Needle in a Haystack: Label-Efficient Evaluation under Extreme Class Imbalance

Neil G. Marchant
nmarchant@unimelb.edu.au
School of Computing and Information Systems
University of Melbourne
Melbourne, Victoria, Australia

Benjamin I. P. Rubinstein
brubinstein@unimelb.edu.au
School of Computing and Information Systems
University of Melbourne
Melbourne, Victoria, Australia

ABSTRACT

Important tasks like record linkage and extreme classification demonstrate extreme class imbalance, with 1 minority instance to every 1 million or more majority instances. Obtaining a sufficient sample of all classes, even just to achieve statistically-significant evaluation, is so challenging that most current approaches yield poor estimates or incur impractical cost. Where importance sampling has been levied against this challenge, restrictive constraints are placed on performance metrics, estimates do not come with appropriate guarantees, or evaluations cannot adapt to incoming labels. This paper develops a framework for online evaluation based on adaptive importance sampling. Given a target performance metric and model for $p(y|x)$, the framework adapts a distribution over items to label in order to maximize statistical precision. We establish strong consistency and a central limit theorem for the resulting performance estimates, and instantiate our framework with worked examples that leverage Dirichlet-tree models. Experiments demonstrate an average MSE superior to state-of-the-art on fixed label budgets.

CCS CONCEPTS

• **General and reference** → **Evaluation**; • **Computing methodologies** → *Machine learning*; • **Mathematics of computing** → **Sequential Monte Carlo methods**.

KEYWORDS

performance evaluation, adaptive importance sampling, Dirichlet tree, central limit theorem

ACM Reference Format:
Neil G. Marchant and Benjamin I. P. Rubinstein. 2021. Needle in a Haystack: Label-Efficient Evaluation under Extreme Class Imbalance. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '21), August 14–18, 2021, Virtual Event, Singapore.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3447548.3467435

1 INTRODUCTION

Evaluation of machine learning systems under extreme class imbalance seems like a hopeless task. When minority classes are exceedingly rare—e.g. occurring at a rate of one in a million—a massive number of examples (1 million in expectation) must be labeled before a single minority example is encountered. It seems nigh impossible to reliably estimate performance in these circumstances, as the level of statistical noise is simply too high. Making matters worse, is the fact that high quality labels for evaluation are rarely available for free. Typically they are acquired manually at some cost—e.g. by employing expert annotators or crowdsourcing workers. Reducing labeling requirements for evaluation, while ensuring estimates of performance are precise and free from bias, is therefore paramount. One cannot afford to waste resources on evaluation if the results are potentially misleading or totally useless.

From a statistical perspective, evaluation can be cast as the estimation of population performance measures using independently-drawn test data. Although *unlabeled* test data is abundant in many applied settings, labels must usually be acquired as part of the evaluation process. To ensure estimated performance measures converge to their population values, it is important to select examples for labeling in a statistically sound manner. This can be achieved by sampling examples *passively* according to the data generating distribution. However, passive sampling suffers from poor label efficiency for some tasks, especially under extreme class imbalance. This impacts a range of application areas including fraud detection [39], record linkage [21], rare diseases [18] and extreme classification [36].

The poor efficiency of passive sampling for some evaluation tasks motivates *active* or *biased sampling* strategies, which improve efficiency by focusing labeling efforts on the "most informative" examples [33]. Previous work in this area is based on variance-reduction methods, such as stratified sampling [1, 11, 14], importance sampling [34, 35] and adaptive importance sampling [21]. However existing approaches suffer from serious limitations, including lack of support for a broad range of performance measures [1, 21, 34, 35, 40], weak theoretical justification [1, 11, 40] and an inability to adapt sampling based on incoming labels [34, 35].

In this paper, we present a general framework for label-efficient online evaluation that addresses these limitations. Our framework supports any performance measure that can be expressed as a transformation of a vector-valued risk functional—a much broader class than previous work. This allows us to target simple scalar measures such as accuracy and F1 score, as well as more complex multi-dimensional measures such as precision-recall curves for the first time. We leverage adaptive importance sampling (AIS) to efficiently select examples for labeling in batches. The AIS proposal is adapted using labels from previous batches in order to approximate the asymptotically-optimal variance-minimizing proposal. This approximation relies on online estimates of $p(y|x)$, which we

KDD '21, August 14–18, 2021, Virtual Event, Singapore

ACM ISBN 978-1-4503-8332-5/21/08…$15.00
https://doi.org/10.1145/3447548.3467435

propose to estimate via a Bayesian Dirichlet-tree [10] model that achieves asymptotic optimality for deterministic labels.

We analyze the asymptotic behavior of our framework under general conditions, establishing strong consistency and a central limit theorem. This improves upon a weak consistency result obtained in a less general setting [21]. We also compare our framework empirically against four baselines: passive sampling, stratified sampling, importance sampling [34], and the stratified AIS method of [21]. Our approach based on a Dirichlet-tree model, achieves superior or competitive performance on all but one of seven test cases. Key proofs are included in appendices and further proofs and extensions are detailed in the full technical report [22]. A Python package implementing our framework has been released open source at https://github.com/ngmarchant/activeeval.

2 PRELIMINARIES

We introduce notation and define the label-efficient evaluation problem in Section 2.1. Then in Section 2.2, we specify the family of performance measures supported by our framework. Section 2.3 presents novel insights into the impracticality of passive sampling relative to class imbalance and evaluation measure, supported by asymptotic analysis.

2.1 Problem formulation

Consider the task of evaluating a set of systems $\mathcal{S}$ which solve a prediction problem on a feature space $\mathcal{X} \subseteq \mathbb{R}^m$ and label space $\mathcal{Y} \subseteq \mathbb{R}^l$. Let $f^{(s)}(x)$ denote the output produced by system $s \in \mathcal{S}$ for a given input $x \in \mathcal{X}$—e.g. a predicted label or distribution over labels. We assume instances encountered by the systems are generated i.i.d. from an unknown joint distribution with density $p(x, y)$ on $\mathcal{X} \times \mathcal{Y}$. Our objective is to obtain *accurate* and *precise* estimates of target performance measures (e.g. F1 score) with respect to $p(x, y)$ at minimal cost.

We consider the common scenario where an *unlabeled* test pool $\mathcal{T} = \{x_1, \ldots, x_M\}$ drawn from $p(x)$ is available upfront. We assume labels are *unavailable* initially and can only be obtained by querying a stochastic *oracle* that returns draws from the conditional $p(y|x)$. We assume the response time and cost of oracle queries far outweigh contributions from other parts of the evaluation process. This is reasonable in practice, since the oracle requires human input—e.g. annotators on a crowdsourcing platform or domain experts. Under these assumptions, minimizing the cost of evaluation is equivalent to minimizing the number of oracle queries required to estimate target performance measures to a given precision.

Remark 1. *A* stochastic oracle *covers the most general case where $p(y|x)$ has support on one or more labels. This may be due to a set of heterogeneous or noisy annotators (not modeled) or genuine ambiguity in the label. We also consider a* deterministic oracle *where $p(y|x)$ is a point mass. This is appropriate when trusting individual judgments from an expert annotator.*

2.2 Generalized measures

When embarking on an evaluation task it is important to select a suitable measure of performance. For some tasks it may be sufficient to measure global error rates, while for others it may be desirable to measure error rates for different classes, sub-populations or parameter configurations—the possibilities are boundless. Since no single measure is suitable for all tasks, we consider a broad family of measures which correspond mathematically to transformations of vector-valued risk functionals.

Table 1: Representations of binary classification measures as generalized measures. Here we assume $\mathcal{Y} = \{0, 1\}$, $f(x)$ denotes the predicted class label, and $\hat{p}_1(x)$ is an estimate of $p(y = 1|x)$ according to the system under evaluation.

Measure	$\ell(x, y)^\intercal$	$g(R)$
Accuracy	$\mathbf{1}_{y \neq f(x)}$	$1 - R$
Balanced accuracy	$[yf(x), y, f(x)]$	$\frac{R_1 + R_2(1-R_2-R_3)}{2R_2(1-R_2)}$
Precision	$[yf(x), f(x)]$	$\frac{R_1}{R_2}$
Recall	$[yf(x), y]$	$\frac{R_1}{R_2}$
F_β score	$\left[yf(x), \frac{\beta^2 y + f(x)}{1+\beta^2}\right]$	$\frac{R_1}{R_2}$
Matthews correlation coefficient	$[yf(x), y, f(x)]$	$\frac{R_1 - R_2 R_3}{\sqrt{R_2 R_3 (1-R_2)(1-R_3)}}$
Fowlkes-Mallows index	$[yf(x), y, f(x)]$	$\frac{R_1}{\sqrt{R_2 R_3}}$
Brier score	$2(\hat{p}_1(x) - y)^2$	R

Definition 1 (Generalized measure). *Let $\ell(x, y; f)$ be a vector-valued loss function that maps instances $(x, y) \in \mathcal{X} \times \mathcal{Y}$ to vectors in $\mathbb{R}^d$ dependent on the system outputs $f = \{f^{(s)}\}$. We suppress explicit dependence on f where it is understood. Assume $\ell(x, y; f)$ is uniformly bounded in sup norm for all system outputs f. Denote the corresponding vector-valued risk functional by $R = \mathbb{E}_{X,Y \sim p}[\ell(X, Y; f)]$. For any choice of ℓ and continuous mapping $g : \mathbb{R}^d \to \mathbb{R}^m$ differentiable at R, the* generalized measure *is defined as $G = g(R)$.*

Although this definition may appear somewhat abstract, it encompasses a variety of practical measures. For instance, when g is the identity and $d = 1$ the family reduces to a scalar-valued risk functional, which includes accuracy and mean-squared error as special cases. Other well-known performance measures such as precision and recall can be represented by selecting a non-linear g and a vector-valued ℓ. For example, Table 1 demonstrates how to recover standard binary classification measures for different settings of g and ℓ. In addition to scalar measures, the family also encompasses vector-valued measures for vector-valued g and ℓ. These can be used to estimate multiple scalar measures simultaneously—e.g. precision and recall of a system, accuracy of several competing systems, or recall of a system for various sub-populations. Below, we demonstrate that a vector-valued generalized measure can represent a precision-recall (PR) curve.

Example 1 (PR curve). *A precision-recall (PR) curve plots the precision and recall of a soft binary classifier on a grid of classification thresholds $\tau_1 < \tau_2 < \cdots < \tau_L$. Let $f(x) \in \mathbb{R}$ denote the classifier score for input x, where a larger score means the classifier is more confident the label is positive (encoded as '1') and a smaller score means the classifier is more confident the label is negative (encoded as '0'). We define a vector loss function that measures whether an instance (x, y) is: (1) a predicted positive for each threshold (the first*

L entries), (2) a true positive for each threshold (the next L entries), and/or (3) a positive (the last entry):

$$\ell(x,y) = \left[\mathbf{1}_{f(x)\geq\tau_1},\ldots,\mathbf{1}_{f(x)\geq\tau_L},y\mathbf{1}_{f(x)\geq\tau_1},\ldots,y\mathbf{1}_{f(x)\geq\tau_L},y\right]^\intercal.$$

A PR curve can then be obtained using the following mapping function:

$$G = g(R) = \left[\frac{R_{L+1}}{R_1},\ldots,\frac{R_{2L}}{R_L},\frac{R_{L+1}}{R_{2L+1}},\ldots,\frac{R_{2L}}{R_{2L+1}}\right]^\intercal,$$

where the first L entries correspond to the precision at each threshold in ascending order, and the last L entries correspond to the recall at each threshold in ascending order.

Remark 2. *We have defined generalized measures with respect to the data generating distribution $p(x,y)$. While this is the ideal target for evaluation, it is common in practice to define performance measures with respect to a sample. We can recover these from our definition by substituting an empirical distribution for $p(x,y)$. For example, the familiar sample-based definition of recall can be obtained by setting $\ell(x,y) = [yf(x),y]^\intercal$, $g(R) = R_1/R_2$ and $p(x,y) = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}_{x_i=x}\mathbf{1}_{y_i=y}$. Then*

$$G_{\text{rec}} = g(R) = \frac{\frac{1}{N}\sum_{i=1}^{N} y_i f(x_i)}{\frac{1}{N}\sum_{i=1}^{N} y_i} = \frac{\text{TP}}{\text{TP}+\text{FN}}.$$

Given our assumption that the test pool $\mathcal{T}$ is drawn from $p(x)$, any consistent sample-based estimator will converge to the population measure.

2.3 Inadequacy of passive sampling

We have previously mentioned passive sampling as an obvious baseline for selecting instances to label for evaluation. In this section, we conduct an asymptotic analysis for two sample evaluation tasks, which highlights the impracticality of passive sampling under extreme class imbalance. This serves as concrete motivation for our interest in label-efficient solutions. We begin by defining an estimator for generalized measures based on passive samples.

Definition 2 (Passive estimator for G). *Let $\mathcal{L} = \{(x_1,y_1),\ldots,(x_N,y_N)\}$ be a labeled sample of size N drawn passively according to $p(x,y)$. In practice, $\mathcal{L}$ is obtained by drawing instances i.i.d. from the marginal $p(x)$ and querying labels from the oracle $p(y|x)$. The* Monte Carlo *or* passive *estimator for a generalized measure G is then defined as follows:*

$$\hat{G}_N^{\text{MC}} = g(\hat{R}_N^{\text{MC}}) \text{ with } \hat{R}_N^{\text{MC}} = \frac{1}{N}\sum_{(x,y)\in\mathcal{L}}\ell(x,y).$$

Note that $\hat{G}_N^{\text{MC}}$ is a biased estimator for G in general, since g may be non-linear. However, it is asymptotically unbiased—that is, $\hat{G}_N^{\text{MC}}$ converges to G with probability one in the limit $N\to\infty$. This property is known as *strong consistency* and it follows from the strong law of large numbers [12, pp. 243–245] and continuity of g. There is also a central limit theorem (CLT) for $\hat{G}_N^{\text{MC}}$, reflecting the rate of convergence: $\mathbb{E}[\|\hat{G}_N^{\text{MC}} - G\|] \leq \|\Sigma\|/\sqrt{N}$ asymptotically where Σ is an asymptotic covariance matrix (see Theorem 2). We shall now use this result to analyse the asymptotic efficiency of the passive estimator for two evaluation tasks.

Example 2 (Accuracy). *Consider estimating the accuracy G_{acc} (row 1 of Table 1) of a classifier. By the CLT, it is straightforward to show that the passive estimator for G_{acc} is asymptotically normal with variance $G_{\text{acc}}(1-G_{\text{acc}})/N$. Thus, to estimate G_{acc} with precision w we require a labeled sample of size $N \propto G_{\text{acc}}(1-G_{\text{acc}})/w^2$. Although this is suboptimal[1] it is not impractical. A passive sample reasonably captures the variance in G_{acc}.*

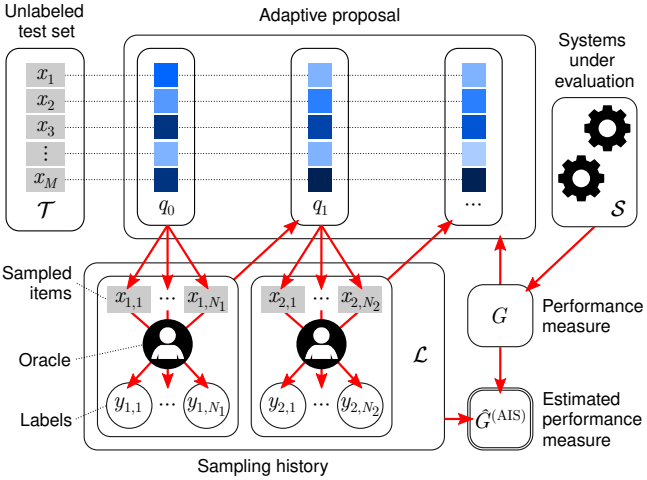

Figure 1: Schematic of the proposed AIS-based evaluation framework.

This example shows that passive sampling is not always a poor choice. It can yield reasonably precise estimates of a generalized measure, so long as the measure is sensitive to regions of the space $\mathcal{X}\times\mathcal{Y}$ with *high density* as measured by $p(x,y)$. However, where these conditions are not satisfied, passive sampling may become impractical, requiring huge samples of labeled data in order to sufficiently drive down the variance. This is the case for the example below, where the measure is is sensitive to regions of $\mathcal{X}\times\mathcal{Y}$ with *low density* as measured by $p(x,y)$.

Example 3 (Recall). *Consider estimating recall G_{rec} (row 4 of Table 1) of a binary classifier. By the CLT, the passive estimator for G_{rec} is asymptotically normal with variance $G_{\text{rec}}(1-G_{\text{rec}})/N\epsilon$ where ϵ denotes the relative frequency of the positive class. Thus we require a labeled sample of size $N \propto G_{\text{rec}}(1-G_{\text{rec}})/w^2\epsilon$ to estimate G_{rec} with precision w. This dependence on ϵ makes passive sampling impractical when $\epsilon \ll 0$—i.e. when the positive class is rare.*

This example is not merely an intellectual curiosity—there are important applications where ϵ is exceedingly small. For instance, in record linkage ϵ scales inversely in the size of the databases to be linked [21].

3 AN AIS-BASED FRAMEWORK FOR LABEL-EFFICIENT EVALUATION

When passive sampling is inefficient[2], as we have seen in the preceding analysis, substantial improvements can often be achieved through biased sampling. In this section, we devise a framework for

[1] Theoretically it is possible to achieve an asymptotic variance of zero.
[2] Unless otherwise specified, we mean *label efficiency* or *sample efficiency* when we use the term "efficiency" without qualification.

Algorithm 1 AIS for generalized measures

Input: generalized measure G; unlabeled test pool $\mathcal{T}$; proposal update procedure; sample allocations $N_1, \ldots, N_T$.

Initialize proposal q_0 and sample history $\mathcal{L} \leftarrow \emptyset$
for $t = 1$ to T **do**
  for $n = 1$ to N_t **do**
    $x_{t,n} \sim q_{t-1}$
    $w_{t,n} \leftarrow \frac{p(x_{t,n})}{q_{t-1}(x_{t,n})}$
    $y_{t,n} \sim \text{Oracle}(x_{t,n})$
    $\mathcal{L} \leftarrow \mathcal{L} \cup \{(x_{t,n}, y_{t,n}, w_{t,n})\}$
  end for
  Update proposal q_t using $\mathcal{L}$
end for
$\hat{R}_N^{\text{AIS}} \leftarrow \frac{1}{N} \sum_{(x,y,w)\in\mathcal{L}} w\, \ell(x,y)$
$\hat{G}_N^{\text{AIS}} \leftarrow g(\hat{R}_N^{\text{AIS}})$
return $\hat{G}_N^{\text{AIS}}$ and history $\mathcal{L}$

efficiently estimating generalized measures that leverages a biased sampling approach called *adaptive importance sampling (AIS)* [3]. AIS estimates an expectation using samples drawn sequentially from a biased proposal distribution, that is adapted in stages based on samples from previous stages. It produces non-i.i.d. samples in general, unlike passive sampling and static (non-adaptive) *importance sampling (IS)* [32]. AIS is a powerful approach because it does not assume an effective proposal distribution is known a priori—instead it is learnt from data. This addresses the main limitation of static IS—that one may be stuck using a sub-optimal proposal, which may compromise label efficiency.

There are many variations of AIS which differ in: (i) the way samples are allocated among stages; (ii) the dependence of the proposal on previous stages; (iii) the types of proposals considered; and (iv) the way samples are weighted within and across stages. Our approach is completely flexible with respect to points (i)–(iii). For point (iv), we use simple importance-weighting as it is amenable to asymptotic analysis using martingale theory [27]. A more complex weighting scheme is proposed in [8] which may have better stability, however its asymptotic behavior is not well understood.[3]

Our framework is summarized in Figure 1. Given a target performance measure G and an unlabeled test pool $\mathcal{T}$, the labeling process proceeds in several stages indexed by $t \in \{1, \ldots, T\}$. In the t-th stage, $N_t \geq 1$ items are drawn i.i.d. from $\mathcal{T}$ according to proposal q_{t-1}. Labels are obtained for each item by querying the oracle, and recorded with their importance weights in $\mathcal{L}$. At the end of the t-th stage, the proposal is updated for the next stage. This update may depend on the weighted samples from all previous stages as recorded in $\mathcal{L}$. At any point during sampling, we estimate G as follows:

$$\hat{G}_N^{\text{AIS}} = g(\hat{R}_N^{\text{AIS}}) \text{ where } \hat{R}_N^{\text{AIS}} = \frac{1}{N} \sum_{(x,y,w)\in\mathcal{L}} w\, \ell(x,y). \quad (1)$$

For generality, we permit the user to specify the sample allocations and the procedure for updating the proposals in Figure 1. In practice, we recommend allocating a small number of samples to each stage, as empirical studies suggest that efficiency improves when the proposal is updated more frequently [27]. However, this must be balanced with practical constraints, as a small sample allocation limits the ability to acquire labels in parallel. In Section 3.2, we recommend a practical procedure for updating the proposals. It approximates the asymptotically-optimal variance-minimizing proposal based on an online model of $p(y|x)$. We present an example model for $p(y|x)$ in Section 4, which can leverage prior information from the systems under evaluation. Further practicalities including sample reuse and confidence intervals are discussed in Section 3.3.

Remark 3 (Constraint on the proposal). *In a standard application of AIS for estimating $\mathbb{E}_{X,Y\sim p}[\phi(X,Y)]$, one is free to select any proposal $q_t(x,y)$ with support on the set $\{(x,y) : \|\phi(x,y)\|p(x,y) \neq 0\}$. However, we have an additional constraint since we cannot bias sampling from the oracle $p(y|x)$. Thus we consider proposals of the form $q_t(x,y) = q_t(x)p(y|x)$.*

Remark 4 (Static Importance Sampling). *Our AIS framework reduces to static importance sampling when $T = 1$ so that all samples are drawn from a single static proposal q_0.*

3.1 Asymptotic analysis

We study the asymptotic behavior of estimates for G produced by the generic framework in Figure 1. Since our analysis does not depend on how samples are allocated among stages, it is cleaner to identify a sample using a single index $j \in \{1, \ldots, N\}$ where $N = \sum_{t=1}^{T} N_t$ is the total number of samples, rather than a pair of indices (t, n). Concretely, we map each (t, n) to index $j = n + \sum_{t'=1}^{t-1} N_{t'}$. With this change, we let $q_{j-1}(x)$ denote the proposal used to generate sample j. It is important to note that this notation conceals the dependence on previous samples. Thus $q_{j-1}(x)$ should be understood as shorthand for $q_{j-1}(x|\mathcal{F}_{j-1})$ where $\mathcal{F}_j = \sigma((X_1, Y_1), \ldots, (X_j, Y_j))$ denotes the filtration.

Our analysis relies on the fact that $Z_N = N(\hat{R}_N^{\text{AIS}} - R)$ is a martingale with respect to $\mathcal{F}_N$. The consistency of $\hat{G}_N^{\text{AIS}}$ then follows by a strong law of large numbers for martingales [13] and the continuous mapping theorem. A proof is included in Appendix A.

Theorem 1 (Consistency). *Let the support of proposal $q_j(x,y) = q_j(x)p(y|x)$ be a superset of $\{x, y \in \mathcal{X} \times \mathcal{Y} : \|\ell(x,y)\|p(x,y) \neq 0\}$ for all $j \geq 0$ and assume*

$$\sup_{j\in\mathbb{N}} \mathbb{E}\left[\left(\frac{p(X_j)}{q_{j-1}(X_j)}\right)^2 \middle| \mathcal{F}_{j-1}\right] < \infty. \quad (2)$$

Then $\hat{G}_N^{\text{AIS}}$ is strongly consistent for G.

We also obtain a central limit theorem (CLT), which is useful for assessing asymptotic efficiency and computing approximate confidence intervals. Our proof (see Appendix B) invokes a CLT for martingales [27] and the multivariate delta method.

Theorem 2 (CLT). *Suppose*

$$V_j := \text{var}\left[\frac{p(X_j)}{q_{j-1}(X_j)}\ell(X_j, Y_j) - R \middle| \mathcal{F}_{j-1}\right] \to V_\infty \quad a.s., \quad (3)$$

[3] Consistency was proved for this weighting scheme in the limit $T \to \infty$ where $\{N_t\}$ is a monotonically increasing sequence [23]. To our knowledge, a CLT remains elusive.

where V_∞ is an a.s. finite random positive semidefinite matrix, and there exists $\eta > 0$ such that

$$\sup_{j\in\mathbb{N}} \mathbb{E}\left[\left(\frac{p(X_j)}{q_{j-1}(X_j)}\right)^{2+\eta}\middle|\mathcal{F}_{j-1}\right] < \infty \quad a.s. \tag{4}$$

Then $\sqrt{N}(\hat{G}_N^{\mathrm{AIS}} - G)$ converges in distribution to a multivariate normal $\mathcal{N}(0,\Sigma)$ with covariance matrix $\Sigma = \mathrm{D}g(R)V_\infty \mathrm{D}g(R)^\intercal$ where $[\mathrm{D}g]_{ij} = \frac{\partial g_i}{\partial R_j}$ is the Jacobian of g.

3.2 Variance-minimizing proposals

In order to achieve optimal asymptotic efficiency, we would like to use the AIS proposal that achieves the minimal asymptotic variance Σ as defined in Theorem 2. We can solve for this optimal proposal using functional calculus, as demonstrated in the proposition below. Note that we must generalize the variance minimization problem for vector-valued G since Σ becomes a covariance matrix. We opt to use the *total variance* (the trace of Σ) since the diagonal elements of Σ are directly related to statistical efficiency, while the off-diagonal elements measure correlations between components of $\hat{G}_N^{\mathrm{AIS}}$ that are beyond our control. This choice also ensures the functional optimization problem is tractable.

Proposition 1 (Asymptotically-optimal proposal). *Suppose the Jacobian $\mathrm{D}_g(R)$ has full row rank and $\mathbb{E}_{X,Y\sim p}\|\mathrm{D}_g(R)\,\ell(X,Y)\|_2^2 > 0$. Then the proposal*

$$q^\star(x) = \frac{v(x)}{\int v(x)\,\mathrm{d}x} \quad \text{with}$$
$$v(x) = p(x)\sqrt{\int \|\mathrm{D}_g(R)\,\ell(x,y)\|_2^2\, p(y|x)\,\mathrm{d}y} \tag{5}$$

achieves the minimum total asymptotic variance of

$$\operatorname{tr}\Sigma = \left(\int v(x)\,\mathrm{d}x\right)^2 - \|\mathrm{D}_g(R)\,R\|_2^2.$$

Appendix D provides sufficient conditions on G and the oracle which ensure $\operatorname{tr}\Sigma = 0$.

We use the above result to design a practical scheme for adapting a proposal for AIS. At each stage of the evaluation process, we approximate the asymptotically-optimal proposal $q^\star(x)$ using an online model for the oracle response $p(y|x)$. The model for $p(y|x)$ should be initialized using prior information if available (e.g. from the systems under evaluation) and updated at the end of each stage using labels received from the oracle. However, we cannot simply estimate $q^\star(x)$ by plugging in estimates of $p(y|x)$ directly, as the resulting proposal may not satisfy the conditions of Theorems 1 and 2. Below we provide estimators for $q^\star(x)$ which do satisfy the conditions, and provide sufficient conditions for achieving asymptotic optimality.

Proposition 2. *If the oracle is* stochastic, *let $\hat{p}_t(y|x)$ be an estimate for $p(y|x)$ whose support includes the support of $p(y|x)$ for all stages $t \geq 0$, and assume $\hat{p}_t(y|x) \xrightarrow{\text{a.s.}} \hat{p}_\infty(y|x)$ pointwise in x. Alternatively, if the oracle is* deterministic, *let $\pi_t(y|x)$ be a posterior distribution for the response $y(x)$ whose support includes $y(x)$ for all $t \geq 0$, and assume $\pi_t(y|x) \xrightarrow{\text{a.s.}} \pi_\infty(y|x)$ pointwise in x. Let ϵ_t be a positive bounded sequence and $\hat{R}_t$ be an estimator for R which converges a.s. to $\hat{R}_\infty$. Assume $\mathcal{X}$ is finite (e.g. a pool of test data) and $\|\mathrm{D}_g(\hat{R}_t)\|_2 \leq K < \infty$ for all $t \geq 0$. Then the proposals*

$$q_t(x) \propto \begin{cases} p(x)\int \max\{\|\mathrm{D}_g(\hat{R}_t)\,\ell(x,y)\|_2, \epsilon_t \mathbf{1}_{\|\ell(x,y)\|\neq 0}\} \\ \quad \pi_t(y|x)\,\mathrm{d}y, \quad \text{for a deterministic oracle,} \\ p(x)[\int \max\{\|\mathrm{D}_g(\hat{R}_t)\,\ell(x,y)\|_2^2, \epsilon_t \mathbf{1}_{\|\ell(x,y)\|\neq 0}\} \\ \quad \hat{p}_t(y|x)\,\mathrm{d}y]^{\frac{1}{2}}, \quad \text{for a stochastic oracle,} \end{cases}$$

satisfy the conditions of Theorems 1 and 2. If in addition $\hat{R}_\infty = R$ and $\hat{p}_\infty(y|x) = p(y|x)$ (alternatively $\pi_\infty(y|x) = \mathbf{1}_{y=y(x)}$) and $\epsilon_t \downarrow 0$, then the proposals approach asymptotic optimality.

3.3 Practicalities

We briefly discuss solutions to two issues that may arise in practical settings: sample reuse and approximate confidence regions.

3.3.1 Sample reuse. Suppose our framework is used to estimate a generalized measure G_1. If the joint distribution $p(x,y)$ associated with the prediction problem has not changed, it may be desirable to *reuse* the weighted samples $\mathcal{L}$ to estimate a different generalized measure G_2. This is possible so long as the sequence of proposals used to estimate G_1 have the required support for G_2. More precisely, the support of $q_j(x,y)$ must include $\{x,y \in \mathcal{X}\times\mathcal{Y} : \|\ell(x,y)\|p(x,y) \neq 0\}$ for the loss functions associated with G_1 and G_2.

If one anticipates sample reuse, the proposals can be made less specialized to a particular measure by mixing with the marginal distribution $p(x)$, i.e. $q(x) \to (1-\delta)q(x) + \delta p(x)$ where $\delta \in (0,1]$ is a hyperparameter that controls the degree of specialization.

3.3.2 Approximate confidence regions. When publishing performance estimates, it may be desirable to quantify statistical uncertainty. An asymptotic $100(1-\alpha)\%$ confidence region for a generalized measure G is given by the ellipsoid

$$\left\{G^\star \in \mathbb{R}^k : (G^\star - \hat{G})^\intercal \hat{\Sigma}^{-1}(G^\star - \hat{G}) \leq \frac{(N-1)k}{N(N-k)} F_{\alpha,k,N-k}\right\},$$

where $\hat{G}$ is the sample mean, $\hat{\Sigma}$ is the sample covariance matrix, and F_{α,d_1,d_2} is the critical value of the F distribution with d_1, d_2 degrees of freedom at significance level α. This region can be approximated using the estimator for G in (1) and the following estimator for Σ:

$$\hat{\Sigma}^{\mathrm{AIS}} = \mathrm{D}_g(\hat{R}^{\mathrm{AIS}})\left(\frac{1}{N}\sum_{j=1}^{N}\frac{p(x_j)^2\ell(x_j,y_j)\ell(x_j,y_j)^\intercal}{q_N(x_j)q_{j-1}(x_j)} - \hat{R}^{\mathrm{AIS}}\hat{R}^{\mathrm{AIS}\intercal}\right)\mathrm{D}_g(\hat{R}^{\mathrm{AIS}})^\intercal.$$

This is obtained from the expression for Σ in Theorem 2, by plugging in AIS estimators for the variance and R, and approximating $q_\infty(x)$ by the most recent proposal $q_N(x)$.

4 A DIRICHLET-TREE MODEL FOR THE ORACLE RESPONSE

In the previous section, we introduced a scheme for updating the AIS proposal which relies on an online model of the oracle response. Since there are many conceivable choices for the model, we left it unspecified for full generality. In this section, we propose a particular model that is suited for evaluating classifiers when the response

from the oracle is deterministic (see Remark 1). Concretely, we make the assumption that the label space $\mathcal{Y} = \{1, \ldots, C\}$ is a finite set and $p(y|x)$ is a point mass at $y(x)$ for all $x \in \mathcal{T}$.

Since we would like to leverage prior information (e.g. classifier scores) from the system(s) under evaluation and perform regular updates as labels are received from the oracle, we opt to use a Bayesian model. Another design consideration is label efficiency. Since labels are scarce and the test pool $\mathcal{T}$ may be huge, we want to design a model that allows for sharing of statistical strength between "similar" instances in $\mathcal{T}$. To this end, we propose a model that incorporates a hierarchical partition of $\mathcal{T}$, where instances assigned to hierarchically neighboring blocks are assumed to elicit a similar oracle response.[4] Various unsupervised methods may be used to learn a hierarchical partition, including hierarchical agglomerative/divisive clustering [30], k-d trees [2], and stratification based on classifier scores (see [22]).

4.1 Generative process

We assume the global oracle response θ (averaged over all instances) is generated according to a Dirichlet distribution, viz.

$$\theta|\alpha \sim \text{Dirichlet}(\alpha)\,,$$

where $\alpha = [\alpha_1, \ldots, \alpha_C] \in \mathbb{R}_+^C$ are concentration hyperparameters. The label y_i for each instance $i \in \{1, \ldots, M\}$ (indexing $\mathcal{T}$) is then assumed to be generated i.i.d. according to θ:

$$y_i|\theta \overset{\text{iid.}}{\sim} \text{Categorical}(\theta)\,, \qquad i \in 1, \ldots, M.$$

We assume a hierarchical partition of the test pool $\mathcal{T}$ is given. The partition can be represented as a tree T, where the leaf nodes of T correspond to the finest partition of $\mathcal{T}$ into disjoint blocks $\{\mathcal{T}_k\}_{k=1}^K$ such that $\mathcal{T} = \bigcup_{k=1}^K \mathcal{T}_k$. We assume each instance i is assigned to one of the blocks (leaf nodes) $k_i \in \{1, \ldots, K\}$ according to a distribution ψ_y with a Dirichlet-tree prior [10, 24]:

$$\psi_y|\beta_y, T \overset{\text{ind.}}{\sim} \text{DirichletTree}\big(\beta_y; T\big)\,, \qquad y \in \mathcal{Y},$$

$$k_i|y_i, \psi_{y_i} \overset{\text{ind.}}{\sim} \text{Categorical}\big(\psi_{y_i}\big)\,, \qquad i \in 1, \ldots, M.$$

The Dirichlet-tree distribution is a generalization of the Dirichlet distribution, which allows for more flexible dependencies between the categories (blocks in this case). Categories that are hierarchically nearby according to the tree T tend to be correlated. The Dirichlet concentration hyperparameters β_y associated with the internal nodes also control the correlation structure.

4.2 Inference

For a deterministic oracle, the response y_i for instance i is either observed (previously labeled) or unobserved (yet to be labeled). It is important to model the observation process in case it influences the values of inferred parameters. To this end, we let $\mathbf{o}_t = (o_{t,1}, \ldots, o_{t,M})$ be observation indicators for the labels $\mathbf{y} = (y_1, \ldots, y_M)$ at the end of stage t of the evaluation process (see Figure 1). We initialize $\mathbf{o}_0 = 0$ and define $\mathbf{o}_t$ in the obvious way: $o_{t,i}$ is 1 if the label for instance i has been observed by the end of stage t and 0 otherwise. From Algorithm 1, we have that the n-th instance selected in stage t depends on the labels of the previously observed instances $\mathbf{y}_{(\mathbf{o}_{t-1})}$ and the block assignments $\mathbf{k} = (k_1, \ldots, k_M)$:

$$i_{t,n}|\mathbf{o}_{t-1}, \mathbf{y}_{(\mathbf{o}_{t-1})}, \mathbf{k} \sim q_{t-1}(\mathbf{y}_{(\mathbf{o}_t)}, \mathbf{k}).$$

Our goal is to infer the unobserved labels (and hence the oracle response) at each stage t of the evaluation process. We assume the block assignments $\mathbf{k} = (k_1, \ldots, k_M)$ are fully observed. Since the observation indicators are *independent* of the unobserved labels conditional on the observed labels, our model satisfies ignorability [16]. This means we can ignore the observation process when conducting inference. Since we do not require a full posterior distribution over all parameters, it is sufficient to conduct inference using the expectation-maximization algorithm. This yields a distribution over the unobserved label for each item and point estimates for the other parameters (ψ_y and θ). Full details are provided in the full technical report [22].

4.3 Asymptotic optimality

Since the Dirichlet-tree model described in this section is consistent for the true oracle (deterministic) response, it can be combined with the proposal updates described in Proposition 2 to yield an asymptotically-optimal AIS algorithm. This result is made precise in the following proposition, which is proved in the technical report.

Proposition 3. *Consider an instantiation of our framework under a deterministic oracle where:*

- *the oracle response is estimated online using the Dirichlet-tree model described in this section via the EM algorithm;*
- *the proposals are adapted using the estimator defined in Proposition 2 with*
- $\epsilon_t = \epsilon_0(1 - \frac{1}{M}\sum_{i=1}^M o_{t,i})$ *for some user-specified* $\epsilon_0 > 0$.

Then Theorems 1 and 2 hold and the estimator is asymptotically-optimal.

5 EXPERIMENTAL STUDY

We conduct experiments to assess the label efficiency of our proposed framework[5] for a variety of evaluation tasks. The tasks vary in terms of the degree of class imbalance, the quality of predictions/scores from the classifier under evaluation, the size of the test pool, and the target performance measure. Where possible, we compare our framework (denoted Ours) with the following baselines:

- Passive: passive sampling as specified in Definition 2.
- IS: static importance sampling as described in Remark 4. We approximate the asymptotically-optimal proposal as in Proposition 2 using estimates of $p(y|x)$ derived from classifier scores.
- Stratified: an online variant of stratified sampling with proportional allocation, as used in [11]. Items are sampled one-at-a-time in proportion to the size of the allocated stratum.
- OASIS: a stratified AIS method for estimating F scores [21].

5.1 Evaluation tasks

We prepare classifiers and test pools for evaluation using publicly-available datasets from various domains, as summarized in Table 2.

[4] This is in contrast to models used in related work [1, 21] which assume the oracle response is independent across blocks of a non-hierarchical partition.

[5] Using the procedure described in Section 3.2 to adapt the AIS proposal, together with the online model for the oracle response presented in Section 4.

For amzn-goog, dblp-acm, abt-buy, restaurant and tweets100k we use the same classifiers and test pools as in [21]. For safedriver and creditcard we prepare our own by randomly splitting the data into train/test with a 70:30 ratio, and training classifiers using supervised learning. In scenarios where labeled data is scarce, semi-supervised or unsupervised methods might be used instead—the choice of learning paradigm has no bearing on evaluation. We consider three target performance measures—F1 score, accuracy and precision-recall curves—as separate evaluation tasks.

5.2 Setup

Oracle. We simulate an oracle using labels included with each dataset. Since the datasets only contain a single label for each instance, we assume a deterministic oracle. Thus, when computing the consumed label budget, we only count the *first* query to the oracle for an instance as a consumed label. If an instance is selected for labeling again, we reuse the label from the first query.

Partitioning. Ours, Stratified and OASIS assume the test pool is partitioned so the oracle response within each block is ideally uniform. We construct partitions by binning instances according to their classifier scores. The bin edges are determined using the *cumulative square-root frequency (CSF) method* [9], which is widely used for stratified sampling. We set the number of bins to $K = 2^8$. Since Ours is capable of exploiting partitions with hierarchical structure, we fill in post-hoc structure by associating the CSF bins with the leaf nodes of an appropriately-size tree in breadth-first order. We consider two trees: a tree of depth 1 with branching factor K (denoted Ours-1, equivalent to a non-hierarchical partition) and a tree of depth 8 with branching factor 2 (denoted Ours-8).

Hyperparameters. We leverage prior information from the classifiers under evaluation to set hyperparameters. Wherever a prior estimate of $p(y|x)$ is required, we use the classifier score $s(y|x)$, applying the softmax function to non-probabilistic scores. For the Dirichlet-tree model we set hyperparameters as follows:

$$\alpha_y = 1 + \sum_{k=1}^{K} s(y|k), \quad \beta_{yv} = \text{depth}(v)^2 + \sum_{k=1}^{K} s(y|k)\delta_v(k)$$

where $s(y|k) = \frac{1}{|\mathcal{T}_k|} \sum_{x_i \in \mathcal{T}_k} s(y|x_i)$ is the mean score over instances in the k-th block, v denotes a non-root node of the tree T, $\text{depth}(v)$ denotes the depth of node v in T, and $\delta_v(k)$ is an indicator equal to 1 if node v is traversed to reach leaf node k and 0 otherwise.

Repeats. Since the evaluation process is randomized, we repeated each experiment 1000 times, computing and reporting the mean behavior with bootstrap confidence intervals.

5.3 Results

We provide a summary of the results here, with a particular focus on the results for F1 score, which is supported by all baselines. Detailed results for accuracy and precision-recall curves are included in the full technical report [22].

F1 score. To assess convergence of the estimated F1 score, we plot the mean-squared error (MSE) as a function of the consumed label budget for all datasets and methods. A plot of this kind is included for abt-buy in Figure 2. It shows that Ours is significantly more efficient than the baseline methods in this case, achieving a lower MSE for all label budgets. The Passive and Stratified methods perform significantly worse, achieving an MSE at least one order of magnitude greater than the biased sampling methods. Figure 2 also plots the convergence of the proposal in terms of the mean KL divergence from the asymptotically-optimal proposal. The results here are in line with expectations: convergence of the F1 score is more rapid when the proposal is closer to asymptotic optimality.

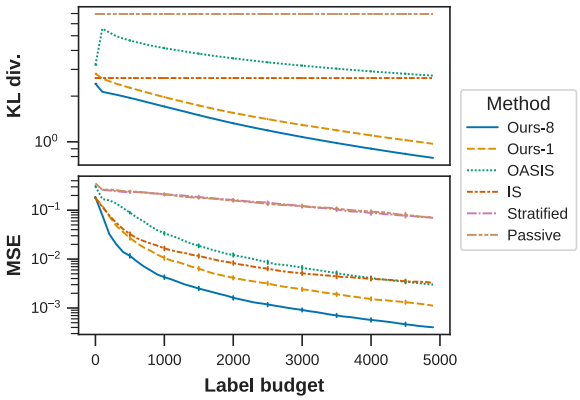

Figure 2: Convergence for abt-buy over 1000 repeats. The upper panel plots the KL divergence from the proposal to the asymptotically-optimal one. The lower panel plots the MSE of the estimated F1 score. 95% bootstrap confidence intervals are included.

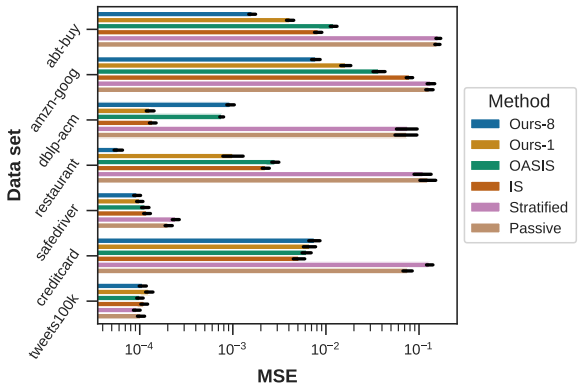

Figure 3: MSE of the estimated F1 score after 2000 label queries over 1000 repeats. The order of the bars (from top to bottom) for each dataset matches the order in the legend. 95% bootstrap confidence intervals are shown in black.

Figure 3 summarizes the convergence plots for the six other data sets (included in Appendix F), by plotting the MSE of the estimated F1 score after 2000 labels are consumed. It shows that Ours achieves best or equal-best MSE on all but one of the datasets (dblp-acm) within the 95% confidence bands. We find the adaptive methods Ours and OASIS perform similarly to IS when the prior estimates for $p(y|x)$ are accurate—i.e. there is less to gain by adapting. Of the two adaptive methods, Ours generally converges more rapidly than

Table 2: Summary of test pools and classifiers under evaluation. The imbalance ratio is the number of positive class instances divided by the number of negative class instances. The true F1 score is assumed unknown.

Source	Domain	Size	Imbalance ratio	Classifier type	F1 score
`abt-buy` [19]	Entity resolution	53,753	1075	Linear SVM	0.595
`amzn-goog` [19]	Entity resolution	676,267	3381	Linear SVM	0.282
`dblp-acm` [19]	Entity resolution	53,946	2697	Linear SVM	0.947
`restaurant` [31]	Entity resolution	149,747	3328	Linear SVM	0.899
`safedriver` [28]	Risk assessment	178,564	26.56	XGBoost [5]	0.100
`creditcard` [29]	Fraud detection	85,443	580.2	Logistic Regression	0.728
`tweets100k` [25]	Sentiment analysis	20,000	0.990	Linear SVM	0.770

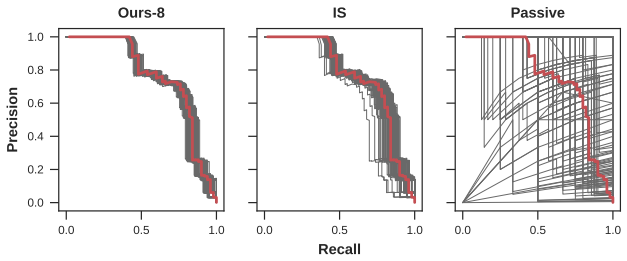

Figure 4: A sample of 100 estimated precision-recall (PR) curves (in dark gray) for three evaluation methods: Ours-8, IS and Passive. The PR curves are estimated for `abt-buy` using a label budget of 5000. The thick red curve is the true PR curve (assuming all labels are known).

OASIS, which might be expected since our procedure for adapting the proposal is asymptotically-optimal. The hierarchical variant of our model Ours-8 tends to outperform the non-hierarchical variant Ours-1, which we expect when $p(y|x)$ varies smoothly across neighboring blocks. Finally, we observe that Passive and Stratified are competitive when the class imbalance is less severe. This agrees with our analysis in Section 2.3. The full technical report observes all of the same trends for accuracy as discussed here for F1 score.

PR curves. We estimate PR curves on a uniformly-spaced grid of classifier thresholds $\tau_1 < \cdots < \tau_L$, where τ_1 is the minimum classifier score, τ_L is the maximum classifier score, and $L = 2^{10}$ (see Example 1). We also use the uniform grid to partition the test pool (in place of the CSF method), associating each block of the partition with four neighboring bins on the grid to yield $K = 2^8$ blocks.

Figure 4 illustrates the vast improvement of the biased sampling methods (Ours-8 and IS) over Passive for this evaluation task. The estimated PR curves shown for Ours-8 and `IS` vary minimally about the true PR curve, and are reliable for selecting an operating threshold. The same cannot be said for the curves produced by Passive, which exhibit such high variance that they are essentially useless.

6 RELATED WORK

Existing approaches to label-efficient evaluation in a machine learning context largely fall into three categories: model-based [40], stratified sampling [1, 11] and importance sampling [21, 34]. The model-based approach in [40] estimates precision-recall curves for binary classifiers. However, it uses inefficient passive sampling to select items to label, and makes strong assumptions about the distribution of scores and labels which can result in biased estimates. Stratified sampling has been used to estimate scalar performance measures such as precision, accuracy and F1 score. Existing approaches [1, 11] bias the sampling of items from strata (akin to blocks) using a heuristic generalization of the optimal allocation principle [6]. However, stratified sampling is considered to be a less effective variance reduction method compared to importance sampling [32], and it does not naturally support stochastic oracles.

Static importance sampling [34] and stratified adaptive importance sampling [21] have been used for online evaluation, and are most similar to our approach. However [21] only supports the estimation of F1 score, and [34] only supports the estimation of scalar generalized risks[6]. Both of these methods attempt to approximate the asymptotically-optimal variance-minimizing proposal, however the approximation used in [34] is non-adaptive and is not optimized for deterministic oracles, while the approximation used in [21] is adaptive but less accurate due to the stratified design.

Novel evaluation methods are also studied in the information retrieval (IR) community (see survey [17]). Some tasks in the IR setting can be cast as prediction problems by treating query-document pairs as inputs and relevance judgments as outputs. Early approaches used relevance scores from the IR system to manage the abundance of irrelevant documents in an ad hoc manner [7]. Recent approaches [20, 35] are based on a statistical formulation similar to ours, however they are specialized to IR systems. Within the IR community, stratified sampling and cluster sampling have also been used to efficiently evaluate knowledge graphs [14].

There are many examples of AIS algorithms [4, 8, 26, 27] for general purpose Monte Carlo integration (see review [3]). However, these methods are ill-suited for our application as they assume a continuous space without constraints on the proposal (see Remark 3).

Finally, we note that label-efficient evaluation may be viewed as a counterpart to *active learning*, as both are concerned with reducing labeling requirements. There is a large body of literature concerning active learning—we refer the reader to surveys [15, 37]. However whereas active learning aims to find a model with low bounded risk using actively-sampled training data, active evaluation aims to estimate risk using actively-sampled test data for *any model*.

[6]These can be viewed as a sub-family of generalized measures with the following correspondence $\ell(x, y) = w(x, y, f(x))[\ell(f(x), y), 1]^\intercal$, and $g(R) = R_1/R_2$.

7 CONCLUSION

We have proposed a framework for online supervised evaluation, which aims to minimize labeling efforts required to achieve precise, asymptotically-unbiased performance estimates. Our framework is based on adaptive importance sampling, with variance-minimizing proposals that are refined adaptively based on an online model of $p(y|x)$. Under verifiable conditions on the chosen performance measure and the model, we proved strong consistency (asymptotic unbiasedness) of the resulting performance estimates and a central limit theorem. We instantiated our framework to evaluate classifiers using deterministic or stochastic human annotators. Our approach based on a hierarchical Dirichlet-tree model, achieves superior or competitive performance on all but one of seven test cases.

ACKNOWLEDGMENTS

N. Marchant acknowledges the support of an Australian Government Research Training Program Scholarship. B. Rubinstein acknowledges the support of Australian Research Council grant DP150103710.

REFERENCES

[1] Paul N. Bennett and Vitor R. Carvalho. 2010. Online Stratified Sampling: Evaluating Classifiers at Web-scale. In *CIKM*. 1581–1584. https://doi.org/10.1145/1871437.1871677

[2] Jon Louis Bentley. 1975. Multidimensional Binary Search Trees Used for Associative Searching. *Commun. ACM* 18, 9 (Sept. 1975), 509–517. https://doi.org/10.1145/361002.361007

[3] M. F. Bugallo, V. Elvira, L. Martino, D. Luengo, J. Miguez, and P. M. Djuric. 2017. Adaptive Importance Sampling: The past, the present, and the future. *IEEE Signal Processing Magazine* 34, 4 (July 2017), 60–79. https://doi.org/10.1109/MSP.2017.2699226

[4] Olivier Cappé, Randal Douc, Arnaud Guillin, Jean-Michel Marin, and Christian P. Robert. 2008. Adaptive importance sampling in general mixture classes. *Statistics and Computing* 18, 4 (01 Dec. 2008), 447–459. https://doi.org/10.1007/s11222-008-9059-x

[5] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. ACM, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785

[6] William G. Cochran. 1977. *Sampling Techniques* (3rd ed.). Wiley, New York.

[7] Gordon V. Cormack, Christopher R. Palmer, and Charles L. A. Clarke. 1998. Efficient Construction of Large Test Collections. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Melbourne, Australia) *(SIGIR '98)*. Association for Computing Machinery, New York, NY, USA, 282–289. https://doi.org/10.1145/290941.291009

[8] Jean-Marie Cornuet, Jean-Michel Marin, Antonietta Mira, and Christian P. Robert. 2012. Adaptive Multiple Importance Sampling. *Scandinavian Journal of Statistics* 39, 4 (2012), 798–812. https://doi.org/10.1111/j.1467-9469.2011.00756.x

[9] Tore Dalenius and Joseph L. Hodges. 1959. Minimum Variance Stratification. *J. Amer. Statist. Assoc.* 54, 285 (March 1959), 88–101. https://doi.org/10.1080/01621459.1959.10501501

[10] Samuel Y. Dennis. 1996. A Bayesian analysis of tree-structured statistical decision problems. *Journal of Statistical Planning and Inference* 53, 3 (1996), 323–344. https://doi.org/10.1016/0378-3758(95)00112-3

[11] Gregory Druck and Andrew McCallum. 2011. Toward Interactive Training and Evaluation. In *CIKM* (New York, NY, USA). 947–956. https://doi.org/10.1145/2063576.2063712

[12] W. Feller. 1968. *An Introduction to Probability Theory and Its Applications, Volume 1* (3rd ed.). Wiley.

[13] W. Feller. 1971. *An Introduction to Probability Theory and Its Applications, Volume 2* (2nd ed.). Wiley.

[14] Junyang Gao, Xian Li, Yifan Ethan Xu, Bunyamin Sisman, Xin Luna Dong, and Jun Yang. 2019. Efficient Knowledge Graph Accuracy Evaluation. *Proc. VLDB Endow.* 12, 11 (July 2019), 1679–1691. https://doi.org/10.14778/3342263.3342642

[15] R. A. Gilyazev and D. Yu. Turdakov. 2018. Active Learning and Crowdsourcing: A Survey of Optimization Methods for Data Labeling. *Programming and Computer Software* 44, 6 (Nov. 2018), 476–491. https://doi.org/10.1134/S0361768818060142

[16] Manfred Jaeger. 2005. Ignorability in Statistical and Probabilistic Inference. *J. Artif. Int. Res.* 24, 1 (Dec. 2005), 889–917.

[17] Evangelos Kanoulas. 2016. *A Short Survey on Online and Offline Methods for Search Quality Evaluation*. Springer International Publishing, Cham, 38–87. https://doi.org/10.1007/978-3-319-41718-9_3

[18] Mohammed Khalilia, Sounak Chakraborty, and Mihail Popescu. 2011. Predicting disease risks from highly imbalanced data using random forest. *BMC Medical Informatics and Decision Making* 11, 1 (2011), 13 pages. https://doi.org/10.1186/1472-6947-11-51

[19] Hanna Köpcke, Andreas Thor, and Erhard Rahm. 2010. Evaluation of Entity Resolution Approaches on Real-world Match Problems. *PVLDB* 3, 1 (2010), 484–493. https://doi.org/10.14778/1920841.1920904

[20] Dan Li and Evangelos Kanoulas. 2017. Active Sampling for Large-scale Information Retrieval Evaluation. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (Singapore, Singapore) *(CIKM '17)*. ACM, New York, NY, USA, 49–58. https://doi.org/10.1145/3132847.3133015

[21] Neil G. Marchant and Benjamin I. P. Rubinstein. 2017. In Search of an Entity Resolution OASIS: Optimal Asymptotic Sequential Importance Sampling. *Proc. VLDB Endow.* 10, 11 (Aug. 2017), 1322–1333. https://doi.org/10.14778/3137628.3137642

[22] Neil G. Marchant and Benjamin I. P. Rubinstein. 2021. Needle in a Haystack: Label-Efficient Evaluation under Extreme Class Imbalance. arXiv:2006.06963 [cs.LG]

[23] Jean-Michel Marin, Pierre Pudlo, and Mohammed Sedki. 2019. Consistency of adaptive importance sampling and recycling schemes. *Bernoulli* 25, 3 (Aug. 2019), 1977–1998. https://doi.org/10.3150/18-BEJ1042

[24] Tom Minka. 1999. *The Dirichlet-tree Distribution*. Technical Report. Justsystem Pittsburgh Research Center.

[25] Barzan Mozafari, Purna Sarkar, Michael Franklin, Michael Jordan, and Samuel Madden. 2014. Scaling up Crowd-Sourcing to Very Large Datasets: A Case for Active Learning. *Proc. VLDB Endow.* 8, 2 (Oct. 2014), 125–136. https://doi.org/10.14778/2735471.2735474

[26] Man-Suk Oh and James O. Berger. 1992. Adaptive importance sampling in monte carlo integration. *Journal of Statistical Computation and Simulation* 41, 3-4 (1992), 143–168. https://doi.org/10.1080/00949659208810398

[27] François Portier and Bernard Delyon. 2018. Asymptotic optimality of adaptive importance sampling. In *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.). Curran Associates, Inc., 3138–3148.

[28] Porto Seguro. 2017. Porto Seguro's Safe Driver Prediction. https://www.kaggle.com/c/porto-seguro-safe-driver-prediction. Accessed: Dec 2019.

[29] A. D. Pozzolo, O. Caelen, R. A. Johnson, and G. Bontempi. 2015. Calibrating Probability with Undersampling for Unbalanced Classification. In *2015 IEEE Symposium Series on Computational Intelligence*. 159–166. https://doi.org/10.1109/SSCI.2015.33

[30] Chandan K. Reddy and Bhanukiran Vinzamuri. 2014. *A Survey of Partitional and Hierarchical Clustering Algorithms* (1st ed.). Chapman & Hall/CRC, 87–110. https://doi.org/10.1201/9781315373515-4

[31] RIDDLE 2003. Duplicate Detection, Record Linkage, and Identity Uncertainty: Datasets. http://www.cs.utexas.edu/users/ml/riddle/data.html. Accessed: Dec 2016.

[32] Reuven Y. Rubinstein and Dirk P. Kroese. 2016. *Simulation and the Monte Carlo Method*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118631980

[33] Christoph Sawade, Niels Landwehr, Steffen Bickel, and Tobias Scheffer. 2010. Active Risk Estimation. In *Proceedings of the 27th International Conference on International Conference on Machine Learning* (Haifa, Israel) *(ICML'10)*. Omnipress, Madison, WI, USA, 951–958.

[34] Christoph Sawade, Niels Landwehr, and Tobias Scheffer. 2010. Active Estimation of F-Measures. In *Advances in Neural Information Processing Systems 23*, J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta (Eds.). Curran Associates, Inc., 2083–2091.

[35] Tobias Schnabel, Adith Swaminathan, Peter I. Frazier, and Thorsten Joachims. 2016. Unbiased Comparative Evaluation of Ranking Functions. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval* (Newark, Delaware, USA) *(ICTIR '16)*. ACM, New York, NY, USA, 109–118. https://doi.org/10.1145/2970398.2970410

[36] Erik Schultheis, Mohammadreza Qaraei, Priyanshu Gupta, and Rohit Babbar. 2020. Unbiased Loss Functions for Extreme Classification With Missing Labels. arXiv:2007.00237 [stat.ML]

[37] Burr Settles. 2009. *Active Learning Literature Survey*. Technical Report. University of Wisconsin-Madison Department of Computer Sciences. http://digital.library.wisc.edu/1793/60660

[38] A. W. van der Vaart. 1998. *Delta Method*. Cambridge University Press, 25—-34. https://doi.org/10.1017/CBO9780511802256.004

[39] Wei Wei, Jinjiu Li, Longbing Cao, Yuming Ou, and Jiahang Chen. 2013. Effective detection of sophisticated online banking fraud on extremely imbalanced data. *World Wide Web* 16, 4 (2013), 449–475. https://doi.org/10.1007/s11280-012-0178-0

[40] P. Welinder, M. Welling, and P. Perona. 2013. A Lazy Man's Approach to Benchmarking: Semisupervised Classifier Evaluation and Recalibration. In *CVPR*. 3262–3269. https://doi.org/10.1109/CVPR.2013.419

A PROOF OF THEOREM 1

First we prove that $\hat{R}_N^{\text{AIS}} \xrightarrow{\text{a.s.}} R$ using a strong law of large numbers (SLLN) for martingales [13, p. 243]. Consider the martingale

$$Z_N = N(\hat{R}_N^{\text{AIS}} - R) = \sum_{j=1}^{N} \left\{ \frac{p(X_j)}{q_{j-1}(X_j)} \ell(X_j, Y_j) - R \right\}$$

and let $\delta_{j,i} = \frac{p(X_j)}{q_{j-1}(X_j)} \ell_i(X_j, Y_j) - R_i$ denote the i-th component of the j-th contribution to Z_N. Since X_j is drawn from $q_{j-1}(x)$ and $q_{j-1}(x) > 0$ wherever $p(x)\|\ell(x,y)\| \neq 0$, it follows that $\mathbb{E}[\delta_{j,i}|\mathcal{F}_{j-1}] = 0$. In addition, we have

$$\sum_{j=1}^{\infty} \frac{\mathbb{E}\left[\delta_{j,i}^2\right]}{j^2} = \sum_{j=1}^{\infty} \frac{1}{j^2} \left\{ \mathbb{E}\left[\frac{p(X_j)^2}{q_{j-1}(X_j)^2} \ell_i(X_j, Y_j)^2 \right] + R_i^2 \right\} < \infty,$$

where the inequality follows from the boundedness of $\ell(x,y)$ and (2). Thus the conditions of the SLLN are satisfied and we have $\frac{1}{N}\sum_{j=1}^{N} \delta_{i,j} \xrightarrow{\text{a.s.}} 0 \implies \hat{R}_N^{\text{AIS}} \xrightarrow{\text{a.s.}} R$. Now the continuous mapping theorem states that

$$\hat{R}_N^{\text{AIS}} \xrightarrow{\text{a.s.}} R \quad \implies \quad g(\hat{R}_N^{\text{AIS}}) \xrightarrow{\text{a.s.}} g(R),$$

provided R is not in the set of discontinuity points of g. This condition is satisfied by assumption.

B PROOF OF THEOREM 2

The CLT of Portier and Delyon [27] implies that $\sqrt{N}(\hat{R}^{\text{AIS}} - R) \Rightarrow \mathcal{N}(0, V_\infty)$, provided two conditions hold. The first condition of their CLT holds by assumption (3) and the second condition holds by the boundedness of the loss function and (4). The multivariate delta method [38] then implies that $\sqrt{N}(g(\hat{R}^{\text{AIS}}) - g(R)) \Rightarrow \mathcal{N}(0, \mathrm{D}g(R) V_\infty \mathrm{D}g(R)^\intercal)$, since g is assumed to be differentiable at R in Definition 1.

C PROOF OF PROPOSITION 1

We want to find the proposal q that minimizes the total asymptotic variance $\operatorname{tr}\Sigma$. We can express this as a functional optimization problem:

$$\begin{aligned} \min_{q} \quad & \int \frac{c(x)}{q(x)} \mathrm{d}x \\ \text{s.t.} \quad & \int q(x)\, \mathrm{d}x = 1, \end{aligned} \tag{6}$$

where $c(x) = p(x)^2 \int \|\mathrm{D}_g(R)\, \ell(x,y)\|_2^2\, p(y|x) \mathrm{d}y$.

Using the method of Lagrange multipliers, Sawade et al. [34] show that the solution to (6) is $q^\star(x) \propto \sqrt{c(x)}$. This yields the required result.

D CONDITIONS FOR ACHIEVING ZERO TOTAL ASYMPTOTIC VARIANCE

In typical applications of IS, the asymptotically-optimal proposal can achieve zero variance. This is not guaranteed in our application, since we do not have complete freedom in selecting the proposal (see Remark 3). Below we provide sufficient conditions under which the total asymptotic variance of $\hat{G}_N^{\text{AIS}}$ can be reduced to zero.

Proposition 4. *Suppose the oracle is deterministic (i.e. $p(y|x)$ is a point mass for all x) and the generalized measure is such that* $\operatorname{sign}(\ell(x,y) \cdot \nabla g_l(R))$ *is constant for all* $(x,y) \in \mathcal{X} \times \mathcal{Y}$ *and* $l \in \{1, \ldots, m\}$. *Then the asymptotically-optimal proposal achieves* $\operatorname{tr}\Sigma = 0$.

Proof. From Proposition 1, the asymptotically optimal proposal achieves a total variance of

$$\operatorname{tr}\Sigma = \left(\int v(x)\, \mathrm{d}x \right)^2 - \|\mathrm{D}_g(R)\, R\|_2^2. \tag{7}$$

We evaluate the two terms in this expression separately. Using the fact that $p(y|x) = \mathbf{1}_{y=y(x)}$, the first term becomes

$$\begin{aligned} \left(\int v(x)\, dx \right)^2 &= \left(\int \|\mathrm{D}_g(R)\, \ell(x, y(x))\|_2\, p(x)\, \mathrm{d}x \right)^2 \\ &\leq \left(\int \|\mathrm{D}_g(R)\, \ell(x, y(x))\|_1\, p(x)\, \mathrm{d}x \right)^2 \\ &= \left(\sum_{l=1}^{m} \int \ell(x, y(x)) \cdot \nabla g_l(R)\, p(x)\, \mathrm{d}x \right)^2. \end{aligned}$$

The second line follows by application of the inequality $\|x\|_2 \leq \|x\|_1$, and the third line follows by assumption. For the second term we have

$$\begin{aligned} \|\mathrm{D}_g(R)\, R\|_2^2 &= \left\| \int \mathrm{D}_g(R)\, \ell(x, y(x))\, p(x)\, \mathrm{d}x \right\|_2^2 \\ &= \sum_{l=1}^{m} \left(\int \ell(x, y(x)) \cdot \nabla g_l(R)\, p(x)\, \mathrm{d}x \right)^2 \\ &\geq \left(\sum_{l=1}^{m} \int \ell(x, y(x)) \cdot \nabla g_l(R)\, p(x)\, \mathrm{d}x \right)^2, \end{aligned}$$

by application of Jensen's inequality. Subtracting the second term from the first, we have $\operatorname{tr}\Sigma \leq 0$. □

E PROOF OF PROPOSITION 2

Before proving the proposition, we establish a useful corollary.

Corollary 1. *Suppose the generalized measure G is defined with respect to a finite input space $\mathcal{X}$ (e.g. a finite pool of test data).*

(i) If the support of proposal $q_j(x,y) = q_j(x)p(y|x)$ is a superset of $\{x, y \in \mathcal{X} \times \mathcal{Y} : p(x,y)\|\ell(x,y)\| \neq 0\}$ for all $j \geq 0$, then Theorem 1 holds.

(ii) If in addition $q_j(x) \xrightarrow{\text{a.s.}} q_\infty(x)$ pointwise in x, then Theorem 2 holds.

Proof. For the first statement, we check conditions (2) and (4) of Theorem 1. Let $\mathcal{Q}_j \subset \mathcal{X}$ be the support of $q_j(x)$ and let $\delta_j = \inf_{x \in \mathcal{Q}_j} q_j(x) > 0$. For $\eta \geq 0$ we have

$$\begin{aligned} \mathbb{E}\left[\left(\frac{p(X_j)}{q_{j-1}(X_j)} \right)^{2+\eta} \middle| \mathcal{F}_{j-1} \right] &= \int \sum_{x \in \mathcal{Q}_{j-1}} \frac{p(x)^{2+\eta} q_{j-1}(x) p(y|x)}{q_{j-1}(x)^{2+\eta}}\, \mathrm{d}y \\ &\leq \left(\frac{1}{\delta_j} \right)^{1+\eta} < \infty. \end{aligned}$$

For the second statement, we must additionally check condition (3) regarding the convergence of V_j. Letting

$$f_j(x,y) = \left(\frac{p(x)}{q_j(x)} \ell(x,y) - R \right) \left(\frac{p(x)}{q_j(x)} \ell(x,y) - R \right)^\intercal q_j(x) p(y|x),$$

we can write $V_j = \int \sum_{x \in \mathcal{Q}_j} f_j(x,y)\,\mathrm{d}y$. By the a.s. pointwise convergence of $q_j(x)$ and the continuous mapping theorem, we have $f_j(x,y) \to f_\infty(x,y)$ a.s. pointwise in x and y. Now observe that

$$\begin{aligned}\|f_j(x,y)\|_2 &= q_j(x,y)\left\|\frac{p(x)}{q_j(x)}\ell(x,y) - R\right\|_2^2 \\ &\le q_j(x,y)\left(\frac{p(x,y)^2}{q_j(x,y)^2}\|\ell(x,y)\|_2^2 + \|R\|_2^2\right) \\ &\le p(y|x)\left(\frac{1}{\epsilon^2}\|\ell(x,y)\|_2^2 + \|R\|_2^2\right) = h(x,y).\end{aligned}$$

It is straightforward to show that $\int \sum_{x \in \mathcal{Q}_j} h(x,y)\,\mathrm{d}y < \infty$ using the boundedness of $\ell(x,y)$ (see Definition 1). Thus we have $V_j \to V_\infty$ by the dominated convergence theorem. □

We can now prove Proposition 2 by showing that the conditions of Corollary 1 hold. We focus on the case of a deterministic oracle—the proof for a stochastic oracle follows by a similar argument.

First we examine the support of the sequence of proposals. At stage t, the proposal can be expressed as

$$q_t(x) = \frac{v_t(x)}{\sum_{x \in \mathcal{X}} v_t(x)} \quad \text{with}$$

$$v_t(x) = p(x)\int \max\left\{\|\mathrm{D}_g(\hat{R}_t)\,\ell(x,y)\|_2, \epsilon_t \mathbf{1}_{\|\ell(x,y)\|\neq 0}\right\}\,\pi_t(y|x)\,\mathrm{d}y.$$

Observe that

$$v_t(x) \ge \epsilon_t\, p(x)\int \mathbf{1}_{\|\ell(x,y)\|\neq 0}\,\pi_t(y|x)\,\mathrm{d}y$$

and

$$\begin{aligned}v_t(x) &\le p(x)\int \{\epsilon_t + \|\mathrm{D}_g(\hat{R}_t)\|_2\,\|\ell(x,y)\|_2\}\pi_t(y|x)\,\mathrm{d}y \\ &\le p(x)\left(\epsilon_t + d^2 K \sup_{x,y \in \mathcal{X}\times\mathcal{Y}} \|\ell(x,y)\|_\infty\right) \\ &\le Cp(x)\end{aligned}$$

where $C < \infty$ is a constant. The upper bound follows from the boundedness of $\ell(x,y)$ (see Definition 1), the boundedness of ϵ_t, and the boundedness of the Jacobian. Since

$$\sum_{x \in \mathcal{X}} v_t(x) \ge \epsilon_t \sum_{x \in \mathcal{X}} p(x)\int \mathbf{1}_{\|\ell(x,y)\|\neq 0}\pi_t(y|x)\,\mathrm{d}y > 0$$

by assumption and $v_t(x)$ is bounded from above, we conclude that $q_t(x)$ is a valid distribution for all $t \ge 0$. The lower bound on $v_t(x)$ implies that the support of $q_t(x,y) = q_t(x)p(y|x)$ is

$$\begin{aligned}&\{(x,y) \in \mathcal{X}\times\mathcal{Y} : p(x,y)\pi_t(y|x)\|\ell(x,y)\| \neq 0\} \\ &\quad \supseteq \{(x,y) \in \mathcal{X}\times\mathcal{Y} : p(x,y)\|\ell(x,y)\| \neq 0\}.\end{aligned}$$

The inequality follows from the fact that the support of $\pi_t(y|x)$ includes the support of $p(y|x) = \mathbf{1}_{y=y(x)}$. Thus $q_t(x)$ has the required support for all $t \ge 0$.

Next we prove that the sequence of proposals converges a.s. pointwise in x. Given that $\hat{R}_t \xrightarrow{\text{a.s.}} \hat{R}_\infty$ and $\pi_t(y|x) \xrightarrow{\text{a.s.}} \pi_\infty(y|x)$ pointwise in x, one can show by application of the continuous mapping theorem and dominated convergence theorem that

$$v_t(x) \xrightarrow{\text{a.s.}} v_\infty(x) = p(x)\int \|\mathrm{D}_g(R_\infty)\,\ell(x,y)\|_2\pi_\infty(y|x)\,\mathrm{d}y$$

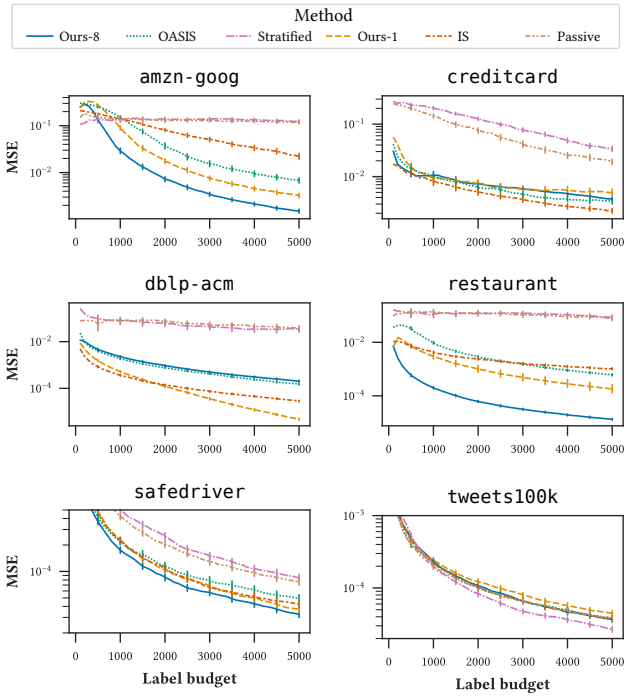

Figure 5: MSE of estimated F1 score over 1000 repeats as a function of consumed label budget. 95% bootstrap confidence intervals are included.

pointwise in x. By application of the continuous mapping theorem, we then have that $q_t(x) \xrightarrow{\text{a.s.}} q_\infty(x) = \frac{v_\infty(x)}{\sum_{x \in \mathcal{X}} v_\infty(x)}$. Thus Theorems 1 and 2 hold. Furthermore, if $\hat{R}_\infty = R$ and $\pi_\infty(y|x) = \mathbf{1}_{y=y(x)}$, then $q_\infty(x)$ is equal to the asymptotically-optimal proposal $q^\star(x)$ as defined in (5). □

F ADDITIONAL CONVERGENCE PLOTS

We provide additional convergence plots for estimating F1 score in Figure 5, which cover six of the seven datasets listed in Table 2. The convergence plot for `abt-buy` is featured in the main paper in Figure 2. The biased sampling methods (`Ours-8`, `Ours-1`, `OASIS` and `IS`) converge significantly more rapidly than `Passive` and `Stratified` for five of the six datasets. `tweets100k` is an exception because it is the only dataset with well-balanced classes. Of the biased sampling methods, `Ours-8` performs best on two of the six datasets (`amzn-goog` and `restaurant`) and equal-best on one (`safedriver`). `Ours-1` performs best on one dataset (`dblp-acm`) and equal-best on one (`safedriver`), while `IS` performs best on one dataset (`creditcard`). In general, we expect `IS` to perform well when the oracle response $p(y|x)$ is already well-approximated by the model under evaluation. When this is not the case, the adaptive methods are expected to perform best as they produce refined estimates of $p(y|x)$ during sampling.