# ML Assisted Sampling for Populations With Large Class Imbalances

Emanuel Strauss
Facebook
1 Hacker Way
Menlo Park, CA 94025
eman@fb.com

Spencer Beecher
Facebook
1 Hacker Way
Menlo Park, CA 94025
spencebeecher@fb.com

Daniel Olmedilla
Facebook
1 Hacker Way
Menlo Park, CA 94025
doc@fb.com

## ABSTRACT

There are many cases in which one would wish to measure the frequency of a rare class of objects or events in a dataset which is large enough that it is impractical to generate labels for the full population. When the minority class is infrequent enough, the number of labels required to achieve the desired precision using a random sampling of the population is also prohibitive, and a common solution is to perform stratified sampling using some scoring function to build strata in which the minority class is enhanced.

In this paper, we will cover a method which extends the idea of stratified sampling by computing sampling weights to uniformly sample across a range of scores or regions. We will show how such weights can be calculated using a machine learning classifier and demonstrate that along with increasing the measurement precision by as much as 50%, and increasing the minority class yield by up to 15x, the method also helps to generalize the dataset. Using a combination of synthetic data and experiences gained testing the method with data from Facebook's Ad Review System, we will show how this generalizeability allows the re-use of the data for a host of applications, from enabling teams performing deep dives to gain further insights, to calibrating classification models.

## CCS Concepts

**•Mathematics of computing → Probability and statistics; •Computing methodologies → Machine learning; •General and reference →** *Cross-computing tools and techniques; Empirical studies;*

## Keywords

class imbalance, sampling, machine learning, data science

## 1. INTRODUCTION



*KDD '16 August 13–17, 2016, San Francisco, CA, USA*



DOI: 10.475/123_4

The more than 1,5 billion monthly (1 billion daily) active people using Facebook have access to content such as pages' updates, group posts, products and ads. Keeping a satisfying experience requires that the content shown to them is of the highest quality. Throughout this paper, we use *"not-suitable"* to describe low quality creatives and content that violate Facebook's policies [1]. The mission of the Integrity team at Facebook consists of identifying and blocking not-suitable content, at scale, before they enter the matching and ranking algorithms to be potentially displayed to people on our platform. We combine Machine Learning models and Human Computation to detect not-suitable content, block its distribution within the platform, and notify the content creator with hints about how to remedy the issue.

Many platforms are exposed to not-suitable content, the type and volumes of which varies per domain. In this paper we will mainly focus on the domain of Advertisements. Given the very high-quality bar Facebook requires, it is one of the most restrictive domains. Although one might think that not-suitable ads are mainly created by bad actors trying to exploit the system, that is not the only source. There are three main types of advertisers generating low-quality content:

- advertisers unintentionally producing not-suitable ads (e.g., new advertisers that have not read or fully understood our policy guidelines [1]),
- advertisers intentionally using not-suitable content in order to increase the performance of their ads (e.g., ads which are click-baity or overly sexually suggestive),
- and malicious advertisers trying to bypass Facebook's review controls in order to get quick benefits (e.g., scam people or promote illegal items)

Not all not-suitable ads are created equal. In order to maximize the impact of our efforts, we focus on the number of impressions (i.e. the number of times ads are shown) accumulated by the not-suitable ads, rather that on the absolute number of not-suitable ads themselves. Essentially, we wish to minimize the number of experiences not-suitable ads generate. For instance, blocking a not-suitable ad that could potentially be shown to millions of people is more important than blocking another equally not-suitable ad that would only be shown to hundreds. For this reason, we bias our sampling by the number of ad impressions. This helps us better track the impact of our work and at the same time, helps provide key insights into what types of content our integrity systems haven't yet caught so that we can retrain

and update our models. The expert reviewers assigned the task of labeling these samples are our most accurate ones, but also the most limited resource in the whole process.

An extreme class imbalance poses a real challenge in constructing the samples. Existing sampling schemes make suboptimal use of our high-accuracy human reviewers, spending most of their review time on good quality ads. At the same time, the small yield in newly identified not-suitable ads limits our ability to identify new trends and therefore reduces our responsiveness in building effective models to automatically identify such patterns.

In this paper, we will propose a method of machine learning assisted sampling, which uses the output of a model predicting whether an ad is not-suitable, as well as the number of impressions these ads are generating, in order to make the sampling process much more effective. We sample 6x more not-suitable samples with ML assisted sampling than with random sampling, 20% more than with stratified sampling with no increase in the error of the estimates. The lift in yield increases to 10-15x when applied to other domains at Facebook where the class imbalance is even more extreme. The benefits include more accurate metrics (we would require 6-15x more reviewers to get the same level of results), better use of our reviewers' time, and much more data to identify new patterns not yet caught by our review system and modeling efforts.

The paper is organized as follows: Section 2 introduces the ads integrity review process followed by Facebook and why having optimal sampling strategies is important. Relevant work related to our sampling scheme is described in Section 3. The main sampling schemes we compare against as well as our ML assisted sampling are explained in detail in Section 4 while Section 5 presents the results. Finally, additional considerations are described in Section 6 and Section 7 summarizes the contribution of this paper and highlights some future work.

# 2. ADS INTEGRITY REVIEW

Facebook's Advertisement system is designed to foster a positive experience in which ads are shown to the people most likely to care about the content. Unfortunately, a small number of the submitted ads are not suitable to be shown to the people using Facebook: some contain low quality creatives, some are spammy or misleading, others may run afoul of local customs and laws, others still may prey on people's emotions or contain excessively shocking content. To minimize the risk of such undesirable content reaching people on our platform, Facebook developed and maintains an Ad Review System which processes each and every ad to identify and block not-suitable content, as well as to notify its creator and help her to fix it.

## 2.1 Facebook Review in a Nutshell

Millions of ads are submitted to Facebook and the Integrity team performs a mix of manual and automatic reviews on each and every one of them. The lion's share of the Integrity team's efforts are dedicated to classify these ads and prevent not-suitable ones from being shown to people on Facebook (similarly to other large ad-serving platforms [2]). However, given such high volumes, it is infeasible to have them all reviewed by human experts, so we use a combination of Machine Learning and Human Computing to scale the process: all ads are scored by hundreds of supervised and unsupervised ML models, classified by a complex set of rule based engines, and only those most likely to contain low-quality content are reviewed by humans (who must ensure the ad complies with the high standards expected from advertisers on Facebook's platform). Given the limited nature of our human review time, we aim to select the right ads to be manually reviewed in order a) to get the cleanest ecosystem of live ads and b) to acquire more labeled data for the models. In order to make the process more efficient, given that ads might reuse creatives (e.g., same image across different ads), the system considers the full ad, as well as each of the components (e.g., title, description, image, video, etc.) individually.

In a nutshell, for each ad submitted at Facebook, a high-level overview of our review system is as follows:

1. The ad is split into each of its components
2. A large number of machine learning models score each of the components and the ad as a whole.
3. Based on the results of the scores, the ad is marked for:
   - Automatic acceptance/rejection
   - Manual review by a human reviewer
4. Depending on the outcome, the ad is approved for delivery and passed to our matching and ranking algorithms, or rejected, at which point we notify its creator and provide guidance on how to improve it.

The top priorities are to improve the recall of our models (reduce to a minimum the amount of not-suitable ads that go live), improve our precision and therefore reduce potential false positives, as well as to shorten overall the time between when an ad is submitted and the notification of review results is sent out.

## 2.2 The Importance of Sampling

To gauge the performance of the Ad Review System, identify new trends in not-suitable ads, measure the impact of new efforts, and estimate where operational resources should be allocated, it is important that we are able to assess the health of the whole review process. This includes how many undesirable ads are slipping through the cracks, the quality level of the ads, how many people are being exposed to them, and how frequently.

We want to track the number of not-suitable ads served to people on Facebook as well as to understand their patterns, but given the high volumes we cannot manually look at all live ads to measure this quantity. Given this limitation, our main metrics and KPIs[1] are measured by performing probability sampling [3] to create a small but representative sample of the ads which are currently being served on Facebook. Those ads are sent to our most accurate and high-quality reviewers for labeling, and typically each ad is seen by more than one reviewer to ensure accuracy. From these labeled data we construct an unbiased estimator [4] of the fraction of impressions served that contained not-suitable content in the parent population. As the labeling is done by expert human reviewers the time they spend on this is especially

[1]Key Performance Indicators

precious, entails a lot of operational effort, and the excessive review of good quality content (as opposed to more not-suitable ads) would be a poor use of resources. Additionally, reviewers may suffer from click fatigue. Their accuracy has been observed to decrease when the overwhelming majority of the content they review is of good quality (e.g., if only one ad out of hundreds of reviews was not-suitable vs. having a more balanced distribution). Ideally, the dataset we generate would:

- produce accurate estimates of the fraction of impressions generated by undesirable ads,
- make improved usage of the limited time of our expert reviewers,
- maintain reasonable measurement uncertainties,
- identify as many positive samples (not-suitable ads) as possible to better understand trends
- and be re-usable for a wide range of related tasks and measurements.

There are many complications introduced by the extreme class imbalance and the volume of live ads. To get a fixed uncertainty on the estimate, the required number of labeled samples grows in quadrature as a function of the size of the class imbalance. These numbers can increase again if we consider plurality (sending each ad to multiple reviewers to improve manual review accuracy). The inefficiencies of the process quickly blow up when operating at the scale of our system.

# 3. SAMPLING SCHEMES AND RELATED WORK

There are several factors to consider when selecting an estimation method. Since we rely on the estimate to understand the fraction of impressions generated by not-suitable ads live and to identify new trends, we require an unbiased estimator with relatively low variance. Without this we would not be able to confidently prioritize our work on areas of greatest impact.

A simple random sampling of ads yields an unnecessarily high variance since impressions per ad is heavy tailed [5]. Sampling with probability proportionate to size (PPS) [6] drastically reduces error associated with the estimate, effectively sampling in impression space. This approach still leaves room for improvement, since not-suitable ads make up a very small fraction of overall impressions. Stratified sampling provides another method for reducing variance if there are differences in sub-populations [3]. One can oversample the sub-populations with the most variance to yield more certain estimates. Importance sampling provides a mechanism to bias our sampling scheme towards not-suitable ads while still fulfilling all our other requirements. Importance sampling is well covered in the literature [7, 8, 9]. In short, the method allows oversampling from an important region while providing an unbiased estimate. Under the right circumstances, importance sampling also reduces the error of the estimate. We will explore this approach further for our domain in Sec. 4.

Recently, Attenberg studied the use of active learning aided by search for adversarial problems with extreme class imbalances [10]. This approach demonstrates how to select new training instances for learning within a similar problem domain. Active learning is not well suited to our problem since it does not provide estimates about the underlying class distribution.

Sawade et. al. propose an extension of importance sampling for understanding model risk [11]. They provide the framework for estimating risk under severe class imbalance across several different loss functions. Their work relies on the model already having a well calibrated estimate of the probability in order to select uncertain instances from the pool. Using model scores, with no assumption on their relationship to the class probability, as we do in this paper, has two primary advantages: First, it allows the use of uncalibrated models, which is particularly important when the labeled data that has been acquired to train the models is not identically distributed with the data on which it will be applied, or if a high-fidelity calibrator is not readily available. Second, the rate of sampling is increased in the low density region where the rate of change between the two classes is greatest. This results in a more accurate measurement of the calibration curves in the cases when the dataset is used to eventually calibrate models.

Finally, Sculley et. al. describe Google's system for detecting adversarial advertisements [2]. They sample from strata over classifier probabilities to obtain an estimate, and we will cover in Sec. 4.2 how ML assisted sampling improves the behavior of the measurement uncertainties relative to this approach.

# 4. ADS INTEGRITY SAMPLING

This section describes some sampling schemes relevant to our problem and presents our ML assisted approach in which a sampling probability is assigned to all ads as a function of the output of a classifier trained to identify potentially not-suitable ads. There are many classes of not-suitable ads, but for the purposes of sampling, we train a one-vs-all classifier.

Given the confidential nature of Facebook's data, and in order to be able to share exact numbers in a reproducible manner, we have used two different datasets when reporting the results in this paper.

To illustrate the different sampling scheme concepts and compare our approach to them, we relied on a synthetic dataset generated to simulate the distribution of model scores one might expect to observe in a population of ads given an extreme class imbalance and imperfect classifiers. For this dataset, we pull "scores" out of an exponentiated Weibull [12] function

$$F(x;\boldsymbol{\theta}) = (1 - e^{-(x/\theta_0)^{\theta_1}})^{\theta_2} \quad (1)$$

where $x$ is the "model score" and $\boldsymbol{\theta}$ are the shape parameters of the function which have been picked to give something approaching the PDF[2] of Ad Review System scores. We then assigned a label, "suitable" or "not-suitable", to each synthetic record so as to achieve a reasonably large class imbalance. There are in actuality many classes of not-suitable ads, but for the purposes of sampling, we build a single one (suitable) vs all (not-suitable) classifier. The class distribution of this synthetic dataset is depicted in a logarithmic plot in Fig. 1, together with the cumulative ratio distribution of not-suitable ads. This dataset was used to demonstrate

[2]Probability Density Function

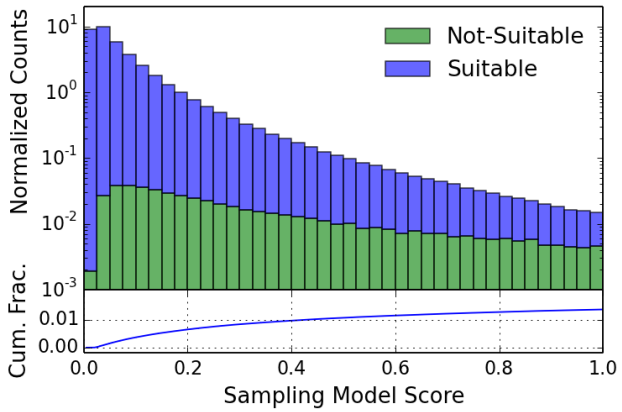


**Figure 1: Distribution of generated model scores and the assigned labels in the synthetic data (top) and the cumulative fraction of minority class elements above the model score (bottom). The data was generated using an exponentiated Weibull function with parameters chosen to reflect the shape of the population in the Ad Review system and the labels were randomly assigned so as to achieve a reasonably large class imbalance.**

the performance of the sampling methods examined in this paper and to share a detailed comparison of each scheme.

In order to evaluate the actual overall impact in our live systems, the methods have also been tested using data and scores of ads taken from parts of the production system. We have provided relative results achieved on real world data to demonstrate the effectiveness of the ML assisted sampling.

## 4.1 Biased Random Sampling and Stratified Sampling

The top-level quantity of interest is the number of Not-Suitable Impressions ($NSI$), the number times a not-suitable ad has been served, and the Fraction of Not-Suitable Impressions ($\widetilde{NSI}$), the number of times a not-suitable ad has been shown divided by the total number of times ads are shown. We wish to use a sampling of live ads with expert labels, where the number of $NSI$ can be estimated as:

$$\mathbb{E}(NSI) = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_i} \tag{2}$$

using a Hansen-Hurwitz [13] estimator with $y_i = I_i l_i$ and $p_i = 1/N$, where $I_i$ is the number of impressions generated by the i$^{\text{th}}$ ad, $l_i$ is the label (0 for suitable, 1 for not-suitable), $n$ is the number of sampled ads, and $N$ is the number of ads in the parent population. The estimator for $\widetilde{NSI}$ is then:

$$\mathbb{E}_{rndm}(\widetilde{NSI}) = \frac{\mathbb{E}_{rndm}(NSI)}{\sum_{i=1}^{N} I_i} = \frac{N}{n}\frac{\sum_{i=1}^{n} I_i l_i}{\sum_{i=1}^{N} I_i} \tag{3}$$

Estimators for the variance of $NSI$ and $\widetilde{NSI}$ are:

$$\mathbb{V}_{rndm}(NSI) = \frac{1}{n}\frac{\sum_{i=1}^{n}(\frac{y_i}{p_i} - \mathbb{E}_{rndm}(NSI))^2}{n-1} = \frac{N^2}{n(n-1)}\sum_{i=1}^{n}(I_i l_i - \langle Il\rangle)^2 \tag{4}$$

$$\mathbb{V}_{rndm}(\widetilde{NSI}) = \frac{\mathbb{V}_{rndm}(NSI)}{\sum_{i=1}^{N} I_i} \tag{5}$$

The variance on the estimator can be reduced by biasing the random sampling such that the sampling probability is proportional to $I_i$. The estimator of $\widetilde{NSI}$ becomes:

$$\mathbb{E}_{bias}(\widetilde{NSI}) = \frac{\sum_{i=1}^{n} l_i}{n} \tag{6}$$

and the estimated variance when sampling this proportion,

$$\mathbb{V}_{bias}(\widetilde{NSI}) = \mathbb{E}_{bias}(\widetilde{NSI})(1 - \mathbb{E}_{bias}(\widetilde{NSI})) \tag{7}$$

Given that $\widetilde{NSI}$ is $\sim 1\%$ in the synthetic data, as explained in previous sections, even this weighted sampling is very inefficient (e.g., reviewers spend the majority of their time reviewing good ads and the sparsity of not-suitable ads makes it hard to slice the sample after labeling to gain further insights into the makeup of not-suitable ads).

In stratified sampling, the parent population is split into non-overlapping groups (strata), which we did here by bining along an ad score related to the likelihood that the ad violates one or more policies. By oversampling from the higher probability strata, we enhanced the number of not-suitable ads collected.

There are methods of optimum allocation [14, 15] which place guarantees on the variance by setting the strata sample size proportional to the variance of the strata, but practically speaking, for large class imbalance problems, the variance is likely to be improved with even a naive uniform allocation.

$$\mathbb{E}_{strata}(\widetilde{NSI}) = \frac{1}{N_{tot}}\sum_{b=1} N_b \mathbb{E}(\widetilde{NSI}_b) \tag{8}$$

With estimated variance,

$$\mathbb{V}_{strata}(\widetilde{NSI}) = \sum_{b=1}(\frac{N_b}{N_{tot}})^2 \frac{N_b - n_b}{N_b}\frac{\mathbb{V}(\widetilde{NSI}_b)}{n_b} \tag{9}$$

where $N_b$ is the number of elements in the b$^{\text{th}}$ strata, $n_b$ is the number of elements sampled from the b$^{\text{th}}$ strata, $N_{tot}$ is the number of elements sampled across all strata, $\mathbb{E}(\widetilde{NSI}_b)$ is $\widetilde{NSI}$ from an estimator within the b$^{\text{th}}$ strata, and $\mathbb{V}(\widetilde{NSI}_b)$ is it's variance.

The impact of stratified sampling depends on several factors, such as:

- the number of strata
- the sample size of each strata,
- and the variance in the quantity per strata.

In this synthetic data, where $\widetilde{NSI}$ is $\sim$1%, the raw number of not-suitable ads with stratified random sampling was increased by a factor of 5 with respect to uniform random sampling.

However, by the nature of the class imbalance and our imperfect classifiers, the PDF of the not-suitable model score is steeply falling. If we define strata along this score, we always find that within any individual strata we are more likely to sample low-likelihood ads. This is highlighted in Fig. 2, which shows the distribution of suitable and not-suitable ads. In the bottom figure we show a histogram of the estimated distribution of suitable and not-suitable ads as a function of the model score, constructed using stratified sampling. In the top figure, we show the raw number of suitable and not-suitable ads for the 7,000 sampled records. It can be easily seen that while the estimate on the top is fairly representative of the true population, the simple random sampling of ads in each strata tends to pick records from the left (i.e., from where there are fewer not-suitables). This results in a saw-toothed distribution, which implies heteroscedasticity in the sampling errors. The variance sometimes changes in large steps, as a function of the not-suitable likelihood so this violates our requirement for low variance.

## 4.2 ML Assisted Sampling

As mentioned in Sec. 2, the experts who label this dataset are a scarce resource, and as such we place a lot of emphasis on a) making the best use of their time, and b) enhancing the potential value and usability of this dataset. Stratified sampling has the benefit of sending more not-suitable ads for human review than pure random sampling, so the experts spend less of their time on the rote task of labeling suitable ads. While powerful when optimized for a specific goal (e.g., minimizing variance of the $\widetilde{NSI}$ measurement), stratified sampling can result in a dataset with limited reusability if, for example, one wanted to slice the dataset into finer strata or make estimates about quantities in the high-variance areas just below a strata edge.

The alternative approach which we propose here is to extend the idea of applying different sampling weights to ads in different strata of the parent population one step further by apply different sampling weights to all ads in the parent population in a continuous manner. By making the sampling probability a smooth function of the inverse of the PDF of the likelihood scores, we can create a sampled dataset that is uniform in that space (see Appendix A for formalism). The estimator becomes the Hansen-Hurwitz estimator of Eqn. 2 where the $y_i$ are again the label $l_i$. To achieve a uniform sampling with respect to the distribution of model scores, the new sampling probabilities $w_i$ are proportional to the impressions of the ads and inversely proportional to the frequency of the model score $s_i$, such that $w_i \propto \frac{1}{\mathrm{PDF}(s_i)}$ and $\sum_{i=1}^{N} w_i = 1$.

$$\mathbb{E}_w(\widetilde{NSI}) = \frac{1}{n}\frac{\sum_{i=1}^{n} I_i l_i w_i^{-1}}{\sum_{i=1}^{N} I_i} \tag{10}$$

$$\mathbb{V}_w(\widetilde{NSI}) = \frac{1}{n(n-1)}\frac{\sum_{i=1}^{n} (I_i l_i w_i^{-1} - \langle Ilw \rangle)^2}{\sum_{i=1}^{N} I_i} \tag{11}$$

As with Eqns. 6 and 7, we can improve the variance of the estimator by biasing the sampling weights proportional to $I_i$, such that $w_i' \propto w_i \frac{I_i}{\sum_{j=1}^{N} I_j}$, and the estimators become:

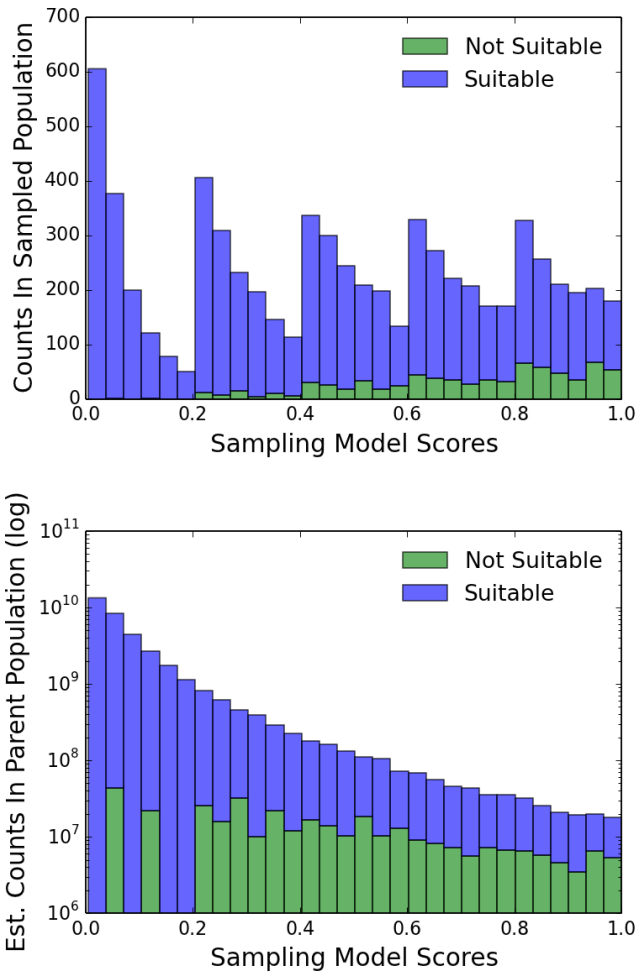


**Figure 2: Stacked histograms of the number of sampled synthetic ads per bin (top) and the number of synthetic ads in the parent population estimated using the sample (bottom) for synthetic ads sampled using stratified sampling with 5 uniformly spaced strata along a not-suitable model score.**

$$\mathbb{E}_{w'}(NSI) = \frac{\sum_{i=1}^{N} I_i}{n}\sum_{i=1}^{n} l_i w_i^{-1} \tag{12}$$

$$\mathbb{E}_{w'}(\widetilde{NSI}) = \frac{\mathbb{E}_{w'}(NSI)}{\sum_{i=1}^{N} I_i} = \frac{1}{n}\sum_{i=1}^{n} l_i w_i^{-1} \tag{13}$$

Note that the impression terms in Eqns. 12 and 13 cancel out and the expectation value takes on a very simple form to compute where the estimator depends on $w_i$, the PDF weighting function, not $w_i'$. The variance becomes:

$$\mathbb{V}_{w'}(NSI) = \frac{(\sum_{j=1}^{N} I_j)^2}{n(n-1)}\sum_{i=1}^{N} (l_i w_i^{-1} - \langle l w^{-1} \rangle)^2 \tag{14}$$

and

$$\mathbb{V}_{w'}(\widetilde{NSI}) = \frac{\mathbb{V}_{w'}(NSI)}{\sum_{i=1}^{N} I_i} \tag{15}$$

There are several reasons to pick this approach over the traditional stratified sampling defined in the previous section. For instance, coming back to one of our stated goals of generalizing the data, having measured the current state of the ecosystem, imagine we now wanted to use this dataset to estimate how many people would be spared a not-suitable ad experience by increasing reviewer head-count. To illustrate this idea, let's say we are currently reviewing 10% (model score of 0.225) of all ads and we wish to know by how much we would reduce $\widetilde{NSI}$ if we reviewed 12% (model score of 0.2) or 15% (model score of 0.16) of all ads. In our synthetic dataset, since 0.2 is right above the edge of one of the strata, and 0.16 is deep into the previous strata, the estimates of the fraction of not-suitable ads missed at the lower threshold have 50% larger uncertainties, as shown in Tab. 1, making it more difficult to justify the need for additional operational resources. In practice, the reliability of such predictions is of course affected by the adversarial nature of the content being classified as bad actors will change their behavior in response to the new ecosystem, but that is outside the scope of this paper.

| | $s \geq 0.225$ | $s \geq 0.2$ | $s \geq 0.16$ |
|---|---|---|---|
| Rand. | $0.074 \pm 0.010$ | $0.069 \pm 0.0090$ | $0.057 \pm 0.0069$ |
| Strat. | $0.074 \pm 0.0047$ | $0.069 \pm 0.0045$ | $0.057 \pm 0.0063$ |
| ML | $0.074 \pm 0.0046$ | $0.068 \pm 0.0047$ | $0.056 \pm 0.0042$ |

**Table 1: Using Randomized, Stratified, and ML Assisted sampling to estimate the $\widetilde{NSI}$ above three thresholds on the model score in the synthetic dataset, with thresholds picked so as to "review" 10%, 12%, or 15% of the ads. Note that while the mean values are quite similar in the three cases, the uncertainty on the estimates vary drastically**

The estimated PDF and raw distribution of sampled ads is shown in Fig. 3, using the same synthetic population of ads. The saw-tooth behavior is gone, replaced by a distribution which is relatively flat as a function of the model score. Furthermore, the estimates of $\widetilde{NSI}$ using ML assisted sampling on the synthetic dataset yielded confidence intervals that were often 20% to 50% tighter than those achieved with stratified sampling using five, equally sampled, strata. The precision of the estimates using stratified sampling outperformed those from ML assisted sampling by 5% to 10% near the left edge of the strata, but on average, the ML assisted sampling produced tighter error bounds.

#### 4.2.1 *Practical Notes On Calculating Sampling Probabilities*

In real world applications, not all distributions can be described by a closed form function. In cases where the functional form under-fits the data, the sampling weights may not result in a sample which is flat with respect to the model scores. This is demonstrated in Fig. 4, where a large number of elements has been injected at 0.3. The goodness of fit using the formula from Eqn. 1 is significantly degraded, and were we to generate sampling weights by evaluating the

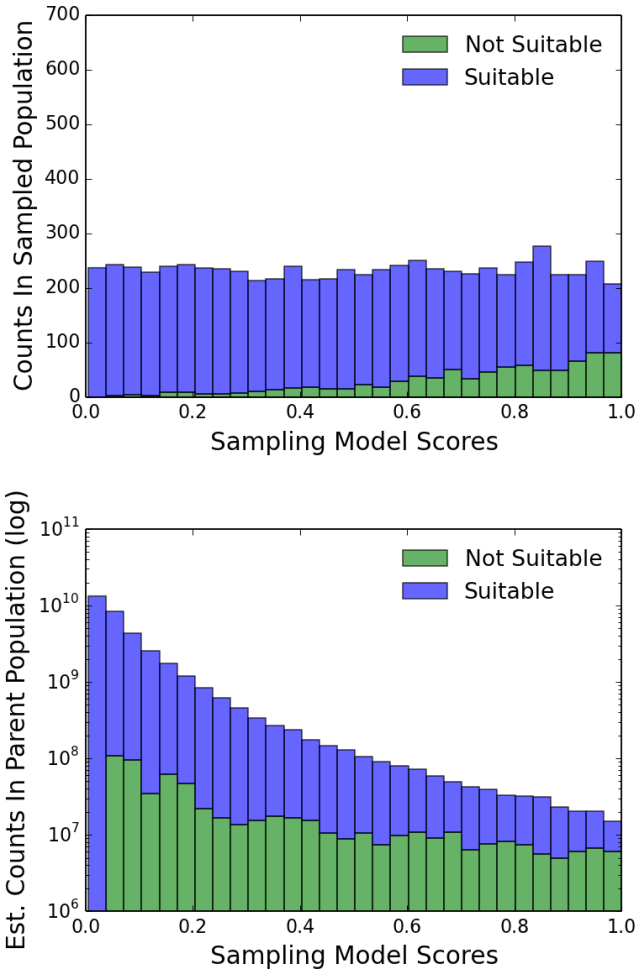


**Figure 3: Stacked histograms of the number of sampled synthetic ads (top) and the number of synthetic ads in the parent population estimated using the sample (bottom) for synthetic ads sampled using ML Assisted sampling along a not-suitable model score.**

function with the coefficients from the fit we would select an excessive number of elements from the peak at 0.3, under-sample the tails, and over-sample the bulk between 0.0 and 0.15.

Such scenarios where a simple function fails to describe the PDF are not particularly far-fetched. It may easily happen in the cases where the distribution is produced by a model which relies heavily on categorical features, the coverage of features is not complete, or the features fed to the models themselves have large discontinuities.

Since we perform batch sampling of the data, a simple solution is to use a non-parametric model to describe the PDF of the parent population for that day. The dashed line in Fig. 4 does a much better job of describing the data, and is achieved by fitting a monotonic cubic spline to the CDF of the parent population [16]. Given the monotonicity of the CDF, a monotonic cubic spline is both fast to fit, and produced robust results for a number of test cases. The derivative of the spline tracks the dominant features of the

PDF, resulting in a sampling which is generally insensitive to distortions in the distribution over time.

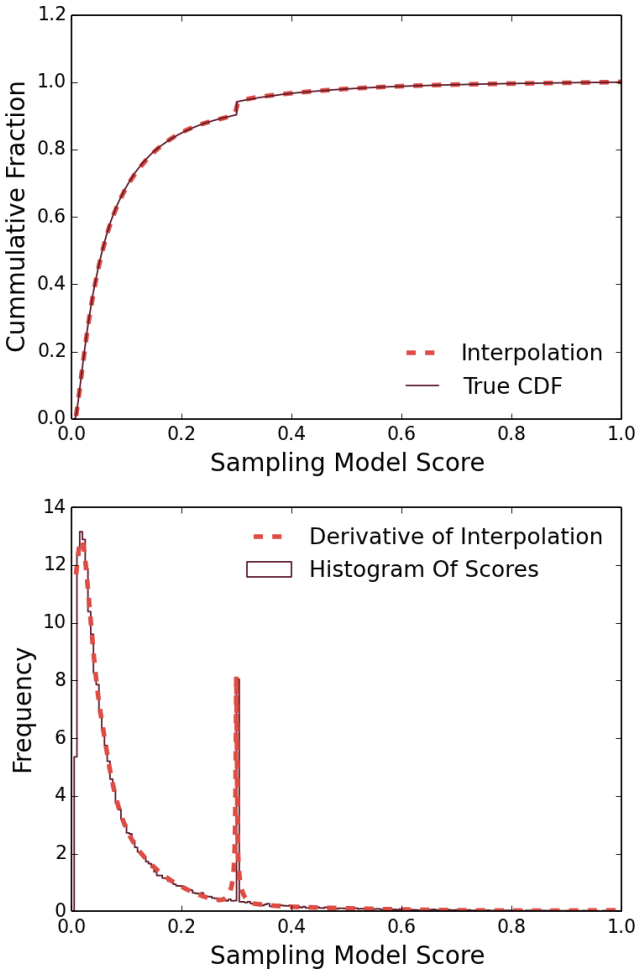


**Figure 4: Describing the PDF by fitting a spline to the CDF of synthetic ad scores (top) and taking it's derivative (bottom) when the distribution cannot be fit by a smooth function. The true distribution is displayed with a solid line, the spline is displayed with a dashed line.**

#### 4.2.2 *Biased ML Sampling*

The $w_i$ weights need not strictly be generated by inverting the PDF. Instead, the $w_i$ can be generated according to a different density function altogether. More generally, if $f(x)$ describes the observed PDF, and $g(x)$ is a normalized PDF which describes the desired sampling density, then we can calculate $w_i = \frac{g(x_i)}{f(x_i)}$. Using weights generated in this way will yield a sample in which the conditional expectation of the measurement is unchanged, though it does have implications on the variance.

This, along with Eqn. 15, gives some insight into designing an optimal allocation schema for ML assisted sampling if calibrated model scores are available, but the reduction in variance may not always be compatible with business goals.

Depending on how the data will be used, it may be the case that a uniform sampling with respect to the PDF is the best approach. For example, if the business needs are more geared towards exploring the low model score region for false negatives, then a $g(x_i)$ with a negative slope which pulls more from the left of the distribution would be preferable. Such decisions must largely be driven by either business decisions, the feasibility of collecting certain data, or a priori knowledge about the properties of the phase space. A uniform sampling with respect to the model score PDF is generally a good place to start and has served us well.

### 4.3 Relative Error Estimate

As an example to better explain some of the improvements of the ML assisted sampling, Fig. 5 shows how ML assisted sampling has the benefit of yielding uncertainties that vary smoothly as a function of the model score. ML assisted sampling offers the lowest errors except right after the strata edges where the stratified sampling is sometimes a bit better. Note that the result of this particular experiment varies with respect to the parameters used for the strata sampling (e.g., number of strata and samples per strata), so it is only intended to illustrate the differences between the sampling schemes, providing a qualitative feel for their properties.

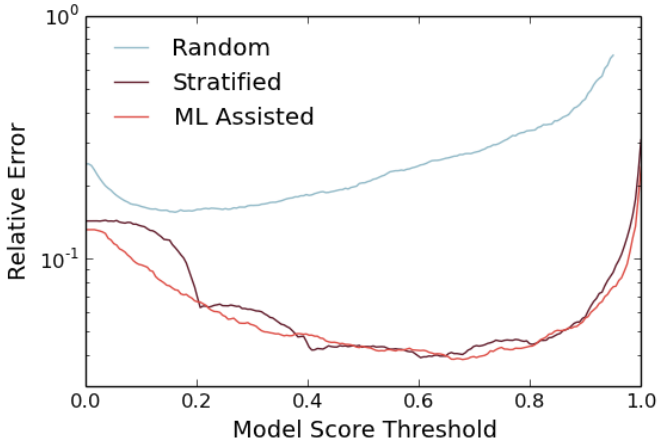


**Figure 5: The relative error on $\widetilde{NSI}$ above a given model threshold for each of the three discussed sampling approaches on our synthetic data: Random, Five Strata, and ML Assisted. The relative uncertainty between the estimators depends on several factors, such as number of strata and strata size.**

Another advantage of having smoothly varying errors is demonstrated in Table 1, where it is easier to re-use this dataset to estimate improvements in $\widetilde{NSI}$ in order to make operational decisions, optimize head count or prioritize efforts.

## 5. RESULTS

As discussed in Section 4.2, Machine Learning assisted sampling offers several quantifiable results with respect to random and stratified sampling. This section describes the results obtained when testing this sampling scheme using data from our live systems, namely:

- Removal of discrete jumps in uncertainty (see Fig. 5), and allowing for a better use for estimating operations

needs and for calibrating models, without loss in precision.

- 6x-15x increase in not-suitable samples selected with respect to random sampling, 20% lift with respect to stratified sampling. These numbers represent an equivalent reduction in required reviewer head-count for equal precision (same results with fewer reviewers).

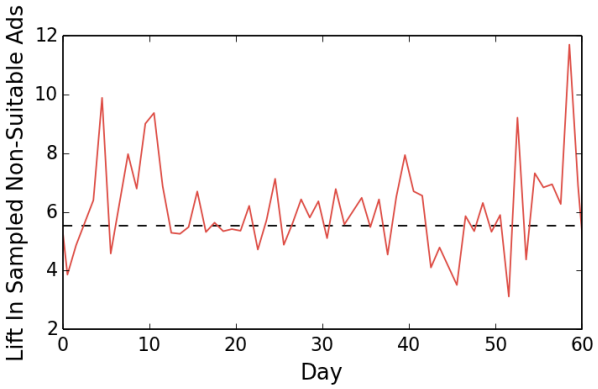


**Figure 6: Fraction of not-suitable ads in an ML assisted sample divided by the fraction in a biased random sample. ML assisted sampling increases the yield by $> 5$x on average.**

## 5.1 ML Assisted vs. Other Schemes

We ran independent tests using randomly sampled ads for review and using ML assisted sampled ads from real data as an important cross-check of the method. Fig. 6 shows the fraction of ads marked as non-suitable by expert reviewers using ML assisted sampling and random sampling on those ads. The fraction of not-suitable ads in the ML assisted sampling has been consistently up to 6 times higher than from the random sampling method, which runs in parallel as a cross-check. This is also a 20% increase in sampled not-suitable ads with respect to the expected yield from the stratified sampling method.

The ratio varies over time, depending on the breakdown of violation types present in the ecosystem at any given time, and on the model's ability to separate these from the suitable-like ads. There is a strong seasonality in the features of suitable and non-suitable ads, coupled with an adversarial component to some types of not-suitable ads. For that reason, as the distribution of feature vectors can change quite rapidly over time, the model losses sensitivity and must be re-trained on a regular basis with updated labeled data. The seasonality and adversarial effects are expected to affect the yield of non-suitable ads, but not the measurements performed on the sample dataset, by virtue of being an unbiased estimator. Inefficiencies or biased assignment of sampling weights may have a large effect on the variance of the estimator, which is another motivation for the frequent re-training of the classifier.

Fig. 7 contains plots of the $\widetilde{NSI}$ estimated using the ML assisted sampling (solid line), the estimate and uncertainty using random sampling (dashed line), and shows that over time, the ML assisted sampling and the random sampling methods have been predicting very similar mean values on real ads. Pearson's R correlation coefficient between the two measures is 76%, and the two are consistent within the measurement uncertainty on real ads.

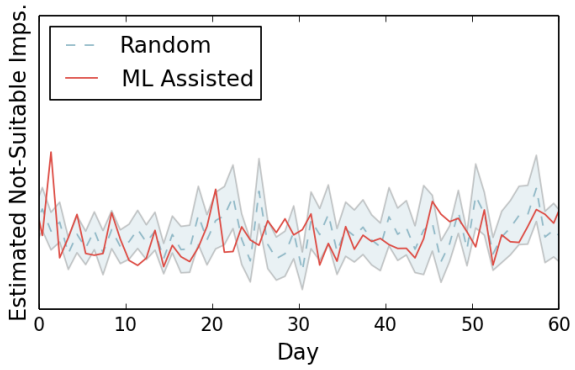


**Figure 7: $\widetilde{NSI}$ measured via biased random sampling (dashed line) vs ML assisted sampling (solid line). The relative errors are approximately 10%.**

## 5.2 Additional Impact in Other Domains

The benefits of the ML assisted sampling depend on both the performance of the model and the size of the class asymmetry. A model which produces a long region of increasing discriminating power works well for generalizing the dataset, while the better the discriminating power of the model, the larger the lift in number of minority class ads selected. Moreover, the larger the class imbalance, the larger the impact of the ML assisted sampling.

An additional test of the method was performed using another Facebook dataset from a different domain where the minority class is an order of magnitude less frequent. After applying ML Assisted Sampling, results showed a 15x lift in the number of not-suitable samples sent for expert review in this domain, leading to a $\sim$3.5x reduction on the uncertainty of our measurements related to the health of that ecosystem, without requiring any additional reviewers.

In addition, ML assisted sample is being tested on data from numerous other domains at Facebook where we have large class imbalance problems (e.g., organic post quality or spam detection among many others).

# 6. OTHER CONSIDERATIONS

## 6.1 Model Calibration

Due to the variety of "not-suitable" classes that we aim to identify, and the range of signals that we have from the various aspects of the ads, there is no unique model type suited for all of our tasks. The Ad Review System is therefore a mix of supervised and unsupervised machine learning models (e.g. spanning Boosted Decision Trees, Clustering, Deep Nets, Logistic Regressions, Outlier Detection, SVMs, etc. . . ). Each of these model types is known to return scores which are biased away from the true probabilities in a distinct and systematic way. In the case where we are performing classification, we desire calibrated models such that scores from different classifiers are comparable. Some typical methods of calibrating model scores include the use of

Platt Scaling, Isotonic Regression [17], or some other functional transformation [18] to correct the inherent distortions in the model score distributions.

Since ML Assisted sampling produces a uniform density as a function of the model score, it tends to over-sample the areas near the decision boundary of the algorithm relative to random sampling and stratified sampling. For classifiers which produce scores (e.g., Decision-Tree based methods, SVMs, Logistic Regressions, and Naive Bayes), the gradient for the ratio of positive and negative examples is often largest towards middle values of the classifier. Comparing Fig. 2 to Fig. 3, you can see that we have increased the density of the samples in the region where additional evaluation samples will improve our ability to measure the rate of change, and hence the quality of the calibrations.

# 7. CONCLUSION AND FUTURE WORK

In this paper we discussed the approach taken at Facebook for enhancing the yield of minority class ads sent to expert reviewers. By leveraging the output of a one-vs-all classifier to calculate per-ad sampling probabilities we increased the yield of the minority class (not-suitable) ads by up to 15x, increased the precision of our KPIs by up to 50%, while at the same time maintained a highly generalizable dataset. We demonstrated that by using non-parametric modeling of the probability density function, the sampling probabilities can be calculated in a way which makes it relatively insensitive to drifts in the distribution of scores over time and to distortions of the density.

Further we gave examples how this generalizability enables the re-use of this valuable and scarce data to

- model the impact of allocating additional reviewers,
- enable deep dives of the data to gain insights into the changing landscape of policy-violating ads,
- assess the performance of the models used in production and calibrate their outputs to allow easy comparison across model types
- train lightweight models based on the results of the sample,
- all while maintaining high-quality measurements of our key metrics.

There are obvious extensions to this method which we have not explicitly covered in this paper, such as designing sampling probability functions which yield an optimal allocation with respect to the variance of the KPIs, or updating the sampling probability function online so as to reduce the latency of the method. We leave such extensions to the imagination of interested parties that may wish to adopt ML assisted sampling to their particular use case.

# References


[1] Facebook, "Advertising policies," 2016.

[2] D. Sculley, M. E. Otey, M. Pohl, B. Spitznagel, J. Hainsworth, and Y. Zhou, "Detecting adversarial advertisements in the wild," in *Proceedings of the 17th ACM SIGKDD International Conference on Data Mining and Knowledge Discovery*, 2011.

[3] C.-E. Särndal, B. Swensson, and J. Wretman, *Model assisted survey sampling*. Springer Science & Business Media, 2003.

[4] L. A. Wasserman, *All Of Statistics*. Springer-Verlag, 2004.

[5] N. Duffield, C. Lund, and M. Thorup, "Priority sampling for estimation of arbitrary subset sums," *J. ACM*, vol. 54, Dec. 2007.

[6] S. L. Lohr, *Sampling : design and analysis*. Brooks/Cole, 2 ed., Dec. 2010.

[7] A. B. Owen, *Monte Carlo theory, methods and examples*. 2013.

[8] S. T. Tokdar and R. E. Kass, "Importance sampling: a review," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 1, pp. 54–60, 2010.

[9] M.-S. Oh and J. O. Berger, "Adaptive importance sampling in monte carlo integration," *Journal of Statistical Computation and Simulation*, vol. 41, no. 3-4, pp. 143–168, 1992.

[10] J. Attenberg and F. Provost, "Why label when you can search?: alternatives to active learning for applying human resources to build classification models under extreme class imbalance," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 423–432, ACM, 2010.

[11] C. Sawade, N. Landwehr, S. Bickel, and T. Scheffer, "Active risk estimation," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 951–958, 2010.

[12] S. Nadarajah and A. K. Gupta, "On the Moments of the Exponentiated Weibull Distribution," *Communications in Statistics: Theory and Methods*, vol. 34, pp. 253–256, Feb. 2005.

[13] W. N. H. Morris H. Hansen, "On the theory of sampling from finite populations," *The Annals of Mathematical Statistics*, vol. 14, no. 4, pp. 333–362, 1943.

[14] P. Étoré and B. Jourdain, "Adaptive Optimal Allocation in Stratified Sampling Methods," *Methodology and Computing in Applied Probability*, vol. 12, no. 3, pp. 335–360, 2010.

[15] L. Kish, *Survey sampling*. Wiley, Feb. 1995.

[16] F. N. Fritsch and R. E. Carlson, "Monotone Piecewise Cubic Interpolation," *SIAM Journal on Numerical Analysis*, vol. 17, pp. 238–246, Apr. 1980.

[17] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Proceedings of the 22Nd International Conference on Machine Learning*, ICML '05, (New York, NY, USA), pp. 625–632, ACM, 2005.

[18] B. Zadrozny and C. Elkan, "Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers," in *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, (San Francisco, CA, USA), pp. 609–616, Morgan Kaufmann Publishers Inc., 2001.

# APPENDIX

## A. ESTIMATING QUANTITIES WITH DISTORTED PDFS

Here we demonstrate that the measurement of the quantity of interest in a sample using ML assisted sampling can trvially be transformed into an estimate of the quantity of interest in the parent population, i.e. that we still have an unbiased estimator.

The quantity we wish to estimate is the number of elements from the minority class in a population, which is the probability of any element belonging to the minority class times the size of the population:

$$N^- = NP_f(\boldsymbol{x} \in \Omega) = N \int f(\boldsymbol{x}) d\boldsymbol{x} \tag{16}$$

where $\boldsymbol{x}$ is a vector describing the state, $P_f(\boldsymbol{x} \in \Omega)$ is the probability of $\boldsymbol{x}$ belonging to the minority class given a density function $f(\boldsymbol{x})$, the integral runs over all of phase space or over a specific acceptance region, and the population has $N$ elements in it.

In cases where the density has been distorted, such that it follows a different form $g(\boldsymbol{x})$, the probability of an element being in the minority class may be re-written as:

$$\begin{aligned} P_f(\boldsymbol{x} \in \Omega) &= \int f(\boldsymbol{x}) d\boldsymbol{x} \\ &= \int \frac{f(\boldsymbol{x})}{g(\boldsymbol{x})} g(\boldsymbol{x}) d\boldsymbol{x} \\ &= \frac{\int \frac{f(\boldsymbol{x})}{g(\boldsymbol{x})} g(\boldsymbol{x}) d\boldsymbol{x}}{\int g(\boldsymbol{x}) d\boldsymbol{x}} \int g(\boldsymbol{x}) d\boldsymbol{x} \end{aligned} \tag{17}$$

In our case, the distortion of the PDF is a known invertible weight function $w(\boldsymbol{x}) = g(\boldsymbol{x})/f(\boldsymbol{x})$ and therefore:

$$P_f(\boldsymbol{x} \in \Omega) = \mathbb{E}_g[w^{-1}(\boldsymbol{x}) | \boldsymbol{x} \in \Omega] P_g(\boldsymbol{x} \in \Omega) \tag{18}$$

When we sample $n$ elements using $g(\boldsymbol{x})$, we can measure the $P_g(\boldsymbol{x} \in \Omega)$ as the number of elements belonging to the minority class in the sample:

$$\hat{P}_g(\boldsymbol{x} \in \Omega) = \frac{n^-}{n} \tag{19}$$

We can plug Eqn. 19 and the average weight associated with minority class elements $\left\langle w^{-1} \right\rangle = \frac{1}{n^-} \sum_{i=1}^{n^-} w_i^{-1}$ into Eqn. 18, and get:

$$\hat{P}_f(\boldsymbol{x} \in \Omega) = \frac{1}{n^-} \sum_{i=1}^{n^-} w_i^{-1} \times \frac{n^-}{n} = \frac{1}{n} \sum_{i=1}^{n^-} w_i^{-1} \tag{20}$$

Thus as long as we know $\left\langle w^{-1} \right\rangle$ and the number of minority class elements in the sample, we can estimate the number of minority class elements in the parent population:

$$\hat{N^-} = N\hat{P}_f(\boldsymbol{x} \in \Omega) = \frac{N}{n} \sum_{i=1}^{n^-} w_i^{-1} \tag{21}$$

Eqn. 21 has the same form as the Hansen-Hurwitz estimator in which inverse probability weighting is used to create an unbiased estimator from a sampled population.