# Statistical Learning from Attribution Sets

**Lorne Applebaum** LAPPLEBAUM@GOOGLE.COM
*Google, New York, USA*

**Robert Busa-Fekete** BUSAROBI@GOOGLE.COM
*Google Research, New York, USA*

**August Chen** AYC74@CORNELL.EDU
*Cornell, Ithaca, USA*

**Claudio Gentile** CGENTILE@GOOGLE.COM
*Google Research, New York, USA*

**Tomer Koren** TKOREN@GOOGLE.COM
*Google and Tel Aviv University, Tel Aviv, Israel*

**Aryan Mokhtari** AMOKHTARI@GOOGLE.COM
*Google Research and UT Austin, New York, USA*

**Editors:** Steve Hanneke and Tor Lattimore

## Abstract

We address the problem of training conversion prediction models in advertising domains under privacy constraints, where direct links between ad clicks and conversions are unavailable. Motivated by privacy-preserving browser APIs and the deprecation of third-party cookies, we study a setting where the learner observes a sequence of clicks and a sequence of conversions, but can only link a conversion to a set of candidate clicks (an attribution set) rather than a unique source. We formalize this as learning from attribution sets generated by an oblivious adversary equipped with a prior distribution over the candidates. Despite the lack of explicit labels, we construct an unbiased estimator of the population loss from these coarse signals via a novel approach. Leveraging this estimator, we show that Empirical Risk Minimization achieves generalization guarantees that scale with the informativeness of the prior and is also robust against estimation errors in the prior, despite complex dependencies among attribution sets. Simple empirical evaluations on standard datasets suggest our unbiased approach significantly outperforms common industry heuristics, particularly in regimes where attribution sets are large or overlapping.

**Keywords:** Weak supervision, privacy, anti-tracking, attribution problem, ERM, loss debiasing.

## 1. Introduction

Web advertising—one of the largest real-world applications of machine learning—has undergone a significant shift in recent years. To power automated bidding, advertisers (or their AdTech partners) train models to predict the probability of a *conversion* (e.g., a product purchase, a sign-up, an app installation, etc.) following an ad click. These predictions are essential for calculating bid prices in real-time online auctions running at the publisher side. Because the initial click occurs on a publisher's site while the conversion happens on the advertiser's site, generating training labels requires linking these two distinct events. This process essentially involves tracking user behavior across different web domains (see, e.g., Wilander, 2019).

While third-party cookies and link decoration have historically made tracking straightforward, a shift toward user privacy has transformed the landscape. Acknowledging the conflict between essential web advertising and the demand for privacy, major browsers have introduced specialized APIs to measure performance without compromising user anonymity. This transition is highlighted by the deprecation of third-party cookies in browsers like Apple's Safari (Wilander, 2019) and Mozilla's Firefox (Crouch and Crawford, 2022). These APIs restrict AdTechs to collecting cross-site data exclusively in some obfuscated form. This creates a challenge for publishers who require precise per-interaction predictions to run effective auctions. Under these privacy constraints, the publisher can see the list of individual ad interactions (clicks) but only receives approximate information about the resulting conversions from the advertiser's side, rather than direct links between specific clicks and sales. In particular, the publisher learns some coarse information about the conversion and click, such as the ad campaign they belonged to and the approximate time of the conversion. Based on this coarse information, we can usually identify a collection of clicks that could have produced the conversion (i.e., clicks from the same ad campaign in a reasonable time interval given the conversion time), but it is not possible to determine exactly which click was responsible. Our goal is to learn conversion prediction models from these weak conversion signals.

### 1.1. Our contributions

We formalize our problem as a novel setting of statistical learning from attribution sets: collections of clicks generated by an oblivious adversary according to a known prior (Section 2). Because the direct association between clicks and conversions is unobserved, we seek to learn this relationship using only these coarse signals.

We provide three main theoretical contributions. Surprisingly, we first show that it is possible to construct an unbiased estimator of the population loss by decomposing the expected loss into moments that can be estimated from the attribution sets (Theorem 1). The core innovation lies in decoupling features from labels by conditioning on the adversary's actions, allowing us to leverage a combinatorial argument to map inaccessible population moments to observable indicators. Second, by minimizing our unbiased surrogate, we establish that Empirical Risk Minimization (ERM) attains strong generalization guarantees despite the statistical dependencies induced by the attribution process. Specifically, Theorem 2 demonstrates that the sample complexity of our method scales with the standard capacity of the hypothesis class, inflated by a factor of $1/\|\pi\|_2^2$, where $\pi$ is the adversary's prior distribution governing the possible locations of the true conversion within an attribution set. This is indeed expected as $\|\pi\|_2^2$ serves as a fundamental measure of the statistical difficulty of the task: more concentrated priors heighten the signal-to-noise ratio, yielding more favorable convergence rates. A limitation of Theorem 2 is that the learner requires exact knowledge of $\pi$. However, as our third theoretical contribution, we show in Theorem 5 that even if we have an estimate $\widehat{\pi}$ of $\pi$, our method is robust to the estimation error.

Finally, to verify our theoretical guarantees, we conduct preliminary experiments on standard datasets, like MNIST, CIFAR-10, and Higgs, showing that our unbiased approach substantially outperforms common industry heuristics–such as random or maximum-prior attribution–particularly when attribution sets are large or overlapping (Section 5).

### 1.2. Related literature

Conversion Rate (CVR) prediction remains a foundational challenge in online advertising, generating a vast body of literature. Central to this field is the *attribution problem*—the assignment of credit

to specific user interactions for subsequent conversions. Established attribution heuristics (such as "last touch," "first touch," or "linear attribution") dictate the mechanisms for label generation and training, which in turn drive automated bidding and traffic allocation strategies. Relevant works include (Borgs et al., 2007; Cai et al., 2017; Zhu et al., 2017; Jin et al., 2018; Wang et al., 2017; Yang et al., 2019; Singal et al., 2019; Liu et al., 2021; Chen et al., 2022; Fan et al., 2025; Chen et al., 2025). While the above list is very far from doing justice, it is fair to say that many of these investigations are mostly experimental in nature.

More theoretically oriented is the related bulk of research on (stochastic) online/bandit algorithms with delayed feedback, with early investigations including (Joulani et al., 2013; Vernade et al., 2017, 2020). These works predominantly address streaming data problems (online prediction), where models are continuously fine-tuned as feedback arrives. Crucially, these frameworks generally assume no privacy-induced label obfuscation; they postulate that a click will eventually yield an observable signal unless it is censored by "freshness" constraints. Typically, this involves setting an observation window $w_0$ (e.g., 48 hours) where a click at time $t_0$ is frozen until $t_0 + w_0$. If a conversion occurs within this window, the click is labeled positive; otherwise, it is treated as a negative sample. Consequently, the primary technical challenge in these streaming settings is optimizing the trade-off between the cost of adaptivity (where larger $w_0$ delays updates) and the bias introduced by censoring (where smaller $w_0$ mislabels valid but delayed conversions). Bias correction mechanisms are often based on importance sampling (see, e.g., Chen et al. (2022), and references therein).

In contrast, our work addresses an emerging landscape defined by privacy preservation. We operate in a setting where, even if deterministically linking a click to a conversion is technically feasible, the association is deliberately obfuscated by anti-tracking APIs mediating between publisher and advertiser data. Furthermore, we depart from the streaming paradigm to focus on a (more practical) *batch* learning setting. Since the complete, albeit obfuscated, dataset is available at the outset, concerns regarding data freshness and update latency are not directly relevant to our approach.

Our work is also related to weak supervision paradigms, specifically Multiple Instance Learning (MIL) (e.g., Maron and Lozano-Pérez (1997); Dietterich et al. (1997); Ilse et al. (2018); Tian et al. (2021); Lv et al. (2023); Javed et al. (2022); Jang and Kwon (2024)) and Learning from Label Proportions (LLP). Early references on LLP include Quadrianto et al. (2008); Patrini et al. (2014), more recent ones are Saket (2021, 2022); Scott and Zhang (2020); Zhang et al. (2022); Busa-Fekete et al. (2023); Brahmbhatt et al. (2023); Li et al. (2024); Busa-Fekete et al. (2025); Applebaum et al. (2026). In these paradigms, attribution sets are referred to as *bags*. Both frameworks focus on learnability at the bag and instance levels. In MIL, a bag is labeled positive if it contains at least one positive instance and negative otherwise. In LLP, the learner observes the proportion of positive labels within each bag. Crucially, the observability structure in these settings is significantly more informative than ours. In both MIL and LLP, every bag conveys a signal to the learner; MIL explicitly includes negative bags (containing zero positive labels), whereas our setting typically only generates attribution sets for positive outcomes (conversions). Furthermore, general statistical analyses of LLP (e.g., Busa-Fekete et al. (2023); Li et al. (2024); Busa-Fekete et al. (2025); Applebaum et al. (2026)) predominantly assume non-overlapping bags—an assumption that need not hold in the context of API-mediated attribution, where user interaction windows are often wide and overlapping.

## 2. Preliminaries and Notation

We move from an idealized physical process to a distilled model that removes temporal dependencies.

### 2.1. The click-conversion process

Consider a stylized advertising setup, illustrated in Figure 1 (Left). This involves two parties: a *publisher*, who observes a stream of click events (interactions with a website by users), and an *advertiser*, who observes a stream of conversion events (e.g., purchases). In order to decide which slot will be assigned to the competing advertisers, the publisher typically runs an auction, which is powered by a *conversion prediction* model. This is a model that takes as input click features and returns the estimated probability that the click will (eventually) lead to a conversion.

Due to privacy constraints (such as those enforced by anti-tracking APIs) and random time delays between clicks and conversions, the publisher cannot deterministically link a specific conversion at time $T_Y$ to its originating click at time $T_X$. Instead, for every observed conversion, the system provides an *attribution set*: a window of candidate clicks that *could* have caused the conversion.

We are facing here a classical *attribution problem* for CVR (aka conversion rate) prediction. Yet, unlike the voluminous literature on the subject (e.g., Borgs et al. (2007); Cai et al. (2017); Zhu et al. (2017); Jin et al. (2018); Wang et al. (2017); Yang et al. (2019); Liu et al. (2021); Chen et al. (2022); Fan et al. (2025); Chen et al. (2025), and references therein), we are dealing with the more practical scenario of *batch* attribution, whereby an offline dataset of clicks and conversions has been recorded and made available to the publisher. The prediction model is built via this set of observations.

To analyze this setting rigorously, we start by viewing this ecosystem as a *click-conversion process* described by two pairs of random variables $\big\langle (X, D_X), (Y, D_Y) \big\rangle$, where: $X \in \mathcal{X}$ describes the click event features; $Y \in \{0, 1\}$ is the corresponding binary label (conversion yes/no); $D_X$ and $D_Y$ are the *event delay* variables (time until the next event). In particular, $D_X$ is the time until the next click and $D_Y$ is the time until the next label.

Data is generated by fixing a total number of events $n$ and drawing $n$ i.i.d. pairs $(X_i, D_{X,i})$ and $(Y_i, D_{Y,i})$ from the joint distributions. The observed timestamps are cumulative sums of these delays: $T_{X,i} = \sum_{j=1}^{i} D_{X,j}$ and $T_{Y,i} = \sum_{j=1}^{i} D_{Y,j}$. The main assumption we make is that the timing variables $(D_X, D_Y)$ are independent of the event variables $(X, Y)$. This independence allows us to separate the temporal dynamics from the feature-label relationship. Our ultimate goal is to train a conversion prediction model using only these coarse, aggregate signals (the attribution sets), without ever observing direct links between individual clicks $(X)$ and labels $(Y)$.

### 2.2. Mathematical formalization

We can now distill the process above into a learning framework on sequences, as depicted in Figure 1 (Right). There exists a hidden (possibly randomized) bijection $b : [n] \to [n]$ that links the click $(X_i, T_{X,i})$ to its corresponding binary label and event delay $(Y_{b(i)}, T_{Y,b(i)})$. This bijection defines a hidden dataset $S = \big\langle (X_1, Y_{b(1)}), \ldots, (X_n, Y_{b(n)}) \big\rangle$. Based on our independence assumption, the pairs in $S$ are i.i.d. draws from a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0, 1\}$, unknown to the learner. For the remainder of the paper, we simplify notation by re-indexing such that $b$ is the identity, denoting the $i$-th pair simply as $(X_i, Y_i)$.

**The observation model.** The learner does not see the labels $Y$. Instead, the learner observes:

- A sequence of feature vectors $X_1, \ldots, X_n$ drawn i.i.d. from the marginal distribution over $\mathcal{X}$;
- A collection of *Attribution Sets* $\mathcal{A} = \{A_1, \ldots, A_M\}$, each attribution set representing the candidate clicks that could have caused the conversion. While the sequence length $n$ is fixed,

the number of observed attribution sets $M = \sum_{i=1}^{n} Y_i$ is a random variable equal to the number of positive labels/conversions.

**The adversary (attribution mechanism).** We model the generation of these sets (or windows of candidate clicks) via an oblivious adversary. Let $i_j(S) \in [n]$ be the index of the $j$-th positive label ($Y_{i_j} = 1$) in $S$. For each conversion index $i_j(S)$, the adversary generates an attribution set $A_j \subseteq [n]$ consisting of $k$ *consecutive*[1] indices that includes $i_j(S)$. Crucially, the position of the true conversion within the set is governed by a *prior distribution* $\pi$ over $[k]$. Specifically, the adversary constructs the window $A_j$ such that the true index $i_j(S)$ appears at the $r$-th position of $A_j$ with probability $\pi[r]$:[2]

$$\mathbb{P}(A_j[r] = X_{i_j(S)}) = \pi[r] \quad \text{for } r \in \{1, \ldots, k\}.$$

This prior $\pi$ captures domain knowledge, such as the "last-touch" heuristic (where $\pi[k]$ is large, encoding the belief that more frequently the last element is the cause of the conversion) or time-decay models. The generation of attribution sets is independent of the feature values $X$ (obliviousness), and the sets may overlap. To streamline notation, we assume a constant set size $k$ and a fixed prior $\pi$ known to the learner. As we discuss later, our results extend to estimated priors. And they also extend to variable set sizes and variable priors, as briefly discussed in Remark 18 (Appendix B.1).

**Learning goal.** Recall we want to learn a model to predict $Y$ from $X$ through such data. When learning a model $p : \mathcal{X} \to [0, 1]$, we operate within a defined hypothesis space $\mathcal{H}$. Each hypothesis $h \in \mathcal{H}$ represents a deterministic mapping from the input space $\mathcal{X}$ to $[0, 1]$, where the output $h(x)$ estimates the probability that $Y = 1$ given $X = x$. We measure the discrepancy between predictions and labels via a loss function $\ell : [0, 1] \times \{0, 1\} \to \mathbb{R}^+$. We assume for simplicity that the loss is *bounded* (e.g., the square loss $\ell(h(x), y) = (y - h(x))^2$).

Given distribution $\mathcal{D}$, hypothesis class $\mathcal{H}$, and loss function $\ell$, the *population loss* (or *statistical risk*) of a hypothesis $h \in \mathcal{H}$ is defined as: $\mathcal{L}(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(h(x), y)]$. We aim to minimize the *excess risk*, also referred to as the *regret*, $Reg(h)$, which quantifies the performance gap between $h$ and the *best-in-class* hypothesis $h^*_{\mathcal{H}}$: $Reg(h) = \mathcal{L}(h) - \mathcal{L}(h^*_{\mathcal{H}})$, where $h^*_{\mathcal{H}} = \arg\min_{h\in\mathcal{H}} \mathcal{L}(h)$. We work in the general non-realizable (agnostic) setting where the true optimal mapping may not lie within $\mathcal{H}$ (i.e., $h^* \notin \mathcal{H}$).

We denote by $\mu$ the joint distribution over the two sources of randomness in our setting: (i) The generation of the dataset $S$ (drawn from $\mathcal{D}^n$); (ii) The adversary's generation of attribution sets $A_1, \ldots, A_M$ (drawn i.i.d. via $\pi$, conditioned on $S$). Our goal is to find an estimator $\widehat{h} \in \mathcal{H}$ such that the population loss $\mathcal{L}(\widehat{h})$ is minimized with high probability over $\mu$. Specifically, we aim to design algorithms for $\widehat{h}$ and quantify $Reg(\widehat{h})$ in terms of prior $\pi$ (which encodes the degree of label obfuscation), the sample size $n$, and the general properties of the loss function $\ell$ and hypothesis space $\mathcal{H}$.

**Further notation.** Let $A_j[i] \in \mathcal{X}$ be the $i$-th feature vector in $A_j$. From here on out, when expectations are not explicitly specified, they are w.r.t. $(X, Y) \sim \mathcal{D}$. We let $\mathcal{D}_X, \mathcal{D}_Y$ denote the $X$ and $Y$-marginals of $\mathcal{D}$ respectively, and $\mathcal{D}_{X|Y=1}$, $\mathcal{D}_{X|Y=0}$ denote conditional laws of $X$ given $Y = 1$ and $Y = 0$ respectively. We let $\mathbb{E}_1[\cdot]$ denote the conditional expectation $\mathbb{E}[\cdot \mid Y = 1]$ and $\mathbb{E}_0[\cdot]$ denote $\mathbb{E}[\cdot \mid Y = 0]$, and define $\mathbb{P}_1$ and $\mathbb{P}_0$ analogously. Finally, recall for $n$ i.i.d. Rademacher

1. The consecutiveness is not a strict requirement here and is only assumed to simplify the subsequent notation.
2. If there are boundary effects, e.g., if $i_j(S) = 1$, the adversary constructs $A_j$ s.t. $\mathbb{P}(A_j[r] = X_{i_j(S)}) \propto \pi[r]$ for valid $r$. This will anyhow not matter, as our algorithm will only consider $j : k \le j \le M - k$, with no boundary effects.

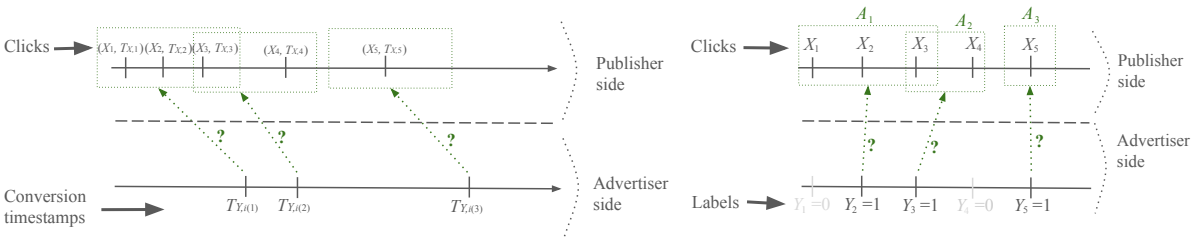


Figure 1: **Left:** The physical process. Publisher clicks $(X_i, T_{X,i})$ generate advertiser conversion timestamps $T_{Y,i(j)}$ with unknown delays. Attribution sets capture this uncertainty; for example, the conversion at $T_{Y,i(2)}$ is attributed to $\{X_3, X_4\}$. While $X_4$ is the more likely cause due to temporal proximity, the sets reflect all candidates defined by the window. Note that the attribution sets may overlap. **Right:** The simplified sequence model used for analysis. Random variables $X_1, \ldots, X_5$ with positive labels at indices $2, 3, 5$ generate the observed attribution sets $A_1 = \{1, 2, 3\}$, $A_2 = \{3, 4\}$, and $A_3 = \{5\}$.

variables $\sigma_i \in \{-1, +1\}$, the quantity

$$R_n(\mathcal{H}) = \frac{1}{n}\,\mathbb{E}_{X_1,\ldots,X_n}\Big[\mathbb{E}_{\sigma_1,\ldots\sigma_n}\Big[\sup_{h\in\mathcal{H}}\Big|\sum_{i=1}^{n}\sigma_i h(X_i)\Big| \mid X_1,\ldots,X_n\Big]\Big]$$

is the (average) Rademacher Complexity of function class $\mathcal{H}$.

## 3. An Unbiased Estimator for $\mathcal{L}(h)$

Consider a hypothesis $h : \mathcal{X} \to [0, 1]$ and a loss function $\ell(h(x), y)$. In stark contrast to standard statistical learning settings, we have only weak partial label information in the form of the attribution sets. Lacking any explicit labels, constructing an unbiased estimator for $\mathcal{L}(h)$ is far from obvious.

To construct an unbiased estimator, our first step is the following decomposition. As the labels $y$ are binary ($y \in \{0, 1\}$), we can decompose the loss into a base term and a label-dependent term:

$$\ell(h(x), y) = \overbrace{\ell(h(x), 0)}^{f_1(h(x))} + y\overbrace{\Big(\ell(h(x), 1) - \ell(h(x), 0)\Big)}^{f_2(h(x))}. \tag{1}$$

Consequently, for suitable functions $f_1, f_2 : [0, 1] \to \mathbb{R}$, any binary loss can be expressed in the affine form: $\ell(h(x), y) = f_1(h(x)) + y f_2(h(x))$. For instance, the square loss is obtained by $f_1(h) = h^2$, and $f_2(h) = 1 - 2h$, where in both cases $h \in [0, 1]$. Thus, estimating the population risk reduces to estimating $\mathbb{E}_{(X,Y)\sim\mathcal{D}}\big[f_1(h(X))\big]$ and $\mathbb{E}_{(X,Y)\sim\mathcal{D}}\big[Y f_2(h(X))\big]$.

The fundamental challenge in our setting is the latent nature of the labels $Y$, which precludes standard techniques to estimate these expectations such as importance sampling–the joint density is never observed directly. Remarkably, we show that the combinatorial structure of the attribution sets–governed by the known prior $\pi$–renders the population loss identifiable. Specifically, we leverage a combinatorial argument to derive an exact mapping between the inaccessible population moment $\mathbb{E}_{(X,Y)\sim\mathcal{D}}\big[Y f_2(h(X))\big]$ and an expectation over the observable attribution signals: $\mathbb{E}_{\mu}\big[f_2(h(A_j[i]))\,\mathbb{1}\{j \le M - k\}\big]$. This leads to our first main result:

**Theorem 1** *Let $\ell(h(x), y) = f_1(h(x)) + y f_2(h(x))$ be an arbitrary loss function for binary labels $y \in \{0, 1\}$, and $\mathcal{D}$ be a distribution over $\mathcal{X} \times \{0, 1\}$ such that $p = \mathbb{P}(Y = 1) \in (0, 1)$. Let*

*$M = \sum_{i=1}^{n} Y_i$ be a random variable denoting the number of conversions (1s) among the labels in the stream $S$. Consider any $j$ with $k \le j \le n$, any $i$ with $1 \le i \le k$, and any $h : \mathcal{X} \to [0,1]$. Let*

$$\widehat{\ell}(h,j,i) = \frac{f_2(h(A_j[i]))}{\beta_1(j,i)} + \frac{\mathbb{E}\big[f_1(h(X))\big]}{B_{n,p,j+k}} - \frac{\beta_0(j,i)\ \mathbb{E}\big[f_2(h(X))\big]}{\beta_1(j,i)\, B_{n,p,j+k}}\,, \tag{2}$$

*where $B_{n,p,k'} := \sum_{i'=k'}^{n} \binom{n}{i'} p^{i'}(1-p)^{n-i'}$ is the Binomial tail, and where*

$$\beta_1(j,i) := \frac{\pi[i]\, B_{n,p,j+k}}{p} + \Big(B_{n-1,p,j+k-1} - \frac{1-p\, B_{n-1,p,j+k-1}}{1-p}\Big)\big(1-\pi[i]\big)\,,$$
$$\beta_0(j,i) := \frac{1-p\, B_{n-1,p,j+k-1}}{1-p}\big(1-\pi[i]\big)\,.$$

*Then, we have*

$$\mathbb{E}_\mu\big[\widehat{\ell}(h,j,i)\,\mathbb{1}\{j \le M-k\}\big] = \mathcal{L}(h)\,.$$

A proof sketch follows; a full proof is in Appendix A. Next in Section 4, we leverage these unbiased estimators from Theorem 1 across different $j$ and $i$ to build an unbiased estimator of the population loss, and study its statistical properties under ERM.

**Proof (of Theorem 1; sketch)** Consider the decomposition (1). To obtain an unbiased estimator of $\mathcal{L}(h)$ having access to features without explicit labels, we need to leverage the coarse attribution set signals to estimate the label-conditional moment $\mathbb{E}\left[Y f_2(h(X))\right]$. $\mathbb{E}\left[f_1(h(X))\right]$ can be estimated just from features. Surprisingly, we show that $\mathbb{E}\left[Y f_2(h(X))\right]$ can be cast in terms of $A_j[i]$, the $i$-th element of the $j$-th attribution set, which we do have access to:

$$\mathbb{E}\big[Y f_2(h(X))\big] = \frac{1}{\beta_1(j,i)}\mathbb{E}_\mu\Big[f_2(h(A_j[i]))\cdot\mathbb{1}\{j \le M-k\}\Big] - \frac{\beta_0(j,i)}{\beta_1(j,i)}\,\mathbb{E}\big[f_2(h(X))\big]\,. \tag{3}$$

Note the $B_{n,p,j+k}$ terms in Theorem 1 arise naturally as $\mathbb{E}_\mu\left[\mathbb{1}\{j \le M-k\}\right] = B_{n,p,j+k}$. Combining (3) with the decomposition (1) now proves Theorem 1.

We will now explain the steps to establish the result in (3). Let $\eta(x) = \mathbb{P}(Y=1|X=x)$. Noting $\mathbb{E}\left[Y f_2(h(X))\right] = \mathbb{E}\left[f_2(h(X))\,\mathbb{E}[Y|X]\right] = \int \mathbb{P}(X=x) f_2(h(x))\eta(x)dx$, (3) follows from integrating the following result:

$$\mathbb{E}_\mu\Big[\mathbb{1}\{A_j[i]=x\}\cdot\mathbb{1}\{j \le M-k\}\Big] = \mathbb{P}(X=x)\Big(\eta(x)\,\beta_1(j,i) + \beta_0(j,i)\Big)\,. \tag{4}$$

Our aim is now to prove (4). To do so, we first simplify the left hand side of (4). By definition of the adversary's action, $A_j[i]$, the $i$-th element of $A_j$, is $X_{i_j(S)+i-r}$ with probability $\pi[r]$. Thus,

$$\mathbb{E}_\mu\Big[\mathbb{1}\{A_j[i]=x\}\cdot\mathbb{1}\{j \le M-k\}\Big] = \sum_{r=1}^{k}\pi[r]\,\mathop{\mathbb{E}}_{S\sim\mathcal{D}^n}\Big[\mathbb{1}\{j \le M-k\}\cdot\mathbb{1}\big\{X_{i_j(S)+i-r}=x\big\}\Big]\,. \tag{5}$$

The summation splits into two cases: $r=i$ and $r \ne i$. We now characterize $\mathbb{E}_{S\sim\mathcal{D}^n}\left[\mathbb{1}\{X_{i_j(S)+i-r} = x\}\cdot\mathbb{1}\{j \le M-k\}\right]$ for $r \ne i$, the argument for $r=i$ is similar (and omitted).

A-priori, this is challenging, as the two events in the expectation are tightly coupled. We simplify this expectation with the following key observation: conditioned on any realization of the labels—in

particular the event $\{j \leq M-k\}$—the law of $X_{i_j(S)+i-r}$ can be readily understood. Specifically, conditioned on $\{Y_{i_j(S)+i-r}=1\,,\, j \leq M-k\}$ we have $X_{i_j(S)+i-r} \sim \mathcal{D}_{X|Y=1}$, and conditioned on $\{Y_{i_j(S)+i-r}=0\,,\, j \leq M-k\}$ we have $X_{i_j(S)+i-r} \sim \mathcal{D}_{X|Y=0}$: these results are in Lemma 8.

Now, the only remaining piece to compute the expectation of $\mathbb{1}\{X_{i_j(S)+i-r}=x\}\cdot\mathbb{1}\{j \leq M-k\}$ is computing the probabilities of the events $\{Y_{i_j(S)+i-r}=1\,,\, j \leq M-k\}$ and $\{Y_{i_j(S)+i-r}=0\,,\, j \leq M-k\}$. Interestingly, by leveraging a combinatorial argument, we can show that the probability of these two events can be simplified to $pB_{n-1,p,j+k-1}$ and $1-pB_{n-1,p,j+k-1}$, respectively; this is proven in Lemma 7. Given these simplifications, and by leveraging the Bayes' rule, we can show that $\mathbb{E}_{S\sim\mathcal{D}^n}\left[\mathbb{1}\{X_{i_j(S)+i-r}=x\}\cdot\mathbb{1}\{j \leq M-k\}\right]$ equals the expression

$$\mathbb{P}(X=x)\,\eta(x)\,B_{n-1,p,j+k-1} + \frac{\mathbb{P}(X=x)\left(1-\eta(x)\right)}{1-p}\cdot\left(1-p\,B_{n-1,p,j+k-1}\right).$$

Given this expression and the definitions of $\beta_1(j,i), \beta_0(j,i)$, the result in (4) follows. ■

## 4. From an Unbiased Estimator to an ERM Algorithm

We now leverage the unbiased estimator from Theorem 1 to create a sample-efficient unbiased estimator that uses a sizeable fraction of the data. Specifically, Theorem 1 implies that for any hypothesis $h$, the quantity $\widehat{\ell}(h,j,i)\,\mathbb{1}\{j \leq M-k\}$ is an unbiased estimator of $\mathcal{L}(h)$, provided we have exact knowledge of the conversion rate $p$ (involved in the expression for $\beta_0(j,i)$ and $\beta_1(j,i)$) and the two expectations $\mathbb{E}[f_1(h(X))]$ and $\mathbb{E}[f_2(h(X))]$.

However, the cost for removing this prior knowledge would in fact be minor (and leading to a negligibly biased estimator). This is because these three quantities can be straightforwardly estimated at a higher resolution than the one allowed by the signals we receive from the adversary. To see this, note we can always split $\{X_1,\ldots,X_n\}$ into two equal-size subsets $\{X_1,\ldots,X_{n/2}\}$ and $\{X_{n/2+1},\ldots,X_n\}$, estimate $\mathbb{E}[f_1(h(X))]$ and $\mathbb{E}[f_2(h(X))]$ via $\{X_1,\ldots,X_{n/2}\}$, estimate $p$ as the fraction of conversions up to time $n/2$, and then build the estimator $\widehat{\ell}(h,j,i)$ on variables in the second half $\{X_{n/2+1},\ldots,X_n\}$, where true expectations are replaced by the estimates constructed on the first half. Now, any uniform guarantee over $h\in\mathcal{H}$ in estimating $\mathbb{E}[f_1(h(X))]$ and $\mathbb{E}[f_2(h(X))]$ will be at a rate $1/\sqrt{n}$, and similarly for $p$. On the other hand, as we shall see below in Theorem 2, the amount of information the adversary releases to the learner can only afford rates at best $1/\sqrt{n}$.

Consequently, with little loss of generality, we assume that $p, \mathbb{E}[f_1(h(X))], \mathbb{E}[f_2(h(X))]$ are known to the learner. Furthermore, for simplicity, we work with a *bounded* loss function: for all $h\in\mathcal{H}$, and all $(x,y)\in\mathcal{X}\times\{0,1\}$ we have $\ell(h(x),y) = f_1(h(x)) + yf_2(h(x))$, with[3] $|f_1(h(x))|\leq F_1$, and $|f_2(h(x))|\leq F_2$, for some $F_1,F_2>0$.

### 4.1. The ERM Algorithm

We now describe the ERM algorithm. Write $S=\big\langle(X_1,Y_1),\ldots,(X_n,Y_n)\big\rangle$, and denote the family of attribution sets by $\mathcal{A}=\{A_1,\ldots,A_M\}$, with $M=\sum_{i=1}^n Y_i$. Define $M_{\text{UPPER}} := \min\{\frac{np}{2}-k, M-k\}$.

3. The boundedness involving $f_1(\cdot)$ will not play any role here. Since we assumed prior knowledge of $\mathbb{E}[f_1(h(X)]$ in Section 3, it suffices to have $\mathbb{E}[f_1(h(X))]<\infty$.

Let $\Sigma = \sum_{i=1}^{k} \pi[i]^2 = \|\pi\|_2^2$; the quantity $\frac{1}{\Sigma}$ is the "effective" set size determined by prior sparsity.[4] Recalling the definition of $\widehat{\ell}(h, j, i)$ from (2), we then consider the estimator

$$\widehat{\ell}_M(h, S, \mathcal{A}) = \frac{1}{\frac{np}{2} - 2k + 1} \sum_{j=k}^{M_{\text{UPPER}}} \widehat{\ell}(h, j) \quad \text{where} \quad \widehat{\ell}(h, j) = \frac{1}{\Sigma} \sum_{i=1}^{k} \pi[i]^2 \, \widehat{\ell}(h, j, i) \,. \tag{6}$$

Note $\widehat{\ell}_M(h, S, \mathcal{A}) = \frac{1}{\frac{np}{2}-2k+1} \sum_{j=k}^{\frac{np}{2}-k} \widehat{\ell}(h, j) \cdot \mathbb{1}\{j \le M - k\}$. Thus by Theorem 1, $\widehat{\ell}_M(h, S, \mathcal{A})$ is unbiased. Finally, define the ERM estimator

$$\widehat{h} = \widehat{h}(S, \mathcal{A}) = \arg\min_{h \in \mathcal{H}} \widehat{\ell}_M(h, S, \mathcal{A}) \,. \tag{7}$$

We next state the following sample complexity guarantee for $\widehat{h}$. This is the main result of this paper.

**Theorem 2** *Let $\ell(h(x), y) = f_1(h(x)) + y f_2(h(x))$ be a bounded and Lipschitz loss function: for some $F_1, F_2, L > 0$, for all $h \in \mathcal{H}$ and $x \in \mathcal{X}$, and $f_2(\cdot)$ is $L$-Lipschitz. Let $\widehat{h} \in \mathcal{H}$ be the hypothesis returned by the ERM estimator (7). Suppose $p$, $\delta \in (0, 1/2]$, $k \le \frac{np}{8}$, and $np = \Omega\left(\log\left(\frac{1}{\delta p} \max_{i \in [k]} \frac{1}{\pi[i]}\right)\right)$. Then with $\mu$-probability at least $1 - \delta$,*

$$Reg(\widehat{h}) = \widetilde{O}\left(\frac{L\, R_n(\mathcal{H})}{\Sigma} + \frac{F_2}{\Sigma}\sqrt{\frac{\log 1/\delta}{n}} + \frac{F_2}{\Sigma}\sqrt{\frac{p \, \min\{PDim(\mathcal{H}), k\}}{n}}\right),$$

*where $PDim(\mathcal{H})$ denotes the pseudo-dimension of $\mathcal{H}$. Here $\widetilde{O}(\cdot)$ hides a logarithmic dependence on $np, L, \Sigma, F_2$ but excluding $1/\delta$.*

Proving Theorem 2 poses several challenges. The attribution sets may overlap, thus the estimators $\widehat{\ell}(h, j)$ are *not* independent across different $j$. Moreover, the unbiased estimators $\widehat{\ell}(h, j, i)$ are *also not* independent across $i$, as the $A_j[i]$ are not independent; for example, for any $1 \le j \le M$, it is known that there is at least one conversion among the labels corresponding to $A_j[1], \ldots, A_j[k]$.

One way partially around the independence issue is to force independence across different $j$ by skipping data; instead of using all attribution sets, only use a largest subsequence of well-separated attribution sets, e.g., $A_k, A_{3k}, A_{5k}, \ldots$. The drawback of this approach is that we are only using $\frac{M}{2k}$-many sets instead of the available $M$, in contrast to the ERM estimator (7) that we consider. This would inevitably lead to suboptimal sample complexity guarantees. We instead eschew an approach based on independence of attribution sets, and as such we are able to use a sizeable fraction of them, as claimed above. The full proof of Theorem 2 is in Appendix B. A proof sketch follows.

**Proof (of Theorem 2; sketch)** We sidestep the potentially complicated dependency structure by rewriting $\widehat{\ell}(h, S, \mathcal{A})$ as a function of the $X_i$, and directly studying the sensitivity of $\widehat{\ell}(h, S, \mathcal{A})$. We then control $Reg(\widehat{h})$ by splitting into two separate uniform convergence guarantees:

1. Convergence of $\mathbb{E}_{\mu|S}[\widehat{\ell}(h, S, \mathcal{A})]$ to its expectation $\mathcal{L}(h)$, where probability is w.r.t. $S \sim \mathcal{D}^n$. Here we use a Rademacher complexity analysis that views $f(S) = \mathbb{E}_{\mu|S}[\widehat{\ell}(h, S, \mathcal{A})]$ as a function of i.i.d. random variables $(X_1, Y_1), \ldots, (X_n, Y_n)$; we control the sensitivity of $f(S)$ to each individual pair $(X_i, Y_i)$. This yields the $\frac{L\, R_n(\mathcal{H})}{\Sigma}$ term in the regret guarantee.

4. We have $1 \ge \Sigma \ge \frac{1}{k}$, where the lower bound is from Cauchy-Schwarz; equality is obtained in the upper and lower bound when $\pi$ is a singleton and uniform, respectively.

2. Convergence of $\widehat{\ell}(h, S, \mathcal{A})$ to its expectation $\mathbb{E}_{\mu|S}[\widehat{\ell}(h, S, \mathcal{A})]$ in the conditional space where $S$ is frozen, where probability is w.r.t. the adversary. This instead relies on a covering argument that views $\widehat{\ell}(h, S, \mathcal{A})$ solely as a function of the attribution sets (since $S$ is frozen), and applies with high probability over the generation of $S$. This generates the term $\frac{F_2}{\Sigma}\sqrt{\frac{p\ \min\{\text{PDim}(\mathcal{H}),k\}}{n}}$.

Finally, the middle term $\frac{F_2}{\Sigma}\sqrt{\frac{\log 1/\delta}{n}}$ is a confidence term that is common to both analyses. ■

**Remark 3** *We note that the condition* $np = \Omega\left(\log\left(\frac{1}{\delta p}\max_{i\in[k]}\frac{1}{\pi[i]}\right)\right)$ *in Theorem 2 forces* $\pi[i]$ *not to be exponentially small in* $np$*. This assumption is only made above for technical convenience, and can be relaxed to* $np \geq \log\left(\frac{\sqrt{2k}}{\delta p}\right)$*. Indeed, we can always construct* $\widehat{\ell}(h, j)$ *by restricting to the* $i$ *such that* $\pi[i] \geq \frac{1}{\delta p}e^{-np}$*. One then follows the exact same proof of Theorem 2 in Appendix B, only considering these* $i$*. Note that, as per Theorem 1, unbiasedness is retained. The difference is that* $\Sigma$ *now is replaced by the slightly smaller quantity* $\Sigma' = \sum_{i:\pi[i]\geq\frac{1}{\delta p}e^{-np}}\pi[i]^2$*. Yet, when* $np \geq \log\left(\frac{\sqrt{2k}}{\delta p}\right)$*, since* $\Sigma \geq \frac{1}{k}$*, we have* $\sum_{i:\pi[i]<\frac{1}{\delta p}e^{-np}}\pi[i]^2 \leq \frac{ke^{-2np}}{\delta^2p^2} \leq \frac{1}{2k} \leq \frac{\Sigma}{2}$ *, implying* $\Sigma' \geq \frac{\Sigma}{2}$*. Thus the regret rate only changes by a constant factor.*

**Remark 4** *It is worth stressing that the weights* $w_i = \pi[i]^2/\Sigma$*, in the definition of* $\widehat{\ell}(h, j)$ *in (6) have been selected just for the sake of minimizing the variance of* $\widehat{\ell}(h, j)$*. Note the random variable in* $\widehat{\ell}(h, j, i)$ *is* $f_2(h(A_j[i]))/\beta_1(j, i)$*, and* $\beta_1(j, i) \approx \pi[i]/p$ *(Lemma 11). Ignoring pairwise correlations, we approximate the variance of* $\widehat{\ell}(h, j) = \sum_{i=1}^{k} w_i\widehat{\ell}(h, j, i)$ *by* $\sum_i w_i^2/\pi[i]^2$*, which is minimized at* $w_i \propto \pi[i]^2$*, as per our choice in (6). Moreover, one can show that within each attribution set, such pairwise correlations are* non-positive*. The proof of this claim is lengthy but otherwise similar to the proof of Theorem 1. Yet, since this is not needed to prove the other results, we did not include it in the paper for brevity. In general, it seems difficult to find the optimal weights to construct the* global *estimator* $\widehat{\ell}_M(h, S, \mathcal{A})$ *from the* $\widehat{\ell}(h, j)$ *in (6), due to possible overlaps among attribution sets.*

We now instantiate Theorem 2 on several concrete examples.

1. **Uniform vs decaying priors.** For a uniform prior $\pi[\cdot]$, our regret bound is of the form

$$k \cdot \frac{LR_n(\mathcal{H}) + F_2\sqrt{\log 1/\delta} + F_2\sqrt{p\ \min\{\text{PDim}(\mathcal{H}), k\}}}{\sqrt{n}}\ .$$

   However, in practice the last-touch heuristic (last element in attribution set triggers conversion) is often approximately true. A heavily decaying prior, e.g., with polynomial or exponential decay, is a more accurate model. For these examples, $\Sigma$ is much larger, being $\Omega(1)$. Our regret in Theorem 2 adapts to such settings gracefully.

2. $\mathcal{H}$ **is a VC-class.** That is, $h(x) \in \{0, 1\}$ for all $x \in \mathcal{X}$ and $h \in \mathcal{H}$, and its VC-dimension $VCdim(\mathcal{H}) = d < \infty$. Our regret bound is of the form

$$\frac{L\sqrt{d} + F_2\sqrt{\log 1/\delta} + F_2\sqrt{p\ \min\{d, k\}}}{\Sigma\sqrt{n}}\ .$$

   This follows from the fact that for VC-classes, $R_n(\mathcal{H}) = \widetilde{O}\left(\sqrt{VCdim(\mathcal{H})/n}\right)$.

3. $|\mathcal{H}|$ **is a finite class.** Our regret bound is of the form

$$\frac{L\sqrt{\log|\mathcal{H}|}+F_2\sqrt{\log 1/\delta}+F_2\sqrt{p\log|\mathcal{H}|}}{\Sigma\sqrt{n}}\,.$$

This follows from Massart's Finite Lemma (Massart, 2000) and the bound $\mathrm{PDim}(\mathcal{H})\le\log|\mathcal{H}|$.

4. **When** $pk=O(1)$**.** Here the attribution sets have minimal overlap with high probability. Our regret bound is of the form

$$\frac{LR_n(\mathcal{H})}{\Sigma}+\frac{F_2}{\Sigma}\sqrt{\frac{\log 1/\delta}{n}}\,.$$

### 4.2. Robustness to errors in the prior

Here we discuss how to extend Theorem 2 to when the learner knows an estimate $\widehat{\pi}$ of the distribution $\pi$ with small error. This is a realistic situation where, e.g., $\pi$ is well-approximated by a parametric form, and we can estimate its parameters, for example with a small source of labeled data in the clear.

The algorithm is similar but implemented in terms of $\widehat{\pi}$. We define $\widehat{\beta}_1(j,i),\widehat{\beta}_0(j,i)$ the same way as $\beta_1(j,i),\beta_0(j,i)$, but using $\widehat{\pi}$ instead of $\pi$, and then define $\widehat{\ell}(h,j,i),\widehat{\ell}(h,j)$ in terms of $\widehat{\beta}_1(j,i),\widehat{\beta}_0(j,i)$ as before. The explicit definition is provided in Appendix B.1. We now let

$$\widehat{\ell}=\widehat{\ell}(h,S,\mathcal{A})=\frac{1}{\frac{np}{2}-2k+1}\sum_{j=k}^{M_{\text{UPPER}}}\widehat{\ell}(h,j)\,,\qquad \widehat{h}=\widehat{h}(S,\mathcal{A})=\arg\min_{h\in\mathcal{H}}\widehat{\ell}(h,S,\mathcal{A})\,.$$

We will establish in Appendix B.1 that we can obtain a regret $Reg(\widehat{h})$ that is the sum of a "Bias" term, plus a term analogous to the regret from Theorem 2. The Bias term decreases in the approximation error between $\pi$ and $\widehat{\pi}$, and does not feature explicit $k$-dependence when the squared $L_2$ distance $\|\pi-\widehat{\pi}\|_2^2\le\frac{\Sigma}{8}$ (recall $\Sigma=||\pi||_2^2$). When $\|\pi-\widehat{\pi}\|_2^2>\frac{\Sigma}{8}$, the explicit $k$-dependence in $Reg(\widehat{h})$ below is unavoidable using our current analysis; see Remark 17 in Appendix B.

**Theorem 5** *Using the same notation as defined above, under the same conditions on the loss function $\ell$ and on $k$ as in Theorem 2, we have the following. If $np=\Omega\Big(\log\big(\frac{\sqrt{2k}}{\delta p}\big)\Big)$, then with $\mu$-probability at least $1-\delta$, the ERM estimator $\widehat{h}$ satisfies*

$$Reg(\widehat{h})=\begin{cases}\widetilde{O}\left(\frac{L\,R_n(\mathcal{H})}{\Sigma}+\frac{F_2}{\Sigma}\sqrt{\frac{\log\frac{1}{\delta}}{n}}+\frac{F_2}{\Sigma}\sqrt{\frac{p\,\min\{PDim(\mathcal{H}),k\}}{n}}\right)+Bias & \text{if } \Sigma\ge 8\|\pi-\widehat{\pi}\|_2^2\,,\\ \widetilde{O}\left(Lk\,R_n(\mathcal{H})+F_2k\sqrt{\frac{\log\frac{1}{\delta}}{n}}+F_2k\sqrt{\frac{p\,\min\{PDim(\mathcal{H}),k\}}{n}}\right)+Bias & \text{if } \Sigma<8\|\pi-\widehat{\pi}\|_2^2\,,\end{cases}$$

*where $PDim(\mathcal{H})$ denotes the pseudo-dimension of $\mathcal{H}$, and where*

$$Bias:=\begin{cases}O\left(pF_2\left(\frac{\|\pi-\widehat{\pi}\|_1}{\Sigma}+\frac{\|\pi-\widehat{\pi}\|_2}{\Sigma^{3/2}}\right)\right) & \text{if } \Sigma\ge 8\|\pi-\widehat{\pi}\|_2^2\,,\\ O\left(\frac{pkF_2\|\pi-\widehat{\pi}\|_2}{\Sigma^{1/2}}\right) & \text{if } \Sigma<8\|\pi-\widehat{\pi}\|_2^2\,.\end{cases}$$

*Again, $\widetilde{O}(\cdot)$ hides a logarithmic dependence on $np,L,\Sigma,F_2$ but excluding $1/\delta$.*

**Remark 6** *As an application of Theorem 5, it is instructive to consider the case where the prior $\pi$ is unknown, but can be estimated by using a smaller set of labeled data. Given such a labeled dataset, an alternative method would be to simply perform ERM directly on that data. We now show that, in regimes of practical interest, our results in Theorem 5 yields a superior rate over ERM just operating on the smaller set of data available in the clear. Most notably, Theorem 5 avoids any dependence on the complexity of $\mathcal{H}$ evaluated on this reduced sample size, unlike ERM.*

*For concreteness, suppose $\mathcal{H}$ is a VC-class of VC-dimension $d$ and that, on top of attribution sets, the labels of $\mathcal{S}$, a set of examples $(x, y)$ covered by $s$ attribution sets of size $k$ are also known. Hence, overall, $|\mathcal{S}| \leq sk$. Assume, as plausible in practice, that $sk \ll n$. Then ERM on $\mathcal{S}$ gives a rate at best of the form $\sqrt{d/(sk)}$. Meanwhile, if we estimate $\pi$ by empirical probabilities over $\mathcal{S}$,* Bias *from Theorem 5 is of the form $pk^2/\sqrt{s}$ (this is when $\pi$ is uniform; for a non-uniform $\pi$, this can only improve). Theorem 5 thus yields a rate of the form $k\sqrt{d/n} + pk^2/\sqrt{s}$, improving on ERM operating solely on $\mathcal{S}$ when $d$ is large relative to $pk^2$ (e.g., $d \geq k(pk^2)^2$), and $n$ is large relative to $s$ (e.g., $n \geq k^3 s$), a setting which is arguably closer to practice.*

## 5. Experiments

We conduct preliminary experiments to validate the estimator constructed from Theorem 2 vs. simple baselines that correspond to industry heuristics (e.g., Ktena et al. (2019)). Our experiments are performed on MNIST (LeCun et al., 2010), CIFAR-10 (Krizhevsky, 2009), and Higgs (Baldi et al., 2014), each modified in a way compatible with our model. More details are in Section C.

**Modifying the datasets.** We binarize each dataset: 1-vs-rest for MNIST, Animal-vs-Machine for CIFAR-10, while Higgs is natively binary. We then shuffle the data. For each positive label, we generate an attribution set by drawing an interval of $k$ adjacent indices that contain the positive label, where the position of the window is drawn according to the prior $\pi$. The algorithms observe only the unlabeled data and the resulting attribution sets. We consider a uniform prior $\pi = \left(\frac{1}{k}, \ldots, \frac{1}{k}\right)$, and an exponential prior $\pi \propto \left(2^{-k}, \ldots, 2^{-2}, 2^{-1}\right)$ motivated by last-touch attribution heuristics.

**Algorithms.** We implement three algorithms. For each algorithm the base loss $\ell(\widehat{y}, y)$ is log loss with prediction $\widehat{y}$ clipped to the interval $[0.01, 0.99]$ for boundedness and numerical stability.

1. Our algorithm (UNBIASED): minimizes the loss estimator in (6), where $p$ in Theorem 1 is estimated by the fraction of positive labels in the training set, and $\mathbb{E}[f_1(h(X)]$, $\mathbb{E}[f_2(h(X)]$ are estimated on each mini-batch (see training details below) by empirical averages. While this estimation introduces a slight bias, our experiments confirm that this effect is negligible.
2. RANDOM baseline: Both this algorithm and the following baseline operate directly on the base loss $\ell(\widehat{y}, y)$ (with the same clipping for $\widehat{y}$) on fully supervised but *hallucinated* labels. RANDOM assigns label 1 to a single position $i$ per attribution set, drawing this position according to $\pi$, and label 0 to all remaining points in the attribution set. The data points in the training set that do not fall into any attribution set are assigned label 0. Overlapping attribution sets may produce duplicate instances with potentially conflicting labels.[5]
3. MAX PRIOR baseline: analogous to RANDOM, but the positive label is placed deterministically at the position $i$ where $\pi[i]$ is maximized.

**Training and evaluation.** For each dataset, we train standard neural architectures known to perform reasonably well: a 3-hidden-layer MLP for MNIST, a 2-layer CNN for CIFAR-10, and a fully

5. A natural alternative would be to generate fractional labels, but we did not explore this solution here.

connected network for Higgs, training with the Adam optimizer for each algorithm (Kingma and Ba, 2015). For RANDOM and MAX PRIOR, we take a minibatch of 128 training examples with the hallucinated labels. For UNBIASED, we estimate the loss from (6) by subsampling a minibatch of 128 attribution sets and another 128 training examples (without labels) directly from the dataset in order to estimate the expectation components $\mathbb{E}[f_1(h(X))]$, $\mathbb{E}[f_2(h(X))]$. For all three algorithms, we use bag sizes $k = 2^i$ for $0 \leq i \leq 8$ (or $0 \leq i \leq 7$ for Higgs), learning rates range in 10 log-spaced values from $10^{-6}$ to $10^{-2}$, and we use 200 training epochs for MNIST and 100 for CIFAR-10 and Higgs. Each experiment (a given dataset, algorithm, attribution set size, and learning rate) is repeated 10 times with randomized data shuffling and model initialization. Performance is measured on the test set with labels in the clear, averaged across repetitions. For each dataset, algorithm, and attribution set size, we report the best average over learning rates.

**Results.** We report test set accuracy (Figure 2), and test set log loss (Figure 3 in Appendix C); for MNIST, we also show test set F1-measure in Figure 3 due to label imbalance. As expected, the performance of all algorithms degrades as $k$ increases. For large enough $k$, performance becomes trivial, for instance with RANDOM and MAX PRIOR on CIFAR-10 when $k \geq 4$. On CIFAR-10, the trivial accuracy performance of 60% is obtained by always predicting "1"; for MNIST, always predicting "0" achieves 88.65% accuracy. In such cases, we report the trivial performance level (with 0 variance) instead of the actual performance in Figure 2. Several observations can be made:

- UNBIASED vs. RANDOM and MAX PRIOR. We can see the clear advantage offered by our theory as opposed to the baselines; the performance gap is striking in all cases.
- Uniform vs. Exponential prior. All algorithms perform better with exponential prior than with uniform, with MAX PRIOR being comparatively better than RANDOM. Note for the exponential prior, $\Sigma \to \frac{1}{3}$ as $k \to \infty$; our Theorem 2 predicts UNBIASED will degrade gracefully as $k$ increases, consistent with our results in Figure 2.
- Small $k$. At $k = 1$, the two baselines both reduce to full supervision with true training labels (no label hallucination); this is not the case for UNBIASED. Thus the baselines' performance at $k = 1$ ($2^0$ in Figures 2 and 3) reflects fully supervised training with the base loss.
- Overlapping ($pk > 1$) vs. non-overlapping ($pk < 1$) regime. The difference here can only be appreciated on MNIST, where $p \approx 0.1$. UNBIASED remains largely unaffected by the attribution set overlap, but both baselines are affected significantly, especially RANDOM.

## 6. Conclusions and Future Work

We introduced a formal framework for statistical learning from attribution sets, addressing the growing challenge of tracking-prevention conversion prediction, where publishers observe clicks but only receive coarse signals about conversions. We derived an unbiased risk estimator, established generalization bounds that scale with the "effective" set size determined by prior sparsity, and proved robustness to prior estimation errors. Given the availability of unbiased (or approximately unbiased) loss estimators, these analyses can be readily adapted to stochastic gradient descent-like algorithms, since unbiasedness of loss estimates translate to unbiasedness of loss gradient estimates. Our preliminary experiments suggest that our method significantly outperforms simple industry heuristics on readily available datasets, particularly when attribution sets are large and/or overlapping. Future work includes the following.

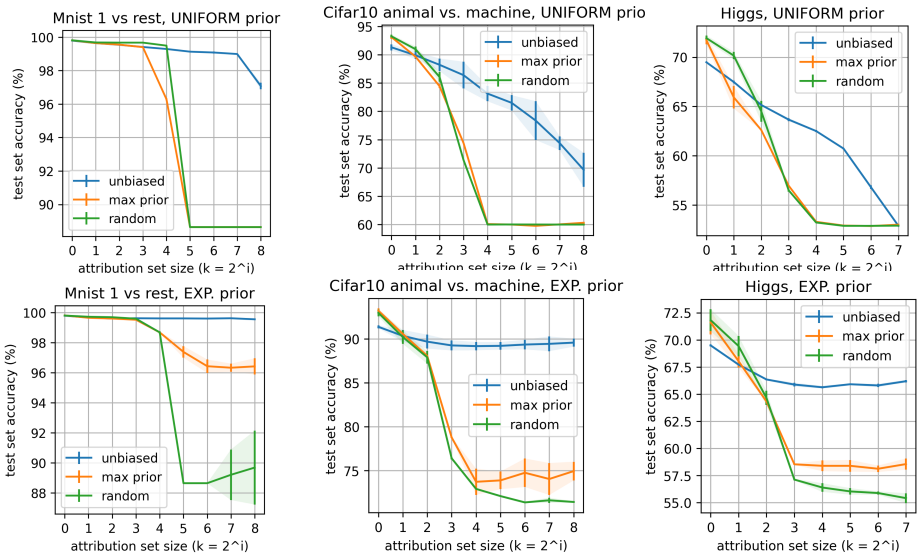


Figure 2: Experiment on MNIST (1-vs-rest), CIFAR-10 (animal vs. machine) and Higgs datasets. We plot test set accuracy vs. attribution set size $k = 2^0, 2^1, \ldots, 2^8$ (or $k = 2^0, 2^1, \ldots, 2^7$ for Higgs), averaged over 10 repetitions. On the top row is the uniform prior, on the bottom is the exponential prior. For MNIST, the trivial accuracy performance is $88.65\%$, for CIFAR-10 it is $60\%$, for Higgs it is $52.87\%$. Standard deviations are also depicted.

- We have assumed an oblivious adversary and a known (or partially known) prior $\pi$ over positions within each attribution set; this structure is critical for identifiability. The estimators in Theorem 2 and Theorem 5 hinge on expressing the inaccessible population moment $\mathbb{E}[Y f_2(h(X))]$ as an expectation over observable quantities, with coefficients that depend explicitly on $\pi$. One may wonder if learning against a non-oblivious adversary is at all possible. In a non-oblivious regime, the above decoupling fails. Since an attribution set merely indicates the presence of at least one positive label, multiple labelings–and thus multiple population risks–remain consistent with the same observations. This lack of identifiability suggests that consistent learning is impossible without additional structure. We provide this as an intuition rather than a formal impossibility result. Characterizing the minimal assumptions required for learning under adaptive adversaries is a compelling direction for future work.
- Our agnostic bounds extend to the realizable setting. However, capturing the fast convergence rates expected under benign losses (like square loss) requires a localized Rademacher complexity analysis, which we leave to future work. Proving matching regret lower bounds to establish the tightness of our results also remains an open problem.
- We plan to extend out framework to more complex attribution logics, like multi-touch attribution (where multiple clicks contribute to a single conversion), and to develop methods to jointly learn the prior distribution and the conversion model from the data stream itself.

## Acknowledgments

We thank the anonymous reviewers for their useful comments that helped improve the presentation of this paper. Also, we thank Haim Kaplan for early discussions related to this paper. This work was done while AC was a student researcher at Google Research, NY.

## References

Lorne Applebaum, Travis Dick, Claudio Gentile, Haim Kaplan, and Tomer Koren. Optimal learning from label proportions with general loss functions, 2026. URL https://arxiv.org/abs/2509.15145.

Pierre Baldi, Peter Sadowski, and Daniel Whiteson. Searching for Exotic Particles in High-Energy Physics with Deep Learning. *Nature Commun.*, 5:4308, 2014. doi: 10.1038/ncomms5308.

Christian Borgs, Jennifer Chayes, Nicole Immorlica, Kamal Jain, Omid Etesami, and Mohammad Mahdian. Dynamics of bid optimization in online advertisement auctions. In *Proceedings of the 16th international conference on World Wide Web*, pages 531–540, 2007.

Anand Brahmbhatt, Rishi Saket, and Aravindan Raghuveer. Pac learning linear thresholds from label proportions. *Advances in Neural Information Processing Systems*, 36:66610–66646, 2023.

Robert Busa-Fekete, Heejin Choi, Travis Dick, Claudio Gentile, and Andres Munoz Medina. Easy learning from label proportions. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NeurIPS 2023. Curran Associates Inc., 2023.

Robert Busa-Fekete, Travis Dick, Claudio Gentile, Haim Kaplan, Tomer Koren, and Uri Stemmer. Nearly optimal sample complexity for learning with label proportions. In *ICML 2025*, 2025.

Han Cai, Kan Ren, Weinan Zhang, Kleanthis Malialis, Jun Wang, Yong Yu, and Defeng Guo. Real-time bidding by reinforcement learning in display advertising. In *Proceedings of the tenth ACM international conference on web search and data mining*, pages 661–670, 2017.

Sishuo Chen, Zhangming Chan, Xiang-Rong Sheng, Lei Zhang, Sheng Chen, Chenghuan Hou, Han Zhu, Jian Xu, and Bo Zheng. See beyond a single view: Multi-attribution learning leads to better conversion rate prediction, 2025. URL https://arxiv.org/abs/2508.15217.

Yu Chen, Jiaqi Jin, Hui Zhao, Pengjie Wang, Guojun Liu, Jian Xu, and Bo Zheng. Asymptotically unbiased estimation for delayed feedback modeling via label correction. In *Proc. WWW 2022*, 2022.

Richard Combes. An extension of Mcdiarmid's inequality. In *2024 IEEE International Symposium on Information Theory (ISIT)*, pages 79–84. IEEE, 2024.

Luke Crouch and Maxx Crawford. Over a decade of anti-tracking work at mozilla. https://blog.mozilla.org/en/privacy-security/mozilla-antitracking-milestones-timeline/, 2022.

Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71, 1997.

Zhikang Fan, Lan Hu, Ruirui Wang, Zhongrui Ma, Yue Wang, Qi Ye, and Weiran Shen. Two-stage auction design in online advertising. In *Proceedings of the ACM on Web Conference 2025*, pages 3571–3585, 2025.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018.

Jaeseok Jang and Hyuk-Yoon Kwon. Are multiple instance learning algorithms learnable for instances? In *Proc. Neurips*, 2024.

Syed Ashar Javed, Dinkar Juyal, Harshith Padigela, Amaro Taylor-Weiner, Limin Yu, and Aaditya Prakash. Additive mil: Intrinsically interpretable multiple instance learning for pathology. *Advances in Neural Information Processing Systems*, 35:20689–20702, 2022.

Junqi Jin, Chengru Song, Han Li, Kun Gai, Jun Wang, and Weinan Zhang. Real-time bidding with multi-agent reinforcement learning in display advertising. In *Proceedings of the 27th ACM international conference on information and knowledge management*, pages 2193–2201, 2018.

Pooria Joulani, Andras Gyorgy, and Csaba Szepesvari. Online learning under delayed feedback. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1453–1461. PMLR, 2013.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, Univ. of Toronto, 2009.

Sofia Ira Ktena, Alykhan Tejani, Lucas Theis, Pranay Kumar Myana, Deepak Dilipkumar, Ferenc Huszár, Steven Yoo, and Wenzhe Shi. Addressing delayed feedback for continuous training with neural networks in ctr prediction. In *Proceedings of the 13th ACM conference on recommender systems*, pages 187–195, 2019.

Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

Michel Ledoux and Michel Talagrand. *Probability in Banach spaces. Classics in Mathematics. Isoperimetry and processes, Reprint of the 1991 edition*. Springer-Verlag, Berlin, 2011, 2011.

G. Li, L. Chen, A. Javanmard, and V. Mirrokni. Optimistic rates for learning from label proportions. In *Proceedings of Machine Learning Research*, volume 247 of *37th Annual Conference on Learning Theory*, pages 1–38, 2024.

Xiangyu Liu, Chuan Yu, Zhilin Zhang, Zhenzhe Zheng, Yu Rong, Hongtao Lv, Da Huo, Yiqing Wang, Dagui Chen, Jian Xu, et al. Neural auction: End-to-end learning of auction mechanisms for e-commerce advertising. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 3354–3364, 2021.

Hui Lv, Zhongqi Yue, Qianru Sun, Bin Luo, Zhen Cui, and Hanwang Zhang. Unbiased multiple instance learning for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8022–8031, 2023.

Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. In *Advances in neural information processing systems*, volume 10, 1997.

Pascal Massart. Some applications of concentration inequalities to statistics. *Annales de la Faculté des Sciences de Toulouse*, IX:245–303, 2000.

Shahar Mendelson and Roman Vershynin. Entropy and the combinatorial dimension. *Inventiones Mathematicae*, 152(1):37–55, 2003.

Giorgio Patrini, Richard Nock, Paul Rivera, and Tiberio Caetano. (almost) no label no cry. *Advances in Neural Information Processing Systems*, 27, 2014.

Novi Quadrianto, Alex J Smola, Tiberio S Caetano, and Quoc V Le. Estimating labels from label proportions. In *Proceedings of the 25th International Conference on Machine learning*, pages 776–783, 2008.

Rishi Saket. Learnability of linear thresholds from label proportions. *Advances in Neural Information Processing Systems*, 34:6555–6566, 2021.

Rishi Saket. Algorithms and hardness for learning linear thresholds from label proportions. *Advances in Neural Information Processing Systems*, 35:1267–1279, 2022.

Clayton Scott and Jianxin Zhang. Learning from label proportions: A mutual contamination framework. *Advances in neural information processing systems*, 33:22256–22267, 2020.

Raghav Singal, Omar Besbes, Antoine Desir, Vineet Goyal, and Garud Iyengar. Shapley meets uniform: An axiomatic framework for attribution in online advertising. In *The world wide web conference*, pages 1713–1723, 2019.

Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W Verjans, and Gustavo Carneiro. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4975–4986, 2021.

Claire Vernade, Olivier Cappé, and Vianney Perchet. Stochastic bandit models for delayed conversions. In *Proc. UAI, 2017*, 2017.

Claire Vernade, Alexandra Carpentier, Tor Lattimore, Giovanni Zappella, Beyza Ermis, and Michael Brueckner. Linear bandits with stochastic delayed feedback. In *Proc. ICML, 2020*, 2020.

Jun Wang, Weinan Zhang, Shuai Yuan, et al. Display advertising with real-time bidding (rtb) and behavioural targeting. *Foundations and Trends® in Information Retrieval*, 11(4-5):297–435, 2017.

John Wilander. Intelligent tracking prevention 2.3. apple. https://webkit.org/blog/9521/intelligent-tracking-prevention-2-3/, 2019.

Xun Yang, Yasong Li, Hao Wang, Di Wu, Qing Tan, Jian Xu, and Kun Gai. Bid optimization by multivariable control in display advertising. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1966–1974, 2019.

Jianxin Zhang, Yutong Wang, and Clay Scott. Learning from label proportions by learning with label noise. *Advances in Neural Information Processing Systems*, 35:26933–26942, 2022.

Han Zhu, Junqi Jin, Chang Tan, Fei Pan, Yifan Zeng, Han Li, and Kun Gai. Optimized cost per click in taobao display advertising. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 2191–2200, 2017.

## Appendix A. Proofs for Section 3

This section contains the proofs that apply to the individual attribution set.

### A.1. Proof of Theorem 1

First we need the following lemma which lets us control the probabilities that labels prior to and after $i_j(S)$ are 1.

**Lemma 7** *For any $t \geq 0$ with $t+1 \leq j$, we have*

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^n}\Big(Y_{i_j(S)\pm t}=1\,,j\leq M-k\Big)=p\,B_{n-1,p,j+k-1}\,,$$

*where $B_{n,p,j}$ is defined in Theorem 1.*

**Proof** Note the event $\{Y_{i_j(S)\pm t}=1\}\cap\{j\leq M-k\}$ can be rewritten as a disjoint union over all $m$ such that $j+k\leq m\leq n$, of the intersections $\{Y_{i_j(S)\pm t}=1\}\cap\{M=m\}$. Now we claim that for fixed $m$, the number of binary strings satisfying $\{Y_{i_j(S)\pm t}=1\}\cap\{M=m\}$ is $\binom{n-1}{m-1}$.

- For $Y_{i_j(S)-t}$: Note $\{Y_{i_j(S)\pm t}=1\}\cap\{M=m\}$ can be written as the disjoint union over all $l$ such that $\max\{t+1,j\}\leq l\leq n$ of the intersection $\{i_j(S)=l\}\cap\{Y_{l-t}=1\}\cap\{M=m\}$. The reason $l\geq j$ is because $i_j(S)$ is the $j$-th occurrence of a sample in the stream $S$ with label 1, thus it must occur at the $j$-th position or after. Moreover we need $l\geq t+1$ for the event $\{Y_{i_j(S)-t}=1\}$ to make sense. This event occurs if and only if $Y_l=Y_{l-t}=1$, if there are exactly $j-2$ ones in $[l-1]-\{l-t\}$, and if there are exactly $m-j$ ones in $\{l+1,\ldots,n\}$. These ones can be chosen in $\binom{l-2}{j-2}\binom{n-l}{m-j}$ ways, where we adopt the convention that $\binom{\cdot}{-1}=0$ (which matches the combinatorial interpretation). Since $j\geq t+1$, it follows that the number of desired strings in this case is

$$\sum_{l=j}^{n}\binom{l-2}{j-2}\binom{n-l}{m-j}=\sum_{k=0}^{n-j}\binom{j-2+k}{j-2}\binom{n-j-k}{m-j}=\binom{n-1}{m-1}\,,$$

  where we apply Vandermonde's identity in the last step.

- For $Y_{i_j(S)+t}$: Similar to the proof of Lemma 7, note the event $\{Y_{i_j(S)+t}=1\}\cap\{M=m\}$ can be rewritten as a disjoint union over all $l$ such that $j\leq l\leq\min\{n-t,n-m+k\}$ of the intersection $\{i_j(S)=l\}\cap\{Y_{l+t}=1\}\cap\{M=m\}$. Here $l\leq\min\{n-t,n-m+j\}$ because we need $i_j(S)+t\in[n]$ and as there are $m-j$ 1s in $\{l+1,\ldots,n\}$. These ones can be chosen in $\binom{l-1}{j-1}\binom{n-l-1}{m-j-1}$ ways. Note $n-m+j\leq n-t$ as $m\geq j+k\geq j+t$. Thus, the total number of desired strings in this case is

$$\sum_{l=j}^{n-m+j}\binom{l-1}{j-1}\binom{n-l-1}{m-j-1}=\sum_{k=0}^{n-m}\binom{k+j-1}{j-1}\binom{n-j-1-k}{m-j-1}=\binom{n-1}{m-1}\,,$$

  where we again apply Vandermonde's identity in the last step.

Each such binary string occurs with probability $p^m(1-p)^{n-m}$. Hence

$$\begin{aligned}\mathop{\mathbb{P}}_{S\sim\mathcal{D}^n}\Big(Y_{i_j(S)\pm t}=1\cap j\le M-k\Big) &= \sum_{m=j+k}^{n}\binom{n-1}{m-1}p^m(1-p)^{n-m}\\ &= p\sum_{m'=j+k-1}^{n-1}\binom{n-1}{m-1}p^{m'}(1-p)^{n-1-m'}\\ &= pB_{n-1,p,j+k-1}\,.\end{aligned}$$

Also observe that this probability is independent of $t$ for $j\ge t+1$, as can be seen by symmetry among the indices $1,2,\dots,i_j(S)-1$. ■

We next control the law of $X_{i_j(S)\pm t}$ conditioned on $Y_{i_j(S)\pm t}$.

**Lemma 8** *Consider any $j$ and $t\ge 0$ such that $j\ge t+1$. Then:*

- *The law of $X_{i_j(S)\pm t}$ conditioned on the event $\{Y_{i_j(S)\pm t}=1\,,\ j\le M-k\}$ is that of a random variable distributed according to $X|Y=1$.*
- *The law of $X_{i_j(S)\pm t}$ conditioned on the event $\{Y_{i_j(S)\pm t}=0\,,\ j\le M-k\}$ is that of a random variable distributed according to $X|Y=0$.*

**Proof** The high-level idea is that the event $\{j\le M-k\}$ is only determined by the labels, and given the value of the labels (thus $Y_{i_j(S)\pm t}=y$), the law of $X_{i_j(S)\pm t}$ is determined as $X|Y=y$ since the data points $(X_i,Y_i)$ are i.i.d.. We will prove the $\{Y_{i_j(S)-t}=1\}$ case, the proofs for the other cases being analogous.

To make this formal, we let $\boldsymbol{Y}(S)$ denote the labels $Y_1,\dots,Y_n$ (we write $\boldsymbol{Y}(S)$ to make it clear that this random variable depends on $S$). We also let $\boldsymbol{s}$ denote a fixed binary string in $\{0,1\}^n$, and let $i_j(\boldsymbol{s})$ denote the index of the $j$-th 1 in the binary string $\boldsymbol{s}$. For a binary string $\boldsymbol{s}\in\{0,1\}^n$, we let $|\boldsymbol{s}|$ denote the number of its 1s. Therefore for $\boldsymbol{s}$ such that its $(l-t)$-th entry is 1, there are $|\boldsymbol{s}|-1$ other 1s in the string $\boldsymbol{s}$, and $n-|\boldsymbol{s}|$ 0s. We have

$$\begin{aligned}&\mathbb{P}_{S\sim\mathcal{D}^n}\Big(X_{i_j(S)-t}=x\,|\,Y_{i_j(S)-t}=1\,,\ j\le M-k\Big)\\ &\quad=\frac{\mathbb{P}_{S\sim\mathcal{D}^n}\Big(X_{i_j(S)-t}=x\,,\,Y_{i_j(S)-t}=1\,,\ j\le M-k\Big)}{\mathbb{P}_{S\sim\mathcal{D}^n}\Big(Y_{i_j(S)-t}=1\,,\ j\le M-k\Big)}\\ &\quad=\frac{1}{\mathbb{P}_{S\sim\mathcal{D}^n}\Big(Y_{i_j(S)-t}=1\,,\ j\le M-k\Big)}\Bigg(\sum_{\boldsymbol{s}\in\{0,1\}^n}\mathop{\mathbb{P}}_{S\sim\mathcal{D}^n}\Big(X_{i_j(S)-t}=x\,\Big|\,\boldsymbol{Y}(S)=\boldsymbol{s}\,,\,Y_{i_j(S)-t}=1\,,\ j\le M-k\Big)\\ &\qquad\qquad\cdot\mathop{\mathbb{P}}_{S\sim\mathcal{D}^n}\Big(\boldsymbol{Y}(S)=\boldsymbol{s}\,,\,Y_{i_j(S)-t}=1\,,\ j\le M-k\Big)\Bigg)\\ &\quad=\frac{1}{\mathbb{P}_{S\sim\mathcal{D}^n}\Big(Y_{i_j(S)-t}=1\,,\ j\le M-k\Big)}\Bigg(\sum_{l=j}^{n}\sum_{\boldsymbol{s}\,:\,i_j(\boldsymbol{s})=l}\mathop{\mathbb{P}}_{S\sim\mathcal{D}^n}\Big(X_{l-t}=x\,\Big|\,\boldsymbol{Y}(S)=\boldsymbol{s}\,,\,Y_{l-t}=1\,,\ j\le M-k\Big)\\ &\qquad\qquad\cdot\mathop{\mathbb{P}}_{S\sim\mathcal{D}^n}\Big(\boldsymbol{Y}(S)=\boldsymbol{s}\,,\,Y_{l-t}=1\,,\ j\le M-k\Big)\Bigg)\,.\end{aligned}$$

The above uses the fact that we must have $i_j(\boldsymbol{s}) \geq j$, so the only possible indices $l$ for $i_j(\boldsymbol{s})$ are $j, j+1, \ldots, n-1, n$. Notice if $|\boldsymbol{s}| \geq j+k$ then $\{\boldsymbol{Y}(S) = \boldsymbol{s}\,,\, Y_{l-t} = 1\,,\, j \leq M-k\} = \{\boldsymbol{Y}(S) = \boldsymbol{s}\,,\, Y_{l-t} = 1\}$. Else $\mathbb{P}_{S\sim\mathcal{D}^n}\Big(\boldsymbol{Y}(S) = \boldsymbol{s}\,,\, Y_{l-t} = 1\,,\, j \leq M-k\Big) = 0$. Thus

$$\begin{aligned}
&\sum_{\boldsymbol{s}\,:\,i_j(\boldsymbol{s})=l} \mathbb{P}_{S\sim\mathcal{D}^n}\Big(X_{l-t} = x \,\Big|\, \boldsymbol{Y}(S) = \boldsymbol{s}\,,\, Y_{l-t} = 1\,,\, j \leq M-k\Big)\, \mathbb{P}_{S\sim\mathcal{D}^n}\Big(\boldsymbol{Y}(S) = \boldsymbol{s}\,,\, Y_{l-t} = 1\,,\, j \leq M-k\Big) \\
&\quad = \sum_{\boldsymbol{s}\,:\,i_j(\boldsymbol{s})=l, |\boldsymbol{s}|\geq j+k} \mathbb{P}_{S\sim\mathcal{D}^n}\Big(X_{l-t} = x \,\Big|\, \boldsymbol{Y}(S) = \boldsymbol{s}\,,\, Y_{l-t} = 1\Big)\, \mathbb{P}_{S\sim\mathcal{D}^n}\Big(\boldsymbol{Y}(S) = \boldsymbol{s}\,,\, Y_{l-t} = 1\Big)
\end{aligned}$$

Then since $S$ is i.i.d., we have

$$\begin{aligned}
&\mathbb{P}_{S\sim\mathcal{D}^n}\Big(X_{l-t} = x \,\Big|\, \boldsymbol{Y}(S) = \boldsymbol{s}\,,\, Y_{l-t} = 1\Big) \\
&\qquad = \frac{\mathbb{P}_{S\sim\mathcal{D}^n}\Big(\boldsymbol{Y}(S) = \boldsymbol{s}\,,\, Y_{l-t} = 1 \,\Big|\, X_{l-t} = x\Big)\, \mathbb{P}_{S\sim\mathcal{D}^n}\Big(X_{l-t} = x\Big)}{\mathbb{P}_{S\sim\mathcal{D}^n}\Big(\boldsymbol{Y}(S) = \boldsymbol{s}\,,\, Y_{l-t} = 1\Big)} \\
&\qquad = \frac{\mathbb{P}(Y = 1 \,|\, X = x) \cdot p^{|\boldsymbol{s}|-1} \cdot (1-p)^{n-|\boldsymbol{s}|} \cdot \mathbb{P}(X = x)}{p \cdot p^{|\boldsymbol{s}|-1} \cdot (1-p)^{n-|\boldsymbol{s}|}} \\
&\qquad = \mathbb{P}_1(X = x)\,,
\end{aligned}$$

where the last step follows from Bayes' Rule.

Combining this with the earlier display, and using our earlier observations, thus gives

$$\begin{aligned}
&\mathbb{E}_{S\sim\mathcal{D}^n}\Big[\mathbb{1}\{X_{i_j(S)-t} = x\} \,|\, y_{i_j(S)-t} = 1\Big] \\
&\quad = \mathbb{P}_{S\sim\mathcal{D}^n}\Big(X_{i_j(S)-t} = x \,|\, Y_{i_j(S)-t} = 1\Big) \\
&\quad = \frac{1}{\mathbb{P}_{S\sim\mathcal{D}^n}\Big(Y_{i_j(S)-t} = 1\,,\, j \leq M-k\Big)} \Bigg(\sum_{l=j}^{n} \sum_{\boldsymbol{s}\,:\,i_j(\boldsymbol{s})=l, |\boldsymbol{s}|\geq j+k} \mathbb{P}_{S\sim\mathcal{D}^n}\Big(X_{l-t} = x \,\Big|\, \boldsymbol{Y}(S) = \boldsymbol{s}\,,\, Y_{l-t} = 1\Big) \\
&\qquad\qquad \cdot \mathbb{P}_{S\sim\mathcal{D}^n}\Big(\boldsymbol{Y}(S) = \boldsymbol{s}\,,\, Y_{l-t} = 1\Big)\Bigg) \\
&\quad = \frac{1}{\mathbb{P}_{S\sim\mathcal{D}^n}\Big(Y_{i_j(S)-t} = 1\,,\, j \leq M-k\Big)} \cdot \mathbb{P}_1(X = x) \cdot \sum_{l=j}^{n} \sum_{\boldsymbol{s}\,:\,i_j(\boldsymbol{s})=l, |\boldsymbol{s}|\geq j+k} \mathbb{P}_{S\sim\mathcal{D}^n}\Big(\boldsymbol{Y}(S) = \boldsymbol{s}\,,\, Y_{l-t} = 1\Big) \\
&\quad = \frac{1}{\mathbb{P}_{S\sim\mathcal{D}^n}\Big(Y_{i_j(S)-t} = 1\,,\, j \leq M-k\Big)} \cdot \mathbb{P}_1(X = x) \cdot \sum_{l=j}^{n} \sum_{\boldsymbol{s}\,:\,i_j(\boldsymbol{s})=l} \mathbb{P}_{S\sim\mathcal{D}^n}\Big(\boldsymbol{Y}(S) = \boldsymbol{s}\,,\, Y_{l-t} = 1\,,\, j \leq M-k\Big) \\
&\quad = \mathbb{P}_1(X = x) \cdot \frac{\mathbb{P}_{S\sim\mathcal{D}^n}\Big(Y_{i_j(S)-t} = 1\,,\, j \leq M-k\Big)}{\mathbb{P}_{S\sim\mathcal{D}^n}\Big(Y_{i_j(S)-t} = 1\,,\, j \leq M-k\Big)} \\
&\quad = \mathbb{P}_1(X = x)\,.
\end{aligned}$$

This proves that $X_{i_j(S)-t}$ is distributed according to $X|Y = 1$. The proof for $X_{i_j(S)+t}$ and the $\{Y_{i_j(S)\pm t} = 0\}$ case is analogous. ■

Using Lemma 7, Lemma 8, and conditioning on $Y_{i_j(S)\pm t}$, we obtain a formula for the law of $X_{i_j(S)\pm t}$.

**Lemma 9** *For any $t \geq 0$ with $t+1 \leq j$, we have*

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^n}\Big(X_{i_j(S)\pm t} = x\,, j \leq M-k\Big) = \mathbb{P}_1(X=x)\cdot p\, B_{n-1,p,j+k-1} + \mathbb{P}_0(X=x)\cdot\Big(1 - p\, B_{n-1,p,j+k-1}\Big)\,.$$

**Proof** We can rewrite

$$\begin{aligned}
&\mathop{\mathbb{P}}_{S\sim\mathcal{D}^n}\Big(X_{i_j(S)\pm t} = x\,, j \leq M-k\Big) \\
&\quad= \mathop{\mathbb{P}}_{S\sim\mathcal{D}^n}\Big(X_{i_j(S)\pm t} = x\,, j \leq M-k \,|\, Y_{i_j(S)\pm t} = 1\,, j \leq M-k\Big)\mathop{\mathbb{P}}_{S\sim\mathcal{D}^n}\Big(Y_{i_j(S)\pm t} = 1\,, j \leq M-k\Big) \\
&\qquad+ \mathop{\mathbb{P}}_{S\sim\mathcal{D}^n}\Big(X_{i_j(S)\pm t} = x\,, j \leq M-k \,|\, Y_{i_j(S)\pm t} = 0\,, j \leq M-k\Big)\mathop{\mathbb{P}}_{S\sim\mathcal{D}^n}\Big(Y_{i_j(S)\pm t} = 0\,, j \leq M-k\Big).
\end{aligned}$$

The result now follows by noting that $\mathbb{P}(A\,, E \,|\, B\,, E) = \mathbb{P}(A \,|\, B\,, E)$ for measurable events $A, B, E$, and then combining with Lemma 7 and Lemma 8. ■

We now use the above to study the law of the $i$-th element of $A_j$, and prove Theorem 1.

**Proof** (of Theorem 1) First note that by the assumed interval structure, we have for all $1 \leq r \leq k$ that with probability $\pi[r]$,

$$A_j[i] = \begin{cases} X_{i_j(S)+i-r} & \text{for} \quad 1 \leq i_j(S)+i-r \leq n \\ X_1 & \text{for} \quad i_j(S)+i-r < 1 \\ X_n & \text{for} \quad i_j(S)+i-r > n\,. \end{cases}$$

For any integer $i'$, we let $i'_{\text{truncate}} = 1$ if $i' < 1$, $n$ if $i' > n$, and $i'$ if $1 \leq i' \leq n$. We thus obtain

$$\mathbb{E}_\pi\Big[\mathbb{1}\{A_j[i] = x\} \,|\, S\Big] = \sum_{r=1}^{k} \pi[r]\,\mathbb{1}\Big\{X_{(i_j(S)+i-r)_{\text{truncate}}} = x\Big\}\,.$$

Note that when $k \leq j \leq M-k \leq n-k$, as $i-r \in \{-(k-1), \ldots, k-1\}$, we have $1 \leq i_j(S)+i-r \in \{1,\ldots,n\}$. Since the event $k \leq j \leq M-k$ is independent of $\pi$, only depending on $S$, we obtain

$$\begin{aligned}
\mathbb{E}_\pi\Big[\mathbb{1}\{A_j[i] = x\}\cdot\mathbb{1}\{j \leq M-k\} \,|\, S\Big] &= \mathbb{E}_\pi\Big[\mathbb{1}\{A_j[i] = x\} \,|\, S\Big]\cdot\mathbb{1}\{j \leq M-k\} \\
&= \Big(\sum_{r=1}^{k} \pi[r]\,\mathbb{1}\Big\{X_{(i_j(S)+i-r)_{\text{truncate}}} = x\Big\}\Big)\cdot\mathbb{1}\{j \leq M-k\} \\
&= \sum_{r=1}^{k} \pi[r]\mathbb{1}\Big\{X_{(i_j(S)+i-r)_{\text{truncate}}} = x\Big\}\cdot\mathbb{1}\{j \leq M-k\} \\
&= \sum_{r=1}^{k} \pi[r]\mathbb{1}\big\{X_{i_j(S)+i-r} = x\big\}\cdot\mathbb{1}\{j \leq M-k\}\,.
\end{aligned}$$

Here the last equality follows as when $j \le M-k$, $(i_j(S)+i-r)_{\text{truncate}} = i_j(S)+i-r$.

Since $\pi$ does not depend on $S$, unfreezing over $S$ now gives

$$\mathbb{E}_\mu\Big[\mathbb{1}\{A_j[i]=x\}\cdot\mathbb{1}\{j\le M-k\}\Big] = \pi[i]\,\mathop{\mathbb{E}}_{S\sim\mathcal{D}^n}\Big[\mathbb{1}\big\{X_{i_j(S)}=x\big\}\cdot\mathbb{1}\{j\le M-k\}\Big] + \sum_{r\neq i}\pi[r]\,\mathop{\mathbb{E}}_{S\sim\mathcal{D}^n}\Big[\mathbb{1}\{X_{i_j(S)+i-r}=x\}\,\cdot\,\mathbb{1}\{j\le M-k\}\Big].$$

The sum is now split into two parts: $r=i$, $r\neq i$. The high-level idea to understand the law $\mathbb{P}_{S\sim\mathcal{D}^n}\Big(X_{i_j(S)+i-r}=x\,,j\le M-k\Big)$ in each of these two cases.

**Case 1, $r=i$ terms:** We rewrite

$$\begin{aligned}\mathop{\mathbb{E}}_{S\sim\mathcal{D}^n}\Big[\mathbb{1}\big\{X_{i_j(S)}=x\big\}\cdot\mathbb{1}\{j\le M-k\}\Big] &= \mathop{\mathbb{P}}_{S\sim\mathcal{D}^n}\Big(X_{i_j(S)}=x\,,j\le M-k\Big)\\ &= \mathop{\mathbb{P}}_{S\sim\mathcal{D}^n}\Big(X_{i_j(S)}=x\,\Big|\,j\le M-k\Big)\mathop{\mathbb{P}}_{S\sim\mathcal{D}^n}\Big(j\le M-k\Big).\end{aligned}$$

Since $Y_{i_j(S)}=1$ is guaranteed to hold, and the event $\{j\le M-k\}$ is measurable based on only the labels $Y_1,\dots,Y_n$, it follows that

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^n}\Big(X_{i_j(S)}=x\,\Big|\,j\le M-k\Big) = \mathbb{P}_1(X=x)\,.$$

Thus the contribution of this term is

$$\mathbb{P}_1(X=x)\mathop{\mathbb{P}}_{S\sim\mathcal{D}^n}\Big(j\le M-k\Big) = \mathbb{P}_1(X=x)\cdot B_{n,p,j+k}\,.$$

**Case 2, $r\neq i$ terms:** Fix any $r\neq i$. By Lemma 9, as $j\ge k\ge t+1$, we obtain

$$\begin{aligned}&\mathop{\mathbb{E}}_{S\sim\mathcal{D}^n}\Big[\mathbb{1}\{X_{i_j(S)+i-r}=x\}\cdot\mathbb{1}\{j\le M-k\}\Big]\\ &\quad= \mathop{\mathbb{P}}_{S\sim\mathcal{D}^n}\Big(X_{i_j(S)+i-r}=x\,,j\le M-k\Big)\\ &\quad= \mathbb{P}_1(X=x)\cdot pB_{n-1,p,j+k-1} + \mathbb{P}_0(X=x)\Big(1-pB_{n-1,p,j+k-1}\Big)\,.\end{aligned}$$

By Bayes' Rule, we have

$$\mathbb{P}_1(X=x) = \frac{\mathbb{P}(Y=1\,|\,X=x)\,\mathbb{P}(X=x)}{\mathbb{P}(Y=1)} = \frac{\eta(x)\;\mathbb{P}(X=x)}{p}\,,$$

where $p=\mathbb{P}(Y=1)$, $\eta(x)=\mathbb{P}(Y=1\,|\,X=x)$ and, analogously,

$$\mathbb{P}_0(X=x) = \frac{\big(1-\eta(x)\big)\;\mathbb{P}(X=x)}{1-p}\,.$$

Thus

$$\begin{aligned}
&\mathop{\mathbb{E}}_{S\sim\mathcal{D}^n}\Big[\mathbb{1}\{X_{i_j(S)+i-r}=x\}\cdot\mathbb{1}\{j\le M-k\}\Big]\\
&\quad=\frac{\eta(x)}{p}\,\mathbb{P}(X=x)\,pB_{n-1,p,j+k-1}\\
&\qquad\qquad+\Big(\frac{1}{1-p}\,\mathbb{P}(X=x)-\frac{\eta(x)}{1-p}\,\mathbb{P}(X=x)\Big)\Big(1-pB_{n-1,p,j+k-1}\Big)\\
&\quad=\left(\eta(x)\cdot\Big(B_{n-1,p,j+k-1}-\frac{1-pB_{n-1,p,j+k-1}}{1-p}\Big)+\frac{1-pB_{n-1,p,j+k-1}}{1-p}\right)\mathbb{P}(X=x)\,.
\end{aligned}$$

Note this expression is independent of $i$.

**Putting it all together:** Combining our work in the above cases yields

$$\begin{aligned}
&\mathbb{E}_\mu\Big[\mathbb{1}\{A_j[i]=x\}\cdot\mathbb{1}\{j\le M-k\}\Big]\\
&\quad=\pi[i]B_{n,p,j+k}\cdot\mathbb{P}_1(X=x)\\
&\qquad+\sum_{r\neq i}\pi[r]\left(\eta(x)\cdot\Big(B_{n-1,p,j+k-1}-\frac{1-pB_{n-1,p,j+k-1}}{1-p}\Big)\right.\\
&\qquad\qquad\left.+\frac{1-pB_{n-1,p,j+k-1}}{1-p}\right)\mathbb{P}(X=x)\\
&\quad=\mathbb{P}(X=x)\Big(\eta(x)\,\beta_1(j,i)+\beta_0(j,i)\Big)\,.
\end{aligned}\tag{8}$$

From here, for an arbitrary (measurable) function $g:\mathcal{X}\to\mathbb{R}$, we obtain

$$\begin{aligned}
\mathbb{E}_\mu\Big[g(A_j[i])\cdot\mathbb{1}\{j\le M-k\}\Big]&=\int g(x)\;\mathbb{P}(X=x)\;\Big(\eta(x)\,\beta_1(j,i)+\beta_0(j,i)\Big)dx\\
&=\beta_1(j,i)\int\mathbb{P}(X=x)\,g(x)\,\eta(x)\,dx+\beta_0(j,i)\;\mathbb{E}\Big[g(X)\Big]\,.
\end{aligned}$$

Now observe that

$$\begin{aligned}
\mathbb{E}\Big[Yg(X)\Big]&=\mathbb{E}\Big[\mathbb{E}\Big[Yg(X)\,|\,X\Big]\Big]\\
&=\mathbb{E}\Big[g(X)\,\mathbb{E}\Big[Y\,|\,X\Big]\Big]\\
&=\mathbb{E}\Big[g(X)\eta(X)\Big]=\int\mathbb{P}(X=x)\,g(x)\,\eta(x)\,dx\,.
\end{aligned}$$

Thus

$$\mathbb{E}_\mu\Big[g(A_j[i])\cdot\mathbb{1}\{j\le M-k\}\Big]=\beta_1(j,i)\;\mathbb{E}\Big[Yg(X)\Big]+\beta_0(j,i)\;\mathbb{E}\Big[g(X)\Big]\,,\tag{9}$$

yielding

$$\mathbb{E}\Big[Yg(X)\Big]=\frac{1}{\beta_1(j,i)}\mathbb{E}_\mu\Big[g(A_j[i])\cdot\mathbb{1}\{j\le M-k\}\Big]-\frac{\beta_0(j,i)}{\beta_1(j,i)}\,\mathbb{E}\Big[g(X)\Big]\,.$$

Hence letting $g = f_2 \circ h$, we obtain

$$\begin{aligned}\mathbb{E}\Big[\ell(h(X),Y)\Big] &= \mathbb{E}\Big[f_1(h(X))\Big] + \mathbb{E}\Big[Y f_2(h(X))\Big] \\ &= \mathbb{E}\Big[f_1(h(X))\Big] - \frac{\beta_0(j,i)}{\beta_1(j,i)}\,\mathbb{E}\Big[f_2(h(X))\Big] + \frac{1}{\beta_1(j,i)}\mathbb{E}_\mu\Big[g(A_j[i])\cdot \mathbb{1}\{j \le M-k\}\Big]\,,\end{aligned}$$

as desired. ■

### A.2. Ancillary Results

Recall $B_{n,p,j+k}$ and $B_{n-1,p,j+k-1}$ from the definitions of $\beta_0(j,i)$ and $\beta_1(j,i)$ in Theorem 1. Standard Chernoff bounds guarantee that in regimes of interest, both quantities are very close to 1.

**Lemma 10** *For all $n \ge 2$, $p \in (0,1]$, all $j$ such that $1 \le j \le np/2 - k$, with $k \le np/2 - 1$ the quantities*

$$B_{n,p,j+k}\,, \qquad B_{n-1,p,j+k-1}$$

*are at least*

$$1 - e^{-\Omega(np)}\,,$$

*the big-$\Omega$ notation to be interpreted "as $n$ grows large".*

**Proof** Simply observe that

$$\begin{aligned}B_{n,p,j+k} &= \mathbb{P}(X_{n,p} \ge j+k) \\ B_{n-1,p,j+k-1} &= \mathbb{P}(X_{n-1,p} \ge j+k-1)\,,\end{aligned}$$

where $X_{n,p}$ is a binomial random variable with parameters $n$ and $p$. Then, if $j + k \le np/2$,

$$\begin{aligned}\mathbb{P}(X_{n,p} \ge j+k) &= 1 - \mathbb{P}(X_{n,p} < j+k) \\ &\ge 1 - \mathbb{P}(X_{n,p} \le j+k) \\ &\ge 1 - \mathbb{P}(X_{n,p} \le np/2) \\ &\ge 1 - e^{-np/8}\,.\end{aligned}$$

The same argument, with the same conditions on $j, k, n, p$, holds for $X_{n-1,p}$. This concludes the proof. ■

Before proceeding, we use a consequence of the above Lemma 10 to approximate $\beta_1(j,i)$ and $\beta_0(j,i)$ occurring in the statement of Theorem 1.

**Lemma 11** *For all $n \ge 2$, $p \in (0, 1/2]$, all $j$ such that $1 \le j \le np/2 - k$ with $k \le np/2 - 1$, we have*

$$\left|\beta_1(j,i) - \frac{\pi[i]}{p}\right| = O\left(\frac{e^{-\Omega(np)}}{p}\right)\,, \qquad \Big|\beta_0(j,i) - \big(1 - \pi[i]\big)\Big| = O(pe^{-\Omega(np)})\,,$$

*where $\Omega(\cdot), O(\cdot)$ hides a universal constant. Thus when*

$$np = \Omega\left(\log\left(\max_{i\in[k]} \frac{1}{\pi[i]}\right)\right),$$

*for a suitable large enough constant hidden in the big-omega notation, we have*

$$\beta_1(j,i) \geq \frac{\pi[i]}{2p}\,,\ 0 \leq \beta_0(j,i) \leq 1\,.$$

**Proof** Let $\delta_1 = 1 - B_{n,p,j+k} \geq 0$ and $\delta_2 = 1 - B_{n-1,p,j+k-1} \geq 0$. By Lemma 10, we have $0 \leq \delta_1, \delta_2 \leq e^{-\Omega(np)}$. Therefore, we can rewrite

$$\begin{aligned}\beta_1(j,i) &= \frac{\pi[i]}{p}(1-\delta_1) + \Big(1 - \delta_2 - \frac{1-p(1-\delta_2)}{1-p}\Big)(1-\pi[i]) \\ &= \frac{\pi[i]}{p} - \frac{\pi[i]\delta_1}{p} - \delta_2\Big(1 + \frac{p}{1-p}\Big)(1-\pi[i])\,.\end{aligned}$$

Since $p \leq \frac{1}{2}$, using that $0 \leq \delta_1, \delta_2 \leq e^{-\Omega(np)}$ the upper bound on $\left|\beta_1(j,i) - \frac{\pi[i]}{p}\right|$ follows. Similarly we can rewrite

$$\beta_0(j,i) = \frac{1-p(1-\delta_2)}{1-p}(1-\pi[i]) = 1 - \pi[i] + \frac{p\delta_2}{1-p}(1-\pi[i])\,,$$

and we use $p \leq \frac{1}{2}$ and $0 \leq \delta_2 \leq e^{-\Omega(np)}$. The final conclusion on the bounds $\beta_1(j,i) \geq \frac{\pi[i]}{2p}$, $0 \leq \beta_0(j,i) \leq 2$ is evident given the second condition on $np$. ■

## Appendix B. Proofs for Section 4

This section contains the proofs that apply to the ERM algorithm defined in Section 4.

Introduce the shorthand

$$\widehat{\ell}_M(h) = \widehat{\ell}_M(h, S, \mathcal{A})\,.$$

Then consider where $\widehat{\ell}(h,j)$ is as in (10). We rewrite

$$\widehat{\ell}_M(h) = \frac{1}{\left(\frac{np}{2} - 2k + 1\right)\Sigma} \sum_{j=k}^{M_{\text{UPPER}}} \left(\sum_{i=1}^{k} r(j,i)\, f_2(h(A_j[i]))\right) + C\,,$$

where we have introduced the short-hand notation $r(j,i) = \frac{\pi[i]^2}{\beta_1(j,i)}$, and

$$\begin{aligned}C = -\frac{1}{\left(\frac{np}{2} - 2k + 1\right)\Sigma} &\left(\sum_{j=k}^{M_{\text{UPPER}}} \frac{1}{B_{n,p,j+k}} \sum_{i=1}^{k} \frac{\pi[i]^2\,\beta_0(j,i)}{\beta_1(j,i)}\right) \mathbb{E}[f_2(h(x))] \\ &+ \frac{(M_{\text{UPPER}} - k + 1)\ \mathbb{E}[f_1(h(x))]}{(\frac{np}{2} - 2k + 1)\, B_{n,p,j+k}}\,.\end{aligned}$$

We note that $\widehat{\ell}_M(h, S, \mathcal{A})$ can always be reformulated in terms of the original variables $X_1, \ldots, X_n$, as specified next. Specifically, denote by $j(i) \in [k]$ the position of variable $X_i$ within attribution set $A_j$ if $X_i \in A_j$, and 0 otherwise.[6] Then we have

$$\widehat{\ell}_M(h) = \frac{1}{\left(\frac{np}{2} - 2k + 1\right)\Sigma} \sum_{i=1}^{n} f_2(h(X_i)) \sum_{j\,:\,k \le j \le M_{\text{UPPER}}\,,\, X_i \in A_j} r(j, j(i)) + C\,. \tag{10}$$

Note that in the above interpretation $|\{j \,:\, k \le j \le M_{\text{UPPER}}\,,\, X_i \in A_j\}|$ is the number of occurrences of variable $X_i$ in estimator $\widehat{\ell}_M(h, S, \mathcal{A})$. We denote by $m_{\text{UPPER}}(S)$ the realization of $M_{\text{UPPER}}$ determined by $S$.

The first observation is that for any $j$, $\mathbb{1}\{X_i \in A_j\}$ equals 1 for exactly $k$ distinct $X_i$. Also, recall the $j(i)$ must all be distinct for different $i$. This allows us to prove the following upper bound which is independent of $k$, $M$ and $n$:

**Lemma 12** *Consider* $np = \Omega\left(\log\left(\max_{i\in[k]} \frac{1}{\pi[i]}\right)\right)$. *Then for any realization of* $S$, *any* $j$ *such that* $k \le j \le m_{\text{UPPER}}(S)$, *and any* $h$ *we have*

$$\left| \sum_{i=1}^{n} f_2(h(X_i))\, r(j, j(i))\, \mathbb{1}\{X_i \in A_j\} \right| \le 2pF_2\,. \tag{11}$$

**Proof** Since $\mathbb{1}\{X_i \in A_j\}$ equals 1 for exactly $k$ items $X_i$, irrespective of the realization of $\pi$, the sum

$$\sum_{i=1}^{n} f_2(h(X_i))\, r(j, j(i))\, \mathbb{1}\{X_i \in A_j\}$$

is only over $k$ distinct data items $X_i$. Also, note the $j(i)$ must all be distinct for different $i$. Furthermore Lemma 11 insures that for $j$ such that $k \le j \le m_{\text{UPPER}}(S) \le \frac{np}{2} - k$, we have

$$r(j, i) \le 2p\pi[i]\,.$$

Hence, as the terms $r(j, j(i))\, \mathbb{1}\{X_i \in A_j\}$ are all non-negative, this yields

$$\begin{aligned}
\left| \sum_{i=1}^{n} f_2(h(X_i))\, r(j, j(i))\, \mathbb{1}\{X_i \in A_j\} \right| &\le \sum_{i=1}^{n} \left|f_2(h(X_i))\right| \cdot r(j, j(i))\, \mathbb{1}\{X_i \in A_j\} \\
&\le F_2 \sum_{i'=1}^{k} r(j, j(i')) \\
&\le 2pF_2 \sum_{i'=1}^{k} \pi[i'] = 2pF_2\,,
\end{aligned}$$

as desired. ■

6. This is clearly well-defined because a given $X_i$ is only in one position for a given attribution set.

It is now convenient to set up some notation. By (10), we can write

$$\widehat{\ell}_M(h) = \sum_{i=1}^{n} f_2(h(x_i)) \, \frac{1}{\left(\frac{np}{2} - 2k + 1\right) \Sigma} \sum_{j \,:\, k \le j \le M_{\text{UPPER}} \,,\, X_i \in A_j} r(j, j(i)) + C \,.$$

Thus

$$\mathbb{E}_{\mu|S}[\widehat{\ell}_M(h)] = \sum_{i=1}^{n} f_2(h(x_i)) \, \mathbb{E}_{\mu|S} \left[ \frac{1}{\left(\frac{np}{2} - 2k + 1\right) \Sigma} \sum_{j \,:\, k \le j \le m_{\text{UPPER}}(S) \,,\, X_i \in A_j} r(j, j(i)) \right] + C \,.$$

Moreover by Lemma 1, we have $\mathcal{L}(h) = \mathbb{E}_S \left[ \mathbb{E}_{\mu|S}[\widehat{\ell}] \right]$. Note that $\mathbb{E}_{\mu|S}[\widehat{\ell}_M(h)]$ is a function of the i.i.d. random variables $\{Z_i\}_{i=1}^n$ where each $Z_i = (X_i, Y_i)$. Thus we can define the function of $S = \langle z_1, \ldots, z_n \rangle = \langle (x_1, y_1)), \ldots, (x_n, y_n) \rangle$

$$f(z_1, \ldots, z_n) = f(z_1, \ldots, z_n; h) := \mathbb{E}_{\mu|S}[\widehat{\ell}_M(h)] \,.$$

We need to upper bound the sensitivity of $f(z_1, \ldots, z_n)$ to each $z_i$. In particular, we now establish the following.

**Lemma 13** *Suppose* $np = \Omega\left(\log\left(\max_{i \in [k]} \frac{1}{\pi[i]}\right)\right)$. *Consider any* $h \in \mathcal{H}$. *Define* $\mathcal{B}$ *as the set of* $\langle z_1, \ldots, z_n \rangle$ *such that* $\frac{np}{2} \le M$. *Consider any* $z = \langle z_1, \ldots, z_{i-1}, z_i, z_{i+1}, \ldots z_n \rangle$, $z' = \langle z_1, \ldots, z_{i-1}, z'_i, z_{i+1}, \ldots z_n \rangle$, *where* $z$ *and* $z'$ *only differ in their* $i$*-th coordinate. Then for all* $z, z' \in \mathcal{B}$ *we have*

$$\left| f(z) - f(z') \right| \le \frac{12 p F_2}{(\frac{np}{2} - 2k + 1)\Sigma} \,.$$

*In particular, if* $k \le \frac{np}{8}$, *then for all* $z, z' \in \mathcal{B}$ *we have*

$$\left| f(z) - f(z') \right| \le \frac{48 F_2}{n\Sigma} \,.$$

**Proof** Recall

$$\begin{aligned} f(Z_1, \ldots, Z_n) &= \mathbb{E}_{\mu|S}\left[\widehat{\ell}_M(h)\right] \\ &= \frac{1}{\left(\frac{np}{2} - 2k + 1\right) \Sigma} \sum_{i=1}^{n} f_2(h(X_i)) \, \mathbb{E}_{\mu|S} \left[ \sum_{j \,:\, k \le j \le M_{\text{UPPER}} \,,\, X_i \in A_j} r(j, j(i)) \right] + C \,, \end{aligned}$$

is a function of i.i.d. random variables $Z_i = (X_i, Y_i)$, $i = 1, \ldots, n$. Consider any $S$. For fixed $i \in [n]$ and $h$, we consider the sensitivity of $f$ as we change $Z_i = z_i = (x_i, y_i)$ to $Z'_i = z'_i = (x'_i, y'_i)$.

First, since $z, z' \in \mathcal{B}$, we have $m_{\text{UPPER}}(S) = \frac{np}{2} - k$, so that $C$ turns to the constant

$$C = -\frac{1}{(\frac{np}{2} - 2k + 1)\Sigma} \left( \sum_{j=k}^{m_{\text{UPPER}}(S)} \frac{1}{B_{n,p,j+k}} \sum_{i=1}^{k} \frac{\pi[i]^2 \, \beta_0(j,i)}{\beta_1(j,i)} \right) \mathbb{E}[f_2(h(x))] + \frac{\mathbb{E}[f_1(h(x))]}{B_{n,p,j+k}} \,,$$

thereby not contributing any sensitivity.

Consider any $j$ such that $k \le j \le m_{\text{UPPER}}(S)$. Let us focus on random variable $j(i)$ in the conditional measure $\mu|S$ where $S$ is given. First, the number of distinct attribution sets that may contain $x_i$ is upper bounded by $2k-1$. This is because there are at most $2k-1$ conversions at distance $\le k-1$ from $x_i$, corresponding to the labels $y_{i-(k-1)}, y_{i-(k-2)}, \ldots, y_{i-1}, y_i, y_{i+1}, \ldots, y_{i+k-2}, y_{i+k-1}$ (if the indices are in bounds), which may or may not be one. Now, suppose $y_{i+b} = 1$ is the $j$th conversion, for some $0 \le b \le k-1$. Then

$$j(i) = \begin{cases} k-b & \text{with prob. } \pi[k] \\ k-b-1 & \text{with prob. } \pi[k-1] \\ \vdots & \vdots \\ 1 & \text{with prob. } \pi[b+1] \\ 0 & \text{with remaining prob.} \end{cases}$$

so that

$$\mathbb{E}_{\mu|S}\Big[r(j, j(i))\Big] = \sum_{i'=b+1}^{k} \pi[i']\, r(j, i'-b)\,.$$

Similarly, suppose $y_{i-b} = 1$ is the $j$th conversion, for some $1 \le b \le k-1$. Then

$$j(i) = \begin{cases} b+1 & \text{with prob. } \pi[1] \\ b+2 & \text{with prob. } \pi[2] \\ \vdots & \vdots \\ k & \text{with prob. } \pi[k-b] \\ 0 & \text{with remaining prob.} \end{cases}$$

with expectation

$$\mathbb{E}_{\mu|S}\Big[r(j, j(i)\Big] = \sum_{i'=1}^{k-b} \pi[i']\, r(j, i'+b)\,.$$

Note these computations only apply for $j$ such that $k \le j \le m_{\text{UPPER}}(S)$.

Denote by

$$0 \le b^+(1) < b^+(2) < \ldots < b^+(r^+)\,, \qquad 0 > -b^-(1) > -b^-(2) > \ldots > -b^-(r^-)$$

the positive and negative offsets of the conversions at distance at most $k-1$ from $i$ within $S$, whose corresponding attribution set's index $j$ is such that $k \le j \le m_{\text{UPPER}}(S)$. Hence both the $b^+(t)$ and the $b^-(t)$ are non-negative. Let the index $j$ of the conversion corresponding to each $b^+(t)$ be $j_{b^+(t)}$ for all $1 \le t \le r^+$, and similarly define $j_{b^-(t)}$ for all $1 \le t \le r^-$. Thus we have for all $1 \le t \le r^+$ that $k \le j_{b^+(t)} \le m_{\text{UPPER}}(S)$, and similarly for all $1 \le t \le r^-$ that $k \le j_{b^-(t)} \le m_{\text{UPPER}}(S)$. Hence

we can write

$$\mathbb{E}_{\mu|S}\left[\sum_{j\,:\,k\le j\le m_{\text{UPPER}}(S)\,,\,x_i\in A_j} r(j,j(i))\right]$$
$$= \sum_{i'=b^+(1)+1}^{k} \pi[i']\, r\Big(j_{b^+(1)},\, i'-b^+(1)\Big) + \ldots + \sum_{i'=b^+(r^+)+1}^{k} \pi[i']\, r\Big(j_{b^+(r^+)},\, i'-b^+(r^+)\Big)$$
$$+ \sum_{i'=1}^{k-b^-(1)} \pi[i']\, r\Big(j_{b^-(1)},\, i'+b^-(1)\Big) + \ldots + \sum_{i'=1}^{k-b^-(r^-)} \pi[i']\, r\Big(j_{b^-(r^-)},\, i'+b^-(r^-)\Big)\,.$$

Note the change $x_i \to x_i'$ does not impact the structure of the attribution sets $A_j$. Thus, the contribution from the change $x_i \to x_i'$ is upper bounded by

$$\frac{2F_2}{(\frac{np}{2}-2k+1)\,\Sigma}\mathbb{E}_{\mu|S}\left[\sum_{j\,:\,k\le j\le m_{\text{UPPER}}(S)\,,\,x_i\in A_j} r(j,j(i))\right].$$

Recall that for all $1 \le t \le r^+$ we have that $k \le j_{b^+(t)} \le m_{\text{UPPER}}(S)$, and similarly for all $1 \le t \le r^-$ we have that $k \le j_{b^-(t)} \le m_{\text{UPPER}}(S)$. By our condition on $n$ and as $k \le j_{b^+(t)}, j_{b^-(t)} \le m_{\text{UPPER}}(S) \le \frac{np}{2} - k$, Lemma 11 gives $\beta_1(j_{b^+(t)}, i'-b^+(t)) \ge \frac{\pi[i'-b^+(t)]}{2p}$ and hence $r(j_{b^+(t)}, i'-b^+(t)) \le 2p\cdot\pi[i'-b^+(t)]$. Similarly we have $r(j_{b^-(t)}, i'+b^-(t)) \le 2p\cdot\pi[i'+b^-(t)]$. Therefore as the entries of $\pi$ are all non-negative, we have

$$\begin{aligned}
&\mathbb{E}_{\mu|S}\left[\sum_{j\,:\,k\le j\le m_{\text{UPPER}}(S)\,,\,x_i\in A_j} r(j,j(i))\right] \\
&= \sum_{t=1}^{r^+}\sum_{i'=b^+(t)+1}^{k} \pi[i']\, r\Big(j_{b^+(t)},\, i'-b^+(t)\Big) + \sum_{t=1}^{r^-}\sum_{i'=1}^{k-b^-(t)} \pi[i']\, r\Big(j_{b^-(t)},\, i'+b^-(t)\Big) \\
&\le 2p\sum_{t=1}^{r^+}\sum_{i'=b^+(t)+1}^{k} \pi[i']\,\pi[i'-b^+(t)] + 2p\sum_{t=1}^{r^-}\sum_{i'=1}^{k-b^-(t)} \pi[i']\,\pi[i'+b^-(t)] \\
&\le 2p\sum_{b=0}^{k-1}\sum_{i'=b+1}^{k} \pi[i']\pi[i'-b] + 2p\sum_{b=1}^{k-1}\sum_{i'=1}^{k-b} \pi[i']\pi[i'+b] \\
&\le 2p\sum_{i'=1}^{k}\pi[i']\sum_{b=1}^{k}\pi[b] + 2p\sum_{i'=1}^{k-1}\pi[i]\sum_{b=1}^{k}\pi[b] \\
&= 2p\sum_{i'=1}^{k}\pi[i'] + 2p\sum_{i'=1}^{k-1}\pi[i'] \\
&\le 4p\,. \qquad (12)
\end{aligned}$$

On the other hand, the contribution from the change $y_i \to y_i'$ amounts to either adding (if $y_i = 0$ and $y_i' = 1$) or subtracting (if $y_i = 1$ and $y_i' = 0$) an attribution set $A_j$ (the one associated with

$y_i$). Therefore this contribution amounts to, respectively, adding or subtracting a term $j$ from the sum $\sum_{j\,:\,k\le j\le m_{\text{UPPER}}(S)\,,\,x_i\in A_j}$ within $\mathbb{E}_{\mu|S}\left[\sum_{j\,:\,k\le j\le m_{\text{UPPER}}(S)\,,\,x_i\in A_j} r(j,j(i))\right]$ for each data index $i$. Moreover, this does not change the other attribution sets, which are generated independently for each conversion. Note if $y_i' = y_i$ then we do not need to take this (non-negative) change into account, so the following bound will hold irrespective of whether $y_i' = y_i$ or $y_i' \neq y_i$.

The idea is that even for this new $j$, as $k \le j \le m_{\text{UPPER}}(S)$, by Lemma 12 we have

$$\begin{aligned}\Big|\sum_{i=1}^{n} f_2(h(x_i))\, r(j,j(i))\, \mathbb{1}\{x_i\in A_j\}\Big| &\le F_2\,\Big|\sum_{i=1}^{n} r(j,j(i))\,\mathbb{1}\{x_i\in A_j\}\,\Big| \\ &\le F_2\,\Big|\sum_{b=1}^{k} r(j,b)\,\Big| \\ &\le 2p\,F_2\,,\end{aligned}$$

allowing us to control the sensitivity.

To see this, denote this new $j$ by $j_0$ and suppose that $y_i = 0, y_i' = 1$ (thus the term corresponding to $j_0$ is *added*). Since we are considering $j$ in the range $[k, m_{\text{UPPER}}(S)]$, we have $k \le j_0 \le m_{\text{UPPER}}(S)$. A very similar argument holds in the case where $y_i = 1, y_i' = 0$ (in which case the term corresponding to $j_0$ is *subtracted*). The above derivations allow us to conclude that the sensitivity of $f$ when turning $z_i$ into $z_i'$ is upper bounded by

$$\begin{aligned}&\Bigg|\frac{1}{(\frac{np}{2}-2k+1)\,\Sigma}\sum_{s=1\,s\neq i}^{n} f_2(h(x_s))\left(\mathbb{E}_{\mu|S}\left[\sum_{j\,:\,k\le j\le m_{\text{UPPER}}(S)\,,\,x_s\in A_j} r(j,j(s))\right]+\mathbb{E}_{\mu|S}\Big[\mathbb{1}\{x_s\in A_{j_0}\}r(j_0,j_0(s))\Big]\right)\\
&\quad+\frac{1}{(\frac{np}{2}-2k+1)\Sigma}\,f_2(h(x_i'))\left(\mathbb{E}_{\mu|S}\left[\sum_{j\,:\,k\le j\le m_{\text{UPPER}}(S)\,,\,x_i\in A_j} r(j,j(i))\right]+\mathbb{E}_{\mu|S}\Big[\mathbb{1}\{x_i'\in A_{j_0}\}r(j_0,j_0(i))\Big]\right)\\
&\quad-\frac{1}{(\frac{np}{2}-2k+1)\,\Sigma}\sum_{s=1\,s\neq i}^{n} f_2(h(x_s))\,\mathbb{E}_{\mu|S}\left[\sum_{j\,:\,k\le j\le m_{\text{UPPER}}(S)\,,\,x_s\in A_j} r(j,j(s))\right]\\
&\quad-\frac{1}{(\frac{np}{2}-2k+1)\,\Sigma}\,f_2(h(x_i))\,\mathbb{E}_{\mu|S}\left[\sum_{j\,:\,k\le j\le m_{\text{UPPER}}(S)\,,\,x_i\in A_j} r(j,j(i))\right]\Bigg|\\
=&\Bigg|\frac{1}{(\frac{np}{2}-2k+1)\,\Sigma}\sum_{s=1\,s\neq i}^{n} f_2(h(x_s))\,\mathbb{E}_{\mu|S}\Big[\mathbb{1}\{x_s\in A_{j_0}\}r(j_0,j_0(s))\Big]\\
&\quad+\frac{1}{(\frac{np}{2}-2k+1)\,\Sigma}\,f_2(h(x_i'))\left(\mathbb{E}_{\mu|S}\left[\sum_{j\,:\,k\le j\le m_{\text{UPPER}}(S)\,,\,x_i\in A_j} r(j,j(i))\right]+\mathbb{E}_{\mu|S}\Big[\mathbb{1}\{x_i\in A_{j_0}\}r(j_0,j_0(i))\Big]\right)\\
&\quad-\frac{1}{(\frac{np}{2}-2k+1)\,\Sigma}\,f_2(h(x_i))\,\mathbb{E}_{\mu|S}\left[\sum_{j\,:\,k\le j\le m_{\text{UPPER}}(S)\,,\,x_i\in A_j} r(j,j(i))\right]\Bigg|.\end{aligned}$$

In turn the above is upper bounded by

$$\underbrace{\frac{1}{(\frac{np}{2}-2k+1)\,\Sigma}\left|\sum_{s=1\,s\neq i}^{n} f_2(h(x_s))\,\mathbb{E}_{\mu|S}\Big[\mathbb{1}\{x_s\in A_{j_0}\}r(j_0,j_0(s))\Big]\right|}_{\text{(I)}}$$
$$+\underbrace{\frac{1}{(\frac{np}{2}-2k+1)\,\Sigma}\Big|f_2(h(x_i'))-f_2(h(x_i))\Big|\cdot\left|\mathbb{E}_{\mu|S}\left[\sum_{j\,:\,k\le j\le m_{\text{UPPER}}(S)\,,\,x_i\in A_j} r(j,j(i))\right]\right|}_{\text{(II)}}$$
$$+\underbrace{\frac{1}{(\frac{np}{2}-2k+1)\,\Sigma}\left|f_2(h(x_i'))\mathbb{E}_{\mu|S}\Big[\mathbb{1}\{x_i'\in A_{j_0}\}r(j_0,j_0(i))\Big]\right|}_{\text{(III)}}.$$

Now, note as the terms $\mathbb{1}\{x_s \in A_{j_0}\}r(j_0, j_0(s))$ are non-negative, we have by the same rationale as in the proof of Lemma 12 that

$$\begin{aligned}
\text{(I)} &= \frac{1}{(\frac{np}{2}-2k+1)\,\Sigma}\left|\mathbb{E}_{\mu|S}\left[\sum_{s=1\,s\neq i}^{n} f_2(h(x_s))\,\mathbb{1}\{x_s\in A_{j_0}\}r(j_0,j_0(s))\right]\right| \\
&\le \frac{1}{(\frac{np}{2}-2k+1)\,\Sigma}\mathbb{E}_{\mu|S}\left[\sum_{s=1\,s\neq i}^{n}\Big|f_2(h(x_s))\Big|\cdot\mathbb{1}\{x_s\in A_{j_0}\}r(j_0,j_0(s))\right] \\
&\le \frac{F_2}{(\frac{np}{2}-2k+1)\Sigma}\mathbb{E}_{\mu|S}\left[\sum_{s=1}^{n}\mathbb{1}\{x_s\in A_{j_0}\}r(j_0,j_0(s))\right] \\
&\le \frac{2p\,F_2}{(\frac{np}{2}-2k+1)\,\Sigma}\,.
\end{aligned}$$

In particular, this is because $\mathbb{1}\{x_s \in A_{j_0}\} = 1$ for at most $k$ indices $s$, and for these $s$, the indices $j_0(s)$ must all be distinct.

Moreover, because of (12),

$$\text{(II)} \le \frac{8p\,F_2}{(\frac{np}{2}-2k+1)\,\Sigma}\,.$$

As for (III), we leverage again our condition on $n$ and that $k \le j_0 \le m_{\text{UPPER}}(S)$. Combining with Lemma 11 this gives us the upper bound $r(j_0, j_0(i)) \le 2p\pi[i] \le 2p$, so that

$$\text{(III)} \le \frac{2p\,F_2}{(\frac{np}{2}-2k+1)\Sigma}\,.$$

Putting these bounds together, this yields the sensitivity bound

$$\frac{12p\,F_2}{(\frac{np}{2}-2k+1)\,\Sigma}\,,$$

holding for all $S$ and $h \in \mathcal{H}$. This completes the proof. ■

Now observe that the sensitivity bound from Lemma 13 depends on the condition $\frac{np}{2} \leq M$, which involves random variable $M$. Thus we use the following variant of McDiarmid's inequality from Combes (2024):

**Theorem 14 (Theorem 3 and Example 3 of Combes (2024))** *Consider a generic metric space $\mathcal{X}$. Let $g(z_1, \ldots, z_n)$ be a function of $n$ i.i.d. random variables $z_1, \ldots, z_n$ with $z_i \in \mathcal{X}$, that satisfies the following property. For some subset $\mathcal{B} \subseteq \mathcal{X}^n$ and any pair of $z = (z_1, \ldots, z_{i-1}, z_i, z_{i+1}, \ldots z_n)$ and $z' = (z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots z_n)$ that only differ in their $i$-th coordinate with $z, z' \in \mathcal{B}$, we have $|g(z) - g(z')| \leq c_i$. Let $\rho = \mathbb{P}(\mathcal{B}^c)$ and $\overline{c} = \sum_{i=1}^n c_i$. Then*

$$\mathbb{P}\Big(g(z) - \mathbb{E}\big[g(z) \,\big|\, z \in \mathcal{B}\big] \geq t + \rho\overline{c}\Big) \leq \rho + \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right).$$

Before moving forward, we need to introduce a simple lemma which is a direct consequence of the contraction lemma for Rademacher Complexity (e.g., Ledoux and Talagrand (2011)).

**Lemma 15** *For a hypothesis space $\mathcal{H}$ and any fixed sample $z = \{z_1, \ldots, z_n\}$, for any $a_1, \ldots, a_n \in \mathbb{R}$ independent of $\sigma_i$ but potentially depending on the samples $z_i$, letting $\sigma_i$ be i.i.d. Rademacher random variables we have*

$$\mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^n \sigma_i a_i h(z_i)\right| \,\middle|\, z\right] \leq \Big(\max_i |a_i|\Big)\mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^n \sigma_i h(z_i)\right| \,\middle|\, z\right].$$

**Proof** For each $1 \leq i \leq n$, consider the function $\phi_i(x) = a_i x$. Thus $a_i h(z_i) = \phi_i\big(h(z_i)\big)$. Each $\phi_i$ is $|a_i|$ Lipschitz, and in particular $\max_i |a_i|$-Lipschitz. Thus the Lemma follows by the Contraction Lemma of Rademacher Complexity. Note the functions used in the Contraction Lemma may depend on the sample $z$, since the entire proof of the Contraction Lemma is done with the sample $z$ fixed. ■

We are now in position to prove Theorem 2.

**Proof** [Proof of Theorem 2] From standard arguments, and the fact that, from Theorem 1, $\mathbb{E}_\mu[\widehat{\ell}_M(h)] = \mathcal{L}(h)$ we can write, for any $\epsilon \geq 0$,

$$\begin{aligned}\mathbb{1}\Big\{Reg(\widehat{h}) \geq \epsilon\Big\} &\leq \mathbb{1}\left\{2\sup_{h\in\mathcal{H}}\Big|\widehat{\ell}_M(h) - \mathcal{L}(h)\Big| \geq \epsilon\right\} \\ &\leq \mathbb{1}\Big\{\sup_{h\in\mathcal{H}}\Big|\mathbb{E}_{\mu|S}[\widehat{\ell}_M(h)] - \mathcal{L}(h)\Big| \geq \epsilon/4\Big\} + \mathbb{1}\Big\{\sup_{h\in\mathcal{H}}\Big|\widehat{\ell}_M(h) - \mathbb{E}_{\mu|S}[\widehat{\ell}_M(h)]\Big| \geq \epsilon/4\Big\}.\end{aligned}$$

We take expectations w.r.t. $\mu$ and consider in turn

$$\mathbb{E}_\mu\left[\mathbb{1}\Big\{\sup_{h\in\mathcal{H}}\Big|\mathbb{E}_{\mu|S}[\widehat{\ell}_M(h)] - \mathcal{L}(h)\Big| \geq \epsilon/4\Big\}\right] = \mathbb{P}_\mu\left(\sup_{h\in\mathcal{H}}\Big|\mathbb{E}_{\mu|S}[\widehat{\ell}_M(h)] - \mathcal{L}(h)\Big| \geq \epsilon/4\right), \quad (13)$$

and

$$\mathbb{E}_\mu\left[\mathbb{1}\Big\{\sup_{h\in\mathcal{H}}\Big|\widehat{\ell}_M(h) - \mathbb{E}_{\mu|S}[\widehat{\ell}_M(h)]\Big| \geq \epsilon/4\Big\}\right] = \mathbb{P}_\mu\left(\sup_{h\in\mathcal{H}}\Big|\widehat{\ell}_M(h) - \mathbb{E}_{\mu|S}[\widehat{\ell}_M(h)]\Big| \geq \epsilon/4\right). \quad (14)$$

**Controlling (13):** We deal first with (13). Recall that

$$f(z_1,\ldots,z_n) = f(z_1,\ldots,z_n;h) = \mathbb{E}_{\mu|S}\Big[\widehat{\ell}_M(h)\Big]$$
$$= \frac{1}{(\frac{np}{2}-2k+1)\,\Sigma}\sum_{i=1}^{n} f_2(h(x_i))\,\mathbb{E}_{\mu|S}\left[\sum_{j\,:\,k\le j\le m_{\text{UPPER}}(S)\,,\,x_i\in A_j} r(j,j(i))\right] + C\,.$$

The function $f(Z_1,\ldots,Z_n)$ is a function of i.i.d. random variables $Z_i=(X_i,Y_i)$, $i=1,\ldots,n$. Let $Z=(Z_1,\ldots,Z_n)$. We also know that, for each $h\in\mathcal{H}$,

$$\mathbb{E}_Z[f(Z;h)] = \mathcal{L}(h)\,.$$

Focus on the random variable $\gamma_i = \mathbb{E}_{\mu|S}\left[\sum_{j\,:\,k\le j\le m_{\text{UPPER}}(S)\,,\,x_i\in A_j} r(j,j(i))\right]$. A closer inspection reveals that this random variable does depend on the sample $Z=z$ only through the labels $y_1,\ldots,y_n$. In particular, $\gamma_i$ depends on index $i$, but does not directly depend on $x_i$ (note that $A_j$ is a set of *indices*, so the condition "$x_i\in A_j$" within the brackets should be interpreted as "$i\in A_j$, for frozen value of $x_i$"), and certainly it does not depend on $h$. So, let us adopt the notation $y=\langle y_1,\ldots,y_n\rangle$, and

$$\gamma_i(y) = \mathbb{E}_{\mu|S}\left[\sum_{j\,:\,k\le j\le m_{\text{UPPER}}(S)\,,\,x_i\in A_j} r(j,j(i))\right]\,,$$

so that now

$$f(z;h) = \frac{1}{(\frac{np}{2}-2k+1)\,\Sigma}\sum_{i=1}^{n}\gamma_i(y)\,f_2(h(x_i)) + C\,.$$

Let $z^{(i)}$ be the same as $z$, with the $i$-th item replaced by $z_i'=(x_i',y_i')$, and introduce independent Rademacher variables $\sigma=(\sigma_1,\ldots,\sigma_n)$. Define

$$\Phi(z) = \sup_{h\in\mathcal{H}}\Big|\mathcal{L}(h) - f(z;h)\Big|\,.$$

We have

$$\Phi(z)-\Phi(z^{(i)}) \le \sup_{h\in\mathcal{H}}\Big|f(z^{(i)};h)-f(z;h)\Big| \le \frac{12pF_2}{(\frac{np}{2}-2k+1)\Sigma} \le \frac{48F_2}{n\Sigma}\,,$$

the second inequality deriving from Lemma 13.

We apply Theorem 14 with $g(z)=\Phi(z)$, $\mathcal{B}=\Big\{z\,:\,\frac{np}{2}\le M\Big\}$, and $c_i=\frac{48F_2}{n\Sigma}$ for all $1\le i\le n$. By Chernoff's bound, $\mathbb{P}(\mathcal{B}^c)\le e^{-\Omega(np)}$. Theorem 14 now gives

$$\mathbb{P}\Big(\Phi(Z)-\mathbb{E}[\Phi(Z)\,|\,Z\in\mathcal{B}]\ge t+\frac{48F_2}{\Sigma}e^{-\Omega(np)}\Big) \le e^{-\Omega(np)} + \exp\Big(-\frac{2n\Sigma^2}{2304F_2^2}\cdot t^2\Big)\,. \qquad (15)$$

We now control $\mathbb{E}[\Phi(Z)\,|\,Z\in\mathcal{B}]-\mathbb{E}[\Phi(Z)]$. By the Law of Total Expectation,

$$\begin{aligned}
&\big|\,\mathbb{E}[\Phi(Z)\,|\,Z\in\mathcal{B}]-\mathbb{E}[\Phi(Z)]\big| \\
&\quad= \big|\,\mathbb{P}(\mathcal{B})\,\mathbb{E}[\Phi(Z)|Z\in\mathcal{B}] + \mathbb{P}(\mathcal{B}^c)\,\mathbb{E}[\Phi(Z)|Z\in\mathcal{B}^c] - \mathbb{E}[\Phi(Z)|Z\in\mathcal{B}]\big| \\
&\quad= \mathbb{P}(\mathcal{B}^c)\big|\,\mathbb{E}[\Phi(Z)|Z\in\mathcal{B}^c] - \mathbb{E}[\Phi(Z)|Z\in\mathcal{B}]\big| \\
&\quad\le 2\,\mathbb{P}(\mathcal{B}^c)\sup_{z\in Z}|\Phi(z)|\,.
\end{aligned}$$

By definition of $\Phi(z) = \sup_{h\in\mathcal{H}}\Big|\mathcal{L}(h) - f(z;h)\Big|$, since the additive term $+C$ in the definition $\mathcal{L}(h)$ and $f(z;h)$ cancel, we obtain

$$\begin{aligned}|\Phi(z)| &\le 2\sup_{h\in\mathcal{H}}\left|\mathbb{E}_{\mu|S}\left[\frac{1}{(\frac{np}{2}-2k+1)\,\Sigma}\sum_{i=1}^{n}\gamma_i(y)\,f_2(h(x_i))\right]\right| \\ &= 2\sup_{h\in\mathcal{H}}\left|\mathbb{E}_{\mu|S}\left[\sum_{j=k}^{m_{\text{UPPER}}(S)}\frac{1}{\big(\frac{np}{2}-2k+1\big)\Sigma}\sum_{i=1}^{n}f_2(h(x_i))\,r(j,j(i))\,\mathbb{1}\{x_i\in A_j\}\right]\right|,\end{aligned}$$

where the last step follows from swapping the order of summation. Thus Lemma 12 gives $|\Phi(z)| \le \frac{4pF_2}{\Sigma}$. Hence by Chernoff's bound we obtain from the above that

$$\big|\,\mathbb{E}[\Phi(Z)\,|\,Z\in\mathcal{B}] - \mathbb{E}[\Phi(Z)]\big| \le \frac{8pF_2}{\Sigma}e^{-\Omega(np)}\,. \tag{16}$$

Combining (15) and (16) gives

$$\mathbb{P}\Big(\Phi(z) - \mathbb{E}[\Phi(z)] \ge t + \frac{40F_2}{\Sigma}e^{-\Omega(np)}\Big) \le e^{-\Omega(np)} + \exp\left(-\frac{n\Sigma^2}{1152F_2^2}\cdot t^2\right). \tag{17}$$

Now, let $Z' = (Z'_1,\ldots,Z'_n)$ be an independent sample, with $Z'_i = (X'_i, Y'_i)$, $i\in[n]$, and set $Y' = \langle Y'_1,\ldots,Y'_n\rangle$. We can write

$$\begin{aligned}\mathbb{E}[\Phi(Z)] &= \mathbb{E}_Z\left[\sup_{h\in\mathcal{H}}\Big|\mathcal{L}(h) - f(Z;h)\Big|\right] \\ &= \mathbb{E}_Z\left[\sup_{h\in\mathcal{H}}\Big|\mathbb{E}_{Z'}[f(Z';h)] - f(Z;h)\Big|\right] \\ &\le \mathbb{E}_{Z,Z'}\left[\sup_{h\in\mathcal{H}}\Big|f(Z';h) - f(Z;h)\Big|\right] \\ &= \frac{1}{(\frac{np}{2}-2k+1)\,\Sigma}\,\mathbb{E}_{Z,Z'}\left[\sup_{h\in\mathcal{H}}\Big|\sum_{i=1}^{n}\Big(\gamma_i(Y')f_2(h(X'_i)) - \gamma_i(Y)f_2(h(X_i))\Big)\Big|\right] \\ &= \frac{1}{(\frac{np}{2}-2k+1)\,\Sigma}\,\mathbb{E}_{Z,Z',\sigma}\left[\sup_{h\in\mathcal{H}}\Big|\sum_{i=1}^{n}\sigma_i\Big(\gamma_i(Y')f_2(h(X'_i)) - \gamma_i(Y)f_2(h(X_i))\Big)\Big|\right] \\ &\le \frac{2}{(\frac{np}{2}-2k+1)\,\Sigma}\,\mathbb{E}_{Z,\sigma}\left[\sup_{h\in\mathcal{H}}\Big|\sum_{i=1}^{n}\sigma_i\gamma_i(Y)f_2(h(X_i))\Big|\right] \\ &\le \frac{2}{(\frac{np}{2}-2k+1)\,\Sigma}\,\mathbb{E}_Z\left[\mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}}\Big|\sum_{i=1}^{n}\sigma_i\gamma_i(Y)f_2(h(X_i))\Big|\,\Big|\,Z\right]\right].\end{aligned}$$

Now, observe that, by (12) and under the conditions of Lemma 11,

$$\gamma_i(y)\in[0,4p]\,,$$

for all $y$, independent of $h$. Thus by the above bound on $\gamma_i(y)$ and Lemma 15, this gives

$$\mathbb{E}[\Phi(Z)] \le \frac{8p}{(\frac{np}{2}-2k+1)\,\Sigma}\,\mathbb{E}_Z\left[\mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^n \sigma_i f_2(h(X_i))\right|\,\middle|\,Z\right]\right] = \frac{8pn}{(\frac{np}{2}-2k+1)\,\Sigma}\,R_n(f_2\circ\mathcal{H})$$

where

$$R_n(\mathcal{H}) = \mathbb{E}_{X_1,\ldots,X_n}\left[\mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^n \sigma_i \frac{h(X_i)}{n}\right|\,\middle|\,X_1,\ldots,X_n\right]\right]$$

is the (average) Rademacher Complexity of function class $\mathcal{H}$, and

$$f_2\circ\mathcal{H} = \{f_2\circ h \,|\, h\in\mathcal{H}\}\,.$$

Since we assumed $k \le \frac{np}{8}$, the above implies

$$\mathbb{E}[\Phi(Z)] \le \frac{32}{\Sigma}R_n(f_2\circ\mathcal{H})\,.$$

Finally, the $L$-lipschitzness of $f_2(\cdot)$, along with Talagrand's contraction lemma (e.g., Ledoux and Talagrand (2011)), gives

$$\mathbb{E}[\Phi(Z)] \le \frac{32L}{\Sigma}\,R_n(\mathcal{H})\,.$$

We plug back into (17) to obtain, as a consequence,

$$\mathbb{P}\left(\Phi(Z) \ge \underbrace{\frac{32L\,R_n(\mathcal{H})}{\Sigma} + t + \frac{40F_2}{\Sigma}e^{-\Omega(np)}}_{\epsilon/4}\right) \le e^{-\Omega(np)} + \exp\left(-\frac{n\Sigma^2}{1152F_2^2}\cdot t^2\right)\,, \tag{18}$$

the left-hand side being a version of (13) once we set

$$\epsilon/4 \ge \frac{32L\,R_n(\mathcal{H})}{\Sigma} + t + \frac{40F_2}{\Sigma}e^{-\Omega(np)}\,.$$

**Controlling (14):** We now turn to (14). We define the random variables

$$V_j = V(h,A_j) = \frac{1}{\big(\frac{np}{2}-2k+1\big)\Sigma}\Big(\sum_{i=1}^n f_2(h(X_i))\;r(j,j(i))\;\mathbb{1}\{X_i\in A_j\}\Big)\,.$$

Observe that in the conditional space where $S$ is given, the adversary operates on each individual conversion independently, hence the random variables $\mathbb{1}\{X_i\in A_j\}\,\pi[j(i)]$ are always independent of each other. Thus, under the measure $\mu|S$, the $f_2(h(X_i))$ are constants, and the variables $\pi[j(i)]\;\mathbb{1}\{X_i\in A_j\}$ are independent for different $j$. Hence the $V_j$ are independent w.r.t. the measure $\mu|S$. We let $m = m(S)$ is the number of conversions (number of positive labels) in $S$, and let $m_{\text{UPPER}}(S) = \min\{\frac{np}{2}-k, m-k\}$.

Note that by swapping the order of summation we can rewrite $\widehat{\ell}_M(h)$ as

$$\widehat{\ell}_M(h) = \sum_{j=k}^{m_{\text{UPPER}}(S)} V_j + C\,,\qquad \mathbb{E}_{\mu|S}\big[\widehat{\ell}_M(h)\big] = \sum_{j=k}^{m_{\text{UPPER}}(S)} \mathbb{E}_{\mu|S}\big[V_j\big] + C\,.$$

Furthermore, by Lemma 12, for $j$ such that $k \leq j \leq \min\{\frac{np}{2}, m\} - k$, and $S \in \mathcal{B}$, with

$$\mathcal{B} = \left\{ S \,:\, \frac{np}{2} \leq M \leq \frac{3np}{2} \right\}$$

we have

$$|V_j| \leq \frac{2pF_2}{(\frac{np}{2} - 2k + 1)\Sigma} \leq \frac{8F_2}{n\Sigma}\,, \tag{19}$$

where we used the condition $k \leq \frac{np}{8}$. Consider, in the conditional space where $S \in \mathcal{B}$,

$$\mathbb{P}_{\mu|S}\left( \sup_{h\in\mathcal{H}} \left| \widehat{\ell}_M(h) - \mathbb{E}_{\mu|S}[\widehat{\ell}_M(h)] \right| \geq \epsilon/4 \right)\,.$$

In this conditional space, denote by $A = \langle A_k, \ldots, A_{m_{\text{UPPER}}(S)} \rangle$ the (random) attribution sets in command of the adversary, and define

$$\Psi(A) = \sup_{h\in\mathcal{H}} \left| \widehat{\ell}_m(h) - \mathbb{E}_{\mu|S}[\widehat{\ell}_m(h)] \right|.$$

Again, define $A^{(j)}$ to be $A$ with the $j$-th attribution set $A_j$ changed to $A'_j$. From (19) we get

$$\Psi(A) - \Psi(A^{(j)}) \leq \frac{16F_2}{n\Sigma}\,.$$

The standard McDiarmid's inequality yields, for any $t \geq 0$,

$$\mathbb{P}_{\mu|S}\left(\Psi(A) - \mathbb{E}_{\mu|S}[\Psi(A)] \geq t\right) \leq \exp\left( -\frac{2t^2\, n^2\, \Sigma^2}{256\,(m-k+1)\,F_2^2} \right) \leq \exp\left( -\frac{t^2\, n\, \Sigma^2}{192\, p\, F_2^2} \right)\,, \tag{20}$$

the last inequality deriving from $m \leq \frac{3np}{2}$ (which is implied by $S \in \mathcal{B}$), and $k \geq 1$.

Consider, for frozen $S \in \mathcal{B}$, the variables $\{V_j\}_{j=k}^{m_{\text{UPPER}}(S)}$. Note that these variables are independent, but they need not have the same distribution. Let $\{V'_j\}_{j=k}^{m_{\text{UPPER}}(S)}$, with $V'_j = V(h, A'_j)$, $j = k, \ldots, m_{\text{UPPER}}(S)$, be an independent sample conditioned on the same $S$, made up of independent random variables, where $V'_j$ has the same distribution as $V_j$. Denote for brevity by $A$ the collection of random variables $\{A_j\}$, and by $A'$ the collection $\{A'_j\}$.

The standard symmetrization lemma still holds:

$$\begin{aligned}
\mathbb{E}_{\mu|S}[\Psi(A)] = \mathbb{E}_A[\Psi(A)] &= \mathbb{E}_A\left[ \sup_{h\in\mathcal{H}} \left| \sum_{j=k}^{m_{\text{UPPER}}(S)} V(h, A_j) - V(h, A'_j) \right| \right] \\
&= \mathbb{E}_{A,A',\sigma}\left[ \sup_{h\in\mathcal{H}} \left| \sum_{j=k}^{m_{\text{UPPER}}(S)} \sigma_j \Big( V(h, A_j) - V(h, A'_j) \Big) \right| \right] \\
&\leq 2\,\mathbb{E}_{A,\sigma}\left[ \sup_{h\in\mathcal{H}} \left| \sum_{j=k}^{m_{\text{UPPER}}(S)} \sigma_j V(h, A_j) \right| \right]\,.
\end{aligned}$$

Now, focus on the quantity

$$\sup_{h\in\mathcal{H}}\left|\sum_{j=k}^{m_{\text{UPPER}}(S)}\sigma_j V(h,A_j)\right| = \frac{1}{\left(\frac{np}{2}-2k+1\right)\Sigma}\sup_{h\in\mathcal{H}}\left|\sum_{j=k}^{m_{\text{UPPER}}(S)}\sigma_j\sum_{i=1}^{n} f_2(h(X_i))\,r(j,j(i))\,\mathbb{1}\{X_i\in A_j\}\right|,$$

which we bound via a covering argument.[7]

For frozen $X=\langle X_1,\ldots,X_n\rangle$, and index set $I\subseteq[n]$, with $|I|=k$, consider $\mathcal{C}_{2,\epsilon}(X,I)$, a minimal $\epsilon$ cover of $\mathcal{H}$ w.r.t. the 2-norm on $X$ projected onto $I$, that is, w.r.t. the (pseudo-)metric

$$||h_1-h_2||_{2,X,I} = \sqrt{\frac{1}{k}\sum_{i\in I}|h_1(X_i)-h_2(X_i)|^2}\,. \tag{21}$$

Denote by $|\mathcal{C}_{2,\epsilon}(X,I)|$ the size of such a cover. For given $h\in\mathcal{H}$, and $j$ such that $k\le j\le m_{\text{UPPER}}(S)$, let $h'_j=h'(h,A_j)\in\mathcal{C}_{2,\epsilon}(X,A_j)$ be such that $||h-h'_j||_{2,X,A_j}\le\epsilon$. Then we can write

$$\begin{aligned}
\sup_{h\in\mathcal{H}}\left|\sum_{j=k}^{m_{\text{UPPER}}(S)}\sigma_j V(h,A_j)\right| &\le \sup_{h\in\mathcal{H}}\left|\sum_{j=k}^{m_{\text{UPPER}}(S)}\sigma_j\Big(V(h,A_j)-V(h'_j,A_j)\Big)\right| + \sup_{h\in\mathcal{H}}\left|\sum_{j=k}^{m_{\text{UPPER}}(S)}\sigma_j V(h'_j,A_j)\right| \\
&\le \sup_{h\in\mathcal{H}}\sum_{j=k}^{m_{\text{UPPER}}(S)}\Big|V(h,A_j)-V(h'_j,A_j)\Big| \\
&\quad + \max_{r\,:\,k\le r\le m_{\text{UPPER}}(S)}\ \sup_{h\in\mathcal{C}_{2,\epsilon}(X,A_r)}\left|\sum_{j=k}^{m_{\text{UPPER}}(S)}\sigma_j V(h,A_j)\right|.
\end{aligned}$$

On the other hand

$$\begin{aligned}
\left(\frac{np}{2}-2k+1\right)\Sigma\Big|V(h,A_j)-V(h'_j,A_j)\Big| &\le \sum_{i=1}^{n}\Big|f_2(h(X_i))-f_2(h'_j(X_i))\Big|\,r(j,j(i))\,\mathbb{1}\{X_i\in A_j\} \\
&\le 2pL\sum_{i=1}^{n}\Big|h(X_i)-h'_j(X_i)\Big|\,\pi[i]\,\mathbb{1}\{X_i\in A_j\} \\
&\quad\text{(by the } L\text{-Lischitzness of } f_2(\cdot) \text{ and Lemma 11)} \\
&\le 2pL\sqrt{\sum_{i\,:\,X_i\in A_j}\Big|h(X_i)-h'_j(X_i)\Big|^2}\sqrt{\sum_{i\,:\,X_i\in A_j}\pi^2[i]} \\
&= 2pL\sqrt{k}\,||h-h'_j||_{2,X,A_j}\sqrt{\Sigma} \\
&\le 2pL\epsilon\sqrt{k\Sigma}\,.
\end{aligned}$$

7. A more refined chaining version of this covering argument can be leveraged here, which leads to replacing pseudo-dimension by a Dudley's integral. We decided not to take this route, as this would not add much to the value of the paper.

Plugging back gives

$$\sup_{h\in\mathcal{H}}\left|\sum_{j=k}^{m_{\text{UPPER}}(S)}\sigma_j V(h,A_j)\right| \le 2pL\epsilon\sqrt{k\Sigma}\,\frac{m_{\text{UPPER}}(S)-k+1}{\left(\frac{np}{2}-2k+1\right)\Sigma} + \max_{r\,:\,k\le r\le m_{\text{UPPER}}(S)}\ \sup_{h\in\mathcal{C}_{2,\epsilon}(X,A_r)}\left|\sum_{j=k}^{m_{\text{UPPER}}(S)}\sigma_j V(h,A_j)\right|$$
$$= \frac{2pL\,\epsilon\sqrt{k}}{\sqrt{\Sigma}} + \max_{r\,:\,k\le r\le \frac{np}{2}-k}\ \sup_{h\in\mathcal{C}_{2,\epsilon}(X,A_r)}\left|\sum_{j=k}^{\frac{np}{2}-k}\sigma_j V(h,A_j)\right|, \quad (22)$$

the equality following from the fact that $S\in\mathcal{B}$ implies $m_{\text{UPPER}}(S)=\frac{np}{2}-k$. Note that the size of

$$\bigcup\nolimits_{r\,:\,k\le r\le\frac{np}{2}-k}\mathcal{C}_{2,\epsilon}(X,A_r)$$

is at most

$$\left(\frac{np}{2}-2k+1\right)\max_{j\,:\,k\le j\le\frac{np}{2}-k}\left|\mathcal{C}_{2,\epsilon}(X,A_j)\right|.$$

We take expectation w.r.t. the Rademacher variables $\sigma$, apply Massart's finite lemma (Massart, 2000) to the second term of (22), and then take an outer expectation w.r.t. $A$. This yields

$$\mathbb{E}_{\mu|S}[\Psi(A)] \le \frac{4pL\,\epsilon\sqrt{k}}{\sqrt{\Sigma}} + 4\,\mathbb{E}_A\left[\sqrt{\sum_{j=k}^{\frac{np}{2}-k}V_j^2}\sqrt{2\log\left(\left(\frac{np}{2}-2k+1\right)\max_{j\,:\,k\le j\le\frac{np}{2}-k}\left|\mathcal{C}_{2,\epsilon}(X,A_j)\right|\right)}\right].$$

But from (19) we have, deterministically,

$$\sqrt{\sum_{j=k}^{m_{\text{UPPER}}(S)}V_j^2} \le \frac{8F_2}{n\,\Sigma}\sqrt{\frac{np}{2}-2k+1}\,.$$

We now find an upper bound on the covering number $|\mathcal{C}_{2,\epsilon}(X,A_j)|$. First, note that, for each $j$, the covering number $|\mathcal{C}_{2,\epsilon}(X,A_j)|$ cannot be bigger than $(\frac{1}{\epsilon})^k$. This is because the functions $h\in\mathcal{H}$ take values in the interval $[0,1]$ and, for the sake of metric (21), they are evaluated only in the $k$ points $\{X_i\}_{i\in A_j}$. Moreover, since the fat-shattering dimension of $\mathcal{H}$ at any scale is always upper bounded by the pseudo-dimension $\text{PDim}(\mathcal{H})$, we have $|\mathcal{C}_{2,\epsilon}(X,A_j)| = O\big((\frac{1}{\epsilon})^{O(\text{PDim}(\mathcal{H}))}\big)$ by, e.g., Theorem 1 of Mendelson and Vershynin (2003).

This implies, using $k\le\frac{np}{8}$, and the inequality $\sqrt{a+b}\le\sqrt{a}+\sqrt{b}$,

$$\mathbb{E}_{\mu|S}[\Psi(A)] = \inf_{\epsilon\ge0}\Bigg(\underbrace{O\left(\frac{pL\sqrt{k}}{\sqrt{\Sigma}}\right)}_{C_1}\epsilon + \underbrace{O\left(\frac{F_2}{\Sigma}\sqrt{\frac{p\,\min\{\text{PDim}(\mathcal{H}),k\}}{n}}\right)}_{C_2}\sqrt{\log\frac{1}{\epsilon}} + \underbrace{O\left(\frac{F_2}{\Sigma}\sqrt{\frac{p\log(1+np)}{n}}\right)}_{C_3}\Bigg),$$

where the $O(\cdot)$ only conceals absolute constants. We have then an expression of the form

$$C_1\epsilon + C_2\sqrt{\log\frac{1}{\epsilon}} + C_3$$

that we want to optimize over $\epsilon \geq 0$. In particular, we set

$$\epsilon = \frac{C_2}{2C_1}$$

to obtain

$$\begin{aligned}\mathbb{E}_{\mu|S}[\Psi(A)] &= O\left(\frac{F_2}{\Sigma}\sqrt{\frac{p\,\min\{\text{PDim}(\mathcal{H}),k\}}{n}}\left[1+\sqrt{\log\left(\frac{pLk\Sigma n}{F_2\min\{\text{PDim}(\mathcal{H}),k\}}\right)}\right]\right.\\ &\qquad\left.+\frac{F_2}{\Sigma}\sqrt{\frac{p\log(1+np)}{n}}\right)\\ &= \widetilde{O}\left(\frac{F_2}{\Sigma}\sqrt{\frac{p\,\min\{\text{PDim}(\mathcal{H}),k\}}{n}}\right),\end{aligned}$$

provided $S \in \mathcal{B}$. Combining with (20) results in

$$\begin{aligned}\mathbb{P}_{\mu|S}\left(\Psi(A) > \underbrace{\mathbb{E}_{\mu|S}[\Psi(A)]+t}_{\epsilon/4}\,\middle|\,S\right) &\leq \mathbb{P}_{\mu|S}\Big(\Psi(A) > \mathbb{E}_{\mu|S}[\Psi(A)]+t\,|\,S\in\mathcal{B}\Big)+\mathbb{1}\{S\notin\mathcal{B}\}\\ &\leq \exp\left(-\frac{t^2\,n\,\Sigma^2}{192\,p\,F_2^2}\right)+\mathbb{1}\{S\notin\mathcal{B}\}\,.\end{aligned}$$

holding for every realization of $S$. Thus, upon setting

$$\epsilon/4 \geq \mathbb{E}_{\mu|S}[\Psi(A)]+t\,,$$

with $\mathbb{E}_{\mu|S}[\Psi(A)]$ bounded as above when $S\in\mathcal{B}$, and noting that $\mathbb{P}(S\notin\mathcal{B}) \leq e^{-\Omega(np)}$ via standard Chernoff bounds, we can write

$$(14) = \mathbb{E}_S\left[\mathbb{P}_{\mu|S}\left(\Psi(A)\geq\epsilon/4\right)\right] \leq \exp\left(-\frac{n\,\Sigma^2}{192\,p\,F_2^2}\cdot t^2\right)+e^{-\Omega(np)}\,.$$

**Finishing the proof:** Combining (13) and (14) we have therefore obtained, for

$$\epsilon/4 \geq\ t\ +\ \max\left\{\frac{32L\,R_n(\mathcal{H})}{\Sigma}+\frac{40F_2}{\Sigma}e^{-\Omega(np)},\,\widetilde{\Omega}\left(\frac{F_2}{\Sigma}\sqrt{\frac{p\,\min\{\text{PDim}(\mathcal{H}),k\}}{n}}\right)\right\},$$

we have

$$\mathbb{P}\left(Reg(\widehat{h})\geq\epsilon\right) \leq e^{-\Omega(np)}+\exp\left(-\frac{n\Sigma^2}{1152F_2^2}\cdot t^2\right)+\exp\left(-\frac{n\,\Sigma^2}{192\,p\,F_2^2}\cdot t^2\right)+e^{-\Omega(np)}\,. \quad (23)$$

Now, the right-hand side of (23) is smaller than $\delta$ if

$$t = \Omega\left(\frac{F_2}{\Sigma}\sqrt{\frac{\log\frac{1}{\delta}}{n}}\right) \qquad \text{and} \qquad np = \Omega\left(\log\frac{1}{\delta}\right).$$

Moreover, the conditions $\delta \le 1/2$ and $np = \Omega\left(\log\frac{1}{p}\right)$ imply

$$\frac{40F_2}{\Sigma}e^{-\Omega(np)} = O\left(\frac{F_2}{\Sigma}\sqrt{\frac{\log\frac{1}{\delta}}{n}}\right).$$

Hence, under these conditions (23) implies $\mathbb{P}\left(Reg(\widehat{h}) \ge \epsilon\right) \le \delta$ for

$$\epsilon = \widetilde{\Omega}\left(\frac{L\,R_n(\mathcal{H})}{\Sigma} + \frac{F_2}{\Sigma}\sqrt{\frac{\log\frac{1}{\delta}}{n}} + \frac{F_2}{\Sigma}\sqrt{\frac{p\,\min\{\mathrm{PDim}(\mathcal{H}), k\}}{n}}\right),$$

and the proof is concluded. ■

### B.1. Robustness to error in prior

First, we explicitly detail the new definitions of $\widehat{\beta}_1, \widehat{\beta}_0, \widehat{\ell}(h,j)$. Specifically, we now let

$$\widehat{\beta}_1(j,i) := \frac{\widehat{\pi}[i]\,B_{n,p,j+k}}{p} + \left(B_{n-1,p,j+k-1} - \frac{1-p\,B_{n-1,p,j+k-1}}{1-p}\right)\left(1-\widehat{\pi}[i]\right),$$
$$\widehat{\beta}_0(j,i) := \frac{1-p\,B_{n-1,p,j+k-1}}{1-p}\left(1-\widehat{\pi}[i]\right).$$

where $B_{n,p,j}$ and $p$ are defined as in Theorem 1. We now define $\widehat{\Sigma} = \sum_{i=1}^k \widehat{\pi}[i]^2 = \|\widehat{\pi}\|_2^2$, and

$$\widehat{\ell}(h,j,i) = \frac{f_2(h(A_j[i]))}{\widehat{\beta}_1(j,i)} + \frac{\mathbb{E}\Big[f_1(h(X))\Big]}{B_{n,p,j+k}} - \frac{\widehat{\beta}_0(j,i)\,\mathbb{E}\Big[f_2(h(X))\Big]}{\widehat{\beta}_1(j,i)B_{n,p,j+k}},$$
$$\widehat{\ell}(h,j) = \left(\frac{1}{\widehat{\Sigma}}\sum_{i=1}^k \frac{\widehat{\pi}[i]^2}{\widehat{\beta}_1(j,i)} f_2(h(A_j[i])) - \frac{1}{\widehat{\Sigma}}\left(\sum_{i=1}^k \frac{\widehat{\pi}[i]^2\cdot\widehat{\beta}_0(j,i)}{\widehat{\beta}_1(j,i)B_{n,p,j+k}}\right)\mathbb{E}[f_2(h(x))] + \frac{1}{B_{n,p,j+k}}\,\mathbb{E}[f_1(h(x))]\right)\mathbb{1}\{j \le M-k\}.$$

Now turning to the proof, the only new step we need to prove Theorem 5 is the following:

**Lemma 16** *For all $h \in \mathcal{H}$, we have*

$$\left|\mathbb{E}_\mu[\widehat{\ell}] - \mathcal{L}(h)\right| = \begin{cases} O\left(pF_2\left(\frac{\|\pi-\widehat{\pi}\|_1}{\Sigma} + \frac{\|\pi-\widehat{\pi}\|_2}{\Sigma^{3/2}}\right) + pk^2F_2e^{-\Omega(np)}\right) & \text{if } \Sigma \ge 8\|\pi-\widehat{\pi}\|_2^2, \\ O\left(\frac{pkF_2\|\pi-\widehat{\pi}\|_2}{\Sigma^{1/2}} + pk^2F_2e^{-\Omega(np)}\right) & \text{if } \Sigma < 8\|\pi-\widehat{\pi}\|_2^2. \end{cases}$$

*Also, when $\Sigma \ge 8\|\pi-\widehat{\pi}\|_1^2$ we have $\widehat{\Sigma} \in [c_1\Sigma, c_2\Sigma]$ for universal constants $c_1, c_2 > 0$.*

Furthermore, as detailed below, it is not possible to avoid explicit $k$-dependence in the last case above using the current strategy of analysis.

**Remark 17** *Notice that up to multiplicative constants, the above analysis in the second case above (when $\Sigma < 8\|\pi - \widehat{\pi}\|_2^2$) is tight. Fix any real parameter $\delta > 0$. For any $k \geq 1$, suppose $\pi = \left(\frac{1}{k} - \frac{\delta}{2}, \frac{1}{k}, \ldots, \frac{1}{k}, \frac{1}{k} + \frac{\delta}{2}\right)$ and $\widehat{\pi} = \left(\frac{1}{k}, \ldots, \frac{1}{k}\right)$. Then $\|\pi - \widehat{\pi}\|_2 = \frac{\delta}{\sqrt{2}}$, $\Sigma \approx \frac{1}{k} + \delta^2$, $\widehat{\Sigma} = \frac{1}{k}$. Hence for $k$ large enough in terms of $\delta$, we have $\Sigma < 8\|\pi - \widehat{\pi}\|_2^2$. Note in this example, we have $\widehat{\Sigma} = \frac{1}{k}$. The resulting regret bound from Theorem 5 comes from replacing $\Sigma$ from Theorem 2 by $\widehat{\Sigma}$ when $\Sigma \geq 8\|\pi - \widehat{\pi}\|_2^2$ and by $\frac{1}{k}$ when $\Sigma < 8\|\pi - \widehat{\pi}\|_2^2$ – see the proof of Theorem 5 below. Since $\widehat{\Sigma} = \frac{1}{k}$ in this example, one cannot refine this current analysis strategy using a larger lower bound on $\widehat{\Sigma}$ in the case $\Sigma < 8\|\pi - \widehat{\pi}\|_2^2$.*

**Proof** [Proof of Theorem 5, given Lemma 16] First, we discuss how to establish the result when we have $np = \Omega\left(\log\left(\frac{1}{\delta p}\max_{i\in[k]}\frac{1}{\widehat{\pi}[i]}\right)\right)$. Then following Remark 3, we will explain how to establish the result in its full generality.

The proof follows the exact same strategy as the proof of Theorem 2. The main change is that in the initial steps of that proof, we now decompose

$$\begin{aligned}
&\mathbb{1}\left\{Reg(\widehat{h}) \geq \epsilon + \left(\text{Bias} + pk^2e^{-\Omega(np)}\right)\right\} \\
&\quad\leq \mathbb{1}\left\{2\sup_{h\in\mathcal{H}}\left|\widehat{\ell}(h,S,\mathcal{A},\alpha) - \mathcal{L}(h)\right| \geq \epsilon + \left(\text{Bias} + pk^2e^{-\Omega(np)}\right)\right\} \\
&\quad\leq \mathbb{1}\left\{2\sup_{h\in\mathcal{H}}\left|\mathbb{E}_\mu[\widehat{\ell}] - \mathcal{L}(h)\right| + 2\sup_{h\in\mathcal{H}}\left|\widehat{\ell}(h,S,\mathcal{A},\alpha) - \mathbb{E}_\mu[\widehat{\ell}]\right| \geq \epsilon + \left(\text{Bias} + pk^2e^{-\Omega(np)}\right)\right\} \\
&\quad\leq \mathbb{1}\left\{\sup_{h\in\mathcal{H}}\left|\widehat{\ell}(h,S,\mathcal{A},\alpha) - \mathbb{E}_\mu[\widehat{\ell}]\right| \geq \epsilon/2\right\}.
\end{aligned}$$

Here we used that $\sup_{h\in\mathcal{H}}\left|\mathbb{E}_\mu[\widehat{\ell}] - \mathcal{L}(h)\right| \leq \frac{1}{2}\left(\text{Bias} + pk^2e^{-\Omega(np)}\right)$ by Lemma 16, and the fact that

$$\sup_{h\in\mathcal{H}}\left|\widehat{\ell}(h,S,\mathcal{A},\alpha) - \mathcal{L}(h)\right| \leq \sup_{h\in\mathcal{H}}\left|\widehat{\ell}(h,S,\mathcal{A},\alpha) - \mathbb{E}_\mu[\widehat{\ell}]\right| + \sup_{h\in\mathcal{H}}\left|\mathbb{E}_\mu[\widehat{\ell}] - \mathcal{L}(h)\right|.$$

The rest of the proof is now identical as that of Theorem 2. In particular, we replace every instantiation of quantities arising in the estimator $\widehat{\ell}$ that formerly depended on $\pi$, now by the analogous quantities depending on $\widehat{\pi}$ (e.g. the $r(j, j(i))$ are now defined analogously as before, but in terms of $\widehat{\pi}$ which defines $\widehat{\beta}_1, \widehat{\beta}_0$). Similarly, the $\Sigma$ are now all replaced by $\widehat{\Sigma}$. The condition $np = \Omega\left(\log\left(\frac{1}{\delta p}\max_{i\in[k]}\frac{1}{\widehat{\pi}[i]}\right)\right)$ enables us to use Lemma 11 to bound $\widehat{\beta}_1, \widehat{\beta}_0$. Note that the attribution sets are constructed as per the adversary's play, which is according to $\pi$. The rationale that for a given $j$, there are at most $k$ indices $i$ such that $x_i \in A_j$ and the corresponding $j(i)$ are all distinct remains exactly the same, so the proof does not change.

The only situation in the proof where a new bound, that does not arise from replacing all the quantities that formerly depended on $\pi$ by the analogous quantities depending on $\widehat{\pi}$, is the proof of

(12). Here, we analogously can derive the exact same bound as follows:

$$
\begin{aligned}
&\mathbb{E}_{\mu|S}\left[\sum_{j\,:\,k\le j\le m_{\text{UPPER}}(S)\,,\,x_i\in A_j} r(j,j(i))\right] \\
&= \sum_{t=1}^{r^+}\sum_{i'=b^+(t)+1}^{k}\pi[i']\,r\Big(j_{b^+(t)},\,i'-b^+(t)\Big) + \sum_{t=1}^{r^-}\sum_{i'=1}^{k-b^-(t)}\pi[i']\,r\Big(j_{b^-(t)},\,i'+b^-(t)\Big) \\
&\le 2p\sum_{t=1}^{r^+}\sum_{i'=b^+(t)+1}^{k}\pi[i']\,\widehat{\pi}[i'-b^+(t)] + 2p\sum_{t=1}^{r^-}\sum_{i'=1}^{k-b^-(t)}\pi[i']\,\widehat{\pi}[i'+b^-(t)] \\
&\le 2p\sum_{b=0}^{k-1}\sum_{i'=b+1}^{k}\pi[i']\widehat{\pi}[i'-b] + 2p\sum_{b=1}^{k-1}\sum_{i'=1}^{k-b}\pi[i']\widehat{\pi}[i'+b] \\
&\le 2p\sum_{i'=1}^{k}\pi[i']\sum_{b=1}^{k}\widehat{\pi}[b] + 2p\sum_{i'=1}^{k-1}\pi[i]\sum_{b=1}^{k}\widehat{\pi}[b] \\
&= 2p\sum_{i'=1}^{k}\pi[i'] + 2p\sum_{i'=1}^{k-1}\pi[i'] \\
&\le 4p\,. \qquad (24)
\end{aligned}
$$

Hence the same proof of Theorem 2 goes through as described above. The claimed regret follows from replacing $\widehat{\Sigma}$ by $\Theta(\Sigma)$ when $\Sigma \ge 8\|\pi-\widehat{\pi}\|_1^2$ as per Lemma 16, and using the worst case bound $\widehat{\Sigma} \ge \frac{1}{k}$ otherwise. We finally upper bound $pk^2F_2e^{-\Omega(np)} \le \frac{F_2}{\Sigma}\sqrt{\frac{1}{n}}$ for $np = \Omega\left(\log\left(\frac{1}{\delta p}\max_{i\in[k]}\frac{1}{\widehat{\pi}[i]}\right)\right)$; note this condition implies that $np = \Omega\Big(\max\Big\{\log\left(\frac{1}{p}\right), \log(pk)\Big\}\Big)$. Thus this extra term $pk^2F_2e^{-\Omega(np)}$ can be subsumed into the pre-existing terms in the regret.

Finally, to prove the Theorem under the condition $np \ge \Omega\Big(\log\left(\frac{\sqrt{2k}}{\delta p}\right)\Big)$, we construct $\widehat{\ell}$ by restricting to the $i$ such that $\widehat{\pi}[i] \ge \frac{1}{\delta p}e^{-np}$, as discussed in Remark 3. We again have

$$
\sum_{i\,:\,\widehat{\pi}[i]<\frac{1}{\delta p}e^{-np}}\widehat{\pi}[i]^2 \le \frac{ke^{-2np}}{\delta^2p^2} \le \frac{1}{2k} \le \frac{\widehat{\Sigma}}{2},\ \text{thus}\ \sum_{i\,:\,\widehat{\pi}[i]\ge\frac{1}{\delta p}e^{-np}}\widehat{\pi}[i]^2 \ge \frac{\widehat{\Sigma}}{2}\,,
$$

and so the rate only changes by a constant factor. Note that $np \ge \Omega\Big(\log\left(\frac{\sqrt{2k}}{\delta p}\right)\Big)$ again implies $np = \Omega\Big(\max\Big\{\log\left(\frac{1}{p}\right), \log(pk)\Big\}\Big)$, allowing us to subsume the extra $pk^2F_2e^{-\Omega(np)}$ term into the pre-existing terms in the regret. ■

**Proof** [Proof of Lemma 16] By Theorem 1 and the same work we did prior to stating Theorem 2 in Section 4, we can write

$$\mathcal{L}(h) = \mathbb{E}_\mu\Bigg[\frac{1}{\frac{np}{2}-2k+1}\sum_{j=k}^{np/2-k}\Bigg(\frac{1}{\Sigma}\sum_{i=1}^{k}\frac{\pi[i]^2}{\beta_1(j,i)}f_2(h(A_j[i])) - \frac{1}{\Sigma}\Bigg(\sum_{i=1}^{k}\frac{\pi[i]^2\cdot\beta_0(j,i)}{\beta_1(j,i)B_{n,p,j+k}}\Bigg)\mathbb{E}[f_2(h(x))] + \frac{1}{B_{n,p,j+k}}\mathbb{E}[f_1(h(x))]\Bigg)\mathbb{1}\{j\le M-k\}\Bigg].$$

By analogous reasoning, we have

$$\mathbb{E}_\mu[\widehat{\ell}] = \mathbb{E}_\mu\Bigg[\frac{1}{\frac{np}{2}-2k+1}\sum_{j=k}^{np/2-k}\Bigg(\frac{1}{\widehat{\Sigma}}\sum_{i=1}^{k}\frac{\widehat{\pi}[i]^2}{\widehat{\beta}_1(j,i)}f_2(h(A_j[i])) - \frac{1}{\widehat{\Sigma}}\Bigg(\sum_{i=1}^{k}\frac{\widehat{\pi}[i]^2\cdot\widehat{\beta}_0(j,i)}{\widehat{\beta}_1(j,i)B_{n,p,j+k}}\Bigg)\mathbb{E}[f_2(h(x))] + \frac{1}{B_{n,p,j+k}}\mathbb{E}[f_1(h(x))]\Bigg)\mathbb{1}\{j\le M-k\}\Bigg].$$

Thus

$$\big|\mathbb{E}_\mu[\widehat{\ell}] - \mathcal{L}(h)\big| \le \big|(\mathrm{I})\big| + \big|(\mathrm{II})\big|,$$

where

$$\begin{aligned}(\mathrm{I}) :=\ & \frac{1}{\big(\frac{np}{2}-2k+1\big)\widehat{\Sigma}}\sum_{j=k}^{np/2-k}\sum_{i=1}^{k}\frac{\widehat{\pi}[i]^2}{\widehat{\beta}_1(j,i)}f_2(h(A_j[i])) \\ & - \frac{1}{\big(\frac{np}{2}-2k+1\big)\Sigma}\sum_{j=k}^{np/2-k}\sum_{i=1}^{k}\frac{\pi[i]^2}{\beta_1(j,i)}f_2(h(A_j[i]))\,, \\ (\mathrm{II}) :=\ & -\frac{1}{\big(\frac{np}{2}-2k+1\big)B_{n,p,j+k}\widehat{\Sigma}}\sum_{j=k}^{np/2-k}\sum_{i=1}^{k}\frac{\widehat{\pi}[i]^2\widehat{\beta}_0(j,i)}{\widehat{\beta}_1(j,i)}\mathbb{E}[f_2(h(x))] \\ & + \frac{1}{\big(\frac{np}{2}-2k+1\big)B_{n,p,j+k}\Sigma}\sum_{j=k}^{np/2-k}\sum_{i=1}^{k}\frac{\pi[i]^2\beta_0(j,i)}{\beta_1(j,i)}\mathbb{E}[f_2(h(x))]\,.\end{aligned}$$

By Lemma 11, we have $\Big|\beta_1(j,i) - \frac{\pi[i]}{p}\Big| \le O\Big(\frac{e^{-\Omega(np)}}{p}\Big)$, and the exact same proof as of Lemma 11 gives $\Big|\widehat{\beta}_1(j,i) - \frac{\widehat{\pi}[i]}{p}\Big| \le O\Big(\frac{e^{-\Omega(np)}}{p}\Big)$. Similarly we have $\Big|\beta_0(j,i) - (1-\pi[i])\Big|, \Big|\widehat{\beta}_0(j,i) - (1-\widehat{\pi}[i])\Big| \le O(pe^{-np})$. Thus letting $\delta = \beta_1(j,i) - \frac{\pi[i]}{p}$, we obtain

$$\Big|\frac{\pi[i]^2}{\pi[i]/p} - \frac{\pi[i]^2}{\beta_1(j,i)}\Big| = \Big|\frac{\pi[i]^2\delta}{\frac{\pi[i]}{p}\cdot\big(\frac{\pi[i]}{p}+\delta\big)}\Big| = \frac{p^2\pi[i]\delta}{\pi[i]+p\delta} \le p^2\delta = O(pe^{-\Omega(np)})\,,$$

and similarly $\left|\frac{\widehat{\pi}[i]^2}{\widehat{\pi}[i]/p} - \frac{\widehat{\pi}[i]^2}{\widehat{\beta}_1(j,i)}\right| = O(pe^{-\Omega(np)})$. Now using the above upper bound, and by our condition on $n$ and Lemma 11, we have

$$\begin{aligned}
&\Big|\frac{\pi[i]^2\beta_0(j,i)}{\beta_1(j,i)} - \frac{\pi[i]^2(1-\pi[i])}{\pi[i]/p}\Big| \\
&\quad= \Big|\frac{\pi[i]^2\beta_0(j,i)}{\beta_1(j,i)} - \frac{\pi[i]^2\beta_0(j,i)}{\pi[i]/p} + \frac{\pi[i]^2\beta_0(j,i)}{\pi[i]/p} - \frac{\pi[i]^2(1-\pi[i])}{\pi[i]/p}\Big| \\
&\quad\le \beta_0(j,i)\Big|\frac{\pi[i]^2}{\beta_1(j,i)} - \frac{\pi[i]^2}{\pi[i]/p}\Big| + p\pi[i]\big|\beta_0(j,i) - (1-\pi[i])\big| \\
&\quad= O(pe^{-\Omega(np)}) + p\cdot O(pe^{-np}) \\
&\quad= O(pe^{-\Omega(np)})\,,
\end{aligned}$$

and similarly for $\left|\frac{\widehat{\pi}[i]^2\widehat{\beta}_0(j,i)}{\widehat{\beta}_1(j,i)} - \frac{\widehat{\pi}[i]^2(1-\widehat{\pi}[i])}{\widehat{\pi}[i]/p}\right|$. Thus applying these bounds, along with $\widehat{\Sigma}, \Sigma \ge \frac{1}{k}$, and $j+k \le \frac{np}{2}$ (so that $B_{n,p,j+k} \ge \frac{1}{2}$), we obtain

$$\begin{aligned}
\big|(\mathrm{I})\big| &\le \frac{p}{np/2-2k+1}\sum_{j=k}^{np/2-k}\sum_{i=1}^{k}\big|f_2(h(A_j[i]))\big|\cdot\Big|\frac{\widehat{\pi}[i]}{\widehat{\Sigma}} - \frac{\pi[i]}{\Sigma}\Big| \\
&\quad+ \frac{k}{\frac{np}{2}-2k+1}\sum_{j=k}^{np/2-k}\sum_{i=1}^{k}\big|f_2(h(A_j[i]))\big|\left(\Big|\frac{\pi[i]^2}{\pi[i]/p} - \frac{\pi[i]^2}{\beta_1(j,i)}\Big| + \Big|\frac{\widehat{\pi}[i]^2}{\widehat{\pi}[i]/p} - \frac{\widehat{\pi}[i]^2}{\widehat{\beta}_1(j,i)}\Big|\right) \\
&\le pF_2\sum_{i=1}^{k}\Big|\frac{\widehat{\pi}[i]}{\widehat{\Sigma}} - \frac{\pi[i]}{\Sigma}\Big| + O\Big(pk^2F_2e^{-\Omega(np)}\Big)\,. \\
\big|(\mathrm{II})\big| &\le \frac{p}{\big(\frac{np}{2}-2k+1\big)B_{n,p,j+k}}\Big|\,\mathbb{E}\Big[f_2(h(x))\Big]\Big|\cdot\sum_{j=k}^{M_{\text{UPPER}}}\left|\sum_{i=1}^{k} -\frac{\widehat{\pi}[i](1-\widehat{\pi}[i])}{\widehat{\Sigma}} + \frac{\pi[i](1-\pi[i])}{\Sigma}\right| \\
&\quad+ \frac{k}{\big(\frac{np}{2}-2k+1\big)B_{n,p,j+k}}\Big|\,\mathbb{E}\Big[f_2(h(x))\Big]\Big|\cdot\sum_{j=k}^{np/2-k}\sum_{i=1}^{k}\Big|\frac{\pi[i]^2\beta_0(j,i)}{\beta_1(j,i)} - \frac{\pi[i]^2(1-\pi[i])}{\pi[i]/p}\Big| \\
&\quad+ \frac{k}{\big(\frac{np}{2}-2k+1\big)B_{n,p,j+k}}\Big|\,\mathbb{E}\Big[f_2(h(x))\Big]\Big|\cdot\sum_{j=k}^{np/2-k}\sum_{i=1}^{k}\Big|\frac{\widehat{\pi}[i]^2\widehat{\beta}_0(j,i)}{\widehat{\beta}_1(j,i)} - \frac{\widehat{\pi}[i]^2(1-\widehat{\pi}[i])}{\widehat{\pi}[i]/p}\Big| \\
&\le 2pF_2\Big|\sum_{i=1}^{k} -\frac{\widehat{\pi}[i](1-\widehat{\pi}[i])}{\widehat{\Sigma}} + \frac{\pi[i](1-\pi[i])}{\Sigma}\Big| + O(pk^2F_2e^{-\Omega(np)}) \\
&= 2pF_2\left|\sum_{i=1}^{k}\Big(-\frac{\widehat{\pi}[i]}{\widehat{\Sigma}} + \frac{\pi[i]}{\Sigma}\Big)\right| + O(pk^2F_2e^{-\Omega(np)}) \\
&= 2pF_2\Big|\frac{1}{\widehat{\Sigma}} - \frac{1}{\Sigma}\Big| + O(pk^2F_2e^{-\Omega(np)})\,.
\end{aligned}$$

It remains to upper bound $\sum_{i=1}^k \left|\frac{\widehat{\pi}[i]}{\widehat{\Sigma}} - \frac{\pi[i]}{\Sigma}\right|$ and $\left|\frac{1}{\widehat{\Sigma}} - \frac{1}{\Sigma}\right|$. To this end note

$$\begin{aligned}\sum_{i=1}^k \left|\frac{\widehat{\pi}[i]}{\widehat{\Sigma}} - \frac{\pi[i]}{\Sigma}\right| &= \sum_{i=1}^k \left|\frac{\widehat{\pi}[i]}{\widehat{\Sigma}} - \frac{\widehat{\pi}[i]}{\Sigma} + \frac{\widehat{\pi}[i]}{\Sigma} - \frac{\pi[i]}{\Sigma}\right| \\ &\le \frac{1}{\Sigma}\sum_{i=1}^k \left|\widehat{\pi}[i] - \pi[i]\right| + \left|\frac{1}{\widehat{\Sigma}} - \frac{1}{\Sigma}\right|\Big(\sum_{i=1}^k \widehat{\pi}[i]\Big) \\ &= \frac{\|\pi - \widehat{\pi}\|_1}{\Sigma} + \left|\frac{1}{\widehat{\Sigma}} - \frac{1}{\Sigma}\right|.\end{aligned}$$

Let $\widehat{\pi}[i] - \pi[i] := \delta_i \in [-1, 1]$, thus $\sum_{i=1}^k \delta_i = 0$, $\sum_{i=1}^k |\delta_i| = \|\pi - \widehat{\pi}\|_1$. We observe that

$$\begin{aligned}\left|\widehat{\Sigma} - \Sigma\right| = \left|\sum_{i=1}^k \big(\pi[i] + \delta_i\big)^2 - \sum_{i=1}^k \pi[i]^2\right| &\le 2\sum_{i=1}^k \left|\pi[i]\right| \cdot \left|\delta_i\right| + \sum_{i=1}^k \delta_i^2 \\ &\le \frac{1}{2}\sum_{i=1}^k \pi[i]^2 + 3\sum_{i=1}^k \left|\delta_i\right|^2 \\ &= \frac{\Sigma}{2} + 3\|\pi - \widehat{\pi}\|_2^2 . \end{aligned} \tag{25}$$

Furthermore, we have

$$\left|\frac{1}{\widehat{\Sigma}} - \frac{1}{\Sigma}\right| = \frac{\left|\Sigma - \widehat{\Sigma}\right|}{\widehat{\Sigma}\,\Sigma} = \frac{\left|\sum_{i=1}^k (\pi[i]^2 - \widehat{\pi}[i]^2)\right|}{\widehat{\Sigma}\,\Sigma} = \frac{\left|\sum_{i=1}^k (\pi[i] - \widehat{\pi}[i])(\pi[i] + \widehat{\pi}[i])\right|}{\widehat{\Sigma}\,\Sigma}.$$

Applying the Cauchy-Schwarz inequality to the numerator:

$$\left|(\pi - \widehat{\pi})^\top(\pi + \widehat{\pi})\right| \le \|\pi - \widehat{\pi}\|_2 \cdot \|\pi + \widehat{\pi}\|_2 .$$

Using the Triangle Inequality for the norm of the sum $\|\pi + \widehat{\pi}\|_2 \le \|\pi\|_2 + \|\widehat{\pi}\|_2$:

$$\|\pi + \widehat{\pi}\|_2 \le \sqrt{\Sigma} + \sqrt{\widehat{\Sigma}} .$$

Substituting this back into the expression yields the result:

$$\left|\frac{1}{\widehat{\Sigma}} - \frac{1}{\Sigma}\right| \le \frac{\|\pi - \widehat{\pi}\|_2(\sqrt{\Sigma} + \sqrt{\widehat{\Sigma}})}{\widehat{\Sigma}\,\Sigma} . \tag{26}$$

To complete the cases, we break into cases depending on whether $\Sigma \ge 8\|\pi - \widehat{\pi}\|_1^2$ or not.

- $\underline{\Sigma \ge 8\|\pi - \widehat{\pi}\|_2^2}$. Eq. (25) yields $\widehat{\Sigma} \in \left[\frac{1}{8}\Sigma, \frac{15}{8}\Sigma\right]$, as originally claimed. Thus by (26),

  $$\left|\frac{1}{\widehat{\Sigma}} - \frac{1}{\Sigma}\right| \le O\Big(\frac{\|\pi - \widehat{\pi}\|_2}{\Sigma^{3/2}}\Big).$$

  Hence $\big|(\mathrm{I})\big| + \big|(\mathrm{II})\big| = pF_2 \cdot O\Big(\frac{\|\pi - \widehat{\pi}\|_1}{\Sigma} + \frac{\|\pi - \widehat{\pi}\|_2}{\Sigma^{3/2}}\Big)$ in this case.

- $\Sigma < 8\|\pi - \widehat{\pi}\|_2^2$. Observe that, as each $\pi[i] - \widehat{\pi}[i] \in [-1, 1]$, Cauchy-Schwarz inequality gives

$$\begin{aligned}\big|\widehat{\Sigma} - \Sigma\big| &= \Big|\sum_{i=1}^{k}(\pi[i] - \widehat{\pi}[i])(\pi[i] + \widehat{\pi}[i])\Big| \\ &\leq \Big(\sum_{i=1}^{k}(\pi[i] + \widehat{\pi}[i])^2\Big)^{1/2}\Big(\sum_{i=1}^{k}(\pi[i] - \widehat{\pi}[i])^2\Big)^{1/2} \\ &\leq 2^{1/2}(\Sigma + \widehat{\Sigma})^{1/2}\Big(\sum_{i=1}^{k}\big|\pi[i] - \widehat{\pi}[i]\big|\Big)^{1/2} = O\Big((\Sigma + \widehat{\Sigma})^{1/2}\|\pi - \widehat{\pi}\|_2\Big)\,.\end{aligned}$$

  Thus

$$\begin{aligned}\Big|\frac{1}{\widehat{\Sigma}} - \frac{1}{\Sigma}\Big| &= O\Big(\frac{(\Sigma + \widehat{\Sigma})^{1/2}\|\pi - \widehat{\pi}\|_2}{\Sigma\widehat{\Sigma}}\Big) \\ &= O\Big(\frac{(1/\Sigma + 1/\widehat{\Sigma})^{1/2}}{(\Sigma\widehat{\Sigma})^{1/2}}\|\pi - \widehat{\pi}\|_2\Big) \\ &= O\Big(\frac{k}{\Sigma^{1/2}}\|\pi - \widehat{\pi}\|_2\Big)\,.\end{aligned}$$

  Hence

$$\big|(\mathrm{I})\big| + \big|(\mathrm{II})\big| = pF_2 \cdot O\Big(\frac{\|\pi - \widehat{\pi}\|_1}{\Sigma} + \frac{k\|\pi - \widehat{\pi}\|_2}{\Sigma^{1/2}}\Big) = O\Big(\frac{k\|\pi - \widehat{\pi}\|_2}{\Sigma^{1/2}}\Big)$$

  in this case. Here in the last step we used the fact that $\frac{\|\pi-\widehat{\pi}\|_1}{\Sigma} \leq \frac{\sqrt{k}\|\pi-\widehat{\pi}\|_2}{\Sigma} \leq \frac{k\|\pi-\widehat{\pi}\|_2}{\Sigma^{1/2}}$, again by the Cauchy-Schwarz inequality.

This concludes the proof. ■

**Remark 18** *Suppose the priors differ for each attribution set $A_j$ (note this encompasses varying set sizes as a special case). Denoting the prior for attribution set $A_j$ as $\pi_j$, Theorem 1 still applies; note the proof of Theorem 1 only considered the $j$-th attribution set $A_j$. Theorem 1 now establishes that $\widehat{\ell}(h, j, i)$ is unbiased, where $\beta_0(j, i), \beta_1(j, i)$ in the definition of $\widehat{\ell}(h, j, i)$ (see (2)) are now defined in terms of $\pi_j$ rather than $\pi$.*

*Letting $k_j$ denote the size of the support of $\pi_j$ and $\Sigma_j = \sum_{i=1}^{k_j}\pi_j[i]^2$, we now define $\widehat{\ell}(h, j) = \frac{1}{\Sigma_j}\sum_{i=1}^{k_j}\pi_j[i]^2\widehat{\ell}(h, j, i)$. We again define $\widehat{\ell}_M(h, S, \mathcal{A}) = \frac{1}{\frac{np}{2}-2k+1}\sum_{j=k}^{M_{\text{UPPER}}}\widehat{\ell}(h, j)$ as in (10). Defining the ERM estimator $\widehat{h}$ in terms of $\widehat{\ell}_M(h, S, \mathcal{A})$ as in (7), we again can study $Reg(\widehat{h})$ as in Theorem 2, following a similar proof as the proof presented above.*

## Appendix C. Further details about the experiments

### C.1. Datasets

**MNIST:** The MNIST dataset (LeCun et al., 2010) is a collection of $28 \times 28$ grayscale handwritten digits containing 60,000 training and 10,000 test examples. To adapt this for binary classification, we

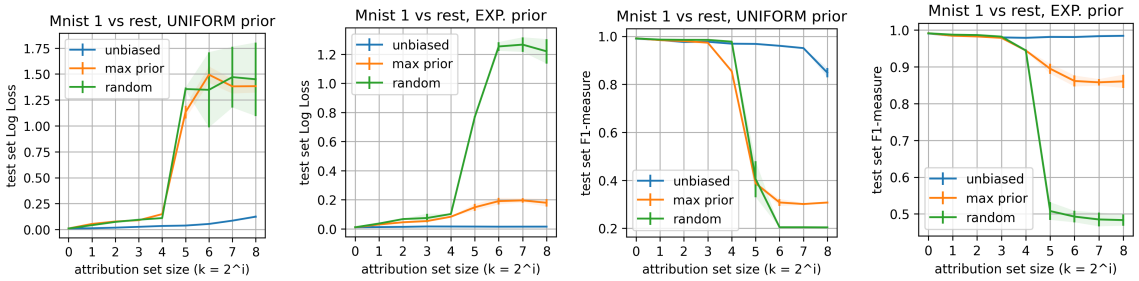


Figure 3: MNIST 1 vs. rest performance measured on the test set via log loss (first two plots from the left) or F1-measure (last two plots) on both the uniform prior and the exponential prior. Standard deviations are also depicted.

labeled digit "1" as the positive class and all other digits as negative. Our model architecture consists of a multilayer perceptron (MLP) with three hidden layers of sizes 512, 512, and 128. We employed ReLU activation functions and a dropout rate of 0.2 at each layer. The network outputs a raw logit via a linear final layer.

**CIFAR-10:** The CIFAR-10 dataset (Krizhevsky, 2009) is a multi-class dataset of 50,000 training and 10,000 test images ($32 \times 32$ color). Each image belongs to one of ten classes: AIRPLANE, AUTOMOBILE, BIRD, CAT, DEER, DOG, FROG, HORSE, SHIP, or TRUCK. For our experiments, we applied an Animal-vs-Machine binarization: Positive Class: BIRD, CAT, DEER, DOG, FROG, and HORSE; Negative Class: AIRPLANE, AUTOMOBILE, SHIP, and TRUCK. The model is a Convolutional Neural Network (CNN) structured as follows:

- Convolutional Layer: 32 filters with ReLU activation.
- Max Pooling: $2 \times 2$ window and stride.
- Convolutional Layer: 64 filters with ReLU activation.
- Dropout Layer: 0.5 rate.
- Fully Connected Layer: Single linear output producing a raw logit.

**Higgs:** The Higgs dataset (Baldi et al., 2014) is a collection of simulated particle physics data used to distinguish between Higgs boson production processes and background noise. While the original dataset contains 11 million examples, we used a subset of 200,000 to accelerate experimentation, allocating the first 10,000 for test, and the subsequent 190,000 for training. Each example comprises 21 features, including both direct physical measurements and hand-crafted high-level features. Our model class is a fully connected model with 4 hidden layers each having 300 neurons and ReLU activations followed by a fully connected layer with 1 output and no activation so that it outputs a logit.

### C.2. Further results

Figure 3 reports average test log loss and F1-measure of the three algorithms on the MNIST dataset.