I}$ which specifies the local partition $\{D[Y = k, Z = z], D[Y = k, Z \neq z]\}$ with the corresponding vector α having a $+1$ coefficient corresponding to a prediction rate of $D[Y = k', Z = z]$.

Many other group fairness constraints exist but they are all reducible to base rate constraints in a similar manner. Note the similarity between Equations (7) and (16), where the only difference is the additional conditioning on ground truth labels Y in equality of odds.

Objective	Formula	Number of Constraints
Demographic Parity	$P_k(D[Z = z]; \theta) - P_k(D[Z \neq z]; \theta) \leq \gamma$	$\forall k \in \mathcal{Y}$ (predicted), $\forall z \in \mathcal{Z}$ (sens.)
Equality of Odds	$P_k(D[Y = k', Z = z]; \theta) - P_k(D[Y = k', Z \neq z]; \theta) \leq \gamma$	$\forall k, k' \in \mathcal{Y}$ (predicted and g.t.) $\forall z \in \mathcal{Z}$ (sens. attr.)
False Negative Rate	$P_k(D[Y \neq k]; \theta) \leq \gamma$	$\forall k \in \mathcal{Y}$ (predicted)

Table 1: **Rate Constraints.** Given a dataset D, $C_k(D)$ is the prediction counts for class k, and $P_k(D) = C_k(D)/|D|$ is the prediction rate. $D[\texttt{pred}]$ indicates the subset of D where predicate `pred` is true, e.g., $D[Y = y, Z = z]$ is the subset of D with sensitive attribute (sen. attr.) $Z = z$ and ground truth (g.t.) labels $Y = y$.

A.2 False Negative Rate

Definition A.3. *(False Negative Rate (FNR)) A classifier's false negative rate (FNR) measures how often it incorrectly predicts negative for samples that are actually positive. More formally, a classifier satisfies a false negative rate constraint if*

$$P_k(D[Y \neq k]; \theta) \leq \gamma \quad \text{for } k \in [|\mathcal{Y}|] \tag{17}$$

Assuming the constraint is well-defined, FNR leads $J = |\mathcal{Y}|$ rate constraints of the form in Eqn. (6), with the global partition of size $Q = |Y|$ with elements $D_y = D[Y = y], y \in \mathcal{Y}$. For the constraint corresponding to a fixed $y \in \mathcal{Y}$, we have $\mathcal{I} = \{\mathcal{Y}/y\}$ with an associated $\alpha_{\{\mathcal{Y}/y\},y} = 1$.

B Proof of Theorem 4.1

Theorem B.1. *Let* $\sigma \geq 10 \max\left\{\frac{C\log(T/\delta)}{r|D|\epsilon}, \frac{C\sqrt{T}\log(T/\delta)}{|D|\epsilon}\right\}$ *and* $b \geq 2\max\left\{\frac{1}{\epsilon}, \frac{r\sqrt{T\log(T/\delta)}}{\epsilon}\right\}$*, then Algorithm 1 is* (ϵ, δ)*-DP.*

Proof. The ℓ_2-sensitivity of $\left(\sum_{x\in B^{(t)}} \text{clip}(g_{x,\theta}^{(t)}, \frac{C}{r|D|})\right)$ is clearly at most $\frac{C}{r|D|}$. Thus the standard guarantees of the Gaussian mechanism ensures $(\frac{1}{2}\epsilon_1, \frac{1}{2}\delta_1)$-DP w.r.t. the minibatch, where $\epsilon_1 \leq \min\{1, \frac{1}{r\sqrt{8T\log(1/\delta)}}\}$ and $\delta_1 = \frac{\delta}{2T}$. Similarly, because $\{D_1, ..., D_Q\}$ is a partition, the ℓ_1-sensitivity of the histogram is at most 1, and so the guarantees of the Laplace mechanism ensure $(\frac{1}{2}\epsilon_1, 0)$-DP w.r.t. to the minibatch. By composition, the combined mechanism is (ϵ_1, δ_1)-DP w.r.t. the minibatch. Since this mechanism acts a Poisson subsampled portion of the dataset and $\epsilon' \leq 1$, the privacy w.r.t. the overall dataset is $(\epsilon_2, \frac{1}{2}\delta_2)$ with $\epsilon_2 = r\epsilon_1 \leq \frac{1}{\sqrt{8T\log(1/\delta)}}$ and $\delta_2 \leq \frac{r\delta}{2T}$. Now applying advanced composition, the overall privacy of Algorithm 1 over T rounds is (ϵ_3, δ_3)-DP with $\epsilon_3 \leq \sqrt{8T\log(1/\delta)\epsilon_2} \leq \epsilon$ and $\delta_3 \leq (T+1)\delta_2 \leq \delta$. □

C Technical Lemmas

Lemma C.1. *Let* X *and* Y *be sums of* k_X *and* k_Y *zero-centered Laplace random variables with scale parameter* b*, respectively, and let* $\mu_X, \mu_Y > 0$*,* $\frac{\mu_x}{\mu_y} \leq 1$*,* $k_X \leq k_Y$*. For any* $\rho \in (0,1)$*, if* $\mu_Y \geq 4k_Y b \ln\left(\frac{1}{2\rho}\right)$ *then it holds that,*

$$P\left[\left|\frac{\mu_X + X}{\mu_Y + Y} - \frac{\mu_X}{\mu_Y}\right| < \frac{4k_Y b}{\mu_y}\ln\left(\frac{8}{\rho}\right)\right] \geq 1 - \rho. \tag{18}$$

Proof. We have,

$$\begin{aligned}
&P\left[\left|\frac{\mu_X + X}{\mu_Y + Y} - \frac{\mu_X}{\mu_Y}\right| < \epsilon\right] \\
&\geq P\left[\left|\frac{X}{\mu_Y + Y}\right| + \left|\frac{\mu_X Y}{\mu_Y(\mu_Y + Y)}\right| < \epsilon, \mu_Y + Y > \frac{\mu_Y}{2}\right] && \text{(triangle inequality)} \\
&\geq P\left[\left|\frac{2X}{\mu_Y}\right| + \left|\frac{2\mu_X Y}{\mu_Y^2}\right| < \epsilon, \mu_Y + Y > \frac{\mu_Y}{2}\right] && \text{(conditioning on } \mu_Y + Y > \frac{\mu_Y}{2}\text{)} \\
&\geq P\left[\left|\frac{2X}{\mu_Y}\right| \leq \frac{\epsilon}{2}, \left|\frac{2Y}{\mu_Y}\right| < \frac{\epsilon}{2}, \mu_Y + Y > \frac{\mu_Y}{2}\right] && \text{(using } \mu_x \leq \mu_y\text{)} \\
&\geq 1 - P\left[|X| \geq \frac{\mu_Y\epsilon}{4}\right] - P\left[|Y| \geq \frac{\mu_Y\epsilon}{4}\right] - P\left[Y \leq -\frac{\mu_Y}{2}\right] && \text{(Negation \& Union Bound)} \\
&\geq 1 - 2\exp(-\frac{\mu_Y\epsilon}{4k_Y b}) - 2\exp(-\frac{\mu_Y\epsilon}{4k_X b}) - \frac{1}{2}\exp(-\frac{\mu_Y}{k_Y b}) && \text{(concentration for Laplace R.Vs \& Laplace CDF)} \\
&\geq 1 - 2\exp(-\frac{\mu_Y\epsilon}{4k_Y b}) - 2\exp(-\frac{\mu_Y\epsilon}{4k_X b}) - \frac{1}{2}\exp(-\frac{\mu_Y}{k_Y b}) && (k_Y \geq k_X) \\
&\geq 1 - 4\exp(-\frac{\mu_Y\epsilon}{4k_Y b}) - \frac{1}{2}\exp(-\frac{\mu_Y}{k_Y b}) \\
&\geq 1 - 4\exp(-\frac{\mu_Y\epsilon}{4k_Y b}) - \frac{\rho}{2}
\end{aligned}$$

the last inequality uses that $\mu_Y \geq 4k_Y b\ln\left(\frac{1}{2\rho}\right)$. Now setting $\epsilon = \frac{4k_Y b}{\mu_y}\ln\left(\frac{8}{\rho}\right)$ yeilds,

$$P\left[\left|\frac{\mu_X + X}{\mu_Y + Y} - \frac{\mu_X}{\mu_Y}\right| < \frac{4k_Y b\ln(8/\rho)}{\mu_y}\right] \geq 1 - \rho. \tag{19}$$

□

Lemma C.2 (Error of sampled rates)**.** *Let* $p = \frac{\sum_{x_i \in X} x_i}{|X|}$ *with* $x_i \in [0,1]$ *and* $p_r = \frac{\sum_{x_i \in X_q} x_i}{r|X|}$ *where* X_r *is obtained by performing Poisson sampling on* X *with probability* r*. If* $|X|\, r \geq \log(1/\rho)$ *then (up to an order):*

$$P\left[|p - p_r| \leq \sqrt{\frac{\log(1/\rho)}{r|X|}}\right] \geq 1 - \rho \tag{20}$$

Proof. We have,

$$\begin{aligned}
&P\left[|p - p_r| \leq \epsilon\right] \\
&= P\left[\left|\frac{\sum_{X_i \in X} X_i}{|X|} - \frac{\sum_{X_i \in X} \text{Bern}(r) X_i}{r\,|X|}\right| \leq \epsilon\right] && (X_r \sim \text{Poisson}(X, r)) \\
&= P\left[\left|r \sum_{X_i \in X} X_i - \sum_{X_i \in X} \text{Bern}(r) X_i\right| \leq \epsilon r\,|X|\right] \\
&\geq 1 - \exp\left(-\frac{\epsilon^2 r^2\,|X|^2}{2\left(\text{Var}(\sum_{X_i \in X} \text{Bern}(r) X_i) + \epsilon r\,|X|\,/3\right)}\right) && \text{(Bernstein Ineq.)} \\
&\geq 1 - \exp\left(-\frac{\epsilon^2 r^2\,|X|}{2\left(r(1-r) + \epsilon r/3\right)}\right) && (X_i \in [0,1]; \text{Var}(\text{Bern}(r)) = r(1-r)) \\
&\geq 1 - \exp\left(-\frac{\epsilon^2 r\,|X|}{2\left(1/4 + \epsilon/3\right)}\right) && r(1-r) \leq 1/4 \\
\implies \epsilon &\geq 2 \max\left\{\frac{\log(1/\rho)}{r|X|}, \sqrt{\frac{\log(1/\rho)}{r|X|}}\right\} \\
&\geq 2\sqrt{\frac{\log(1/\rho)}{r|X|}} && \text{(using the assumption that } |X|r \geq \log(1/\rho))
\end{aligned}$$

□

D Missing Details from Section 5

D.1 Additional Background on Stationary Point Definition

For a smooth function $f : \theta \mapsto \mathbb{R}$, a standard notion of (first order) stationarity would involve bounding the norm of the gradient. However, for non-smooth functions, this notion does not accurately capture convergence. For example, if $f(\theta) = \|\theta\|$, a point may be arbitrarily close to the minimum, but still have gradient norm 1. To address this discrepancy, alternative notions of stationarity for non-smooth functions have been introduced, such as Definition 5.1. In the example where $f(\theta) = \|\theta\|$, this relaxation allows points which are *close* to the cusp at $\theta = \mathbf{0}$, whereas a bound on the gradient norm would allow *only* the point $\theta = \mathbf{0}$ for any non-trivial bound on the gradient. In fact, our convergence proof yields a slightly stronger notion of stationarity known as proximal near stationarity [18]. We elect to present Definition 5.1 as it requires less background information.

D.2 Proof of Theorem 5.2

In this section we will detail the proof of Theorem 5.2 and provide a more precise theorem statement. Before doing so, we will introduce some important notation.

For any $I \subseteq [Q]$, let $D_I = \bigcup_{i \in I} D_i$. Given a subset $B \subset D$, define $B_{\cap I} = B \cap (\bigcup_{i \in I} D_i)$ and recall that $\mathcal{I}_j$ and $\alpha_j \in \mathbb{R}^{|\mathcal{I}_j| \times K}$ denote the corresponding family of susbets of $[Q]$ and weight vector associated with constraint Γ_j. Let $n = \min_{q \in [Q]} \{|D_q|\}$.

Let $\|\Lambda\|_1$ be the ℓ_1 diameter of Λ. Let G_ℓ and G_h be the ℓ_2 Lipschitz constants w.r.t. θ of h and ℓ respectively. Similarly, let β_ℓ and β_h be the corresponding ℓ_2-smoothness constants. We recall the temperature parameter

of the softmax is denoted as τ. Let $c_1 = \max_{j\in[J]} \|\alpha_j\|$. Note that many rate constraint only compare two prediction rates, and so c_1 is typically at most 2. Define $\hat{\Phi}(\theta) = \min_{\theta'} \left\{\Phi(\theta') + \beta\|\theta - \theta'\|^2\right\}$ and $\hat{\Phi}_0 = \hat{\Phi}(\theta_0) - \min_\theta \left\{\hat{\Phi}(\theta)\right\}$ and $\mathcal{L}_0 = \mathcal{L}(\theta_0, \lambda_0) - \min_{\lambda,\theta} \{\mathcal{L}(\lambda, \theta)\}$ (see Section D.6 for more details on these quantities). We can now present the more complete version of convergence result.

Theorem D.1. *Assume* $n \geq \frac{1}{r}\max\{\ln\left(\frac{2c_1JT}{\rho}\right), 8\frac{Kb}{r}\ln\left(\frac{2T}{\rho}\right), 8\log(J|D|TK/\rho\}$. *Then, under appropriate choices of parameters, Algorithm 1 run without clipping is* (ϵ, δ)*-DP and with probability at least* $1-\rho$ *there exists* $t \in [T]$ *s.t.* θ_t *is an* $(\alpha, \alpha/[2\beta])$*-stationary point of* Φ *with*

$$\alpha = O\Bigg(\left(\left(\hat{\Phi}_0\beta G^2\right)^{1/4} + \sqrt{\beta\mathcal{L}_0} + G\right)\left(\left(\frac{\sqrt{d}\log(n/\delta)\sqrt{\log(JKn/\rho)}}{n\epsilon}\right)^{1/3} + \frac{K^{1/4}\sqrt{\beta\|\Lambda\|_1}\sqrt{\log(n/\delta)}(\log(JKn/\rho))^{1/4}}{(n\epsilon)^{1/4}}\right)\Bigg),$$

where $G = G_\ell + c\tau G_h\|\Lambda\|_1$ *and* $\beta = \beta_\ell + 2c\tau \cdot \max\left\{G_h\sqrt{J}, \|\Lambda\|_1(2G_h + \tau\beta_h)\right\}$

Proving this statement will involve several major steps. First, in Section D.4 we derive the necessary noise levels needed to ensure that Algorithm 1 is private. Second, in Section D.5 we bound the error in the gradients at each time step. Next, in Section D.6 we give a general convergence rate for SGDA under the condition that the gradients have bounded error. Finally, we derive the overall Lipschitz and smoothness constants of $\mathcal{L}$ based on the base smoothness and Lipschitz constants in Section D.7. These results are then combined in Section D.3 to obtain the final result.

In one final remark, we note the following fact will be used in several places.

Lemma D.2. *Let* $n \geq 4\log(J|D|/\rho)$ *and* $t \in [T]$. *With probability at least* $1-\rho$ *it holds for every* $j \in [J]$ *and* $I \in \mathcal{I}_j$ *that* $|B_{\cap I}^{(t)}| \geq \frac{1}{2}r|I|n$.

Proof. By Lemma C.2 we have for any $j \in [J]$ and $I \in \mathcal{I}_j$,

$$P\left[r|D_I| - |B_{\cap I}^{(t)}| \geq \sqrt{r|D_I|\log(1/\gamma)}\right] \leq \gamma.$$

Thus since $|D_I| \geq n \geq \log(J|D|/\rho)$, it holds with probability at least $1-\rho$, for every $j \in [J]$ and $I \in \mathcal{I}_j$ that $|B_{\cap I}| \geq r|D_I| - \sqrt{r|D_I|\log(1/\rho)} \geq 0.5r|D_I| \geq 0.5r|I|n$. □

D.3 Proof of Theorem D.1

With the previously established results, we can now verify a setting of parameters which proves the theorem statement. Specifically, we set,

$$T = \min\left\{\left(\frac{n\epsilon}{\sqrt{d}}\right)^{4/3}, \frac{n\epsilon}{K}\right\}, \qquad \sigma = \frac{G\sqrt{T\log(T/\delta)}}{n\epsilon}, \qquad b = \frac{r\sqrt{T\log(T/\delta)}}{\epsilon}.$$

Note that by Theorem D.3, this ensures that Algorithm 1 is (ϵ, δ)-DP so long as $r \geq \frac{1}{\sqrt{T}}$.

Using Lemma D.4 can now instantiate Theorem D.5. For τ_θ we have,

$$\tau_\lambda = O\left(\frac{c_1 K b \log(JKn/\rho)}{rn} + c_1\sqrt{\frac{\log\left(\frac{JKn}{\rho}\right)}{rn}}\right)$$
$$= O\left(\frac{c_1 K\sqrt{T}\log(T/\delta)\log(JKn/\rho)}{n\epsilon} + c_1\sqrt{\frac{\log\left(\frac{JKn}{\rho}\right)}{rn}}\right).$$

Setting τ_λ as the quantity above, we can write the bound on τ_θ as,

$$\tau_\theta = O\left(\sqrt{d}\sigma\sqrt{\log(4T/\rho)} + \frac{4G_\ell\sqrt{\log(4/\rho)}}{\sqrt{r|D|}} + \frac{4G_\ell\sqrt{\log(4/\rho)}}{\sqrt{r|D|}} + |\Lambda\|_1\tau G_h\tau_\theta\right)$$
$$= O\left(G\left(\frac{\sqrt{dT\log(T/\rho)}\log(T/\delta)}{n\epsilon} + \frac{\sqrt{\log(1/\rho)}}{\sqrt{r|D|}}\right) + \|\Lambda\|_1\tau G_h\tau_\theta\right).$$

Now, in the non-trivial regime where $\tau_\theta \le G$, Theorem D.5 implies that for r large enough,

$$\alpha = O\left(\frac{\left(\hat{\Phi}_0\beta(G^2+\tau_\theta^2)\right)^{1/4}}{T^{1/4}} + \tau_\theta + \sqrt{\beta\|\Lambda\|_1\tau_\lambda} + \frac{\sqrt{\beta\mathcal{L}_0}}{\sqrt{T}}\right)$$
$$= O\Bigg(\left(\left(\hat{\Phi}_0\beta G^2\right)^{1/4} + \sqrt{\beta\mathcal{L}_0} + G\right)\left(\left(\frac{\sqrt{d}\log(T/\delta)\sqrt{\log(JKn/\rho)}}{n\epsilon}\right)^{1/3} + \frac{K^{1/4}\sqrt{\beta\|\Lambda\|_1}\sqrt{\log(T/\delta)}(\log(JKn/\rho))^{1/4}}{(n\epsilon)^{1/4}}\right)\Bigg).$$

D.4 Privacy of Algorithm 1 under Lipschitzness

Theorem D.3. *Assume h and ℓ are G_ℓ and G_h Lipschitz. Then for some universal constant c and $\sigma \ge c(c_1\|\Lambda\|_1\tau G_h + G_\ell)\max\left\{\frac{\log(T/\delta)}{rn\epsilon}, \frac{C\sqrt{T}\log(T/\delta)}{n\epsilon}\right\}$ and $b \ge 2\max\left\{\frac{1}{\epsilon}, \frac{r\sqrt{T\log(T/\delta)}}{\epsilon}\right\}$, then Algorithm 1 is $(\epsilon, 3\delta)$-DP.*

Proof. First, by Lemma D.2 and the conditions of Theorem D.1, probability at least $1-\delta$, for every $t \in [T]$, $j \in [J]$ and $I \in [I]$ it holds that $|B_{\cap I}| \ge 0.5r|I|\cdot n$. Consequently, the concentration of Laplace noise and the conditions of Theorem D.1 imply $\sum_{i\in I}\sum_{k\in[K]}\hat{H}_{i,k}^{(t)} \ge 0.25rn$ with probability at least $1-2\delta$. Conditional on this event, the ℓ_2-sensitivity of $\left(\sum_{x\in B^{(t)}} g_{x,\theta}^{(t)}\right)$ is at most $\frac{G_\ell}{r|D|} + \frac{c_1\tau G_h}{0.25rn}$ since then,

$$\|\nabla\hat{R}(\theta,\lambda;H,x)\| \le \sum_{j=1}^{J}\sum_{I\in\mathcal{I}_j}\sum_{k\in K}\frac{\lambda_j\alpha_{j,I,k}}{0.25rn}\|\nabla_\theta[\sigma_\tau(h(\theta;x))_k]\| \le \frac{c_1\|\Lambda\|\tau G_h}{0.25rn}.$$

Thus, the scale of Gaussian noise implies that the releasing the primal gradient is $(\frac{1}{2}\epsilon_1, \frac{1}{2}\delta_1)$-DP w.r.t. the minibatch, where $\epsilon_1 \le \min\{1, \frac{1}{r\sqrt{T\log(1/\delta)}}\}$ and $\delta_1 = \frac{\delta}{T}$. From here one can follow the same steps as in the proof of Theorem 4.1 to obtain an overall privacy of (ϵ,δ)-DP conditional on the previously mentioned event that each $B_{\cap I}$ is large. Since this event happens with probability at least $1-2\delta$, we obtain a final overall privacy guarantee of $(\epsilon, 3\delta)$-DP. □

D.5 Bounding Gradient Error

Lemma D.4. *Let $\rho \in [0,1]$ and $t \in [T]$. Under the assumptions of Theorem D.1, conditional on $\theta^{(t)}, \lambda^{(t)}$, it holds with probability at least $1-2\rho$ that,*

$$\|g_\theta^{(t)} - \nabla_\theta \mathcal{L}(\theta^{(t)}, \lambda^{(t)})\|_2 \le \sigma\sqrt{d\log(4/\rho)} + \frac{4G_\ell\sqrt{\log(4/\rho)}}{\sqrt{r|D|}} + \|\Lambda\|_1 \tau G_h \left(\frac{8c_1 K b \log(64JKn/\rho)}{rn} + c_1\sqrt{\frac{\log\left(\frac{8JKn}{\rho}\right)}{rn}} \right),$$

$$\|g_\lambda^{(t)} - \nabla_\lambda \mathcal{L}(\theta^{(t)}, \lambda^{(t)})\|_\infty \le \frac{8c_1 K b \log(64JKn/\rho)}{rn} + c_1\sqrt{\frac{\log\left(\frac{8JKn}{\rho}\right)}{rn}}.$$

Proof. We will bound each error term separately.

Error of the Dual Gradient. We start with the following bound,

$$P\left[\|\hat{g}_\lambda^{(t)} - \nabla_\lambda \mathcal{L}(\theta^{(t)}, \lambda^{(t)})\|_\infty \ge \epsilon\right] = P\left[\max_j \left\{ \left| \sum_{I\in\mathcal{J}_j}\sum_{k=1}^K \alpha_{j,I,k} P_k(D_I) - \sum_{I\in\mathcal{J}_j}\sum_{k_1\in K} \frac{\alpha_{j,I,k_1}\sum_{i\in I}\hat{H}_{i,k}^{(t)}}{\sum_{i\in I}\sum_{k_2\in K}\hat{H}_{i,k_2}^{(t)}} \right| \right\} \ge \epsilon \right]$$
$$\le \sum_j \sum_{I\in\mathcal{J}_j}\sum_{k=1}^K \mathbb{1}\left[|\alpha_{j,I,k} \ne 0|\right] P\left[\left| P_k(D_I) - \frac{\sum_{i\in I}\hat{H}_{i,k_1}^{(t)}}{\sum_{i\in I}\sum_{k_2\in K}\hat{H}_{i,k_2}^{(t)}} \right| \ge \frac{\epsilon}{c_1} \right]. \tag{21}$$

We thus have for any $\epsilon_1 + \epsilon_2 = \epsilon$ that,

$$P\left[\left| P_k(D_I) - \frac{\sum_{i\in I}\hat{H}_{i,k_1}^{(t)}}{\sum_{i\in I}\sum_{k_2\in K}\hat{H}_{i,k_2}^{(t)}} \right| \ge \frac{\epsilon}{c_1} \right] \le P\left[|P_k(D_I) + P_k(B_{\cap I})| + \left| P_k(B_{\cap I}) - \frac{\sum_{i\in I}\hat{H}_{i,k_1}^{(t)}}{\sum_{i\in I}\sum_{k_2\in K}\hat{H}_{i,k_2}^{(t)}} \right| \ge \frac{\epsilon}{c_1} \right]$$
$$\le 1 - P\left[|P_k(D_I) + P_k(B_{\cap I})| + \left| P_k(B_{\cap I}) - \frac{\sum_{i\in I}\hat{H}_{i,k_1}^{(t)}}{\sum_{i\in I}\sum_{k_2\in K}\hat{H}_{i,k_2}^{(t)}} \right| \le \frac{\epsilon}{c_1} \right]$$
$$\le 1 - P\left[|P_k(D_I) + P_k(B_{\cap I})| \le \frac{\epsilon_1}{c_1}, \left| P_k(B_{\cap I}) - \frac{\sum_{i\in I}\hat{H}_{i,k_1}^{(t)}}{\sum_{i\in I}\sum_{k_2\in K}\hat{H}_{i,k_2}^{(t)}} \right| \le \frac{\epsilon_2}{c_1} \right]$$
$$\le P\left[|P_k(D_I) + P_k(B_{\cap I})| \ge \frac{\epsilon_1}{c_1} \right] + P\left[\left| P_k(B_{\cap I}) - \frac{\sum_{i\in I}\hat{H}_{i,k_1}^{(t)}}{\sum_{i\in I}\sum_{k_2\in K}\hat{H}_{i,k_2}^{(t)}} \right| \ge \frac{\epsilon_2}{c_1} \right].$$

We will start by bounding $P\left[\left| P_k(B_{\cap I}) - \frac{\sum_{i\in I}\hat{H}_{i,k}^{(t)}}{\sum_{i\in I}\sum_{k_2\in K}\hat{H}_{i,k_2}^{(t)}} \right| \ge \frac{\epsilon_1}{c_1}\right]$ for any fixed I and k. Observe that conditional on $|B_{\cap I}|$, the sampling process is equivalent to drawing $|B_{\cap I}|$ samples uniformly at random from $|D_I|$ without replacement. Therefore by Lemma C.1 we have,

$$P\left[\left| P_k(B_{\cap I}) - \frac{\sum_{i\in I}\hat{H}_{i,k}^{(t)}}{\sum_{i\in I}\sum_{k_2\in K}\hat{H}_{i,k_2}^{(t)}} \right| \ge \frac{4K|I|b\log(16JKn/\rho)}{|B_{\cap I}|} \middle| |B_{\cap I}| \right] \le \frac{\rho}{4JKn}.$$

Now by Lemma D.2, $P\left[|B_{\cap}I| \leq \frac{r}{2}|I|n\right] \leq \frac{\rho}{4JKn}$, and so,

$$P\left[\left|P_k(B_{\cap I}) - \frac{\sum_{i\in I}\hat{H}^{(t)}_{i,k}}{\sum_{i\in I}\sum_{k_2\in K}\hat{H}^{(t)}_{i,k_2}}\right| \geq \frac{8Kb\log(16JKn/\rho)}{rn}\right] \leq \frac{\rho}{2JKn}.$$

Thus it suffices to have $\epsilon_1 = \frac{8c_1 b\log(16JKn/\rho)}{rn}$.

Looking now at the statistical error term and applying Lemma C.2 we obtain:

$$P\left[|P_k(B_{\cap I}) - P_k(D_I)| \geq \sqrt{\frac{\log\left(\frac{2JKn}{\rho}\right)}{r|D_I|}}\right] \leq \frac{\rho}{2JKn}. \tag{22}$$

Observing that $\sqrt{\frac{\log\left(\frac{2JKn}{\rho}\right)}{r|D_I|}} \leq \sqrt{\frac{\log\left(\frac{2JKn}{\rho}\right)}{rn}}$, one can see it suffices for $\epsilon_2 = c_1\sqrt{\frac{\log\left(\frac{2JKn}{\rho}\right)}{rn}}$.

Plugging $\epsilon = \epsilon_1 + \epsilon_2$ back into Eqn. (21) we obtain

$$P\left[\|\hat{g}^{(t)}_{\lambda} - \nabla_{\lambda}\mathcal{L}(\theta^{(t)},\lambda^{(t)})\|_{\infty} \geq \frac{8c_1 b\log(16JKn/\rho)}{rn} + c_1\sqrt{\frac{\log\left(\frac{2JKn}{\rho}\right)}{rn}}\right]$$
$$\leq \sum_{j}^{J}\sum_{I\in\mathcal{J}_j}\sum_{k=1}^{K}\mathbb{1}\left[|\alpha_{j,I,k}\neq 0|\right]\frac{\rho}{JKn} \leq \rho.$$

This proves the claim.

Error in Primal Gradient. First observe that

$$\|\hat{g}_{\theta} - \nabla_{\theta}\mathcal{L}(\theta^{(t)},\lambda^{(t)})\| \leq \|Z^{(t)}\| + \left\|\nabla_{\theta}\ell(\theta^{(t)}) - \frac{1}{r|D|}\sum_{x\in B^{(t)}}\nabla_{\theta}\ell(\theta^{(t)};x)\right\|$$
$$+ \left\|\nabla_{\theta}R(\theta^{(t)},\lambda^{(t)}) - \sum_{x\in B^{(t)}}\hat{R}(\theta^{(t)},\lambda^{(t)};\hat{H},x)\right\|$$

For $\hat{g}_{\theta}\in\mathbb{R}^d$, using the concentration of Gaussian noise we obtain,

$$P[\|Z^{(t)}\| \geq \sigma\sqrt{d\log(4/\rho)}] \leq \frac{\rho}{4}.$$

For the second term, by Bernstein's inequality we have

$$P\left[\left\|\left(\sum_{i\in B^{(t)}}\nabla_{\theta}\ell(\theta,\lambda;x_i)\right) - r|D|\nabla_{\theta}\ell(\theta_t,\lambda_t)\right\| \geq \alpha\right] \leq \exp\left(-\frac{\alpha^2/2}{\alpha G_{\ell}/3 + |D|G_{\ell}^2\max\{r,1/2\}}\right).$$

Thus with probability at least $1-\frac{\rho}{4}$ one has that,

$$\left\|\left(\frac{1}{r|D|}\sum_{i\in B^{(t)}}\nabla_{\theta}\ell(\theta,\lambda,x_i)\right) - \nabla_{\theta}\ell(\theta_t,\lambda_t)\right\| \leq 4\max\left\{\frac{G_{\ell}\sqrt{\log(4/\rho)}}{\sqrt{r|D|}}, \frac{G_{\ell}\log(4/\rho)}{r|D|}\right\}.$$

And so if $r|D| \geq \log(4/\rho)$ we have with probability at least $1 - \rho/4$ that,

$$\left\|\nabla_\theta \ell(\theta^{(t)}) - \frac{1}{r|D|}\sum_{x\in B^{(t)}} \nabla_\theta \ell(\theta^{(t)}; x)\right\| \leq \frac{4G_\ell\sqrt{\log(4/\rho)}}{\sqrt{r|D|}}.$$

For the regularizer we have,

$$\begin{aligned}
&\left\|\nabla_\theta R(\theta^{(t)}, \lambda^{(t)}) - \sum_{x\in B^{(t)}} \hat{R}(\theta^{(t)}, \lambda^{(t)}; \hat{H}, x)\right\| \\
&\leq \left\|\sum_{x\in D}\sum_{j=1}^{J}\sum_{I\in\mathcal{I}_j}\sum_{k_1\in[K]} \frac{\lambda_j\ \alpha_{j,I,k_1}\ \mathbb{1}_{[x\in D_I]}}{\sum_{i\in I}\sum_{k_2\in[K]} H^{(t)}_{i,k_2}}\nabla_\theta[\sigma_\tau(h(\theta;x))_{k_1}] - \sum_{x\in D}\sum_{j=1}^{J}\sum_{I\in\mathcal{I}_j}\sum_{k_1\in[K]} \frac{\lambda_j\ \alpha_{j,I,k_1}\ \mathbb{1}_{[x\in B_{\cap I}]}}{\sum_{i\in I}\sum_{k_2\in[K]} \hat{H}^{(t)}_{i,k_2}}\nabla_\theta[\sigma_\tau(h(\theta;x))_{k_1}]\right\| \\
&\leq \left|\sum_{x\in D}\sum_{j=1}^{J}\sum_{I\in\mathcal{I}_j}\sum_{k_1\in[K]} \frac{\lambda_j\ \alpha_{j,I,k_1}\ \mathbb{1}_{[x\in D_I]}}{\sum_{i\in I}\sum_{k_2\in[K]} H^{(t)}_{i,k_2}} - \sum_{x\in D}\sum_{j=1}^{J}\sum_{I\in\mathcal{I}_j}\sum_{k_1\in[K]} \frac{\lambda_j\ \alpha_{j,I,k_1}\ \mathbb{1}_{[x\in B_{\cap I}]}}{\sum_{i\in I}\sum_{k_2\in[K]} \hat{H}^{(t)}_{i,k_2}}\right|\tau G_h \\
&\leq \left|\sum_{j=1}^{J}\lambda_j\sum_{I\in\mathcal{I}_j}\sum_{k_1\in[K]}\alpha_{j,I,k_1}\left(\frac{|D_I|}{\sum_{i\in I}\sum_{k_2\in[K]} H^{(t)}_{i,k_2}} - \frac{|B_{\cap I}|}{\sum_{i\in I}\sum_{k_2\in[K]} \hat{H}^{(t)}_{i,k_2}}\right)\right|\tau G_h \\
&\leq \left(\sum_{j=1}^{J}\lambda_j\left|\sum_{I\in\mathcal{I}_j}\sum_{k_1\in[K]} \frac{\alpha_{j,I,k_1}\ |D_I|}{\sum_{i\in I}\sum_{k_2\in[K]} H^{(t)}_{i,k_2}} - \sum_{I\in\mathcal{I}_j}\sum_{k_1\in[K]} \frac{\alpha_{j,I,k_1}\ |B_{\cap I}|}{\sum_{i\in I}\sum_{k_2\in[K]} \hat{H}^{(t)}_{i,k_2}}\right|\right)\tau G_h
\end{aligned}$$

The term inside the absolute value can be bounded using the same analysis used in bounding the dual gradient error. Thus we have with probability at least $1 - \frac{\rho}{4}$ that,

$$\left\|\nabla_\theta R(\theta^{(t)}, \lambda^{(t)}) - \sum_{x\in B^{(t)}} \hat{R}(\theta^{(t)}, \lambda^{(t)}; \hat{H}, x)\right\| \leq \|\Lambda\|_1\tau G_h\left(\frac{8c_1 Kb\log(64JKn/\rho)}{rn} + c_1\sqrt{\frac{\log\left(\frac{8JKn}{\rho}\right)}{rn}}\right).$$

Combining the above bounds yields the claimed bound on the primal gradient error.

□

D.6 Convergence of SGDA

The overall structure of our convergence proof is similar to that of [41], but with several significant modifications. Most significantly, our proof explicitly leverages the linear structure of the dual to improve the convergence rate for our application. This linear structure also allows our analysis to depend on an $\|\cdot\|_\infty$ bound on the gradient error when Λ has bounded $\|\cdot\|_1$ diameter. This in contrast to previously existing analysis which depend on the $\|\cdot\|_2$ error of the dual gradient, which could be much worse in our case due to the noise added for privacy. Separately, our analysis also differs from [41] in that it accounts for potential bias in the gradient estimates and tracks the disparate impact the scale of noise in g_θ and g_λ may have on the convergence rate.

In order to present our proof, we start with some necessary preliminaries. Let $\Phi(\theta) = \max_{\lambda\in\Lambda}\{\mathcal{L}(\theta,\lambda)\}$. Let $\hat{\Phi}$ denote the Moreau envelope of Φ with parameter 2β. That is, $\hat{\Phi}(\theta) = \min_{\theta'}\left\{\Phi(\theta') + \beta\|\theta - \theta'\|^2\right\}$. Let $\Delta^{(t)} = \Phi(\theta^{(t)}) - \mathcal{L}(\theta^{(t)}, \lambda^{(t)})$ for all $t \in \{0, ..., T\}$. Further, we define $\lambda^*(\theta) = \arg\max_{\lambda\in\Lambda}\{\mathcal{L}(\theta,\lambda)\}$. We denote the proximal operator of a function f as $\text{prox}_f(\theta) = \arg\min_\theta\left\{f(\theta') + \frac{1}{2}\|\theta - \theta'\|^2\right\}$. It is known that under the condition that f is β-smooth and Λ is bounded, that $\hat{\Phi}$ is differentiable with $\nabla\hat{\Phi}(\theta) = 2\beta(\theta - \text{prox}_{\Phi/[2\beta]}(\theta))$, and that any point θ for which which $\|\nabla\hat{\Phi}(\theta)\| \leq \alpha$ is an $(\alpha, \alpha/[2/\beta])$-stationary point with respect to Definition 5.1; see Lin, Jin, and Jordan [41, Lemma 3.8]. Also, under these conditions, Φ is G-Lipschitz. We defer the reader towards [41] for more details on these statements.

We present the following statement, which gives a convergence rate for Algorithm 1 in terms of the amount of noise added.

Theorem D.5. *Define* $\hat{\Phi}_0 = \hat{\Phi}(\theta_0) - \min_\theta \left\{\hat{\Phi}(\theta)\right\}$ *and* $\mathcal{L}_0 = \mathcal{L}(\theta_0, \lambda_0) - \min_{\lambda,\theta}\{\mathcal{L}(\lambda, \theta)\}$. *Assume* $\mathcal{L}(\cdot,\cdot)$ *is* β*-smooth,* $\mathcal{L}(\cdot, \lambda)$ *is* G*-Lipschitz for all* $\lambda \in \Lambda$*, and* $\mathcal{L}(\theta, \cdot)$ *is linear for all* $\theta \in \mathbb{R}^d$. *Conditional on the event that for all* $t \in \{0, ..., T-1\}$, $\|g_{\theta,t} - \nabla_\theta \mathcal{L}(\theta^{(t)}, \lambda^{(t)})\|_2 \leq \tau_\theta$ *and* $\|g_\lambda^{(t)} - \nabla_\lambda \mathcal{L}(\theta^{(t)}, \lambda^{(t)})\|_\infty \leq \tau_\lambda$, *when Algorithm 1 is run with* $\eta_\lambda \geq \left(\frac{\eta_\theta G(G+\tau_\theta)}{\tau_\lambda}\right)$ *and* $\eta_\theta = \sqrt{\frac{\hat{\Phi}_0}{2T\beta(G^2+\tau_\theta^2)}}$ *there exists* $t \in \{0, ..., T-1\}$ *such that* θ_t *is an* $(\alpha, \alpha/[2\beta])$*-stationary point with*

$$\alpha = O\left(\frac{\left(\hat{\Phi}_0\beta(G^2+\tau_\theta^2)\right)^{1/4}}{T^{1/4}} + \tau_\theta + \sqrt{\beta\|\Lambda\|_1\tau_\lambda} + \frac{\sqrt{\beta\mathcal{L}_0}}{\sqrt{T}}\right).$$

We will prove this statement by showing that Algorithm 1 finds a point where the gradient of $\hat{\Phi}$ is small. Note this is sufficient as Lin, Jin, and Jordan [41, Lemma 3.8] implies that a point, θ, for which $\|\nabla\hat{\Phi}(\theta)\| \leq \alpha$ is an $(\alpha, \alpha/[2\beta])$-stationary point with respect to Definition 5.1.

We will break the majority of the proof into three distinct lemmas. The first lemma gives a bound on the decrease in $\hat{\Phi}$.

Lemma D.6. *Under the assumptions of Theorem D.5, the iterates of Algorithm 1 satisfy for any* $t \in [T-1]$,

$$\hat{\Phi}(\theta^{(t)}) - \hat{\Phi}(\theta^{(t-1)}) \leq -\frac{\eta_\theta}{4}\|\nabla\hat{\Phi}(\theta^{(t-1)})\|^2 + 2\beta\eta_\theta\Delta^{(t-1)} + 2\beta\eta_\theta^2(G^2+\tau_\theta^2) + 2\eta_\theta\tau_\theta^2. \tag{23}$$

Proof. Let $\hat{\theta}^{(t-1)} = \text{prox}_{\hat{\Phi}}(\theta^{(t-1)})$. By the definition of the Moreau envelope we have

$$\hat{\Phi}(\theta^{(t)}) \leq \Phi(\hat{\theta}^{(t-1)}) + \beta\|\hat{\theta}^{(t-1)} + \theta^{(t)}\|^2. \tag{24}$$

Using the update rule we have

$$\begin{aligned}
\|\hat{\theta}^{(t-1)} - \theta^{(t)}\|^2 &= \|\hat{\theta}^{(t-1)} - \theta^{(t-1)} + \eta_\theta g_\theta^{(t-1)}\|^2 \\
&= \|\hat{\theta}^{(t-1)} - \theta^{(t-1)}\|^2 + 2\eta_\theta\left\langle\hat{\theta}^{(t-1)} - \theta^{(t-1)}, g_\theta^{(t-1)}\right\rangle + \eta_\theta^2\|g_\theta^{(t-1)}\|^2 \\
&\leq \|\hat{\theta}^{(t-1)} - \theta^{(t-1)}\|^2 + 2\eta_\theta\left\langle\hat{\theta}^{(t-1)} - \theta^{(t-1)}, \nabla_\theta\mathcal{L}(\theta^{(t-1)}, \lambda^{(t-1)})\right\rangle \\
&\quad + 2\eta_\theta\left\langle\hat{\theta}^{(t-1)} - \theta^{(t-1)}, g_\theta^{(t-1)} - \nabla_\theta\mathcal{L}(\theta^{(t-1)}, \lambda^{(t-1)})\right\rangle + 2\eta_\theta^2(G^2+\tau_\theta^2)
\end{aligned}$$

Plugging this back into Eqn. (24) and using the definition of the Moreau envelope we obtain,

$$\begin{aligned}
\hat{\Phi}(\theta^{(t)}) \leq \hat{\Phi}(\hat{\theta}^{(t-1)}) &+ 2\beta\eta_\theta\left\langle\hat{\theta}^{(t-1)} - \theta^{(t-1)}, \nabla_\theta\mathcal{L}(\theta^{(t-1)}, \lambda^{(t-1)})\right\rangle \\
&+ 2\beta\eta_\theta\left\langle\hat{\theta}^{(t-1)} - \theta^{(t-1)}, g_\theta^{(t-1)} - \nabla_\theta\mathcal{L}(\theta^{(t-1)}, \lambda^{(t-1)})\right\rangle + 2\beta\eta_\theta^2(G^2+\tau_\theta^2)
\end{aligned} \tag{25}$$

By way of bounding the third term on the RHS above, we use Young's inequality to derive,

$$2\beta\eta_\theta\left\langle\hat{\theta}^{(t-1)} - \theta^{(t-1)}, g_\theta^{(t-1)} - \nabla_\theta\mathcal{L}(\theta^{(t-1)}, \lambda^{(t-1)})\right\rangle \leq \frac{\beta^2\eta_\theta}{2}\|\hat{\theta}^{(t-1)} - \theta^{(t-1)}\|^2 + 2\eta_\theta\|g_\theta^{(t-1)} - \nabla_\theta\mathcal{L}(\theta^{(t-1)}, \lambda^{(t-1)})\|^2.$$

Plugging this into the above, we now have the following derivation,

$$
\begin{aligned}
&\hat{\Phi}(\theta^{(t)}) - \hat{\Phi}(\hat{\theta}^{(t-1)}) \\
&\leq 2\beta\eta_\theta \left\langle \hat{\theta}^{(t-1)} - \theta^{(t-1)}, \nabla_\theta \mathcal{L}(\theta^{(t-1)}, \lambda^{(t-1)}) \right\rangle + \frac{\beta^2\eta_\theta}{2}\|\hat{\theta}^{(t-1)} - \theta^{(t-1)}\|^2 + 2\beta\eta_\theta^2(G^2 + \tau_\theta^2) + 2\eta_\theta\tau_\theta^2 \\
&\overset{(i)}{\leq} 2\beta\eta_\theta \left(\mathcal{L}(\hat{\theta}^{(t-1)}, \lambda^{(t-1)}) - \mathcal{L}(\theta^{(t-1)}, \lambda^{(t-1)}) + \frac{3\beta}{4}\|\hat{\theta}^{(t-1)} - \theta^{(t-1)}\|^2 \right) + 2\beta\eta_\theta^2(G^2 + \tau_\theta^2) + 2\eta_\theta\tau_\theta^2 \\
&\leq 2\beta\eta_\theta \left(\Phi(\hat{\theta}^{(t-1)}) - \mathcal{L}(\theta^{(t-1)}, \lambda^{(t-1)}) + \frac{3\beta}{4}\|\hat{\theta}^{(t-1)} - \theta^{(t-1)}\|^2 \right) + 2\beta\eta_\theta^2(G^2 + \tau_\theta^2) + 2\eta_\theta\tau_\theta^2 \\
&= 2\beta\eta_\theta \left(\Phi(\hat{\theta}^{(t-1)}) + \Phi(\theta^{(t-1)}) - \Phi(\theta^{(t-1)}) - \mathcal{L}(\theta^{(t-1)}, \lambda^{(t-1)}) + \frac{3\beta}{4}\|\hat{\theta}^{(t-1)} - \theta^{(t-1)}\|^2 \right) \\
&\quad + 2\beta\eta_\theta^2(G^2 + \tau_\theta^2) + 2\eta_\theta\tau_\theta^2 \\
&\overset{(ii)}{\leq} 2\beta\eta_\theta \left(-\beta\|\hat{\theta}^{(t-1)} - \theta^{(t-1)}\|^2 + \Delta^{(t-1)} + \frac{3\beta}{4}\|\hat{\theta}^{(t-1)} - \theta^{(t-1)}\|^2 \right) + 2\beta\eta_\theta^2(G^2 + \tau_\theta^2) + 2\eta_\theta\tau_\theta^2 \\
&= -\frac{\eta_\theta\beta^2}{2}\|\hat{\theta}^{(t-1)} - \theta^{(t-1)}\|^2 + 2\beta\eta_\theta\Delta^{(t-1)} + 2\beta\eta_\theta^2(G^2 + \tau_\theta^2) + 2\eta_\theta\tau_\theta^2 \\
&\overset{(iii)}{=} -\frac{\eta_\theta}{2}\|\nabla\hat{\Phi}((\theta^{(t-1)})\|^2 + 2\beta\eta_\theta\Delta^{(t-1)} + 2\beta\eta_\theta^2(G^2 + \tau_\theta^2) + 2\eta_\theta\tau_\theta^2
\end{aligned}
$$

Above, (i) uses the fact that $\hat{\theta}^{(t-1)}$ is generated by the proximal operator. Inequality (ii) uses the fact the definitions of the Moreau envelope and $\Delta^{(t-1)}$, i.e. $\|\hat{\theta}^{(t-1)} - \theta^{(t-1)}\|^2 = \frac{1}{4\beta^2}\|\nabla\hat{\Phi}(\theta^{(t-1)})\|^2$. Equality (iii) uses properties of the Moreau envelope. □

The next two lemmas pertain to bounding the $\Delta^{(t)}$ terms.

Lemma D.7. *Under the conditions of Theorem D.5, for any $t \in [T]$ and $s \leq t-1$ one has,*

$$
\begin{aligned}
\Delta^{(t-1)} \leq \eta_\theta G(G + \tau_\theta)(2t - 2s - 1) + \frac{1}{2\eta_\lambda}(\|\lambda^*(\theta^{(s)}) - \lambda^{(t-1)}\|^2 - \|\lambda^*(\theta^{(s)}) - \lambda^{(t)}\|^2) + 2\|\Lambda\|_1\tau_\lambda \\
+ [\mathcal{L}(\theta^{(t)}, \lambda^{(t)}) - \mathcal{L}(\theta^{(t-1)}, \lambda^{(t-1)})].
\end{aligned} \tag{26}
$$

Proof. Let $s \leq t-1$. By adding and subtracting terms we have

$$
\begin{aligned}
\Delta^{(t-1)} &= [\mathcal{L}(\theta^{(t-1)}, \lambda^*(\theta^{(t-1)})) - \mathcal{L}(\theta^{(t-1)}, \lambda^*(\theta^{(s)}))] + [\mathcal{L}(\theta^{(t)}, \lambda^{(t)}) - \mathcal{L}(\theta^{(t-1)}, \lambda^{(t-1)})] \\
&\quad + [\mathcal{L}(\theta^{(t-1)}, \lambda^{(t)}) - \mathcal{L}(\theta^{(t)}, \lambda^{(t)})] + [\mathcal{L}(\theta^{(t-1)}, \lambda^*(\theta^{(s)})) - \mathcal{L}(\theta^{(t-1)}, \lambda^{(t)}))] \\
&\leq [\mathcal{L}(\theta^{(t-1)}, \lambda^*(\theta^{(t-1)})) - \mathcal{L}(\theta^{(t-1)}, \lambda^*(\theta^{(t-1)})] + [\mathcal{L}(\theta^{(t-1)}, \lambda^*(\theta^{(s)}) - \mathcal{L}(\theta^{(t-1)}, \lambda^*(\theta^{(s)}))] \\
&\quad + [\mathcal{L}(\theta^{(t)}, \lambda^{(t)}) - \mathcal{L}(\theta^{(t-1)}, \lambda^{(t-1)})] + [\mathcal{L}(\theta^{(t-1)}, \lambda^{(t)}) - \mathcal{L}(\theta^{(t)}, \lambda^{(t)})] + [\mathcal{L}(\theta^{(t-1)}, \lambda^*(\theta^{(s)})) - \mathcal{L}(\theta^{(t-1)}, \lambda^{(t)}))] \\
&\leq G(G + \tau_\theta)[2\|\theta^{(t-1)} - \theta^{(s)}\| + \|\theta^{(t-1)} - \theta^{(t)}\|] + [\mathcal{L}(\theta^{(t)}, \lambda^{(t)}) - \mathcal{L}(\theta^{(t-1)}, \lambda^{(t-1)})] \\
&\quad + [\mathcal{L}(\theta^{(t-1)}, \lambda^*(\theta^{(s)})) - \mathcal{L}(\theta^{(t-1)}, \lambda^{(t)}))] \\
&\leq \eta_\theta G(G + \tau_\theta)(2t - 2s - 1) + [\mathcal{L}(\theta^{(t)}, \lambda^{(t)}) - \mathcal{L}(\theta^{(t-1)}, \lambda^{(t-1)})] + [\mathcal{L}(\theta^{(t-1)}, \lambda^*(\theta^{(s)})) - \mathcal{L}(\theta^{(t-1)}, \lambda^{(t)}))].
\end{aligned}
$$

To complete the lemma, we will bound the loss difference $\mathcal{L}(\theta^{(t-1)}, \lambda^*(\theta^{(s)})) - \mathcal{L}(\theta^{(t-1)}, \lambda^{(t)})$. Since $\mathcal{L}(\theta^{(t-1)}, \cdot)$ is linear we have,

$$
\begin{aligned}
\mathcal{L}(\theta^{(t-1)}, \lambda^{(t-1)}) - \mathcal{L}(\theta^{(t-1)}, \lambda^{(t)}) &= \left\langle \lambda^{(t-1)} - \lambda^{(t)}, \nabla_\lambda \mathcal{L}(\theta^{(t-1)}, \lambda^{(t)}) \right\rangle \\
&\leq \left\langle \lambda^{(t-1)} - \lambda^{(t)}, g_\lambda^{(t)} \right\rangle + \left\langle \lambda^{(t-1)} - \lambda^{(t)}, \nabla_\lambda \mathcal{L}(\theta^{(t-1)}, \lambda^{(t)}) - g_\lambda^{(t)}) \right\rangle \\
&\leq \left\langle \lambda^{(t-1)} - \lambda^{(t)}, g_\lambda^{(t)} \right\rangle + \|\lambda^{(t-1)} - \lambda^{(t)}\|_1 \cdot \|\nabla_\lambda \mathcal{L}(\theta^{(t-1)}, \lambda^{(t)}) - g_\lambda^{(t)})\|_\infty \\
&\leq \left\langle \lambda^{(t-1)} - \lambda^{(t)}, g_\lambda^{(t)} \right\rangle + \|\Lambda\|_1\tau_\lambda
\end{aligned}
$$

Now a standard analysis using the fact that $\lambda^{(t-1)} + g_\lambda^{(t)}$ is projected orthogonaly onto Λ we have

$$0 \leq \|\lambda^{(t)} - \lambda^{(t-1)}\|^2 \leq \frac{1}{2\eta_\lambda}\|\lambda^*(\theta^{(t-1)}) - \lambda^{(t-1)}\|^2 - \frac{1}{2\eta_\lambda}\|\lambda^*(\theta^{(t-1)}) - \lambda^{(t)}\|^2 + \lambda\left\langle g_\lambda^{(t)}, \lambda^{(t)} - \lambda^*\right\rangle$$

Now by plugging back into the above yields,

$$\begin{aligned}
&\mathcal{L}(\theta^{(t-1)}, \lambda^{(t-1)}) - \mathcal{L}(\theta^{(t-1)}, \lambda^{(t)}) \\
&\quad \leq \left\langle \lambda^*(\theta^{(s)}) - \lambda^{(t)}, g_\lambda^{(t)}\right\rangle + \|\Lambda\|_1\tau_\lambda + \frac{1}{2\eta_\lambda}\|\lambda^*(\theta^{(t-1)}) - \lambda^{(t-1)}\|^2 - \frac{1}{2\eta_\lambda}\|\lambda^*(\theta^{(t-1)}) - \lambda^{(t)}\|^2.
\end{aligned}$$

Using concavity we obtain,

$$\begin{aligned}
\mathcal{L}(\theta^{(t-1)}, \lambda^*(\theta^{(s)})) - \mathcal{L}(\theta^{(t-1)}, \lambda^{(t)}) &\leq \left\langle \lambda^*(\theta^{(s)}) - \lambda^{(t)}, g_\lambda^{(t)} - \nabla_\lambda\mathcal{L}(\theta^{(t-1)}, \lambda^{(t)})\right\rangle + \|\Lambda\|_1\tau_\lambda \\
&\quad + \frac{1}{2\eta_\lambda}\|\lambda^*(\theta^{(t-1)}) - \lambda^{(t-1)}\|^2 - \frac{1}{2\eta_\lambda}\|\lambda^*(\theta^{(t-1)}) - \lambda^{(t)}\|^2 \\
&\leq \frac{1}{2\eta_\lambda}\|\lambda^*(\theta^{(t-1)}) - \lambda^{(t-1)}\|^2 - \frac{1}{2\eta_\lambda}\|\lambda^*(\theta^{(t-1)}) - \lambda^{(t)}\|^2 + 2\|\Lambda\|_1\tau_\lambda.
\end{aligned}$$

Plugging this back into the starting inequality achieves the claimed bound.

□

Lemma D.8. *Under the conditions of Theorem D.5 it holds that,*

$$\frac{1}{T}\sum_{t=0}^{T-1}\Delta^{(t)} \leq \|\Lambda\|_1\sqrt{\frac{\eta_\theta G(G+\tau_\theta)}{\eta_\lambda}} + 2\|\Lambda\|_1\tau_\lambda + \frac{\mathcal{L}(\theta_T, \lambda_T) - \mathcal{L}(\theta_0, \lambda_0)}{T}.$$

Proof. For any $s \in [T]$ and $M \in [T]$ one has by analyzing the telescoping sum created from Eqn.(26) that

$$\begin{aligned}
\sum_{t=s}^{s+M-1}\Delta^{(t)} &\leq \eta_\theta G(G+\tau_\theta)M^2 + \frac{1}{2\eta_\lambda}(\|\lambda^{(s)} - \lambda^*(\theta^{(s)})\|^2 + \|\lambda^{(s+M)} - \lambda^*(\theta^{(s)})\|^2) + 2M\|\Lambda\|_1\tau_\lambda \\
&\quad + [\mathcal{L}(\theta^{(s+M)}, \lambda^{(s+M)}) - \mathcal{L}(\theta^{(s)}, \lambda^{(s)})] \\
&\leq \eta_\theta G(G+\tau_\theta)M^2 + \frac{1}{\eta_\lambda}\|\Lambda\|_2^2 + 2M\|\Lambda\|_1\tau_\lambda + [\mathcal{L}(\theta^{(s+M)}, \lambda^{(s+M)}) - \mathcal{L}(\theta^{(s)}, \lambda^{(s)})] \\
&\leq \eta_\theta G(G+\tau_\theta)M^2 + \frac{1}{\eta_\lambda}\|\Lambda\|_1^2 + 2M\|\Lambda\|_1\tau_\lambda + [\mathcal{L}(\theta^{(s+M)}, \lambda^{(s+M)}) - \mathcal{L}(\theta^{(s)}, \lambda^{(s)})].
\end{aligned}$$

By applying this inequality over disjoint "blocks" of iterates, of which there are at most T/M, we can use this to obtain,

$$\frac{1}{T}\sum_{t=0}^{T-1}\Delta^{(t)} \leq \eta_\theta G(G+\tau_\theta)M + \frac{1}{M\eta_\lambda}\|\Lambda\|_1^2 + 2\|\Lambda\|_1\tau_\lambda + \frac{\mathcal{L}_0}{T}.$$

We can now set $M = \frac{\|\Lambda\|_1}{\sqrt{\eta_\theta\eta_\lambda G(G+\tau_\theta)}}$ to obtain the desired inequality. □

We can now prove the main theorem statement.

Proof of Theorem D.5. Recall $\hat{\Phi}_0 = \hat{\Phi}(\theta_0) - \min_\theta \left\{\hat{\Phi}(\theta)\right\}$ and $\mathcal{L}_0 = \mathcal{L}(\theta_0, \lambda_0) - \min_{\lambda,\theta}\{\mathcal{L}(\lambda, \theta)\}$. Summing over Eqn. (23) obtains,

$$\hat{\Phi}(\theta_{T-1}) \leq \hat{\Phi}(\theta_0) + 2\beta\eta_\theta \left(\sum_{t=0}^{T-1} \Delta^{(t)}\right) + 2T[\beta\eta_\theta^2(G^2 + \tau_\theta^2) + \eta_\theta\tau_\theta^2] - \frac{\eta_\theta}{4}\left(\sum_{t=0}^{T-1} \|\nabla\hat{\Phi}(\theta^{(t)})\|^2\right).$$

Which implies for any $M \in [T]$,

$$\begin{aligned}
\frac{1}{T}\left(\sum_{t=0}^{T-1} \|\nabla\hat{\Phi}(\theta^{(t)})\|^2\right) &\leq \frac{4\hat{\Phi}_0}{\eta_\theta T} + \frac{8\beta}{T}\left(\sum_{t=0}^{T-1}\Delta^{(t)}\right) + 4\beta\eta_\theta(G^2 + \tau_\theta^2) + 4\tau_\theta^2 \\
&\overset{(i)}{\leq} \frac{4\hat{\Phi}_0}{\eta_\theta T} + 8\beta\eta_\theta(G^2 + \tau_\theta^2) + 4\tau_\theta^2 \\
&\quad + 8\beta\|\Lambda\|_1\sqrt{\frac{\eta_\theta G(G + \tau_\theta)}{\eta_\lambda}} + 8\beta\|\Lambda\|_1\tau_\lambda + \frac{8\beta\mathcal{L}_0}{T}.
\end{aligned}$$

Inequality (i) above uses Eqn. (26). Setting $\eta_\lambda \geq \left(\frac{\eta_\theta G(G+\tau_\theta)}{\tau_\lambda}\right)$ yields,

$$\frac{1}{T}\left(\sum_{t=0}^{T-1} \|\nabla\hat{\Phi}(\theta^{(t)})\|^2\right) \leq \frac{4\hat{\Phi}_0}{\eta_\theta T} + 8\beta\eta_\theta(G^2 + \tau_\theta^2) + 4\tau_\theta^2 + 16\beta\|\Lambda\|_1\tau_\lambda + \frac{8\beta\mathcal{L}_0}{T}.$$

Setting $\eta_\theta = \sqrt{\frac{\hat{\Phi}_0}{2T\beta(G^2+\tau_\theta^2)}}$ yields,

$$\frac{1}{T}\left(\sum_{t=0}^{T-1} \|\nabla\hat{\Phi}(\theta^{(t)})\|^2\right) \leq \frac{16\sqrt{\hat{\Phi}_0\beta(G^2 + \tau_\theta^2)}}{\sqrt{T}} + 4\tau_\theta^2 + 16\beta\|\Lambda\|_1\tau_\lambda + \frac{16\beta\mathcal{L}_0}{T}.$$

Finally, this implies the claimed convergence.

$$\frac{1}{T}\left(\sum_{t=0}^{T-1} \|\nabla\hat{\Phi}(\theta^{(t)})\|\right) = O\left(\frac{\left(\hat{\Phi}_0\beta(G^2 + \tau_\theta^2)\right)^{1/4}}{T^{1/4}} + \tau_\theta + \sqrt{\beta\|\Lambda\|_1\tau_\lambda} + \frac{\sqrt{\beta\mathcal{L}_0}}{\sqrt{T}}\right).$$

□

D.7 Regularity Properties of Lagrangian

We will use the following standard fact about composing Lipschitz and/or smooth functions.

Lemma D.9. *Let $h : \mathbb{R}^d \mapsto \mathbb{R}^k$ be G_h-Lipschitz and β_h-smooth and $g : \mathbb{R}^k \mapsto \mathbb{R}$ be G_g-Lipschitz and β_g-smooth. Then $g \circ h$ is (G_hG_g)-Lipschitz and $(G_h\beta_g + G_g^2\beta_h)$-smooth.*

Proof. Let $\mathbf{J}_h(\theta)$ denote the Jacobian of h at θ. Since h is G_h Lipschitz, the spectral norm of the Jacobian is at most G_h, $\|\mathbf{J}_h(\theta)\|_2 \leq G_h$. Observe $\nabla_\theta g(h(\theta)) = \nabla g(h(\theta))^\top \mathbf{J}_h(\theta)$. Thus $\|\nabla_\theta g(h(\theta))\| \leq \|\nabla g(h(\theta))\| \cdot \|\mathbf{J}_h(\theta)\|_2 \leq G_gG_h$.

For the second part of the claim, observe that,

$$\begin{aligned}
\|\nabla_\theta g(h(\theta)) - \nabla_\theta g(h(\theta'))\| &\leq \|\nabla g(h(\theta))\mathbf{J}_h(\theta) - \nabla g(h(\theta'))\mathbf{J}_h(\theta')\| \\
&= \|[\nabla g(h(\theta)) - \nabla g(h(\theta'))]\mathbf{J}_h(\theta') + \nabla g(h(\theta))[\mathbf{J}_h(\theta) - \mathbf{J}_h(\theta')]\| \\
&\leq (G_h^2\beta_g + G_g\beta_h)\|\theta - \theta'\|.
\end{aligned}$$

□

In the following, we assume the predictor h is G_h-Lipschitz and β_h-smooth with respect to h, and similarly for ℓ with parameters G_ℓ and β_ℓ. Our aim is to derive regularity parameters for the Lagrangian given these base parameters. Note that the function which outputs one coordinate the tempered soft max is τ-Lipschitz and τ-smooth.

Lemma D.10. *Let $\Lambda \subset (\mathbb{R}^+)^{K\times Q}$ be a bounded set of diameter at most $\|\Lambda\|_1$ w.r.t. $\|\cdot\|_1$. Assume for any $j \in [J]$ that the vector α associated with rate constraint j satisfies $\|\alpha\|_1 \le c_1$ for some constant c. Then $\mathcal{L}(\cdot,\lambda)$ is G-Lipschitz with $G = G_\ell + c\tau G_h\|\Lambda\|_1$ for any $\lambda \in \Lambda$ and $\mathcal{L}$ is β-smooth with $\beta = \beta_\ell + 2c\tau \cdot \max\left\{G_h\sqrt{J}, \|\Lambda\|_1(2G_h + \tau\beta_h)\right\}$.*

Before presenting the proof, we note that G_ℓ and β_ℓ could also be further decomposed using the Lipschitz/smoothness constants of ℓ and h via Lemma D.9. However, as these parameters are not affected by our approach in the way the regularity parameters of the regularizer are, we omit these more specific details.

Proof. Let for $I \subseteq [Q]$ let $D_I = \cup_{q\in I} D_q$. To establish Lipschitzness w.r.t. θ, we have for any θ, λ that

$$\begin{aligned}\|\nabla_\theta \mathcal{L}(\theta,\lambda)\| &\le \|\nabla_\theta \ell(\theta)\| + \sum_{j=1}^{J}\lambda_j \sum_{I\in\mathcal{J}_j}\sum_{k}^{K}\alpha_{j,I,k}\left\|\nabla_\theta P_k(D_I;\theta)\right\|_2 \\ &\le G_\ell + \|\lambda\|_1 c_1 \cdot \max_{S\subseteq D, k\in[K]}\{\|\nabla_\theta P_k(S;\theta)\|\}\end{aligned}$$

Note that for any $S \subseteq D$ and $k \in [K]$ that $\|\nabla_\theta P_k(S;\theta,\tau)\| = \|\frac{1}{|S|}\sum_{x\in S}\nabla_\theta[\sigma_\tau(h(\theta;x))_k]\| \le \tau G_h$. Plugging this into the above achieved the claimed Lipschitz parameter.

To prove $\mathcal{L}$ is smooth, we have for any $\theta, \theta' \in \mathbb{R}^d$ and $\lambda, \lambda' \in \Lambda$,

$$\begin{aligned}\|\nabla\mathcal{L}(\theta,\lambda) - \nabla\mathcal{L}(\theta',\lambda')\|^2 &\le 2\|\nabla\mathcal{L}(\theta,\lambda) - \nabla\mathcal{L}(\theta,\lambda')\|^2 + 2\|\nabla\mathcal{L}(\theta,\lambda') - \nabla\mathcal{L}(\theta',\lambda')\|^2 \\ &= 2\|\nabla_\theta\mathcal{L}(\theta,\lambda) - \nabla_\theta\mathcal{L}(\theta,\lambda')\|^2 + 2\|\nabla_\lambda\mathcal{L}(\theta,\lambda) - \nabla_\lambda\mathcal{L}(\theta,\lambda')\|^2 \\ &\quad + 2\|\nabla_\theta\mathcal{L}(\theta,\lambda') - \nabla_\theta\mathcal{L}(\theta',\lambda')\|^2 + 2\|\nabla_\lambda\mathcal{L}(\theta,\lambda') - \nabla_\lambda\mathcal{L}(\theta',\lambda')\|^2.\end{aligned}$$

Bounding each term we have,

$$\begin{aligned}\|\nabla_\theta \mathcal{L}(\theta,\lambda) - \nabla_\theta \mathcal{L}(\theta,\lambda')\| &\leq \left\| \sum_{j=1}^{J} (\lambda_j - \lambda'_j) \sum_{I \in \mathcal{J}_j} \sum_{k}^{K} \alpha_{j,I,k} \nabla_\theta P_k(D_I;\theta) \right\|_2 \\ &\leq c_1 \tau G_h \|\lambda - \lambda'\|_1 \\ &\leq c_1 \tau G_h \sqrt{J} \|\lambda - \lambda'\|_2.\end{aligned}$$

$$\|\nabla_\lambda \mathcal{L}(\theta,\lambda) - \nabla_\lambda \mathcal{L}(\theta,\lambda')\| = 0.$$

$$\begin{aligned}\|\nabla_\theta \mathcal{L}(\theta,\lambda') - \nabla_\theta \mathcal{L}(\theta',\lambda')\| &\leq \|\nabla_\theta \ell(\theta) - \nabla_\theta \ell(\theta')\| + \left\| \sum_{j=1}^{J} \lambda'_j \sum_{I \in \mathcal{J}_j} \sum_{k}^{K} \alpha_{j,I,k} (\nabla_\theta P_k(D_I;\theta) - \nabla_\theta P_k(D_I;\theta')) \right\|_2 \\ &\overset{(i)}{\leq} \beta_\ell \|\theta - \theta'\| + \|\Lambda\|_1 c_1 (G_h \tau + \tau^2 \beta_h) \|\theta - \theta'\| \\ &\leq (\beta_\ell + c_1 \|\Lambda\|_1 \tau (G_h + \tau \beta_h)) \|\theta - \theta'\|\end{aligned}$$

$$\begin{aligned}\|\nabla_\lambda \mathcal{L}(\theta,\lambda') - \nabla_\lambda \mathcal{L}(\theta',\lambda')\| &\leq \left\| \sum_{j=1}^{J} \lambda'_j \sum_{I \in \mathcal{J}_j} \sum_{k=1}^{K} \alpha_{j,q,k} \left(P_k(h(\theta;x) - P_k(h(\theta';x)) \right) \right\| \\ &\overset{(ii)}{\leq} c_1 \tau \|\Lambda\|_1 G_h \|\theta - \theta'\|\end{aligned}$$

Above, in (i) we have used the fact that P_k is the composition of a G_h-Lipschitz and β_h-smooth function with a τ-Lipschitz and τ-smooth function, resulting in a $(G_h\tau + \tau^2\beta)$-smooth function. Similarly, in (ii), we have used the fact that P_k is the composition of two Lipschitz functions, resulting in another Lipschitz function.

Ultimately we obtain that $\|\nabla \mathcal{L}(\theta,\lambda) - \nabla \mathcal{L}(\theta',\lambda')\|^2$ is bounded by,

$$\max\left\{ (c_1 \tau G_h \sqrt{J})^2, [\beta_\ell + \|\Lambda\|_1 c_1 (G_h \tau + \tau^2 \beta_h)]^2 + [c_1 \tau \|\Lambda\|_1 G_h]^2 \right\} (\|\theta - \theta'\|^2 + \|\lambda - \lambda\|^2).$$

This implies that f is β-smooth with $\beta = \beta_\ell + 2c_1\tau \cdot \max\left\{ G_h\sqrt{J}, \|\Lambda\|_1 (2G_h + \tau\beta_h) \right\}$. □

E Additional Experimental Details

E.1 Dataset and Pre-Processing Details

We evaluate RaCO-DP on tabular fairness and privacy benchmark datasets from [43], namely, `Adult` [10], German `Credit Card` [34], and `Parkinsons` [42]. The classification task for `Credit Card` and `Parkinsons` is "whether the user will default payment the next month", and "whether the total UPDRS score of the patient is greater than the median or not," respectively. For `Adult`, the task is "whether the individual will make more than $50K." In all tasks, the sensitive attribute is gender.

To evaluate on more diverse subgroups, we also evaluate RaCO-DP on `folkstables` which is the 2018 yearly American Community Survey. We use the python package 'folktables' [19] to download and process the data for the Alabama ("AL") state in the US. We choose a "5"-year horizon and choose survey option to be 'person.' We adopt the experimental setup of Lowy, Gupta, and Razaviyayn [43] (including classification task, pre-processing, etc.) and report the baselines results directly from the official repository [31].

We also present results on the Heart Disease Health Indicators dataset [3] (21 risk factors).

Dataset	SGD	DP-SGD	RaCO-DP	DP-FERMI
Adult	0.018 ± 0.001 ms	0.037 ± 0.001 ms	0.064 ± 0.010 ms	85 ± 10 ms
CreditCard	0.020 ± 0.005 ms	0.035 ± 0.004 ms	0.055 ± 0.003 ms	88 ± 14 ms

Table 2: **Computational overhead comparison in terms of wall-time clock.**

E.2 Hyperparameter Tuning

Our hyperparameter selection process follows a two-phase approach. In the first phase, we run a hyperparameter search over predefined ranges: Gaussian noise variance $\sigma \in [3, 6]$, Laplace parameter $b \in [0.1, 0.5]$, learning rates $\eta_\theta, \eta_\lambda \in [10^{-4}, 0.1]$, mini-batch size $B \in [256, 1256]$, and softmax temperature $\tau \in [1, 10]$. We constrain the dual variables λ to be non-negative by setting the projection set $\Lambda = (\mathbb{R}^+)^J$. For each configuration, we target a specific constraint value γ and evaluate performance across five different seeds, selecting the hyperparameters that achieve the best validation accuracy while satisfying the constraint. In the second phase, we use the best hyperparameters identified through 200 optimization runs to train 20 new models. We report test accuracy and constraint satisfaction for each model based on the checkpoint that achieved the highest validation accuracy while satisfying the constraints on the train set.

E.3 Regularization-Privacy-Accuracy Trade-offs

Demographic Parity. In Figure 4 we demonstrate the trade-off curve comparisons with the baseline methods (as discussed in Section 6) for the `Credit-Card` and `Parkinsons` datasets. As with `adult.`, RaCO-DP closes the optimality gap with non-private models on these datasets, Pareto dominating the baselines.

False negative Rates. In Figure 5 we show the versatility of our framework beyond fairness mitigation by training models with performance constraints on false negative rate (1-recall). High recall (low FNR) is critical in medical contexts; XGBoost reaches 90% accuracy but has a 90% FNR. Non-private SGDA lowers FNR to 58% at 87.5% accuracy, and DP-RaCO nearly matches it at 60% FNR with the same accuracy, see Figure 5b.

E.4 Hard vs. Soft constraints

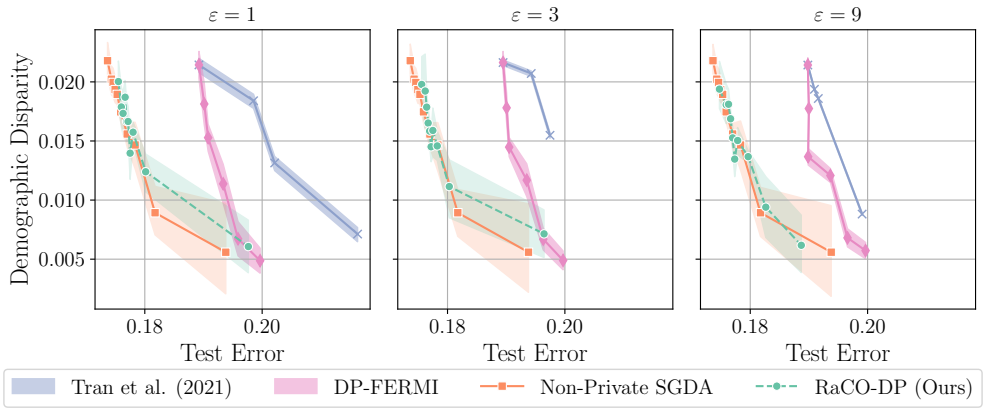
Figure 7 compares soft and hard constraints for the dual update. Notably, the soft constraint (with tuned softmax temperature τ) achieves similar performance compared to its hard constraint counterpart (solid dots) for most target values while maintaining similar levels of constraint satisfaction. This suggests that using soft constraints for the dual update does not significantly impact utility or constraint enforcement.

E.5 Compute Performance

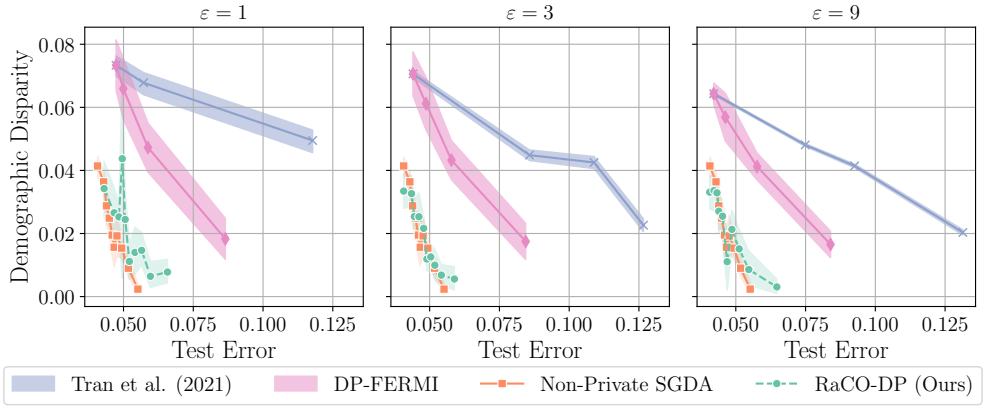
We provide a computational comparison between methods in Table 2, where we report the mean time of computing an SGD step, compared to a DP-SGD step and a RaCO-DP step on a 8 core CPU machine on Adult and Credit-Card on a batch size of 512. For reference, we also report the mean time of a DP-FERMI step using the publicly-available implementation.

RaCO-DP is 3 orders of magnitude faster to train than DP-FERMI on the same machine. Our algorithm builds on DP-SGD with the only additional overhead being computing the dual updates, which scales linearly in the number of constraints.

We note that our method's extra cost over standard DP-SGD is computing the dual update, which, if implemented naively, scales linearly in the number of constraints Q, implying Q extra backward passes to compute the gradients in the worst case. However, in practice, we can compute the gradient only for the active constraints ($\lambda_{(i)} > 0$), which can significantly reducing the computational costs.

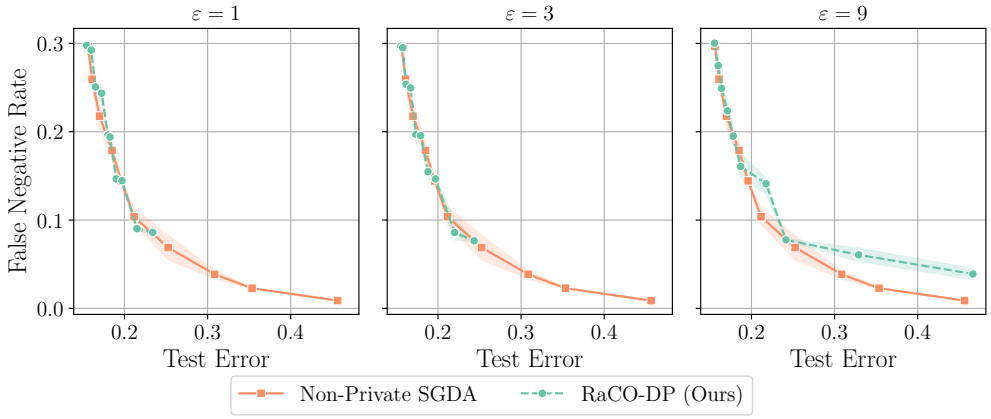


(a) `Credit-Card`

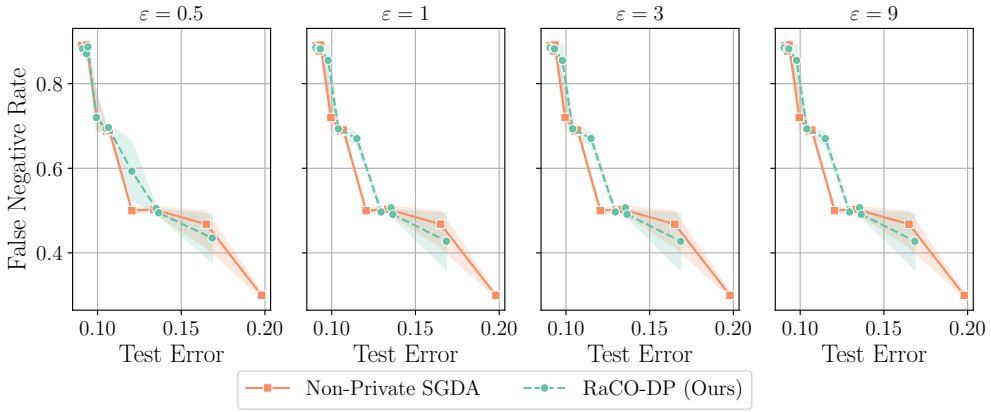


(b) `Parkinsons`

Figure 4: Demographic Parity Constraint



(a) `Adult`



(b) `Heart`

Figure 5: False Negative Rate Constraint

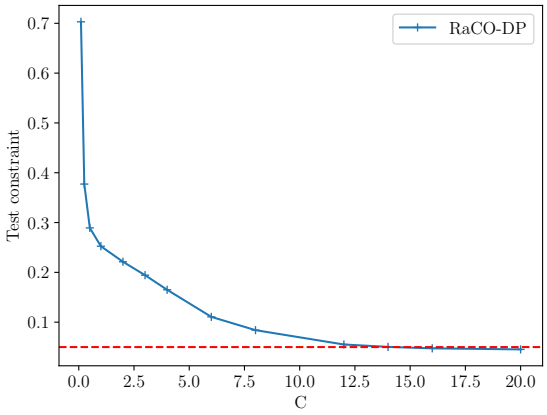


Figure 6: **False Negative Rate-Constrained Classification on Adult.** We show how the clipping norm C plays a critical role in satisfying a pessimistic constraint ($\gamma = 0$), even without noise related to differential privacy ($\sigma = 0, b = \infty$).

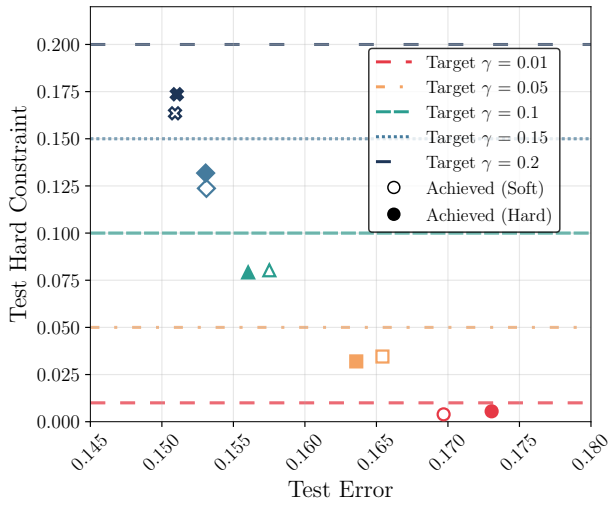


Figure 7: **Hard vs Soft Constraints on Adult.** Trade-off between test error and demographic parity on Adult dataset. Dashed lines show target constraints, with soft (hollow circles) and hard (solid dots) constraint implementations achieving similar performance across different target values.

E.6 Impact of the clipping norm

Figure 6 shows how the clipping threshold C affects RaCO-DP when we enforce a pessimistic constraint of FNR < 0 on the Adult dataset, in a non-private setting ($\sigma = 0,\ b = \infty$). With a small clipping norm ($C \leq 2$) the empirical FNR violation is still above 0.6, confirming that the bias introduced by clipping alone can drive the iterates far outside the feasible set. As the threshold increases, this bias shrinks rapidly; once $C \geq 12.5$, the FNR aligns with the target (red line), and the constraint is consistently satisfied.

These results demonstrate that an obstacle to satisfying the chosen constraints for our RaCO-DP is the bias from clipping and *not* the DP noise. Therefore, we stress that tuning C is an important aspect when applying RaCO-DP in practice.

F Broader Impact

An important takeaway from our work is that privacy and robustness criteria (such as the absence of performance disparities for underrepresented groups) are not inherently at odds with each other. This realization calls into question the practice of broadening the concept of privacy-utility trade-offs to include trade-offs with other robustness criteria. In high-risk decision-making systems that require both privacy and robustness, the responsibility for achieving such robustness falls on the beneficiaries of automated decision-making systems (governments and private institutions), as well as algorithm designers. These stakeholders must take care not to mistakenly attribute a lack of robustness to privacy mitigations, or a lack of privacy to robustness requirements.

Our work contributes to the existing literature in algorithmic fairness and privacy, and as such, adopts and further formalizes their computational interpretations of these human values. It is important to note that these interpretations, while useful in the contexts we have explored, are by no means collectively exhaustive. Specifically, the use of our algorithm does not ensure privacy in the broad sense, but rather in the limited sense of differential privacy, which protects the privacy of individuals whose data has been collected for training. Given the technical complexities of correctly implementing differential privacy, inappropriate tuning of model parameters or use outside its intended context can lead to a false sense of privacy—and, worse, may be exploited for privacy-washing by malicious actors in charge.