Differentially Private Post-Processing for Fair Regression

Ruicheng Xian [1] **Qiaobo Li** [1] **Gautam Kamath** [2] **Han Zhao** [1]

Abstract

This paper describes a differentially private post-processing algorithm for learning fair regressors satisfying statistical parity, addressing privacy concerns of machine learning models trained on sensitive data, as well as fairness concerns of their potential to propagate historical biases. Our algorithm can be applied to post-process any given regressor to improve fairness by remapping its outputs. It consists of three steps: first, the output distributions are estimated privately via histogram density estimation and the Laplace mechanism, then their Wasserstein barycenter is computed, and the optimal transports to the barycenter are used for post-processing to satisfy fairness. We analyze the sample complexity of our algorithm and provide fairness guarantee, revealing a trade-off between the statistical bias and variance induced from the choice of the number of bins in the histogram, in which using less bins always favors fairness at the expense of error.

1. Introduction

Prediction and forecasting models trained from machine learning algorithms are ubiquitous in real-world applications, whose performance hinges on the availability and quality of training data, often collected from end-users or customers. This reliance on data has raised ethical concerns including fairness and privacy. Models trained on past data may propagate and exacerbate historical biases against disadvantaged demographics, and producing less favorable predictions (Bolukbasi et al., 2016; Buolamwini & Gebru, 2018), resulting in unfair treatments and outcomes especially in areas such as criminal justice, healthcare, and finance (Barocas & Selbst, 2016; Berk et al., 2021). Models also have the risk of leaking highly sensitive private information in the training data collected for these applications (Dwork & Roth, 2014).

While there has been significant effort at addressing these concerns, few treats them in combination, i.e., designing algorithms that train fair models in a privacy-preserving manner. A difficulty is that privacy and fairness may not be compatible: exactly achieving group fairness criterion such as *statistical parity* or *equalized odds* requires precise (estimates of) group-level statistics, but for ensuring privacy, only noisy statistics are allowed under the notion of *differential privacy*. Resorting to approximate fairness, prior work has proposed private learning algorithms for reducing disparity, but the focus has been on the classification setting (Jagielski et al., 2019; Xu et al., 2019; Mozannar et al., 2020; Tran et al., 2021).

In this paper, we propose and analyze a differentially private post-processing algorithm for learning *attribute-aware* fair regressors under the squared loss, with respect to the fairness notion of statistical parity. It can take any (privately pre-trained) regressor and remaps its outputs (with minimum deviations) to improve fairness. At a high-level, our algorithm consists of three steps: estimating the output distributions of the regressor from data, computing their Wasserstein barycenter, and the optimal transports (Chzhen et al., 2020; Le Gouic et al., 2020; Xian et al., 2023). To make this process differentially private, we use private histogram density estimates (HDE) for the distributions via the Laplace mechanism (Diakonikolas et al., 2015; Xu et al., 2012), followed by re-normalization, which introduces additional complexity in our analysis. The choice of the number of bins in the HDE induces a trade-off between the statistical bias and variance for the cost of privacy and fairness. Our theoretical analysis and experiments show that using less bins always improves fairness at the expense of higher error.

Paper Organization. In Section 2, we introduce definitions, notation, and the problem setup. Section 3 describes our private post-processing algorithm for learning fair regressors, with finite sample analysis of the accuracy-privacy-fairness trade-offs in Section 3.4. Finally in Section 4, we empirically explore the trade-offs achieved by our post-processing algorithm on Law School and Communities & Crime datasets.

[1]University of Illinois Urbana-Champaign [2]University of Waterloo and Vector Institute. Correspondence to: Ruicheng Xian <rxian2@illinois.edu>, Han Zhao <hanzhao@illinois.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

1.1. Related Work

Differential Privacy. Under the notion of differential privacy (Dwork et al., 2006) and subsequent refinements such as Rényi DP (Mironov, 2017), a variety of private learning algorithms are proposed for settings including regression, classification, and distribution learning. There are private variants of logistic regression (Papernot et al., 2017), decision trees (Fletcher & Islam, 2019), and linear regression (Wang, 2018; Covington et al., 2021; Alabi et al., 2022). The problem of private distribution learning has been studied for finite-support distributions (Xu et al., 2012; Diakonikolas et al., 2015) and parameterized families (Bun et al., 2019; Aden-Ali et al., 2021). Lastly, private optimization algorithms including those based on objective perturbation (Chaudhuri et al., 2011; Kifer et al., 2012) and DP-SGD (Song et al., 2013; Bassily et al., 2014; Abadi et al., 2016) are proposed for solving convex and general optimization problems.

Algorithmic Fairness. In parallel, the study of algorithmic fairness revolves around the formalization of fairness notions and design of bias mitigation methods (Barocas et al., 2023). Fairness criteria include those that focus on disparate impact of the model, such as individual fairness (Dwork et al., 2012), which asks the model to treat similar inputs similarly, or group-level statistical parity (Calders et al., 2009) and equalized odds (Hardt et al., 2016); and, those on performance inequality, e.g., accuracy parity (Buolamwini & Gebru, 2018; Chi et al., 2021), predictive parity (Chouldechova, 2017), and multi-calibration (Hébert-Johnson et al., 2018). Fair algorithms are categorized into pre-processing, by removing biased correlation in the training data (Calmon et al., 2017); in-processing, that turns the original learning problem into a constrained one (Kamishima et al., 2012; Zemel et al., 2013; Agarwal et al., 2018; 2019); and post-processing, which remaps the predictions of a trained model post-hoc to meet the fairness criteria (Hardt et al., 2016; Pleiss et al., 2017; Chzhen et al., 2020; Zhao & Gordon, 2022; Xian et al., 2023).

Fairness and Privacy. Private fair learning and bias mitigation algorithm are proposed and studied in prior work (Jagielski et al., 2019; Xu et al., 2019; Mozannar et al., 2020; Tran et al., 2021), but they have so far been focused on the classification setting. It remains an open question on how to achieve private fair regression, and what are the trade-offs between privacy, fairness, and accuracy.

Cummings et al. (2019) and Agarwal (2020) showed that fairness and privacy are incompatible in the sense that no ε-DP algorithm can generally guarantee group fairness on the training set (from which population-level guarantees can be derived via generalization), unless the hypothesis class is restricted to constant predictors. The argument is that a predictor f that is fair on S may not be fair on its neighbor S', so an ε-DP algorithm that outputs f on S with nonzero probability may also output f on S', which is unfair. Work on private fair algorithms circumvent this incompatibility by relaxing to high probability guarantees for fairness. Bagdasaryan et al. (2019) showed that the performance impact of privacy may be more severe on underrepresented groups, resulting in accuracy disparity.

2. Preliminaries

A regression problem is defined by a joint distribution μ of the observations $X \in \mathcal{X}$, sensitive attribute $A \in \mathcal{A}$ (which has finite support), and the response $Y \in \mathbb{R}$. We will use upper case X, A, Y to denote the random variables, and lower case x, a, y instances of them. The goal of this paper is to develop a *privacy-preserving post-processing* algorithm for learning (randomized) *fair* regressors. We consider the attribute-aware setting, i.e., the sensitive attribute A is available explicitly during both training and prediction and can be taken as input by the regressor, $f : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$.

The risk of a regressor is defined to be mean squared error, and its excess risk is defined with respect to the Bayes regressor, $f^*(x, a) := \mathbb{E}[Y \mid X = x, A = a]$:

$$R(f) := \mathbb{E}[(f(X, A) - Y)^2],$$
$$ER(f) := R(f) - R(f^*) = \mathbb{E}[(f(X, A) - f^*(X, A))^2]$$

by the orthogonality principle, where $\mathbb{E}$ is taken with respect to μ (and the randomness of f).

Given a (training) dataset S consisting of n i.i.d. samples of (X, A, Y), a private learning algorithm minimizes the leakage of any individual's information in its output. We use the notion of differential privacy (Dwork et al., 2006), which limits the influence of any single training sample:

Definition 2.1 (Differential Privacy)**.** A randomized (learning) algorithm $\mathscr{A}$ is ε-differentially private (DP) if for all pairs of nonempty neighboring datasets S, S',

$$\mathbb{P}(\mathscr{A}(S) \in O) \le e^{\varepsilon}\, \mathbb{P}(\mathscr{A}(S') \in O), \quad \forall O \subseteq \mathrm{range}(\mathscr{A}),$$

where $\mathbb{P}$ is taken with respect to the randomness of $\mathscr{A}$.

We say two datasets are neighboring if they differ in one entry by substitution (our result also covers insertion and deletion operations, which may have lower sensitivity). Note that privacy is guaranteed with respect to both A and X.

For fairness, we consider statistical parity (Calders et al., 2009), which requires the output distributions of f conditioned on each group to be similar. The similarity between distributions is measured in Kolmogorov–Smirnov distance, defined for probability measures p, q supported on $\mathbb{R}$ by

$$D_{\mathrm{KS}}(p, q) = \sup_{t \in \mathbb{R}} \left| \int_{-\infty}^{t} (p(x) - q(x))\, \mathrm{d}x \right|.$$

Definition 2.2 (Statistical Parity)**.** A (randomized) regressor $f : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ satisfies α-approximate statistical parity (SP) if

$$\Delta_{\text{SP}}(f) \coloneqq \max_{a,a' \in \mathcal{A}} D_{\text{KS}}(r_a, r_{a'}) \leq \alpha,$$

where r_a is the distribution of regressor output $f(X, A)$ conditioned on $A = a$.

3. Fair and Private Post-Processing

We describe a private post-processing algorithm such that, given an unlabeled training dataset $S = \{(x_i, a_i)\}_{i \in [n]}$ sampled from μ, and a (privately) pre-trained regressor f (e.g., learned using the algorithms mentioned in Section 1.1), the algorithm learns a mapping g that transforms the output of f (a.k.a. post-processing) so that $\bar{f} \coloneqq g \circ f$ is fair, while preserving differential privacy with respect to S.

Compared to in-processing approaches for fairness, post-processing decouples the conflicting goals of fairness and error minimization, since the regressor f can be trained to optimality without any constraint, and then post-processed to satisfy fairness. We show that this decoupling does not affect the optimality of the resulting fair regressor $\bar{f}$—the optimal fair regressor can be recovered via post-processing if the algorithm in the pre-training stage learns the Bayes regressor f^* (Theorem 3.1). Moreover, post-processing has low sample complexity and only requires unlabeled data (Theorem 3.3), so labeled data (that may be scarce in some domains) can be dedicated entirely to error minimization, which is typically a more difficult problem.

3.1. Fair Post-Processing with Wasserstein-Barycenters

Exact statistical parity ($\alpha = 0$) requires all groups to have identical output distributions, so finding the fair post-processing mapping g that incurs minimum deviations from the original outputs of f amounts to the following steps (Chzhen et al., 2020; Le Gouic et al., 2020):

1. Learn the output distributions of f conditioned on each group, r_a, for all $a \in \mathcal{A}$.
2. Find a common (fair) distribution q that is close to the original distributions (i.e., their barycenter).
3. Compute the optimal transports g_a from r_a to the barycenter q.

The (randomized) transports are then applied to post-process f to obtain a fair attribute-aware regressor $\bar{f}(x, a) = g_a \circ f(x)$, so that every group has the same output distribution equal to q.

Formally, q is called the Wasserstein barycenter of the r_a's, which is a distribution supported on $\mathbb{R}$ with minimum total distances to the r_a's as measured in Wasserstein distance:

$$q \in \operatorname*{arg\,min}_{q' : \text{supp}(q') \subseteq \mathbb{R}} \sum_{a \in \mathcal{A}} w_a W_2^2(r_a, q'), \tag{1}$$

where $w_a \coloneqq \mathbb{P}(A = a)$,

$$W_2^2(r_a, q) = \min_{\pi_a \in \Pi(r_a, q)} \int_{\mathbb{R} \times \mathbb{R}} (y - y')^2 \, \mathrm{d}\pi_a(y, y'), \tag{2}$$

and $\Pi(p, q) = \{\pi : \text{supp}(\pi) \subseteq \mathbb{R} \times \mathbb{R}, \int \pi(x, y') \, \mathrm{d}y' = p(x), \int \pi(x', y) \, \mathrm{d}x' = q(y), \forall x, y\}$ is the collection of probability couplings of p, q. The value of $W_2^2(r_a, q)$ naturally represents the squared cost of transforming r_a into q, i.e., the minimum amount of output deviations (in squared distance) required to post-process f for group a so that its output distribution is transformed to q (which is a consequence of Lemma A.4); specifically, this optimal post-processor is given by the optimal transport from r_a to q.

3.2. Generalization to Approximate Statistical Parity

To handle approximate statistical parity ($\alpha > 0$), in step 2, we first replace the barycenter in Equation (1) by a KS ball of radius $\alpha/2$ and relax the problem to finding target output distributions q_a for each group a inside the ball with minimum W_2^2 distance to the original distribution r_a. Then in step 3, we find the optimal transports g_a from r_a to q_a. So, the problem in Equation (1) is generalized to the approximate SP case with

$$\mathrm{P}(\{r_a\}_{a \in \mathcal{A}}, \{w_a\}_{a \in \mathcal{A}}, \alpha) : \operatorname*{arg\,min}_{\substack{q : \text{supp}(q) \subseteq \mathbb{R} \\ \{q_a\}_{a \in \mathcal{A}} \subset B_{\text{KS}}(q, \alpha/2)}} \sum_{a \in \mathcal{A}} w_a W_2^2(r_a, q_a),$$

where $B_{\text{KS}}(q, \alpha/2) = \{p : D_{\text{KS}}(p, q) \leq \alpha/2\}$ is a KS ball centered at q on the space of probability distributions supported on $\mathbb{R}$.

When learning from finite samples S (non-privately), we replace r_a, w_a in P with the respective empirical distributions/estimates. Computationally, solving P on the empirical distributions is as hard as the finite-support Wasserstein barycenter problem ($\alpha = 0$), which can be formulated by a linear program of exponential size in $|\mathcal{A}|$ (Anderes et al., 2016; Altschuler & Boix-Adserà, 2021). To reduce the complexity, we restrict the support of the barycenter q via discretization; this fixed-support approximation is common in prior work (Cuturi & Doucet, 2014; Staib et al., 2017).

The validity of the post-processing algorithm described above is supported by the fact that, if the regressor being post-processed is Bayes optimal, then the resulting fair regressor is also optimal. This has been established for the exact SP case in prior work (e.g., Le Gouic et al., 2020, Theorem 3), and here we provide a more general result for approximate SP:

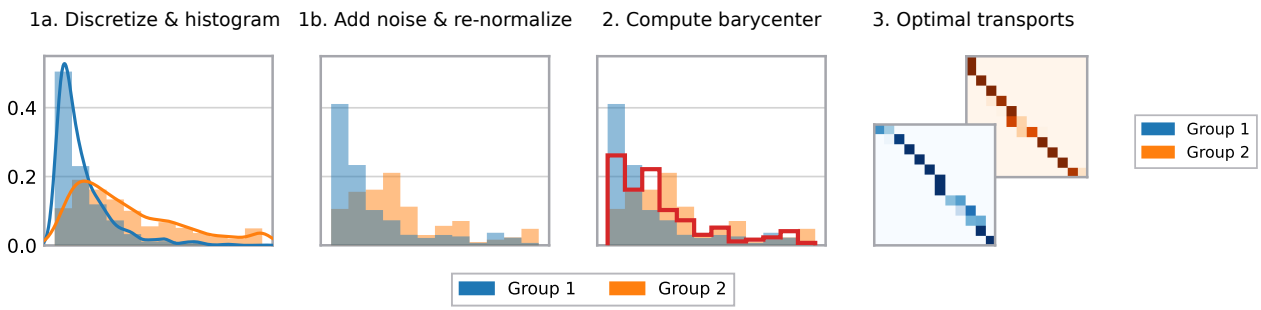

Figure 1. Illustration of the private fair post-processing steps 1–3 performed in Algorithm 1. The (randomized) transports to the barycenter are represented by (sparse) $k \times k$ matrices, and the value at the (i, j)-th entry is the probability of transporting to bin j given bin i.

Theorem 3.1. *Let a regression problem be given by a joint distribution μ of (X, A, Y), denote $w_a = \mathbb{P}(A = a)$, the Bayes regressor by $f^*(x, a) = \mathbb{E}[Y \mid X = x, A = a]$, and its output distribution conditioned on group a by r_a^*. Then*

$$\min_{f:\Delta_{\text{SP}}(f)\le\alpha} ER(f) = \min_{\substack{q:\text{supp}(q)\subseteq\mathbb{R} \\ \{q_a\}_{a\in\mathcal{A}}\subset B_{\text{KS}}(q,\alpha/2)}} \sum_{a\in\mathcal{A}} w_a W_2^2(r_a^*, q_a).$$

This shows that the excess risk of the optimal fair regressor is indeed the value of the Wasserstein barycenter problem (P) discussed above, and can be achieved by f^* post-processed using the optimal transports to the barycenter. All proofs are deferred to Appendices A and B; this result follows from an equivalence between learning regressors and learning post-processings of f^*.

3.3. Privacy via Discretization and Private PMF Estimation

To make the fair post-processing algorithm in Section 3.1 private with respect to S, it suffices to perform step 1 of estimating the output distributions privately. This is because the subsequent steps 2 and 3 of computing the barycenter (including the target distributions) and the optimal transports only depend on the estimated distributions, so privacy is preserved by the post-processing immunity of DP.

To estimate the distributions, one could construct a family of distributions (a.k.a. hypotheses) and (privately) choose the most likely one (Bun et al., 2019; Aden-Ali et al., 2021), or use non-parametric estimators. For generality, we adopt the latter approach and use (private) histogram density estimator (Diakonikolas et al., 2015; Xu et al., 2012).

Altogether, our fair and private post-processing algorithm is detailed in Algorithm 1 and illustrated in Figure 1. It takes as inputs the regressor f being post-processed, the samples S, the fairness tolerance α, and the privacy budget ε. For performing HDE, it also requires specifying an interval[1] (using prior/domain knowledge) that contains the image of the regressor $[s, t] \supseteq f[\text{supp}(\mu_{X,A})]$, and the number of bins k. We describe the algorithm in details below:

Step 1 (Estimate Output Distributions)**.** On line 3, we compute the empirical joint distribution of A and $h \circ f(X, A)$, the discretized regressor output (the support is $v := (\{(j-1/2)(t-s)/k\}_{j\in[k]})$). Then, we make this statistics private via the Laplace mechanism on line 4, noting that the L^1-sensitivity to $\hat{p}$ is at most $2/n$ (Remark B.1).[2]

With the private joint distribution, we get a private estimate of the group marginal distribution $\tilde{w}_a$'s (clipping negative values to zero) on line 6, and the private group conditional discretized output distribution $\tilde{p}_a$'s (with re-normalization) on lines 7–9. The re-normalization on $\tilde{p}_a$ (could also be applied to $\tilde{w}_a$), required for defining the optimal transport problem in the subsequent step, is done by performing isotonic regression on the partial sums (so that its values are non-decreasing) with clipping to get a valid CDF.

We analyze the accuracy of the private estimates:

Theorem 3.2. *Let $p_a(v_j) = \mathbb{P}(h \circ f(X, A) = v_j \mid A = a)$ and $w_a = \mathbb{P}(A = a)$ for all $a \in \mathcal{A}$, $j \in [k]$, and let $\tilde{p}_a$, $\tilde{w}_a$ denote their privately estimated counterparts in Algorithm 1. Then for all $n \ge \Omega(\max_a 1/w_a \ln 1/\beta)$, with probability at least $1 - \beta$, for all $a \in \mathcal{A}$,*

$$|w_a - \tilde{w}_a| \le O\left(\frac{\sqrt{k}}{n\varepsilon} \ln \frac{k|\mathcal{A}|}{\beta}\right),$$

and

$$\|p_a - \tilde{p}_a\|_1 \le O\left(\sqrt{\frac{k}{nw_a} \ln \frac{k|\mathcal{A}|}{\beta}} + \frac{k}{nw_a\varepsilon} \ln \frac{k|\mathcal{A}|}{\beta}\right),$$

$$\|p_a - \tilde{p}_a\|_\infty \le O\left(\sqrt{\frac{1}{nw_a} \ln \frac{|\mathcal{A}|}{\beta}} + \frac{\sqrt{k}}{nw_a\varepsilon} \ln \frac{k|\mathcal{A}|}{\beta}\right),$$

[1] This is necessary for pure $(\varepsilon, 0)$-DP, due to a lower bound by Hardt & Talwar (2010) using a packing argument.

[2] The Laplace mechanism could be replaced by, e.g., the Gaussian mechanism, to relax the pure $(\varepsilon, 0)$-DP guarantee to approximate (ε, δ)-DP.

Algorithm 1 Fair and Private Post-Processing (Attribute-Aware)

Require: Regressor $f : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$, samples $\{x_i, a_i\}_{i\in[n]}$, interval $[s, t]$, number of bins k, fairness tolerance α, privacy budget ε

1: Let $v_j = {}^{(j-1/2)(t-s)}/_{k}$ for all $j \in [k]$ ▷ *midpoints of the bins*
2: Define $h(y) = \arg\min_{j\in[k]} |y - v_j|$ ▷ *discretizer*
3: Define $\hat{p}(a, v_j) = {}^{1}/_{n} \sum_{i\in[n]} \mathbb{1}[h \circ f(x_i, a_i) = v_j]$ ▷ *empirical joint distribution*
4: Define $\check{p}(a, v_j) = \hat{p}(a, v_j) + \mathrm{Laplace}(0, {}^{2}/_{n\varepsilon})$ ▷ *Laplace mechanism to get private joint distribution*
5: **for** $a \in \mathcal{A}$ **do**
6: $\quad \tilde{w}_a \leftarrow \max(\sum_{j=1}^{k} \check{p}(a, v_j), 0)$ ▷ *private group marginal distribution*
7: $\quad$ Define $\check{F}_a(v_j) = {}^{1}/_{\tilde{w}_a} \sum_{\ell=1}^{j} \check{p}(a, v_\ell)$ ▷ *scaled partial sums*
8: $\quad$ Define $\widetilde{F}_a(v_j) = \begin{cases} \mathrm{proj}_{[0,1]}({}^{1}/_{2}(\check{F}_a(v_{l_j}) + \check{F}_a(v_{r_j}))) & \text{if } j < k \\ 1 & \text{else} \end{cases}$ where $(l_j, r_j) = \arg\max_{l\le j\le r}(\check{F}_a(v_l) - \check{F}_a(v_r))$
▷ *L^∞ isotonic regression and clipping to get private CDF*
9: $\quad$ Define $\tilde{p}_a(v_j) = \widetilde{F}_a(v_j) - \widetilde{F}_a(v_{j-1})$ ▷ *private group conditional PMF*
10: $(\{\pi_a\}_{a\in\mathcal{A}}, \tilde{q}, \{\tilde{q}_a\}_{a\in\mathcal{A}}) \leftarrow \mathrm{LP}(\{\tilde{p}_a\}_{a\in\mathcal{A}}, \{\tilde{w}_a\}_{a\in\mathcal{A}}, \alpha)$ ▷ *compute barycenter and get optimal transports*
11: **for** $a \in \mathcal{A}$ **do**
12: $\quad$ Define $g_a(v_j) = \begin{cases} v_\ell \text{ w.p. } {}^{\pi_a(v_j, v_\ell)}/_{\tilde{p}_a(v_j)}, \quad \forall \ell \in [k] & \text{if } \tilde{p}_a(v_j) > 0 \\ v_j & \text{else} \end{cases}$ ▷ *post-processing mappings*
13: **return** $(x, a) \mapsto g_a \circ h \circ f(x, a)$ ▷ *privately post-processed fair regressor*

$$D_{\mathrm{KS}}(p_a, \tilde{p}_a) \le O\left(\sqrt{\frac{k}{nw_a} \ln \frac{k|\mathcal{A}|}{\beta}} + \frac{\sqrt{k}}{nw_a\varepsilon} \ln \frac{k|\mathcal{A}|}{\beta}\right).$$

The private estimation of PMFs with the Laplace mechanism has been analyzed in prior work (Diakonikolas et al., 2015; Vadhan, 2017), but our algorithm performs an extra re-normalization (lines 7–9) after adding Laplace noise to ensure that the PMF returned is valid. This makes the analysis of Theorem 3.2 more involved, because the noise added to each bin can interact during re-normalization.

The rate of $\widetilde{O}(\sqrt{{}^{k}/_{n}} + {}^{k}/_{n\varepsilon})$ for TV distance is consistent with existing results (Diakonikolas et al., 2015). For KS distance, the rate is improved by a $\sqrt{k}$ factor, which is not surprising as KS is a weaker metric than TV. We note that our analysis in attaining this improvement is made easier with the L^∞ isotonic regression CDF re-normalization scheme, because KS distance is also defined via the CDF.

Steps 2 and 3 (Compute Barycenter and Optimal Transports)**.** With the output distributions $\tilde{p}_a$ estimated privately, we now compute their barycenter and the optimal transports to obtain the fair post-processing mappings g_a, i.e., solving $\mathrm{P}(\{\tilde{p}_a\}_{a\in\mathcal{A}}, \{\tilde{w}_a\}_{a\in\mathcal{A}}, \alpha)$. Since the $\tilde{p}_a$'s are distributions supported on v, by restricting the support of the barycenter to v,[3] the barycenter and the optimal transports can be computed by solving a linear program with $O(k^2|\mathcal{A}|)$ variables and constraints (cf. P and Equation (2)):

$$\mathrm{LP}(\{p_a\}_{a\in\mathcal{A}}, \{w_a\}_{a\in\mathcal{A}}, \alpha):$$

$$\begin{aligned} \underset{\substack{\{\pi_a\}_{a\in\mathcal{A}}\ge 0 \\ q, \{q_a\}_{a\in\mathcal{A}}\ge 0}}{\arg\min} \quad & \sum_{a\in\mathcal{A}} w_a \sum_{j,\ell\in[k]} (v_j - v_\ell)^2 \pi_a(v_j, v_\ell) \\ \text{s.t.} \quad & \sum_{\ell\in[k]} \pi_a(v_j, v_\ell) = p_a(v_j), \quad \forall a \in \mathcal{A},\ j \in [k], \\ & \sum_{j\in[k]} \pi_a(v_j, v_\ell) = q_a(v_\ell), \quad \forall a \in \mathcal{A},\ \ell \in [k], \\ & \left|\sum_{j\le\ell} (q_a(v_j) - q(v_j))\right| \le \frac{\alpha}{2}, \forall a \in \mathcal{A},\ \ell \in [k]. \end{aligned}$$

The constraints enforce that the π_a's are couplings, and the target output distributions $\tilde{q}_a$'s are valid PMFs satisfying the fairness constraint of $\max_{a,a'} D_{\mathrm{KS}}(\tilde{q}_a, \tilde{q}_{a'}) \le \alpha$.

Post-Processed Fair Regressor. On Lines 11 to 13, the (randomized) optimal transports g_a from $\tilde{p}_a$ to $\tilde{q}_a$ are extracted by reading off from the optimal couplings π_a (represented by $k \times k$ matrices), and are used to construct the fair regressor, $\bar{f}(x, a) = g_a \circ h \circ f(x, a)$.

Given an input (x, a), the fair regressor $\bar{f}$ obtained from Algorithm 1 calls f to make a prediction $y = f(x, a)$, discretizes it to get $\tilde{y} = h(y)$, and then uses the optimal transport g_a of the respective group to post-process and return a fair prediction, $\bar{y} = g_a(\tilde{y})$.

[3] This approximation reduces the size of the barycenter problem to polynomial, at $O({}^{t-s}/_{k})$ error, the same as that from discretizing the outputs.

3.4. Statistical Analysis

Algorithm 1 is ε-DP by the post-processing theorem of DP (Dwork & Roth, 2014, Proposition 2.1), because its output depends on S only via the private statistics $\tilde{p}$ on line 4 that is ε-DP.[2]

We analyze the suboptimality (or *fair excess risk*) of the post-processed fair regressor and provide fairness guarantee:

Theorem 3.3. *Let regressor f be given along with samples $S = \{(x_i, a_i)\}_{i\in[n]}$ of $\mu_{X,A}$. Denote $L = t - s$, assume $L \leq 1$, and $Y, f(X, A) \in [s, t]$ almost surely. Let $\bar{f}$ denote the fair regressor returned from Algorithm 1 on $(f, S, [s, t], k, \alpha, \varepsilon)$. Then for all $n \geq \Omega(\max_a 1/w_a \ln |\mathcal{A}|/\beta)$, with probability at least $1 - \beta$,*

$$R(\bar{f}) - R(\bar{f}^*) \leq O\left(\sqrt{\frac{k|\mathcal{A}|}{n} \ln \frac{k|\mathcal{A}|}{\beta}} + \frac{k|\mathcal{A}|}{n\varepsilon} \ln \frac{k|\mathcal{A}|}{\beta}\right) + \frac{8}{k} + 5\,\mathbb{E}[|f(X, A) - f^*(X, A)|]$$

where $\bar{f}^* = \arg\min_{f':\Delta_{\mathrm{SP}}(f')\leq\alpha} R(f')$ is the optimal fair regressor, f^* is the (unconstrained) Bayes regressor, and

$$\Delta_{\mathrm{SP}}(\bar{f}) \leq \alpha + \max_{a\in\mathcal{A}} O\left(\frac{\sqrt{k}}{nw_a\varepsilon} \ln \frac{k|\mathcal{A}|}{\beta} + \sqrt{\frac{k}{nw_a} \ln \frac{k|\mathcal{A}|}{\beta}}\right).$$

The risk bound reflects four potential sources of error: the first term is finite sample estimation error, the second term is due to the noise added for ε-DP, the third term is the error introduced by discretization, and the last term is the L^1 excess risk of the regressor f being post-processed (carrying over the error of the pre-trained regressor).

The first three terms of the risk bound (and last two in the fairness bound) are attributed to the accuracy of the private distribution estimate using HDE (cf. Theorem 3.2). In particular, a trade-off between the statistical bias and variance is incurred by the choice of k, the number of bins:

$$\widetilde{O}\left(\underbrace{\sqrt{\frac{k}{n}} + \frac{k}{n\varepsilon}}_{\text{variance}} + \underbrace{\frac{1}{k}}_{\text{bias}}\right).$$

Using too few bins leads to a large discretization error (statistical bias), whereas using too many suffers from large variance due to data sampling and the noise added for privacy.

The cost of privacy is dominated by the estimation error as long as $n \gtrsim \max_a k/w_a\varepsilon^2 \geq k|\mathcal{A}|/\varepsilon^2$. In which case, the choice of $k = \widetilde{\Theta}(n^{1/3})$ is optimal for MSE, which is consistent with classical non-parametric estimation results of HDE (Rudemo, 1982).

The fairness bound, on the other hand, only contains variance terms but not the statistical bias, therefore using fewer bins is always more favorable for fairness at the expense of MSE (e.g., in the extreme case of $k = 1$, the post-processor outputs a constant value, which is exactly fair). This suggests that when n is small, k can be decreased to reduce variance for higher levels of fairness.

The optimal choice of k (in combination with α) is data-dependent, and is typically tuned on a validation split, however; extra care should be taken as this practice could, in principle, violate differential privacy. In our experiments, we sweep α and k to empirically explore the trade-offs between error, privacy and fairness attainable with our Algorithm 1. This is common practice in differential privacy research (Mohapatra et al., 2022). Regarding the selection of hyperparameters while preserving privacy, we refer readers to Liu & Talwar (2019); Papernot & Steinke (2022).

4. Experiments

In this section, we evaluate the private and fair post-processing algorithm described in Algorithm 1.

We do not compare to other algorithms because we are not aware of any existing private algorithms for learning fair regressors. Although it may be possible to adopt some existing algorithms to this setting, e.g., using DP-SGD as in (Tran et al., 2021), making them practical and competitive requires care, and is hence left to future work. Our post-processing algorithm is based on (Chzhen et al., 2020); their paper is referred to for empirical comparisons to in-processing algorithms under the non-private setting ($\varepsilon = \infty$).

Setup. Since the excess risk can be decomposed according to the pre-training and post-processing stages (Theorem 3.3, where $\mathbb{E}[|f - f^*|]$ is carried over from the suboptimality of the regressor f trained in the pre-training stage), we will simplify our experiment setup to isolate and focus on the performance and trade-offs induced by post-processing alone. This means we will access the ground-truth responses and directly apply Algorithm 1 (i.e., $X = Y$ and $f = \mathrm{Id}$, whereby $\mathbb{E}[|f - f^*|] = 0$).[4]

Datasets. The datasets are randomly split 70-30 for training (i.e., post-processing) and testing.

Communities & Crime (Redmond & Baveja, 2002). It contains the socioeconomic and crime data of communities in US, and the task is to predict the rate of violent crimes per 100k population ($Y \in [0, 1]$). The sensitive attribute is an indicator for whether the community has a significant presence of a minority population ($|\mathcal{A}| = 2$). The total size of the dataset is 1,994, and the number of training examples for the smallest group is 679 ($\approx nw_a$).

[4]Code is available at https://github.com/rxian/fair-regression.

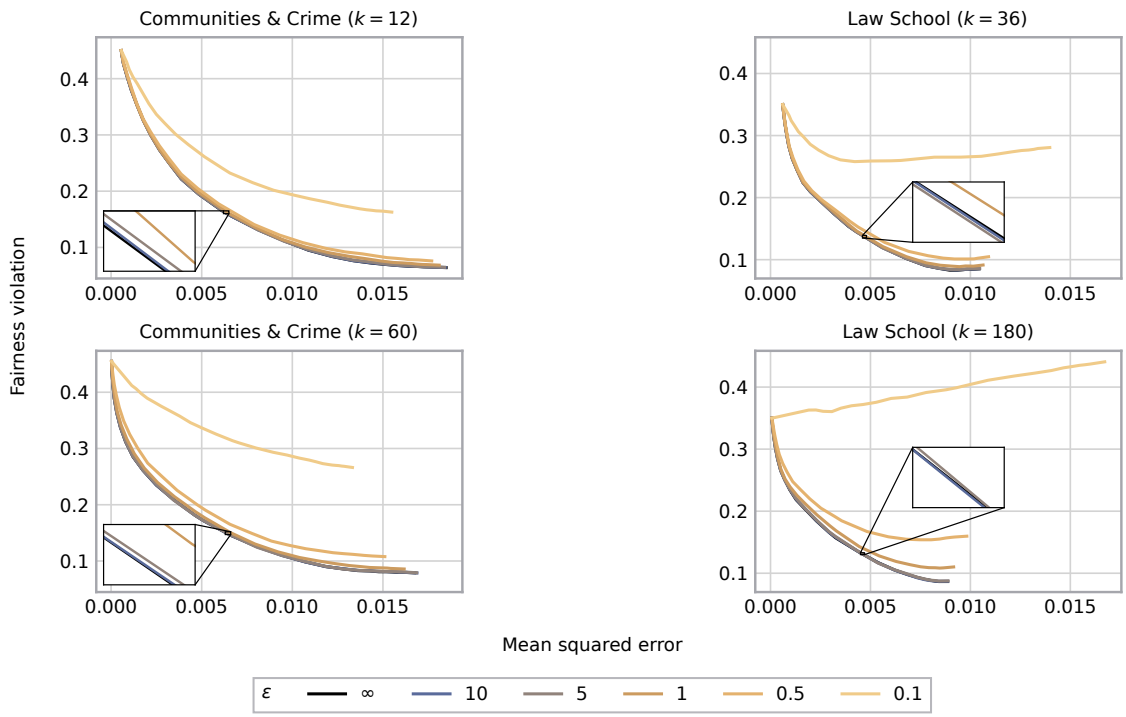

Figure 2. Error-privacy-fairness trade-offs achieved by Algorithm 1 by sweeping α under the indicated number of bins k, with different privacy budgets ε. Fairness violation is measured in KS distance as defined in Definition 2.2 (Δ_{SP}). Average of 50 random seeds (also for Figures 3 and 4).

Law School (Wightman, 1998). This dataset contains the academic performance of law school applicants, and the task is to predict the student's undergraduate GPA ($Y \in [1, 4]$). The sensitive attribute is race ($|\mathcal{A}| = 4$), and the dataset has size 21,983. The smallest group has 628 training examples.

4.1. Results

In Figure 2, we show the MSE and fairness trade-offs (in KS distance; Δ_{SP}, Definition 2.2) attained with Algorithm 1 under various settings of α. The main observations are:

1. The cost of discretization (indicated by the horizontal distance from 0 to the top-left starting points of the curves) can be expected to be insignificant compared to fairness, unless the model is already very fair without post-processing.
2. Although the amount of data available for post-processing is small by modern standards, the results are very insensitive to the privacy budget until the highest levels of DP are demanded or very large k is used.

   This is because according to Theorem 3.3, the error attributed to DP noise is dominated by estimation error when $n \gtrsim \max_a {}^{k}/_{w_a \varepsilon^2}$. Note that in our experiments, we have $n \gg {}^{k|\mathcal{A}|}/_{\varepsilon^2}$ on both datasets except for $\varepsilon = 0.1$, and $\varepsilon = 0.5$ when $k = 180$. Using more bins increases the cost of privacy due to variance, as we will show in Figure 3.
3. The right end of each line is obtained with setting $\alpha = 0$. Relaxing it to larger values could give better trade-offs, especially when n, ε are small, because in these cases the estimated distributions can be inaccurate due to estimation error and the noise added, so aiming for exact SP may fit to data artifacts or the noise rather than the true signal, causing MSE to increase without actual improvements to fairness.

Trade-off Between Statistical Bias and Variance. Recall from the analysis and discussion in Section 3.4 that the MSE of Algorithm 1 exhibits a trade-off between the statistical bias and variance from the choice of k: using more bins lowers discretization error but suffers from larger estimation error and more noise added for DP, and vice versa. On the other hand, the fairness only depends on the variance. Hence, one can decrease k to achieve higher levels of fairness at the expense of MSE.

In Figure 3, we plot the trade-offs on the Law School dataset for $\varepsilon = 0.1$ under different settings of k. On the right extreme, exact SP ($\Delta_{\text{SP}} = 0$) is achieved by setting $k = 1$,

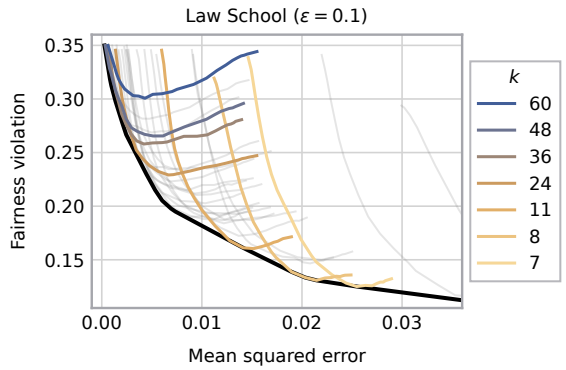

Figure 3. Error-fairness trade-offs achieved by Algorithm 1 on the Law School dataset by sweeping α and k for $\varepsilon = 0.1$. The black line is the lower envelope, and ends on the right at $(0.6772, 0)$ (outside the cropped figure).

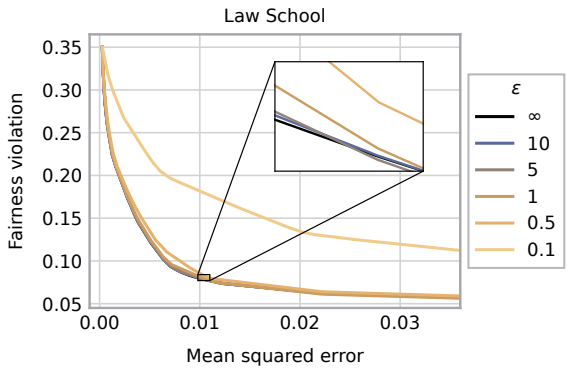

Figure 4. Error-privacy-fairness trade-offs achieved by Algorithm 1 on the Law School dataset by sweeping α and k and taking the lower envelope. The line for $\varepsilon = 0.1$ is the black line in Figure 3. All lines meet and end on the right at $(0.6772, 0)$.

but with a high 0.6772 MSE. As expected:

1. Smaller settings of k lead to smaller Δ_{SP}, but it comes with higher discretization error, as reflected in the rightward-shifting starting points (i.e., results with $\alpha = \infty$, i.e., not post-processed for fairness).
2. Using more bins may result in worse trade-offs compared to less bins due to large variance. Note that the trade-offs with $k = 60$ bins are almost completely dominated by those with $k = 36$ bins.
3. The general trend is that, to achieve the best trade-offs (i.e., the lower envelope), use smaller (α, k) when aiming for higher levels of fairness, and larger (α, k) for smaller MSE (but less fairness).

Error-Privacy-Fairness Trade-Off. The black line in Figure 3 shows the optimal error-fairness trade-offs attainable with Algorithm 1 for $\varepsilon = 0.1$ from sweeping α and k and taking the lower envelope (segments of the line not reached by any k can be achieved by combining two regressors via randomization, although obtaining them via post-processing requires double the privacy budget).

We repeat this experiment for all settings of ε on the Law School dataset, and plot their lower envelopes in Figure 4. This illustrates the Pareto front of the error-privacy-fairness trade-offs achieved by Algorithm 1. Demanding stricter privacy degrades the trade-offs between error and fairness.

Lastly, we remark that while Figures 2 to 4 illustrate the trade-offs that can be possibly attained with our post-processing algorithm from varying the hyperparameters α and k, selecting the desired trade-off requires tuning them (privately) on a validation set. Readers are referred to (Liu & Talwar, 2019; Papernot & Steinke, 2022) on the topic of differentially private hyperparameter tuning.

5. Conclusion and Future Work

In this paper, we described and analyzed a private post-processing algorithm for learning attribute-aware fair regressors, in which privacy is achieved by performing histogram density estimates of the distributions privately, and fairness by computing the optimal transports to their Wasserstein barycenter. We evaluated the error-privacy-fairness trade-offs attained by our algorithm on two datasets.

Although we only studied the attribute-aware setting, that is, we have explicit access to A during training and when making predictions, our post-processing algorithm could be extended to the attribute-blind setting, where A is only available in the training data. This requires training an extra predictor for the sensitive attribute, $\widehat{P}_A = \widehat{\mathbb{P}}(A \mid X) \in \Delta^{|\mathcal{A}|-1}$, to estimate the joint distribution of $(\widehat{Y}, \widehat{P}_A)$ (vs. the joint of $(\widehat{Y}, A)$ estimated in Algorithm 1 for the attribute-aware setting), and modifying P (and LP) to use predicted group membership $\widehat{P}_A$ rather than the true A. This has a higher sample complexity, and the fairness of the post-processed regressor will additionally depend on the accuracy of $\widehat{P}_A$. We leave the implementation and analysis of this extension to future work.

Acknowledgements

GK is supported by a Canada CIFAR AI Grant, an NSERC Discovery Grant, and an unrestricted gift from Google. HZ is partially supported by a research grant from the Amazon-Illinois Center on AI for Interactive Conversational Experiences (AICE) and a Google Research Scholar Award.

Broader Impacts

This paper continues the study of privacy and fairness in machine learning, and fills the gap in prior work on private and fair regression. Since the setting is well-established, and we have theoretically analyzed the risk of our algorithm, we do not find outstanding societal consequences for discussion.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep Learning with Differential Privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016.

Aden-Ali, I., Ashtiani, H., and Kamath, G. On the Sample Complexity of Privately Learning Unbounded High-Dimensional Gaussians. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, 2021.

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. A Reductions Approach to Fair Classification. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Agarwal, A., Dudík, M., and Wu, Z. S. Fair Regression: Quantitative Definitions and Reduction-Based Algorithms. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Agarwal, S. Trade-Offs between Fairness and Privacy in Machine Learning. In *IJCAI 2020 Workshop on AI for Social Good*, 2020.

Alabi, D., McMillan, A., Sarathy, J., Smith, A., and Vadhan, S. Differentially Private Simple Linear Regression. *Proceedings on Privacy Enhancing Technologies*, 2022.

Altschuler, J. M. and Boix-Adserà, E. Wasserstein Barycenters can be Computed in Polynomial Time in Fixed Dimension. *Journal of Machine Learning Research*, 22(44), 2021.

Anderes, E., Borgwardt, S., and Miller, J. Discrete Wasserstein Barycenters: Optimal Transport for Discrete Data. *Mathematical Methods of Operations Research*, 84(2), 2016.

Bagdasaryan, E., Poursaeed, O., and Shmatikov, V. Differential Privacy Has Disparate Impact on Model Accuracy. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Barocas, S. and Selbst, A. D. Big Data's Disparate Impact. *California Law Review*, 104(3), 2016.

Barocas, S., Hardt, M., and Narayanan, A. *Fairness and Machine Learning: Limitations and Opportunities*. The MIT Press, 2023.

Bassily, R., Smith, A., and Thakurta, A. Private Empirical Risk Minimization: Efficient Algorithms and Tight Error Bounds. In *IEEE 55th Annual Symposium on Foundations of Computer Science*, 2014.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research*, 50 (1), 2021.

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., and Kalai, A. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Bun, M., Kamath, G., Steinke, T., and Wu, S. Z. Private Hypothesis Selection. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Buolamwini, J. and Gebru, T. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency*, 2018.

Calders, T., Kamiran, F., and Pechenizkiy, M. Building Classifiers with Independency Constraints. In *2009 IEEE International Conference on Data Mining Workshops*, 2009.

Calmon, F. P., Wei, D., Vinzamuri, B., Ramamurthy, K. N., and Varshney, K. R. Optimized Pre-Processing for Discrimination Prevention. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Chaudhuri, K., Monteleoni, C., and Sarwate, A. D. Differentially Private Empirical Risk Minimization. *Journal of Machine Learning Research*, 12(29), 2011.

Chi, J., Tian, Y., Gordon, G. J., and Zhao, H. Understanding and Mitigating Accuracy Disparity in Regression. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Chizat, L., Roussillon, P., Léger, F., Vialard, F.-X., and Peyré, G. Faster Wasserstein Distance Estimation with the Sinkhorn Divergence. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Chouldechova, A. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, 5(2), 2017.

Chzhen, E., Denis, C., Hebiri, M., Oneto, L., and Pontil, M. Fair Regression with Wasserstein Barycenters. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Covington, C., He, X., Honaker, J., and Kamath, G. Unbiased Statistical Estimation and Valid Confidence Intervals Under Differential Privacy, 2021. *arxiv:2110.14465 [stat.ME]*.

Cummings, R., Gupta, V., Kimpara, D., and Morgenstern, J. On the Compatibility of Privacy and Fairness. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, 2019.

Cuturi, M. and Doucet, A. Fast Computation of Wasserstein Barycenters. In *Proceedings of the 31st International Conference on Machine Learning*, 2014.

Diakonikolas, I., Hardt, M., and Schmidt, L. Differentially Private Learning of Structured Discrete Distributions. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Dwork, C. and Roth, A. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4), 2014.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating Noise to Sensitivity in Private Data Analysis. In *Theory of Cryptography*, 2006.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness Through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 2012.

Fletcher, S. and Islam, M. Z. Decision Tree Classification with Differential Privacy: A Survey. *ACM Computing Surveys*, 52(4), 2019.

Hardt, M. and Talwar, K. On the Geometry of Differential Privacy. In *Proceedings of the Forty-Second ACM Symposium on Theory of Computing*, 2010.

Hardt, M., Price, E., and Srebro, N. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Hébert-Johnson, Ú., Kim, M. P., Reingold, O., and Rothblum, G. N. Multicalibration: Calibration for the (Computationally-Identifiable) Masses. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Jagielski, M., Kearns, M., Mao, J., Oprea, A., Roth, A., Sharifi-Malvajerdi, S., and Ullman, J. Differentially Private Fair Learning. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J. Fairness-Aware Classifier with Prejudice Remover Regularizer. In *Machine Learning and Knowledge Discovery in Databases*, 2012.

Kifer, D., Smith, A., and Thakurta, A. Private Convex Empirical Risk Minimization and High-dimensional Regression. In *Proceedings of the 25th Annual Conference on Learning Theory*, 2012.

Le Gouic, T., Loubes, J.-M., and Rigollet, P. Projection to Fairness in Statistical Learning, 2020. *arxiv:2005.11720 [cs.LG]*.

Li, J. and Tkocz, T. Tail Bounds for Sums of Independent Two-Sided Exponential Random Variables. In *High Dimensional Probability IX*, 2023.

Liu, J. and Talwar, K. Private Selection from Private Candidates. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, 2019.

Mironov, I. Rényi Differential Privacy. In *IEEE 30th Computer Security Foundations Symposium*, 2017.

Mohapatra, S., Sasy, S., He, X., Kamath, G., and Thakkar, O. The Role of Adaptive Optimizers for Honest Private Hyperparameter Selection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(7), 2022.

Mozannar, H., Ohannessian, M. I., and Srebro, N. Fair Learning with Private Demographic Data. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Papernot, N. and Steinke, T. Hyperparameter Tuning with Renyi Differential Privacy. In *The Tenth International Conference on Learning Representations*, 2022.

Papernot, N., Abadi, M., Erlingsson, Ú., Goodfellow, I., and Talwar, K. Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data. In *The Fifth International Conference on Learning Representations*, 2017.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. On Fairness and Calibration. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Redmond, M. and Baveja, A. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research*, 141(3), 2002.

Rudemo, M. Empirical Choice of Histograms and Kernel Density Estimators. *Scandinavian Journal of Statistics*, 9 (2), 1982.

Song, S., Chaudhuri, K., and Sarwate, A. D. Stochastic gradient descent with differentially private updates. In *IEEE Global Conference on Signal and Information Processing*, 2013.

Staib, M., Claici, S., Solomon, J., and Jegelka, S. Parallel Streaming Wasserstein Barycenters. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Stout, Q. F. L_∞ Isotonic Regression for Linear, Multidimensional, and Tree Orders, 2017. *arxiv:1507.02226 [cs.DS]*.

Tran, C., Fioretto, F., and Hentenryck, P. V. Differentially Private and Fair Deep Learning: A Lagrangian Dual Approach. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(11), 2021.

Vadhan, S. The Complexity of Differential Privacy. In *Tutorials on the Foundations of Cryptography: Dedicated to Oded Goldreich*. Springer International Publishing, 2017.

Wang, Y.-X. Revisiting differentially private linear regression: Optimal and adaptive prediction & estimation in unbounded domain. In *Proceedings of the 34th Uncertainty in Artificial Intelligence Conference*, 2018.

Weissman, T., Ordentlich, E., Seroussi, G., Verdu, S., and Weinberger, M. J. Inequalities for the L_1 Deviation of the Empirical Distribution. Technical Report HPL-2003-97R1, Hewlett-Packard Laboratories, 2003.

Wightman, L. F. *LSAC National Longitudinal Bar Passage Study*. Law School Admission Council, 1998.

Xian, R., Yin, L., and Zhao, H. Fair and Optimal Classification via Post-Processing. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Xu, D., Yuan, S., and Wu, X. Achieving Differential Privacy and Fairness in Logistic Regression. In *Companion Proceedings of the 2019 World Wide Web Conference*, 2019.

Xu, J., Zhang, Z., Xiao, X., Yang, Y., and Yu, G. Differentially Private Histogram Publication. In *IEEE 28th International Conference on Data Engineering*, 2012.

Zemel, R., Wu, Y. L., Swersky, K., Pitassi, T., and Dwork, C. Learning Fair Representations. In *Proceedings of the 30th International Conference on Machine Learning*, 2013.

Zhao, H. and Gordon, G. J. Inherent Tradeoffs in Learning Fair Representations. *Journal of Machine Learning Research*, 23(57), 2022.

A. Excess Risk of the Optimal Fair Regressor

This section proves Theorem 3.1, that the excess risk of the optimal attribute-aware fair regressor can be expressed as the sum of Wasserstein distances from the output distributions r_a^* of the Bayes regressor f^* conditioned on each group $a \in \mathcal{A}$ to their barycenter q. We will cover the approximate fairness setting using the same analysis in (Xian et al., 2023).

The result is a direct consequence of Lemma A.4—but before stating which, because the fair regressor is randomized, to make the discussions involving randomized functions rigorous, we provide a formal definition for them with the Markov kernel.

Definition A.1 (Markov Kernel)**.** A Markov kernel from a measurable space $(\mathcal{X}, \mathcal{S})$ to $(\mathcal{Y}, \mathcal{T})$ is a mapping $\mathcal{K} : \mathcal{X} \times \mathcal{T} \to [0, 1]$, such that $\mathcal{K}(\cdot, T)$ is $\mathcal{S}$-measurable, $\forall T \in \mathcal{T}$, and $\mathcal{K}(x, \cdot)$ is a probability measure on $(\mathcal{Y}, \mathcal{T})$, $\forall x \in \mathcal{X}$.

Definition A.2 (Randomized Function)**.** A randomized function $f : (\mathcal{X}, \mathcal{S}) \to (\mathcal{Y}, \mathcal{T})$ is associated with a Markov kernel $\mathcal{K} : \mathcal{X} \times \mathcal{T} \to [0, 1]$, such that $\forall x \in \mathcal{X}, T \in \mathcal{T}, \mathbb{P}(f(x) \in T) = \mathcal{K}(x, T)$.

Definition A.3 (Push-Forward Distribution)**.** Let p be a measure on $(\mathcal{X}, \mathcal{S})$ and $f : (\mathcal{X}, \mathcal{S}) \to (\mathcal{Y}, \mathcal{T})$ a randomized function with Markov kernel $\mathcal{K}$. The push-forward of p under f, denoted by $f\sharp p$, is a measure on $\mathcal{Y}$ given by $f\sharp p(T) = \int_{\mathcal{X}} \mathcal{K}(x, T)\, \mathrm{d}p(x)$, $\forall T \in \mathcal{T}$.

Now, we state the lemma of which Theorem 3.1 is a direct consequence; it says that given any (randomized) regressor f with a particular shape q, one can derive a regressor $g \circ f^*$ from the Bayes regressor f^* that has the same shape and excess risk (g is a randomized function with Markov kernel $\mathcal{K}(y, T) = \pi(y,T)/\pi(y,\mathbb{R})$ where π is given in Equation (3)):

Lemma A.4. *Let a regression problem be given by a joint distribution μ of (X, Y), denote the Bayes regressor by $f^* : \mathcal{X} \to \mathbb{R}$ and $r^* = f^*\sharp\mu_X$, and let q be an arbitrary distribution on $\mathbb{R}$. Then, for any randomized regressor f with Markov kernel $\mathcal{K}$ satisfying $f\sharp\mu_X = q$,*

$$\pi(y^*, y) = \int_{f^{*-1}(y^*)} \mathcal{K}(x, y)\, \mathrm{d}\mu_X(x) \tag{3}$$

(where $f^{*-1}(y^*) \coloneqq \{x \in \mathcal{X} : f^*(x) = y^*\}$) is a coupling $\pi \in \Pi(r^*, q)$ that satisfies

$$ER(f) = \int (y^* - y)^2\, \mathrm{d}\pi(y^*, y). \tag{4}$$

Conversely, for any $\pi \in \Pi(r^*, q)$, the randomized regressor f with Markov kernel

$$\mathcal{K}(x, T) = \frac{\pi(f^*(x), T)}{\pi(f^*(x), \mathbb{R})}$$

satisfies $f\sharp\mu_X = q$ and Equation (4).

Proof. For the first direction, note that

$$\begin{aligned}
ER(f) &= \mathbb{E}[(f^*(X) - f(X))^2] \\
&= \int_{\mathbb{R}\times\mathbb{R}} (y^* - y)^2\, \mathbb{P}(f^*(X) = y^*, f(X) = y)\, \mathrm{d}(y^*, y) \\
&= \int_{\mathbb{R}\times\mathbb{R}} (y^* - y)^2 \left(\int_{\mathcal{X}} \mathbb{P}(f^*(X) = y^*, f(X) = y, X = x)\, \mathrm{d}x \right) \mathrm{d}(y^*, y) \\
&= \int_{\mathbb{R}\times\mathbb{R}} (y^* - y)^2 \left(\int_{f^{*-1}(y^*)} \mathbb{P}(f(X) = y, X = x)\, \mathrm{d}x \right) \mathrm{d}(y^*, y) \\
&= \int_{\mathbb{R}\times\mathbb{R}} (y^* - y)^2 \left(\int_{f^{*-1}(y^*)} \mathbb{P}(f(X) = y \mid X = x)\, \mathrm{d}\mu_X(x) \right) \mathrm{d}(y^*, y) \\
&= \int_{\mathbb{R}\times\mathbb{R}} (y^* - y)^2 \pi(y^*, y)\, \mathrm{d}(y^*, y)
\end{aligned} \tag{5}$$

as desired, where line 4 is because $\mathbb{P}(f^*(X) = y^*, f(X) = y, X = x) = \mathbb{1}[f^*(x) = y^*]\,\mathbb{P}(f(X) = y, X = x)$ as f^* is deterministic. We also verify that the constructed π is a coupling:

$$\begin{aligned}
\int_{\mathbb{R}} \pi(y^*, y)\,\mathrm{d}y &= \int_{\mathbb{R}} \int_{f^{*-1}(y^*)} \mathcal{K}(x, y)\,\mathrm{d}\mu_X(x)\,\mathrm{d}y \\
&= \int_{f^{*-1}(y^*)} \int_{\mathbb{R}} \mathcal{K}(x, y)\,\mathrm{d}y\,\mathrm{d}\mu_X(x) \\
&= \int_{f^{*-1}(y^*)} \mathrm{d}\mu_X(x) \\
&= \mathbb{P}(f^*(X) = y^*) \\
&= r^*(y^*)
\end{aligned}$$

by Definitions A.1 and A.3, and

$$\begin{aligned}
\int_{\mathbb{R}} \pi(y^*, y)\,\mathrm{d}y^* &= \int_{\mathbb{R}} \int_{f^{*-1}(y^*)} \mathcal{K}(x, y)\,\mathrm{d}\mu_X(x)\,\mathrm{d}y^* \\
&= \int_{\mathcal{X}} \mathcal{K}(x, y)\,\mathrm{d}\mu_X(x) \\
&= \int_{\mathcal{X}} \mathbb{P}(f(X) = y \mid X = x)\,\mathrm{d}\mu_X(x) \\
&= \mathbb{P}(f(X) = y) \\
&= q(y)
\end{aligned}$$

by Definition A.2 and the assumption that $f\sharp\mu_X = q$.

For the converse direction, it suffices to show that the Markov kernel constructed for f satisfies the equality in Equation (3), which would immediately imply $f\sharp\mu_X = q$, and Equation (4) with the same arguments in Equation (5). Let $y, y \in \mathbb{R}$ and $z \in f^{*-1}(y^*)$ be arbitrary, then

$$\begin{aligned}
\pi(y^*, y) &= \frac{\pi(y^*, y)}{\pi(y^*, \mathbb{R})}\pi(y^*, \mathbb{R}) \\
&= \frac{\pi(f^*(z), y)}{\pi(f^*(z), \mathbb{R})}\pi(y^*, \mathbb{R}) \\
&= \mathcal{K}(z, y)\pi(y^*, \mathbb{R}) \\
&= \mathcal{K}(z, y)r^*(y^*) \\
&= \mathcal{K}(z, y)\int_{f^{*-1}(y^*)} \mathrm{d}\mu_X(x) \\
&= \int_{f^{*-1}(y^*)} \mathcal{K}(z, y)\,\mathrm{d}\mu_X(x) \\
&= \int_{f^{*-1}(y^*)} \mathcal{K}(x, y)\,\mathrm{d}\mu_X(x),
\end{aligned}$$

where line 3 is by construction of $\mathcal{K}$, line 4 by the assumption that $\pi \in \Pi(r^*, q)$, and the last line is because $\mathcal{K}(x, y) = \mathcal{K}(z, y)$ for all $x \in f^{*-1}(y^*)$, also by construction. □

This lemma allows us to formulate the problem of finding the optimal regressor under a shape constraint q as that of finding the optimal coupling $\pi \in \Pi(r^*, q)$ with the squared cost (and π can be used to derive the regressor $g \circ f^*$ that achieves the minimum of the original problem). Because statistical parity is a shape constraint on the regressors, we can leverage this lemma to prove Theorem 3.1:

Proof of Theorem 3.1. Because we are finding an attribute-aware fair regressor, $f : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$, we can optimize the components corresponding to each group independently, i.e., $f_a \coloneqq f(\cdot, a), \forall a \in \mathcal{A}$. Denote the excess risk conditioned on

group a by $ER_a(f_a) = \mathbb{E}[(f_a(X) - Y)^2 \mid A = a]$, and the marginal distribution of X conditioned on $A = a$ by $\mu_{X|a}$, then

$$\begin{aligned}
\min_{f:\Delta_{\text{SP}}(f)\le\alpha} ER(f) &= \min_{f:D_{\text{KS}}(f_a\sharp\mu_{X|a}, f_{a'}\sharp\mu_{X|a})\le\alpha} ER(f) \\
&= \min_{f:D_{\text{KS}}(r_a(f), r_{a'}(f))\le\alpha} \sum_{a\in\mathcal{A}} w_a\, ER_a(f_a) \\
&= \min_{\{f_a\}_{a\in\mathcal{A}}:D_{\text{KS}}(f_a\sharp\mu_{X|a}, f_{a'}\sharp\mu_{X|a'})\le\alpha} \sum_{a\in\mathcal{A}} w_a\, ER_a(f_a) \\
&= \min_{\{q_a\}_{a\in\mathcal{A}}:D_{\text{KS}}(q_a,q_{a'})\le\alpha} \sum_{a\in\mathcal{A}} w_a \min_{f_a:f_a\sharp\mu_{X|a}=q_a} ER_a(f_a) \\
&= \min_{\substack{q:\text{supp}(q)\subseteq\mathbb{R} \\ \{q_a\}_{a\in\mathcal{A}}\subset B_{\text{KS}}(q,\alpha/2)}} \sum_{a\in\mathcal{A}} w_a \min_{f_a:f_a\sharp\mu_{X|a}=q_a} ER_a(f_a)
\end{aligned}$$

(the proof that $D_{\text{KS}}(\mu,\nu) \le \alpha \iff \exists q$ s.t. $\mu,\nu \in B_{\text{KS}}(q, \alpha/2)$ is omitted; a hint for the forward direction is that such a q can be constructed by averaging the CDFs of μ,ν), where, by Lemma A.4 and the definition of Wasserstein distance,

$$\min_{f_a:f_a\sharp\mu_{X|a}=q_a} ER_a(f_a) = \min_{\pi_a\in\Pi(r_a^*,q_a)} \int (y^* - y)^2 \,\mathrm{d}\pi_a(y^*, y) = W_2^2(r_a^*, q_a),$$

and the theorem follows by plugging this back into the previous equation. □

B. Proofs for Section 3

We analyze the L^1 sensitivity for our notion of neighboring datasets described in Section 2, and provide the proofs to Theorems 3.2 and 3.3, in that order.

Remark B.1. For nonempty neighboring datasets S, S' that differ in at most one entry by insertion, deletion or substitution, the L^1 sensitivity to the empirical PMF is at most $2/n$.

Let $\hat{p}, \hat{p}'$ denote the empirical PMFs of S and S', respectively, assume w.l.o.g. that they have two coordinate, and the insertion/deletion takes place in the first coordinate. Denote $n = |S|$, $n_1 = n\hat{p}_1$, and $n_2 = n\hat{p}_2$.

- (Insertion). The sensitivity is

$$\begin{aligned}
\|\hat{p} - \hat{p}'\|_1 &= \left|\frac{n_1}{n} - \frac{n_1+1}{n+1}\right| + \left|\frac{n_2}{n} - \frac{n_2}{n+1}\right| \\
&= \left|\frac{n_1(n+1) - (n_1+1)n}{n(n+1)}\right| + \left|\frac{n_2(n+1) - n_2 n}{n(n+1)}\right| \\
&= \left|\frac{n_1 - n}{n(n+1)}\right| + \left|\frac{n_2}{n(n+1)}\right| \\
&= 2\left|\frac{n_2}{n(n+1)}\right| \\
&\le \frac{2}{n}.
\end{aligned}$$

- (Deletion). Similarly, $\|\hat{p} - \hat{p}'\|_1 = |n_1/n - n_1-1/n-1| + |n_2/n - n_2/n-1| = 2|n_2/n(n-1)| \le 2/n$, because $n_2 \le n-1$.
- (Substitution). $\|\hat{p} - \hat{p}'\|_1 = |n_1/n - n_1-1/n| + |n_2/n - n_2+1/n| = |1/n| + |1/n| = 2/n$.

For the proofs of the theorems, several technical results are required. First, a concentration bound of i.i.d. sum of Laplace random variables based on the following fact (Li & Tkocz, 2023), which is due to $\sqrt{2Y_j}Z \sim \text{Laplace}(0,1)$ and $\sum_{j=1}^k a_j Z_j \sim \mathcal{N}(0, \sum_{j=1}^k a_j^2)$ for $Z_1,\dots,Z_k \sim \mathcal{N}(0,1)$ and $a_1,\dots,a_k \ge 0$:

Proposition B.2. *Let $X_1,\dots,X_k \sim \text{Laplace}(0,1)$, $Y_1,\dots,Y_k \sim \text{Exponential}(1)$, and $Z \sim \mathcal{N}(0,1)$ all be independent, then $\sum_{j=1}^k X_j$ has the same distribution as $(2\sum_{j=1}^k Y_j)^{1/2} Z$.*

Lemma B.3. *Let independent $X_1,\dots,X_k \sim \text{Laplace}(0,1)$, then for all $t \ge 0$, with probability at least $1-\beta$, $|\sum_{j=1}^k X_j| \le 2\sqrt{k}\ln 2k/\beta$.*

Proof. Using Proposition B.2, we bound $\sum_{j=1}^k X_j$ by analyzing $\sqrt{\sum_{j=1}^k Y_j}|Z|$. For all $t \geq 0$,

$$\mathbb{P}\left(\sum_{j=1}^k Y_j \geq t\right) \leq \mathbb{P}\left(\exists j \text{ s.t. } Y_j \geq \frac{t}{k}\right) \leq k\,\mathbb{P}\left(Y_1 \geq \frac{t}{k}\right) \leq k\exp\left(-\frac{t}{k}\right).$$

On the other hand, the Chernoff bound implies that $\mathbb{P}(|Z| \geq t) \leq 2\exp(-t^2/2)$. With a union bound, with probability at least $1-\beta$,

$$\left|\sqrt{2\sum_{j=1}^k Y_j Z}\right| = \sqrt{2\sum_{j=1}^k Y_j}|Z| \leq 2\sqrt{k}\ln\frac{2k}{\beta}. \qquad \square$$

Next, an L^1 (TV) convergence result of empirical distributions with finite support, which follows from the concentration of i.i.d. sum of Multinoulli random variables:

Theorem B.4 (Weissman et al., 2003)**.** *Let $p \in \Delta^{d-1}$ the $(d-1)$ simplex and $\hat{p}_n \sim {}^1\!/_n\mathrm{Multinomial}(n,p)$, then with probability at least $1-\beta$, $\|p-\hat{p}_n\|_1 \leq \sqrt{2d/n\ln 2/\beta}$.*

Lemma B.5. *Let independent $x_1,\ldots,x_n \sim p$ with finite support $\mathcal{X}$, and denote the empirical distribution by $\hat{p}_n = {}^1\!/_n\sum_{i=1}^n \beta_{x_i}$, then with probability at least $1-\beta$, $\|p-\hat{p}_n\|_1 \leq \sqrt{2|\mathcal{X}|/n\ln 2/\beta}$.*

For the proof of Theorem 3.2, we need two technical results:

Lemma B.6. *Let independent $x_1,\ldots,x_n \sim p$ supported on $[k]$, and denote the empirical PMF by $\hat{p}_j = {}^1\!/_n\sum_i \mathbb{1}[x_i = j]$, for all $j \in [k]$. Let $E \subseteq [k]$ be a subset of size ℓ, denote $w_E = \mathbb{P}(x \in E)$, the PMF conditioned on the event $x \in E$ by $p_{|E}$, and its empirical counterpart by $\hat{p}_{j|E} = {}^1\!/_{n_E}\sum_i \mathbb{1}[x_i = j]$, where $n_E = \sum_i \mathbb{1}[x_j \in E]$. Then for all $n \geq {}^8\!/_{w_E}\ln 8/\beta$, with probability at least $1-\beta$,*

$$\|p_{|E} - \hat{p}_{|E}\|_\infty \leq \sqrt{\frac{1}{nw_E}\ln\frac{8\ell}{\beta}},$$

$$D_{\mathrm{TV}}(p_{|E},\hat{p}_{|E}) = \frac{1}{2}\|p_{|E}-\hat{p}_{|E}\|_1 \leq \sqrt{\frac{\ell}{4nw_E}\ln\frac{8}{\beta}},$$

$$D_{\mathrm{KS}}(p_{|E},\hat{p}_{|E}) \leq \sqrt{\frac{1}{nw_E}\ln\frac{8\ell}{\beta}}.$$

Proof. By Chernoff bound on the Binomial distribution, for all $n \geq {}^8\!/_{w_E}\ln 2/\beta$, with probability at least $1-\beta$,

$$\frac{nw_E}{2} \leq n_E \leq 2nw_E. \tag{6}$$

Order the samples so that the ones with $x_i \in E$ are at the front (or, consider first sample n_E, then sample the first n_E samples from $p_{|E}$). Conditioned on Equation (6), by Hoeffding's inequality and a union bound, for all $j \in E$, with probability at least $1-\beta$,

$$\begin{aligned}
|p_{j|E} - \hat{p}_{j|E}| &= \left|\frac{1}{w_E}\mathbb{P}(X=j) - \frac{1}{n_E}\sum_{i=1}^{n_E}\mathbb{1}[x_i = j]\right| \\
&= \frac{1}{w_E}\left|\mathbb{P}(X=j) - \frac{1}{n_E}\sum_{i=1}^{n_E} w_E\,\mathbb{1}[x_i = j]\right| \\
&\leq \frac{1}{w_E}\sqrt{\frac{w_E^2}{2n_E}\ln\frac{2\ell}{\beta}} \\
&\leq \sqrt{\frac{1}{nw_E}\ln\frac{2\ell}{\beta}}.
\end{aligned}$$

Next, by Lemma B.5, with probability at least $1-\beta$,

$$\begin{aligned}
D_{\text{TV}}(p_{|E}, \hat{p}_{|E}) &= \frac{1}{2}\sum_{j\in E}\left|p_{j|E} - \hat{p}_{j|E}\right| \\
&= \frac{1}{2}\sum_{j\in E}\left|\frac{1}{w_E}P(X=j) - \frac{1}{n_E}\sum_{i=1}^{n_E}\mathbb{1}[x_i = j]\right| \\
&\le \frac{1}{2}\sqrt{\frac{\ell}{2n_E}\ln\frac{2}{\beta}} \\
&\le \sqrt{\frac{\ell}{4nw_E}\ln\frac{2}{\beta}}.
\end{aligned}$$

Lastly, D_{KS} computes the L^∞-distance between two CDFs, so similar to the ℓ_∞ bound, by Hoeffding's inequality and a union bound, with probability at least $1-\beta$,

$$\begin{aligned}
D_{\text{KS}}(p_{|E}, \hat{p}_{|E}) &= \max_{j\in E}\left|\frac{1}{w_E}\,\mathbb{P}(X\le j) - \frac{1}{n_E}\sum_{i=1}^{n_E}\mathbb{1}[x_i \le j]\right| \\
&= \frac{1}{w_E}\left|\mathbb{P}(X\le j) - \frac{1}{n_E}\sum_{i=1}^{n_E} w_E\,\mathbb{1}[x_i \le j]\right| \\
&\le \frac{1}{w_E}\sqrt{\frac{w_E^2}{2n_E}\ln\frac{2\ell}{\beta}} \\
&\le \sqrt{\frac{1}{nw_E}\ln\frac{2\ell}{\beta}}.
\end{aligned}$$

The result follows by taking a final union bound over the four events during the analysis and rescaling $\beta \leftarrow \beta/4$. □

Lemma B.7. *Let constants $a_1, \ldots, a_k \ge 0$, and independent $\xi_1, \ldots, \xi_k \sim \text{Laplace}(0, b)$. Denote $S_j = \sum_{\ell=1}^{j} a_j$, $s = S_k$, and let $t \ge 0$. Define for all $j \in [k]$,*

(add noise) $\quad x_j \coloneqq a_j + \xi_j, \qquad F_j = \sum_{k=1}^{j} x_k,$

(isotonic regression) $\quad y_j = G_j - G_{j-1}, \qquad G_j \coloneqq \frac{1}{2}\big(F_{l_j} + F_{r_j}\big),$

(clipping) $\quad z_j = H_j - H_{j-1}, \qquad H_j \coloneqq \begin{cases} \text{proj}_{[0,t]}\, G_j & \textit{if } j < k \\ t & \textit{else}, \end{cases}$

where

$$(l_j, r_j) = \arg\max_{l\le j\le r}(F_l - F_r).$$

Then with probability at least $1-\beta$,

$$\begin{aligned}
\|a - z\|_1 &\le 3|s-t| + 74bk\ln\frac{4k}{\beta}, \\
\|a - z\|_\infty &\le 2|s-t| + 32b\sqrt{k}\ln\frac{4k}{\beta}, \\
\|S - H\|_\infty &\le |s-t| + 12b\sqrt{k}\ln\frac{4k}{\beta}.
\end{aligned}$$

Proof. Our analysis for proceeds by using the triangle inequality to decompose into and bounding each of the following terms (shown here for $\|\cdot\|_1$, analogously for $\|\cdot\|_\infty$ and the partial sums):

$$\|a-z\|_1 \le \|a-x\|_1 + \|x-y\|_1 + \|y-z\|_1. \tag{7}$$

We will use the following concentration result of Laplace random variables: by Lemma B.3, with probability at least $1-\beta$, for all $0 \le \ell \le m \le k$,

$$\left|\sum_{j=\ell+1}^{m} \xi_j\right| \le 4b\sqrt{m-\ell}\ln\frac{2(m-\ell)k^2}{\beta} \le 12b\sqrt{m-\ell}\ln\frac{2k}{\beta}. \tag{8}$$

First Term in Equation (7). By the CDF of the exponential distribution (of which the Laplace distribution is the two-sided version), with a union bound, with probability at least $1-\beta$,

$$|a_j - x_j| = |\xi_j| \le 2b\ln\frac{k}{\beta}, \quad \forall j \in [k],$$

and it follows that

$$\|a-x\|_1 = \sum_{j=1}^{k}|a_j - x_j| \le 2bk\ln\frac{k}{\beta}.$$

For the partial sums, by Equation (8),

$$\|S-H\|_\infty = \max_j |S_j - F_j| = \max_j\left|\sum_{\ell=1}^{j}\xi_\ell\right| \le \max_j 12b\sqrt{j}\ln\frac{2k}{\beta} = 12b\sqrt{k}\ln\frac{2k}{\beta}.$$

Second Term in Equation (7). Note that for any $\ell \le m$ such that $F_\ell \ge F_m$ (i.e., a violating pair for isotonic regression),

$$G_\ell - G_m = F_\ell - F_\ell - \sum_{j=\ell+1}^{m} a_\ell - \sum_{j=\ell+1}^{m}\xi_\ell \le -\sum_{j=\ell+1}^{m}\xi_\ell \le 12b\sqrt{m-\ell}\ln\frac{2k}{\beta}, \tag{9}$$

because $a_j \ge 0$. So for all $j \in [k]$,

$$\begin{aligned}
0 \le |x_j - y_j| &= |G_j - G_{j-1} - (F_j - F_{j-1})| \\
&\le |G_j - F_j| + |G_{j-1} - F_{j-1}| \\
&= \begin{Bmatrix} G_j - F_j & \text{if } G_j > F_j \\ F_j - G_j & \text{else} \end{Bmatrix} + \begin{Bmatrix} G_{j-1} - F_{j-1} & \text{if } G_{j-1} > F_{j-1} \\ F_{j-1} - G_{j-1} & \text{else} \end{Bmatrix} \\
&= \begin{Bmatrix} G_j - \dfrac{G_{l_j} + G_{r_j}}{2} & \text{if } G_j > F_j \\ \dfrac{G_{l_j} + G_{r_j}}{2} - G_j & \text{else} \end{Bmatrix} + \begin{Bmatrix} G_{j-1} - \dfrac{G_{l_{j-1}} + G_{r_{j-1}}}{2} & \text{if } G_{j-1} > F_{j-1} \\ \dfrac{G_{l_{j-1}} + G_{r_{j-1}}}{2} - G_{j-1} & \text{else} \end{Bmatrix} \\
&\le \begin{Bmatrix} G_j - \dfrac{G_j + G_{r_j}}{2} & \text{if } G_j > F_j \\ \dfrac{G_{l_j} + G_j}{2} - G_j & \text{else} \end{Bmatrix} + \begin{Bmatrix} G_{j-1} - \dfrac{G_{j-1} + G_{r_{j-1}}}{2} & \text{if } G_{j-1} > F_{j-1} \\ \dfrac{G_{l_{j-1}} + G_{j-1}}{2} - G_{j-1} & \text{else} \end{Bmatrix} \\
&= \frac{1}{2}\begin{Bmatrix} G_j - G_{r_j} & \text{if } G_j > F_j \\ G_{l_j} - G_j & \text{else} \end{Bmatrix} + \frac{1}{2}\begin{Bmatrix} G_{j-1} - G_{r_{j-1}} & \text{if } G_{j-1} > F_{j-1} \\ G_{l_{j-1}} - G_{j-1} & \text{else} \end{Bmatrix} \\
&\le 12b\sqrt{\max(r_j - j, j - l_j)}\ln\frac{2k}{\beta} \le 12b\sqrt{\max(k-j, j)}\ln\frac{2k}{\beta},
\end{aligned}$$

where line 5 is because $G_{r_m} \le G_m \le G_{l_m}$ for all j, and line 7 by Equation (9). It then follows that

$$\|x-y\|_1 = \sum_{j=1}^{k}|x_j - y_j| \le 24bk\ln\frac{2k}{\beta}.$$

Lastly, using the fact that the L^∞ error of L^∞ isotonic regression is $1/2 \max_{\ell \le m}(G_\ell - G_m)$ (Stout, 2017), and by Equation (9) again,

$$0 \le \|F - G\|_\infty = \frac{1}{2} \max_{\ell \le m}(G_\ell - G_m) \le 6b\sqrt{k} \ln \frac{2k}{\beta}.$$

Third Term in Equation (7). Note that because $a_j \in [0, s]$, by Equation (8),

$$\min_j G_j = \min_j \frac{1}{2}(F_{l_j} + F_{r_j}) \ge \min_j F_j = \min_j \sum_{m=1}^{j} (a_m + \xi_m) \ge \min_j \sum_{m=1}^{j} \xi_m \ge -\max_j \left| \sum_{j=1}^{k} \xi_j \right| \ge -12b\sqrt{k} \ln \frac{2k}{\beta}, \quad (10)$$

and similarly,

$$\max_j G_j \le \max_j \sum_{m=1}^{j} (a_m + \xi_m) \le s + 12b\sqrt{k} \ln \frac{2k}{\beta}. \quad (11)$$

Since G is nondecreasing after isotonic regression, clipping only affects its prefix and/or suffix. For the prefix, let $l = \max\{j \in [k] : H_j = 0\}$. If l does not exist, then no clipping to zero has occurred. Otherwise, for all $j \le l$, by Equation (10),

$$|G_j - H_j| = \max(-G_j, 0) \le 12b\sqrt{k} \ln \frac{2k}{\beta},$$

and

$$|y_j - z_j| = |G_j - G_{j-1} - (H_j - H_{j-1})| \le -G_j - G_{j-1} \le 2 \max\left(-\min_j G_j, 0\right) \le 24b\sqrt{k} \ln \frac{2k}{\beta}.$$

For the suffix, let $r = \min\{j \in [k] : H_j = t\}$, then for all $r \le j < k$, by Equation (11),

$$|G_j - H_j| = G_j - t \le |s - t| + 12b\sqrt{k} \ln \frac{2k}{\beta},$$

and

$$|y_j - z_j| \le (G_j - t) + \max(G_{j-1} - t, 0) \le 2 \max\left(\max_j G_j - t, 0\right) \le 2|s - t| + 24b\sqrt{k} \ln \frac{2k}{\beta};$$

for $j = k$,

$$|G_k - H_k| \le \begin{cases} G_k - t & \text{if } G_k \ge t \\ t - G_k & \text{else} \end{cases} = \begin{cases} G_k - t & \text{if } G_k \ge t \\ t - \left(s + \sum_{m=1}^{k} \xi_m\right) & \text{else} \end{cases} \le |s - t| + 12b\sqrt{k} \ln \frac{2k}{\beta},$$

and

$$|y_j - z_j| \le |G_j - t| + \max(G_{j-1} - t, 0) \le 2|s - t| + 24b\sqrt{k} \ln \frac{2k}{\beta}.$$

Finally, for $\|\cdot\|_1$,

$$\begin{aligned}
\|y - z\|_1 &= \sum_{j=1}^{l} |G_j - G_{j-1} - (H_j - H_{j-1})| + \sum_{j=r}^{k} |G_j - G_{j-1} - (H_j - H_{j-1})| \\
&= \sum_{j=1}^{l} (G_j - G_{j-1}) + |y_r - z_r| + \sum_{j=r+1}^{k-1} (G_j - G_{j-1}) + |y_k - z_k| \\
&= G_l - G_1 + |y_r - z_r| + G_{k-1} - G_r + |y_k - z_k| \\
&\le -G_1 + |y_r - z_r| + G_{k-1} - t + |y_k - z_k| \\
&\le 12b\sqrt{k} \ln \frac{2k}{\beta} + 2\left(|s - t| + 12b\sqrt{k} \ln \frac{2k}{\beta}\right) + \left(s + 12b\sqrt{k} \ln \frac{2k}{\beta}\right) - t
\end{aligned}$$

$$\leq 3|s-t| + 48b\sqrt{k}\ln\frac{2k}{\beta},$$

keep in mind that $1 \leq l$ and $r \leq k-1, k$ on line 3 and onward.

The result follows by taking a final union bound over the two events above and rescaling $\beta \leftarrow \beta/2$. □

Proof of Theorem 3.2. Because $w_a \geq 0$, by Lemma B.3, with probability at least $1-\beta$,

$$\begin{aligned}
|w_a - \tilde{w}_a| &= \left| w_a - \max\left(w_a + \sum_{j=1}^{k} \text{Laplace}(0, {}^{2}/_{n\varepsilon}), 0 \right) \right| \\
&= \left| \max\left(\sum_{j=1}^{k} \text{Laplace}(0, {}^{2}/_{n\varepsilon}), -w_a \right) \right| \\
&\leq \left| \sum_{j=1}^{k} \text{Laplace}(0, {}^{2}/_{n\varepsilon}) \right| \\
&\leq O\left(\frac{\sqrt{k}}{n\varepsilon} \ln\frac{k|\mathcal{A}|}{\beta} \right). \qquad (12)
\end{aligned}$$

Next, define

$$\hat{p}_a(v_j) = \frac{1}{n_a}\sum_{i=1}^{n} \mathbb{1}[h \circ f(x_i, a_i), a_i = a], \quad \check{p}_a(v_j) = \frac{1}{\tilde{w}_a}\tilde{p}(a, v_j),$$

where $n_a = {}^{1}/_{n}\sum_{i=1}^{n}\mathbb{1}[a_i = a]$. Note that $\check{F}_a(v_j) = \sum_{\ell=1}^{j}\check{p}_a(v_\ell)$. By triangle inequality,

$$\begin{aligned}
&\|p_a - \tilde{p}_a\|_\infty \\
&\quad\leq \|\tilde{p}_a - \check{p}_a\|_\infty + \|\check{p}_a - \hat{p}_a\|_\infty + \|\hat{p}_a - p_a\|_\infty \\
&\quad= \left\| \tilde{p}_a - \frac{1}{\tilde{w}_a}\check{p}(a,\cdot) \right\|_\infty + \left\| \frac{1}{\tilde{w}_a}\check{p}(a,\cdot) - \frac{1}{\hat{w}_a}\hat{p}(a,\cdot) \right\|_\infty + \|\hat{p}_a - p_a\|_\infty \\
&\quad\leq \left\| \tilde{p}_a - \frac{1}{\hat{w}_a}\check{p}(a,\cdot) \right\|_\infty + 2\left\| \frac{1}{\tilde{w}_a}\check{p}(a,\cdot) - \frac{1}{\hat{w}_a}\check{p}(a,\cdot) \right\|_\infty + \left\| \frac{1}{\hat{w}_a}\check{p}(a,\cdot) - \frac{1}{\hat{w}_a}\hat{p}(a,\cdot) \right\|_\infty + \|\hat{p}_a - p_a\|_\infty \\
&\quad= \frac{1}{\hat{w}_a}\|\hat{w}_a\tilde{p}_a - \check{p}(a,\cdot)\|_\infty + \frac{2}{\hat{w}_a}|\hat{w}_a - \tilde{w}_a| + \frac{1}{\hat{w}_a}\|\check{p}(a,\cdot) - \hat{p}(a,\cdot)\|_\infty + \|\hat{p}_a - p_a\|_\infty \\
&\quad\leq \frac{1}{\hat{w}_a}\|\tilde{w}_a\tilde{p}_a - \hat{w}_a\tilde{p}_a\|_\infty + \frac{1}{\hat{w}_a}\|\tilde{w}_a\tilde{p}_a - \check{p}(a,\cdot)\|_\infty + \frac{2}{\hat{w}_a}|\hat{w}_a - \tilde{w}_a| + \frac{1}{\hat{w}_a}\|\check{p}(a,\cdot) - \hat{p}(a,\cdot)\|_\infty + \|\hat{p}_a - p_a\|_\infty \\
&\quad\leq \frac{1}{\hat{w}_a}\|\tilde{w}_a\tilde{p}_a - \check{p}(a,\cdot)\|_\infty + \frac{3}{\hat{w}_a}|\hat{w}_a - \tilde{w}_a| + \frac{1}{\hat{w}_a}\|\check{p}(a,\cdot) - \hat{p}(a,\cdot)\|_\infty + \|\hat{p}_a - p_a\|_\infty \\
&\quad\leq O\left(\frac{1}{\hat{w}_a}\left(|\hat{w}_a - \tilde{w}_a| + \frac{\sqrt{k}}{n\varepsilon}\ln\frac{k|\mathcal{A}|}{\beta} + \frac{1}{n\varepsilon}\ln\frac{k|\mathcal{A}|}{\beta} \right) + \sqrt{\frac{1}{nw_a}\ln\frac{k|\mathcal{A}|}{\beta}} \right) \\
&\quad\leq O\left(\frac{1}{\hat{w}_a}\left(\frac{\sqrt{k}}{n\varepsilon}\ln\frac{k|\mathcal{A}|}{\beta} + \frac{\sqrt{k}}{n\varepsilon}\ln\frac{k|\mathcal{A}|}{\beta} + \frac{1}{n\varepsilon}\ln\frac{k|\mathcal{A}|}{\beta} \right) + \sqrt{\frac{1}{nw_a}\ln\frac{k|\mathcal{A}|}{\beta}} \right) \\
&\quad\leq O\left(\frac{\sqrt{k}}{n\hat{w}_a\varepsilon}\ln\frac{k|\mathcal{A}|}{\beta} + \sqrt{\frac{1}{nw_a}\ln\frac{k|\mathcal{A}|}{\beta}} \right) \\
&\quad\leq O\left(\frac{\sqrt{k}}{nw_a\varepsilon}\ln\frac{k|\mathcal{A}|}{\beta} + \sqrt{\frac{1}{nw_a}\ln\frac{k|\mathcal{A}|}{\beta}} \right)
\end{aligned}$$

by Equation (12) and Lemmas B.6 and B.7; the bound on the $\|\check{p}(a,\cdot) - \hat{p}(a,\cdot)\|_\infty$ follows the same analysis in the proof of Lemma B.7 for the first term.

Similarly,

$$\begin{aligned}\|p_a - \tilde{p}_a\|_1 &\le \frac{1}{\hat{w}_a}\|\tilde{w}_a\tilde{p}_a - \check{p}(a,\cdot)\|_1 + \frac{3}{\hat{w}_a}|\hat{w}_a - \tilde{w}_a| + \frac{1}{\hat{w}_a}\|\check{p}(a,\cdot) - \hat{p}(a,\cdot)\|_1 + \|\hat{p}_a - p_a\|_1 \\ &\le O\left(\frac{k}{nw_a\varepsilon}\ln\frac{k|\mathcal{A}|}{\beta} + \sqrt{\frac{k}{nw_a}\ln\frac{|\mathcal{A}|}{\beta}}\right),\end{aligned}$$

and

$$\begin{aligned}&D_{\text{KS}}(p_a, \tilde{p}_a) \\ &\quad\le \frac{1}{\hat{w}_a}\max_j\left|\sum_{\ell=1}^{j}(\tilde{w}_a\tilde{p}_a(v_\ell) - \check{p}(a,v_\ell))\right| + \frac{3}{\hat{w}_a}|\hat{w}_a - \tilde{w}_a| + \frac{1}{\hat{w}_a}\max_j\left|\sum_{\ell=1}^{j}(\check{p}(a,\cdot) - \hat{p}(a,\cdot))\right| + D_{\text{KS}}(\hat{p}_a, p_a) \\ &\quad\le O\left(\frac{\sqrt{k}}{nw_a\varepsilon}\ln\frac{k|\mathcal{A}|}{\beta} + \sqrt{\frac{k}{nw_a}\ln\frac{k|\mathcal{A}|}{\beta}}\right).\end{aligned}$$

□

For the proof of Theorem 3.3, we need the following technical result for the difference of W_2^2 distances:

Lemma B.8 (Chizat et al., 2020)**.** *Let μ, μ', ν, ν' be distributions whose supports are contained in the centered ball of radius R in $\mathbb{R}^d$, then*

$$\begin{aligned}\left|W_2^2(\mu,\nu) - W_2^2(\mu',\nu')\right| &\le \left|\int \|x\|_2^2 \,\mathrm{d}(\mu-\mu')(x)\right| + \left|\int \|x\|_2^2 \,\mathrm{d}(\nu-\nu')(x)\right| \\ &\quad + 2R\sup_{\text{convex } f\in\text{Lip}(1)}\left|\int f(x)\,\mathrm{d}(\mu-\mu')(x)\right| + 2R\sup_{\text{convex } g\in\text{Lip}(1)}\left|\int g(x)\,\mathrm{d}(\nu-\nu')(x)\right| \\ &\le 4RW_1(\mu,\mu') + 4RW_1(\nu,\nu').\end{aligned}$$

The last line follows from the dual representation of W_1 distance for distributions with bounded support:

$$W_1(\mu,\nu) = \sup_{f\in\text{Lip}(1)}\left|\int f(x)\,\mathrm{d}(\mu-\nu)(x)\right|,$$

and the fact that $x \mapsto \|x\|_2^2$ is $2R$-Lipschitz on the centered ball of radius R.

Also, recall the fact that the W_1 distance of distributions supported on a ball of radius R can be upper bounded by total variation distance:

$$\begin{aligned}W_1(\mu,\nu) &= \inf_{\pi\in\Pi(\mu,\nu)}\int d(x,y)\,\mathrm{d}\pi(x,y) \\ &\le 2R\inf_{\pi\in\Pi(\mu,\nu)}\int \mathbb{1}[x\neq y]\,\mathrm{d}\pi(x,y) \\ &= 2R\left(1 - \sup_{\pi\in\Pi(\mu,\nu)}\int \mathbb{1}[x=y]\,\mathrm{d}\pi(x,y)\right) \\ &= 2R\left(1 - \int \min(\mu(x),\nu(x))\,\mathrm{d}x\right) \\ &= 2R\int \max(0,\nu(x)-\mu(x))\,\mathrm{d}x \\ &= R\int |\nu(x)-\mu(x)|\,\mathrm{d}x \\ &=: R\|\mu-\nu\|_1\end{aligned}$$

$$=: 2RD_{\mathrm{TV}}(\mu, \nu), \tag{13}$$

where line 6 is because $\int(\mu(x) - \nu(x))\,\mathrm{d}x = 0$.

And, note the following simple fact regarding optimal transports $T^* : \mathbb{R} \to \mathbb{R}$ under the squared cost; in the special case where T^* is a Monge transportation plan, the lemma is equivalent to saying that T^* is a nondecreasing function (see the last panel of Figure 1 for a picture):

Lemma B.9. *Let p, q be two distributions supported on $[k]$, and $\pi \in \arg\min_{\pi' \in \Pi(p,q)} \sum_{m,\ell}(m-\ell)^2\pi(m,\ell)$, then for all $j \in [k]$,*

$$f(m) := \sum_{\ell=1}^{j} \frac{\pi(m,\ell)}{p(m)} \begin{cases} = 1 & \text{if } m < l_j \\ \in [0,1] & \text{if } m = l_j \\ = 0 & \text{if } m > l_j, \end{cases} \quad \forall m \text{ s.t. } p(m) > 0, \tag{14}$$

for some $l_1, \ldots, l_k \in [k]$.

Proof. Let $j \in [k]$ be arbitrary. Suppose to the contrary that $\nexists l_j$ such that Equation (14) holds, then either f as a function of m is not non-increasing, i.e., $f(m+1) > f(m)$ for some m, or $0 < f(m+1) \le f(m) < 1$. We show that either of these contradicts the optimality of π.

In both cases, there must exists $l \le j < r$ such that $\pi(m,r), \pi(m+1,l) \ge q$ for some $q > 0$, because $\sum_{\ell=1}^{j}\pi(m+1,\ell) = p(m+1)f(m+1) > 0$ and $\sum_{\ell=j+1}^{k}\pi(m,\ell) = p(m)(1-f(m)) > 0$. Then a coupling γ with a lower cost can be constructed by (partially) exchanging the two entries:

$$\gamma(i,\ell) = \begin{cases} \pi(m,r) - q & \text{if } i = m, \ell = r \\ \pi(m,l) + q & \text{if } i = m, \ell = l \\ \pi(m+1,r) + q & \text{if } i = m+1, \ell = r \\ \pi(m+1,l) - q & \text{if } i = m+1, \ell = l \\ \pi(i,\ell) & \text{else.} \end{cases}$$

We verify that it has a lower cost than π:

$$\sum_{i,\ell}(i-\ell)^2(\gamma-\pi)(i,\ell) = -q(m-r)^2 + q(m-l)^2 + q(m+1-r)^2 - q(m+1-l)^2 = 2q(l-r) < 0. \qquad \square$$

Proof of Theorem 3.3. This proof also relies on properly applying the triangle inequality to decompose into comparable terms. We list the terms that will be compared here:

- Denote the Bayes regressor by f^*, and recall that f is the regressor being post-processed. Denote the output distribution of f^* conditioned on group a by $r_a^* := f^*(\cdot, a)\sharp\mu_{X|a}$, which will be compared to that of f, $r_a := f(\cdot, a)\sharp\mu_{X|a}$.
- Given a discretizer h, the discretized conditional output distribution of f^* is denoted by $p_a^* := h\sharp r_a^*$, and that of f by $p_a := h\sharp r_a$. We will compare r_a^* to its discretized version p_a^*, and p_a to $\tilde{p}_a$, the empirical conditional discretized output distributions of f estimated privately.
- Denote the private group marginal distribution estimated from the samples by $\tilde{w}_a$, which will be compared to the ground-truth $w_a := \mathbb{P}(A = a)$.
- Let $(\cdot, \{\tilde{q}'_a\}_{a\in\mathcal{A}}) \leftarrow \mathrm{P}(\{\tilde{p}_a\}_{a\in\mathcal{A}}, \{\tilde{w}_a\}_{a\in\mathcal{A}}, \alpha)$ and $(\cdot, \cdot, \{\tilde{q}_a\}_{a\in\mathcal{A}}) \leftarrow \mathrm{LP}(\{\tilde{p}_a\}_{a\in\mathcal{A}}, \{\tilde{w}_a\}_{a\in\mathcal{A}}, \alpha)$. The difference between $\tilde{q}'_a, \tilde{q}_a$ is that the support of the latter is restricted to v.

  Recall that the fair regressor returned from Algorithm 1 has the form $\bar{f} = g_a \circ h \circ f$, where g_a is the optimal transport from $\tilde{p}_a$ to $\tilde{q}_a$. The $\tilde{q}_a$'s will be compared to the $\tilde{q}'_a$'s, which are in turn compared to the output distributions q_a^* of an optimal α-fair regressor, denoted by $\bar{f}^*$ (note that $(\cdot, \{q_a^*\}_{a\in\mathcal{A}}) \leftarrow \mathrm{P}(\{r_a^*\}_{a\in\mathcal{A}}, \{w_a\}_{a\in\mathcal{A}}, \alpha)$).

Error Bound. Note that $R(\bar{f}) - R(\bar{f}^*) = ER(\bar{f}) - ER(\bar{f}^*)$. We begin with analyzing the first term. By the orthogonality principle,

$$ER(\bar{f}) = \mathbb{E}\Big[\big(\bar{f}(X,A) - f^*(X,A)\big)^2\Big]$$

$$
\begin{aligned}
&= \sum_{a\in\mathcal{A}} w_a \, \mathbb{E}_{X\sim\mu_{X|a}}\Big[\big(\bar{f}(X,a) - f^*(X,a)\big)^2\Big] \\
&= \sum_{a\in\mathcal{A}} w_a \, \mathbb{E}_{X\sim\mu_{X|a}}\Big[\big(\bar{f}(X,a) - f(X,a) + (f(X,a) - f^*(X,a))\big)^2\Big] \\
&= \sum_{a\in\mathcal{A}} w_a \Big(\mathbb{E}_{X\sim\mu_{X|a}}\Big[\big(\bar{f}(X,a) - f(X,a)\big)^2 + (f(X,a) - f^*(X,a))^2 \\
&\qquad\qquad + 2\big(\bar{f}(X,a) - f(X,a)\big)(f(X,a) - f^*(X,a))\Big]\Big) \\
&\le \sum_{a\in\mathcal{A}} w_a \, \mathbb{E}_{X\sim\mu_{X|a}}\Big[\big(\bar{f}(X,a) - f(X,a)\big)^2\Big] + \underbrace{3\,\mathbb{E}[|f(X,A) - f^*(X,A)|]}_{\mathcal{E}_1},
\end{aligned}
$$

where line 5 is because of the assumption that the images of $\bar{f}, f$ are contained in $[0,1]$. The second term on the last line is the L^1 excess risk of f; for the first term,

$$
\begin{aligned}
&\mathbb{E}_{X\sim\mu_{X|a}}\Big[\big(\bar{f}(X,a) - f(X,a)\big)^2\Big] \\
&\quad= \mathbb{E}_{X\sim\mu_{X|a}}\Big[(g_a \circ h \circ f(X,a) - f(X,a))^2\Big] \\
&\quad= \mathbb{E}_{X\sim\mu_{X|a}}\Big[(g_a \circ h \circ f(X,a) - h \circ f(X,a) + (h \circ f(X,a) - f(X,a)))^2\Big] \\
&\quad\le \mathbb{E}_{X\sim\mu_{X|a}}\Big[(g_a \circ h \circ f(X,a) - h \circ f(X,a))^2 + 3|h \circ f(X,a) - f(X,a)|\Big] \\
&\quad\le \mathbb{E}_{X\sim\mu_{X|a}}\Big[(g_a \circ h \circ f(X,a) - h \circ f(X,a))^2\Big] + \frac{3}{2k} \\
&\quad= \sum_{j=1}^{k} p_a(v_j)\, \mathbb{E}\Big[(g_a(v_j) - v_j)^2\Big] + \frac{3}{2k} \\
&\quad= \sum_{j=1}^{k} \tilde{p}_a(v_j)\, \mathbb{E}\Big[(g_a(v_j) - v_j)^2\Big] + \sum_{j=1}^{k} (p_a(v_j) - \tilde{p}_a(v_j))\, \mathbb{E}\Big[(g_a(v_j) - v_j)^2\Big] + \frac{3}{2k} \\
&\quad\le W_2^2(\tilde{p}_a, \tilde{q}_a) + \|p_a - \tilde{p}_a\|_1 + \frac{3}{2k} \\
&\quad\le W_2^2(\tilde{p}_a, \tilde{q}_a) + O\left(\sqrt{\frac{k}{nw_a} \ln \frac{k|\mathcal{A}|}{\beta}} + \frac{k}{nw_a\varepsilon} \ln \frac{k|\mathcal{A}|}{\beta}\right) + \frac{3}{2k},
\end{aligned}
$$

where line 4 is because h discretizes the input to the midpoint of the bin that it falls in, which displaces it by up to $L/2k = 1/2k$; line 5 is because $p_a(v_j) = \mathbb{P}(h \circ f(X,a) = v_j)$; the first term on line 7 is because g_a is the optimal transport from $\tilde{p}_a$ to $\tilde{q}_a$. Then, combining the above, by Theorem 3.2, with probability at least $1-\beta$,

$$
\begin{aligned}
ER(\bar{f}) &\le \sum_{a\in\mathcal{A}} \left(w_a W_2^2(\tilde{p}_a, \tilde{q}_a) + O\left(\sqrt{\frac{kw_a}{n} \ln \frac{k|\mathcal{A}|}{\beta}} + \frac{k}{n\varepsilon} \ln \frac{k|\mathcal{A}|}{\beta}\right)\right) + \mathcal{E}_1 + \frac{3}{2k} \\
&\le \sum_{a\in\mathcal{A}} w_a W_2^2(\tilde{p}_a, \tilde{q}_a) + \mathcal{E}_1 + \underbrace{O\left(\sqrt{\frac{k|\mathcal{A}|}{n} \ln \frac{k|\mathcal{A}|}{\beta}} + \frac{k|\mathcal{A}|}{n\varepsilon} \ln \frac{k|\mathcal{A}|}{\beta}\right)}_{\mathcal{E}_2} + \frac{3}{2k} \\
&\le \sum_{a\in\mathcal{A}} \big(\tilde{w}_a W_2^2(\tilde{p}_a, \tilde{q}_a) + |\tilde{w}_a - w_a| W_2^2(\tilde{p}_a, \tilde{q}_a)\big) + \mathcal{E}_1 + \mathcal{E}_2 + \frac{3}{2k} \\
&\le \sum_{a\in\mathcal{A}} \big(\tilde{w}_a W_2^2(\tilde{p}_a, \tilde{q}_a) + |\tilde{w}_a - w_a|\big) + \mathcal{E}_1 + \mathcal{E}_2 + \frac{3}{2k} \\
&\le \sum_{a\in\mathcal{A}} \tilde{w}_a W_2^2(\tilde{p}_a, \tilde{q}_a) + \mathcal{E}_1 + \mathcal{E}_2 + \frac{3}{2k}
\end{aligned}
$$

$$
\begin{aligned}
&\leq \sum_{a\in\mathcal{A}} \tilde{w}_a W_2^2(\tilde{p}_a, h\sharp\tilde{q}'_a) + \mathcal{E}_1 + \mathcal{E}_2 + \frac{3}{2k} \\
&\leq \sum_{a\in\mathcal{A}} \tilde{w}_a (W_2(\tilde{p}_a, \tilde{q}'_a) + W_2(\tilde{q}'_a, h\sharp\tilde{q}'_a))^2 + \mathcal{E}_1 + \mathcal{E}_2 + \frac{3}{2k} \\
&= \sum_{a\in\mathcal{A}} \tilde{w}_a \big(W_2^2(\tilde{p}_a, \tilde{q}'_a) + 2W_2(\tilde{p}_a, \tilde{q}'_a)W_2(\tilde{q}'_a, h\sharp\tilde{q}'_a) + W_2^2(\tilde{q}'_a, h\sharp\tilde{q}'_a)\big) + \mathcal{E}_1 + \mathcal{E}_2 + \frac{3}{2k} \\
&\leq \sum_{a\in\mathcal{A}} \tilde{w}_a W_2^2(\tilde{p}_a, \tilde{q}'_a) + \mathcal{E}_1 + \mathcal{E}_2 + \frac{3}{k} \\
&\leq \sum_{a\in\mathcal{A}} \tilde{w}_a W_2^2(\tilde{p}_a, q^*_a) + \mathcal{E}_1 + \mathcal{E}_2 + \frac{3}{k},
\end{aligned}
$$

where line 6 follows by noting that $\{h\sharp\tilde{q}'_a\}_{a\in\mathcal{A}}$ is a feasible solution to LP, as it can be verified that $D_{\text{KS}}(h\sharp\tilde{q}'_a, h\sharp\tilde{q}'_{a'}) \leq \alpha$ given that $D_{\text{KS}}(\tilde{q}'_a, \tilde{q}'_{a'}) \leq \alpha, \forall a, a' \in \mathcal{A}$ (hence restricting the support of the barycenter to v introduces an additional error of $3/2k$ as discussed in footnote 3), and the last line is because $\{\tilde{q}'_a\}_{a\in\mathcal{A}}$ is a minimizer of $\mathrm{P}(\{\tilde{p}_a\}_{a\in\mathcal{A}}, \{\tilde{w}_a\}_{a\in\mathcal{A}}, \alpha)$.

So for the suboptimality of $\bar{h}$, by Theorem 3.1,

$$
\begin{aligned}
ER(\bar{f}) - ER(\bar{f}^*) &\leq \sum_{a\in\mathcal{A}} \tilde{w}_a \big(W_2^2(\tilde{p}_a, q^*_a) - W_2^2(r^*_a, q^*_a)\big) + \mathcal{E}_1 + \mathcal{E}_2 + \frac{3}{k} \\
&\leq \sum_{a\in\mathcal{A}} \tilde{w}_a \big(W_2^2(\tilde{p}_a, q^*_a) - (W_2(q^*_a, p^*_a) - W_2(r^*_a, p^*_a))\big)^2 + \mathcal{E}_1 + \mathcal{E}_2 + \frac{3}{k} \\
&\leq \sum_{a\in\mathcal{A}} \tilde{w}_a \big(W_2^2(\tilde{p}_a, q^*_a) - W_2^2(p^*_a, q^*_a) + 2W_2(p^*_a, q^*_a)W_2(p^*_a, r^*_a)\big) + \mathcal{E}_1 + \mathcal{E}_2 + \frac{3}{k} \\
&\leq \sum_{a\in\mathcal{A}} \tilde{w}_a \big(W_2^2(\tilde{p}_a, q^*_a) - W_2^2(p^*_a, q^*_a)\big) + \frac{1}{k} + \mathcal{E}_1 + \mathcal{E}_2 + \frac{3}{k},
\end{aligned}
$$

where the last line is because h is a transport from r^*_a to p^*_a with displacements of at most $1/2k$; for the first term, by Lemma B.8 and Equation (13),

$$
\begin{aligned}
\sum_{a\in\mathcal{A}} \tilde{w}_a \big(W_2^2(\tilde{p}_a, q^*_a) - W_2^2(p^*_a, q^*_a)\big) &\leq 4 \sum_{a\in\mathcal{A}} \tilde{w}_a W_1(\tilde{p}_a, p^*_a) \\
&\leq 4 \sum_{a\in\mathcal{A}} \tilde{w}_a (W_1(\tilde{p}_a, p_a) + W_1(p_a, r_a) + W_1(r_a, r^*_a) + W_1(r^*_a, p^*_a)) \\
&\leq 4 \sum_{a\in\mathcal{A}} \tilde{w}_a \left(\|\tilde{p}_a - p_a\|_1 + \frac{1}{2k} + \mathbb{E}_{X\sim\mu_{X|a}}[|f^*(X,a) - f(X,a)|] + \frac{1}{2k} \right) \\
&\leq 4 \sum_{a\in\mathcal{A}} \tilde{w}_a \, \mathbb{E}_{X\sim\mu_{X|a}}[|f^*(X,a) - f(X,a)|] + \mathcal{E}_2 + \frac{4}{k} \\
&\leq 4 \sum_{a\in\mathcal{A}} (w_a - w_a + \tilde{w}_a) \, \mathbb{E}_{X\sim\mu_{X|a}}[|f^*(X,a) - f(X,a)|] + \mathcal{E}_2 + \frac{4}{k} \\
&\leq 4 \sum_{a\in\mathcal{A}} w_a \, \mathbb{E}_{X\sim\mu_{X|a}}[|f^*(X,a) - f(X,a)|] + 4 \sum_{a\in\mathcal{A}} |\tilde{w}_a - w_a| + 4\mathcal{E}_2 + \frac{4}{k} \\
&\leq 4\mathcal{E}_1 + \mathcal{E}_2 + \frac{4}{k},
\end{aligned}
$$

the third inequality is because the joint distribution of $(f(X,a), f^*(X,a))$ is a valid coupling belonging to $\Pi(r_a, r^*_a)$ that incurs a transportation cost of

$$
\mathbb{E}[|f^*(X,a) - f(X,a)|] = \int |y - y^*| \, \mathrm{d}\,\mathbb{P}(f(X,a) = y, f^*(X,a) = y^*) \geq W_1(r_a, r^*_a).
$$

Putting everything together, and with a union bound over the two events above, gives the result in the theorem statement.

Fairness Guarantee. Let $\bar{p}_a$ denote the output distribution of the post-processed regressor $\bar{f}$ conditioned on group a. Using triangle inequality, for any $a, a' \in \mathcal{A}$,

$$D_{\text{KS}}(\bar{p}_a, \bar{p}_{a'}) \leq D_{\text{KS}}(\tilde{q}_a, \tilde{q}_{a'}) + D_{\text{KS}}(\bar{p}_a, \tilde{q}_a) + D_{\text{KS}}(\bar{p}_{a'}, \tilde{q}_{a'}) \leq \alpha + D_{\text{KS}}(\bar{p}_a, \tilde{q}_a) + D_{\text{KS}}(\bar{p}_{a'}, \tilde{q}_{a'}),$$

where

$$\begin{aligned}
D_{\text{KS}}(\bar{p}_a, \tilde{q}_a) &= \max_j \left| \sum_{\ell=1}^{j} (\bar{p}_a(v_\ell) - \tilde{q}_a(v_\ell)) \right| \\
&= \max_j \left| \sum_{\ell=1}^{j} \big(\mathbb{P}(\bar{f}(X, a) = v_\ell \mid A = a) - \tilde{q}_a(\ell) \big) \right| \\
&= \max_j \left| \sum_{\ell=1}^{j} \left(\sum_{m=1}^{k} p_a(v_m)\, \mathbb{P}(g_a(v_m) = v_\ell \mid A = a) - \sum_{m=1}^{k} \tilde{p}_a(v_m)\, \mathbb{P}(g_a(v_m) = v_\ell \mid A = a) \right) \right| \\
&= \max_j \left| \sum_{m=1}^{k} (p_a(v_m) - \tilde{p}_a(v_m)) \sum_{\ell=1}^{j} \mathbb{P}(g_a(v_m) = v_\ell \mid A = a) \right| \\
&\leq \max_j \left(\left| \sum_{m=1}^{v_{\iota_j - 1}} (p_a(v_m) - \tilde{p}_a(v_m)) \right| + \left| \big(p_a(v_{\iota_j}) - \tilde{p}_a(v_{\iota_j}) \big)\, \mathbb{P}(g_a(v_{\iota_j}) \leq v_j) \right| \right) \\
&\leq D_{\text{KS}}(p_a, \tilde{p}_a) + \max_j \left| p_a(v_{\iota_j}) - \tilde{p}_a(v_{\iota_j}) \right| \\
&\leq O\left(\frac{\sqrt{k}}{n w_a \varepsilon} \ln \frac{k|\mathcal{A}|}{\beta} + \sqrt{\frac{k}{n w_a} \ln \frac{k|\mathcal{A}|}{\beta}} \right);
\end{aligned}$$

line 3 is because g_a is a transport from $\tilde{p}_a$ to $\tilde{q}_a$, line 5 uses Lemma B.9 and the fact that $\sum_{\ell=1}^{j} {\pi_a(v_m, v_\ell)}/{\tilde{p}_a(v_m)} = \mathbb{P}(g_a(v_m) \leq v_j)$, and line 7 is by Theorem 3.2. □