# Learning from Survey Training Samples:
# Rate Bounds for Horvitz-Thompson Risk Minimizers

**Stephan Clémençon** STEPHAN.CLEMENCON@TELECOM-PARISTECH.FR
**Guillaume Papa** GUILLAUME.PAPA@TELECOM-PARISTECH.FR
*LTCI, Télécom ParisTech, Université Paris-Saclay*
*Paris, France*


**Patrice bertail** PATRICE.BERTAIL@U-PARIS10.FR
*Université Paris Ouest-Nanterre-La Défense, MODALX*
*Nanterre, France*

**Editors:** Robert J. Durrant and Kee-Eung Kim

## Abstract

The generalization ability of minimizers of the empirical risk in the context of binary classification has been investigated under a wide variety of complexity assumptions for the collection of classifiers over which optimization is performed. In contrast, the vast majority of the works dedicated to this issue stipulate that the training dataset used to compute the empirical risk functional is composed of i.i.d. observations and involve sharp control of uniform deviation of i.i.d. averages from their expectation. Beyond the cases where training data are drawn uniformly without replacement among a large i.i.d. sample or modelled as a realization of a weakly dependent sequence of r.v.'s, statistical guarantees when the data used to train a classifier are drawn by means of a more general sampling/survey scheme and exhibit a complex dependence structure have not been documented in the literature yet. It is the main purpose of this paper to show that the theory of empirical risk minimization can be extended to situations where statistical learning is based on survey samples and knowledge of the related (first order) inclusion probabilities. Precisely, we prove that minimizing a (possibly biased) weighted version of the empirical risk, refered to as the (approximate) Horvitz-Thompson risk (HT risk), over a class of controlled complexity lead to a rate for the excess risk of the order $O_{\mathbb{P}}((\kappa_N(\log N)/n)^{1/2})$ with $\kappa_N = (n/N)/\min_{i \leq N} \pi_i$, when data are sampled by means of a rejective scheme of (deterministic) size $n$ within a statistical population of cardinality $N \geq n$, a generalization of basic *sampling without replacement* with unequal probability weights $\pi_i > 0$. Extension to other sampling schemes are then established by a coupling argument. Beyond theoretical results, numerical experiments are displayed in order to show the relevance of HT risk minimization and that ignoring the sampling scheme used to generate the training dataset may completely jeopardize the learning procedure.

**Keywords:** Empirical risk minimization, Horvitz-Thompson estimator, learning rate bound, negatively associated random variables, probabilistic theory of pattern recognition, sampling method, survey scheme

## 1. Introduction

Whereas statistical learning techniques crucially exploit data that can serve as examples to train a decision rule, they may also make use of weights individually assigned to the observations, resulting from survey sampling stratification. Such weights could correspond either to true inclusion probabilities or else to calibrated or post-stratification weights, minimizing some discrepancy under certain margin constraints for the inclusion probabilities. In the context of statistical inference based on survey data, the asymptotic properties of specific statistics such as Horvitz-Thompson estimators (*cf* Horvitz and Thompson (1951)), whose computation involves not only the observations but also the weights, have been widely investigated: in particular, mean estimation and regression have been the subject of much attention, refer to Hajek (1964), Robinson (1982), Berger (1998) for instance, and a comprehensive functional limit theory for distribution function estimation is progressively documented in the statistical literature, see Breslow and Wellner (2007), Breslow et al. (2009), Saegusa and Wellner (2011). At the same time, the last decades have witnessed a rapid development of the field of machine-learning. Revitalized by different breakout algorithms (*e.g.* SVM, boosting methods), its practice is now supported by a sound probabilistic theory based on recent non asymptotic results in the study of empirical processes, see Koltchinskii (2006), Boucheron et al. (2005). However, most papers dedicated to theoretical results grounding the *Empirical Risk Minimization* approach (ERM in short), the main paradigm of statistical learning, assume that the training of a decision rule is based on a dataset formed of independent replications of a generic random vector $Z$, a collection of $N \geq 1$ i.i.d. observations $Z_1$, ..., $Z_N$ namely. In contrast, few results are available in situations where the training dataset is generated by a more complex sampling scheme. One may refer to Bardenet and Maillard (2015) for concentration inequalities permitting to study the generalization ability of empirical risk minimizers when the training data are obtained by standard *sampling without replacement* (SWOR in abbreviated form) or to Steinwart et al. (2009) in the case where the decision rule is learnt from a path of a *weakly dependent stochastic* process.

It is the goal of this paper to extend the ERM theory to situations where the training dataset is generated by means of a more general sampling scheme, with possibly unequal probability weights. We first consider the case of *rejective* sampling (sometimes refered to as *conditional Poisson* sampling), an important generalization of basic SWOR. The rate bound results obtained by means of properties of so-termed *negatively related random variables* in this case are next shown to extend to a class of more general sampling schemes by a coupling argument. In addition, numerical experiments have been carried out in order to provide empirical evidence of the approach developed. They show in particular that statistical accuracy of the ERM approach may go down the drain if the sampling scheme underlying the training dataset is ignored.

The paper is organized as follows. In section 2, the probabilistic framework of the present study is described at length and basic results of the probabilistic theory of classification are briefly recalled, together with some important notions of survey theory. The main theoretical results are stated in section 3, while illustrative numerical experiments are presented in section 4. Technical proofs are deferred to the Appendix section, as well as additional technical details related to the sampling schemes under study.

## 2. Background and Preliminaries

As a first go, we start with recalling key concepts pertaining to the theory of empirical risk minimization in binary classification, the flagship problem in statistical learning. A few notions related to survey theory are next described, which will be involved in the subsequent analysis. Throughout the article, the indicator function of any event $\mathcal{E}$ is denoted by $\mathbb{I}\{\mathcal{E}\}$, the Dirac mass at any point $a$ by $\delta_a$, the power set of any set $E$ by $\mathcal{P}(E)$, the cardinality of any finite set $A$ by $\#A$.

### 2.1. Binary Classification - Empirical Risk Minimization Theory

The binary classification problem is considered as a running example all along the paper. Because it can be easily formulated, it is undeniably the most documented statistical learning problem in the literature and certain results extend to more general frameworks (*e.g.* multiclass classification, regression, ranking). Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $(X, Y)$ a random pair defined on $(\Omega, \mathcal{A}, \mathbb{P})$, taking its values in some measurable product space $\mathcal{X} \times \{-1, +1\}$, with common distribution $P(dx, dy)$: the r.v. $X$ models some observation, hopefully useful for predicting the binary label $Y$. The distribution $P$ can also be described by the pair $(F, \eta)$ where $F(dx)$ denotes the marginal distribution of the input variable $X$ and $\eta(x) = \mathbb{P}\{Y = +1 \mid X = x\}$, $x \in \mathcal{X}$, is the *posterior distribution*. The objective is to build, based on the training dataset at disposal, a measurable mapping $g : \mathcal{X} \mapsto \{-1, +1\}$, called a *classifier*, with minimum risk:

$$L(g) \overset{def}{=} \mathbb{P}\{g(X) \neq Y\}. \tag{1}$$

It is well-known folklore in the probabilistic theory of pattern recognition that the *Bayes classifier* $g^*(x) = 2\mathbb{I}\{\eta(x) \geq 1/2\} - 1$ is a solution of the risk minimization problem $\inf_g L(g)$, where the infimum is taken over the collection of all classifiers defined on the input space $\mathcal{X}$. The minimum risk is denoted by $L^* = L(g^*)$. Since the distribution $P$ of the data is unknown, one substitutes the true risk with its empirical estimate

$$\widehat{L}_n(g) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{g(X_i) \neq Y_i\}, \tag{2}$$

based on a sample $(X_1, Y_1)$, …, $(X_n, Y_n)$ of independent copies of the generic random pair $(X, Y)$. The true risk minimization is then replaced by the empirical risk minimization

$$\min_{g \in \mathcal{G}} \widehat{L}_n(g), \tag{3}$$

where the minimum is taken over a class $\mathcal{G}$ of classifier candidates, supposed rich enough to include the naive Bayes classifier (or a reasonable approximation of the latter). Considering a solution $\widehat{g}_n$ of (3), a major problem in statistical learning theory is to establish upper confidence bounds on the *excess of risk* $L(\widehat{g}_n) - L^*$ in absence of any distributional assumptions and taking into account the complexity of the class $\mathcal{G}$ (*e.g.* described by geometric or combinatorial features such as the VC dimension) and some measure of accuracy

of approximation of $P$ by its empirical counterpart $P_n = (1/n) \sum_{i=1}^{n} \delta_{(X_i, Y_i)}$ over the class $\mathcal{G}$. Indeed, one typically bounds the excess of risk of the empirical risk minimizers as follows

$$L(\widehat{g}_n) - L^* \leq 2 \sup_{g \in \mathcal{G}} |\widehat{L}_n(g) - L(g)| + \left( \inf_{g \in \mathcal{G}} L(g) - L^* \right).$$

The second term on the right hand side is referred to as the *bias* and depends on the richness of the class $\mathcal{G}$, while the first term, called the *stochastic error*, is controlled by means of results in empirical process theory, see Boucheron et al. (2005).

**Remark 1** (ON RISK SURROGATES) *Although its study is of major interest from a theoretical perspective, the problem* (3) *is generally NP-hard. For this reason, the cost function* $\mathbb{I}\{-Yg(X) > 0\}$ *is replaced in practice by a nonnegative convex cost function* $\phi(Yg(X))$, *turning empirical risk minimization to a tractable convex optimization problem. Typical choices include the exponential cost* $\phi(u) = \exp(u)$ *used in boosting algorithms, the hinge loss* $\phi(u) = (1 + u)_+$ *in the case of support vector machines and the logit cost* $\phi(u) = \log(1 + \exp(u))$ *for Neural Networks, see Bartlett et al. (2006) and the references therein. Extension of the results established in the present paper to such risk surrogates are straightforward and left to the reader.*

In this paper, we consider the situation where the training data used to compute of the empirical risk (2) is not an i.i.d. sample but the product of a more general sampling plan of fixed size $n \geq 1$.

## 2.2. Sampling Schemes and Horvitz-Thompson Estimation

Let $N \geq 1$. In the standard *superpopulation* framework we consider, $(X_1, Y_1)$, ..., $(X_N, Y_N)$ is a sample of independent copies of $(X, Y)$ observed on a finite population $\mathcal{I}_N := \{1, \ldots, N\}$. We call a *survey sample* of (possibly random) size $n \leq N$ of the population $\mathcal{I}_N$, any subset $s := \{i_1, \ldots, i_{n(s)}\} \in \mathcal{P}(\mathcal{I}_N)$ with cardinality $n =: n(s)$ less that $N$. A sampling design is determined by a conditional probability distribution $R_N$ on the set of all possible samples $s \in \mathcal{P}(\mathcal{I}_N)$ given the original data $\mathcal{D}_N = \{(X_i, Y_i) : i \in \mathcal{I}_N\}$. For any $i \in \{1, \ldots, N\}$, the first order *inclusion probability*, $\pi_i = \mathbb{P}_{R_N}\{i \in S\}$ is the probability that the unit $i$ belongs to a random sample $S$ drawn from the conditional distribution $R_N$. We set $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)$. The second order inclusion probabilities are denoted by $\pi_{i,j} = \mathbb{P}_{R_N}\{(i, j) \in S^2\}$, for any $i \neq j$ in $\{1, \ldots, N\}^2$. The information related to the observed sample $S \subset \{1, \ldots, N\}$ is fully enclosed in the r.v. $\boldsymbol{\epsilon}_N = (\epsilon_1, \ldots, \epsilon_N)$, where $\epsilon_i = \mathbb{I}\{i \in S\}$ for $1 \leq i \leq N$. The 1-d marginal conditional distributions of the sampling scheme $\boldsymbol{\epsilon}_N$ given $\mathcal{D}_N$ are the Bernoulli distributions $\mathcal{B}(\pi_i) = \pi_i \delta_1 + (1 - \pi_i)\delta_0$, $1 \leq i \leq N$, and the covariance matrix $\Gamma_N$ of the r.v. $\boldsymbol{\epsilon}_N$ has entries given by $\Gamma_N(i, j) = \pi_{i,j} - \pi_i \pi_j$, with $\pi_{i,i} = \pi_i$ by convention, for $1 \leq i, j \leq N$. Observe that, equipped with the notations above, $\sum_{1 \leq i \leq N} \epsilon_i = n(S)$. One may refer to Deville (1987) for accounts of survey sampling techniques. Notice also that, in many applications, the inclusion probabilities are built using some extra information, typically by means of *auxiliary random variables* $W_1, \ldots, W_N$ defined on $(\Omega, \mathcal{A}, \mathbb{P})$ and taking their values in some measurable space $\mathcal{W}$: $\forall i \in \{1, \ldots, N\}$, $\pi_i = nh(W_i)/\sum_{1 \leq j \leq N} h(W_j)$, where $n \max_{1 \leq i \leq n} h(W_i) \leq \sum_{1 \leq i \leq N} h(W_i)$ almost-surely and $h : \mathcal{W} \to ]0, +\infty[$ is a measurable *link function*. The $(X_i, Y_i, W_i)$'s are generally supposed to be i.i.d. copies of a generic r.v.

$(X, Y, W)$. See Särndall and B. Swensson (2003) for more details. For simplicity, the $\pi_i$'s are supposed to be deterministic in the subsequent analysis, which boils down to carrying out the study conditionally upon the $W_i$'s in the example aforementioned.

**Horvitz-Thompson risk.** As defined in Horvitz and Thompson (1951), the Horvitz-Thompson version of the (not available) empirical risk $\widehat{L}_N(g) = N^{-1} \sum_{1 \leq i \leq N} \mathbb{I}\{Y_i \neq g(X_i)\}$ of any classifier candidate $g$ based on the sampled data $\{(X_i, Y_i) : i \in S\}$ with $S \sim R_N$ is given by:

$$\overline{L}_{\boldsymbol{\epsilon}_N}(g) = \frac{1}{N} \sum_{i \in S} \frac{1}{\pi_i} \mathbb{I}\{g(X_i) \neq Y_i\} = \frac{1}{N} \sum_{i=1}^{N} \frac{\epsilon_i}{\pi_i} \mathbb{I}\{g(X_i) \neq Y_i\} \tag{4}$$

with the convention that $0/0 = 0$ and where the subscript $\boldsymbol{\epsilon}_N = (\epsilon_1, \ldots, \epsilon_N)$ denotes the vector in correspondence with the sample $S$. Observe that, conditionally upon the $(X_i, Y_i)$'s, the quantity (4), that shall be referred to as the *empirical Horvitz-Thompson risk* (empirical HT risk in short) throughout the paper, is an unbiased estimate of the empirical risk $\widehat{L}_N(g)$. Its (pointwise) consistency and asymptotic normality are established in Robinson (1982) and Berger (1998) for a variety of sampling schemes.

This article is devoted to investigating the statistical performance of minimizers $\bar{g}_N$ of the HT risk (4) over the class $\mathcal{G}$ under adequate assumptions for the sampling scheme $R_N$ used to generate the training dataset. We point out that such an analysis is far from straightforward due to the possible depence structure of the terms involved in the summation (4): except in the Poisson case (recalled below), concentration results for empirical processes cannot be directly applied to control maximal deviations of the type

$$\sup_{g \in \mathcal{G}} |\overline{L}_{\boldsymbol{\epsilon}_N}(g) - L(g)|.$$

**Conditional Poisson sampling.** One of the simplest sampling plan is undeniably the *Poisson survey scheme* (without replacement), a generalization of *Bernoulli sampling* originally proposed in Goodman (1949) for the case of unequal weights: the $\epsilon_i$'s are independent and the sampling distribution is thus entirely determined by the first order inclusion probabilities $\mathbf{p}_N = (p_1, \ldots, p_N) \in ]0, 1[^N$:

$$\forall s \in \mathcal{P}(\mathcal{I}_N), \quad P_N(s) = \prod_{i \in S} p_i \prod_{i \notin S} (1 - p_i). \tag{5}$$

Observe in addition that the behavior of the quantity (4) can be then investigated by means of results established for sums of independent random variables. However, the major drawback of this sampling plan lies in the random nature of the corresponding sample size, impacting significantly the variability of (4). The variance of the Poisson sample size is given by $d_N = \sum_{i=1}^{N} p_i(1 - p_i)$, while the conditional variance of (4) is in this case: $\sum_{i=1}^{n} ((1 - p_i)/p_i) \mathbb{I}\{g(X_i) \neq Y_i\}$. For this reason, *rejective sampling*, a sampling design $R_N$ of fixed size $n \leq N$, is often preferred in practice. It generalizes the *simple random sampling without replacement* (where all samples with cardinality $n$ are equally likely to be chosen, with probability $(N-n)!/n!$, all the corresponding first and second order probabilities being thus equal to $n/N$ and $n(n-1)/(N(N-1))$ respectively). Denoting by $\boldsymbol{\pi}_N = (\pi_1, \ldots, \pi_N)$

146

its first order inclusion probabilities and by $\mathcal{S}_n = \{s \in \mathcal{P}(\mathcal{I}_N) : \#s = n\}$ the subset of all possible samples of size $n$, it is defined by:

$$\forall s \in \mathcal{S}_n, \;\; R_N(s) = C \prod_{i \in s} p_i \prod_{i \notin s} (1 - p_i), \tag{6}$$

where $C = 1/\sum_{s \in \mathcal{S}_n} \prod_{i \in s} p_i \prod_{i \notin s} (1 - p_i)$ and the vector $\mathbf{p}_N = (p_1, \ldots, p_N) \in ]0, 1[^N$ yields first order inclusion probabilities equal to the $\pi_i$'s and is such that $\sum_{i \leq N} p_i = n$. Under this latter additional condition, such a vector $\mathbf{p}_N$ exists and is unique (see Dupacova (1979)) and the related representation (6) is then said to be *canonical*[1]. Comparing (6) and (5) reveals that rejective $R_N$ sampling of fixed size $n$ can be viewed as Poisson sampling given that the sample size is equal to $n$. It is for this reason that rejective sampling is usually referred to as *conditional Poisson sampling*. One must pay attention not to get the $\pi_i$'s and the $p_i$'s mixed up: the latter are the first order inclusion probabilities of $P_N$, whereas the former are those of its conditional version $R_N$. However they can be related by means of the results stated in Hajek (1964) (see Theorem 5.1 therein): $\forall i \in \{1, \ldots, N\}$,

$$\pi_i(1 - p_i) = p_i(1 - \pi_i) \times (1 - (\tilde{\pi} - \pi_i)/d_N^* + o(1/d_N^*)), \tag{7}$$

$$p_i(1 - \pi_i) = \pi_i(1 - p_i) \times (1 - (\tilde{p} - p_i)/d_N + o(1/d_N)), \tag{8}$$

where $d_N^* = \sum_{i=1}^N \pi_i(1 - \pi_i)$, $\tilde{\pi} = (1/d_N^*) \sum_{i=1}^N \pi_i^2(1 - \pi_i)$ and $\tilde{p} = (1/d_N) \sum_{i=1}^N (p_i)^2(1 - p_i)$.

More examples of sampling schemes with fixed size are given in Appendix C.

## 3. Main Results

We first consider the case where statistical learning is based on the observation of a sample drawn by means of a rejective scheme. As shall be seen below, the main argument underlying the results obtained relies on the fact that the related scheme form a collection of *negatively associated* (binary) random variables, a rather tractable type of dependence structure. This property being shared by many other sampling schemes of deterministic size, the same argument can be thus naturally applied to carry out a similar rate analysis for training data produced by such plans. Extensions of these results to more general sampling schemes are also considered by means of a *coupling* technique.

### 3.1. Horvitz-Thompson Empirical Risk Minimization in the Rejective Case

For clarity, we first recall the definition of *negatively associated random variables*, see Joag-Dev and Proschan (1983).

**Definition 2** *Let $Z_1, \ldots, Z_n$ be random variables defined on the same probability space, valued in a measurable space $(E, \mathcal{E})$. They are said to be negatively associated iff for any pair of disjoint subsets $A_1$ and $A_2$ of the index set $\{1, \ldots, n\}$*

$$Cov \left( f((Z_i)_{i \in A_1}), \; g((Z_j)_{j \in A_2}) \right) \leq 0, \tag{9}$$

*for any real valued measurable functions $f : E^{\#A_1} \to \mathbb{R}$ and $g : E^{\#A_2} \to \mathbb{R}$ that are both increasing in each variable.*

---

1. Notice that any vector $\mathbf{p}'_N \in ]0, 1[^N$ such that $p_i/(1 - p_i) = c p'_i/(1 - p'_i)$ for all $i \in \{1, \ldots, n\}$ for some constant $c > 0$ can be used to write a representation of $R_N$ of the same type as (6)

The theorem stated below reveals that any rejective scheme $\boldsymbol{\epsilon}_N$ forms a collection of negatively associated r.v.'s. The proof is given in Appendix A.

**Theorem 3** *Let $N \geq 1$ and $\boldsymbol{\epsilon}_N = (\epsilon_1, \ldots, \epsilon_N)$ be the vector of indicator variables related to a rejective plan on $\mathcal{I}_N$. Then, the binary random variables $\epsilon_1, \ldots, \epsilon_N$ are negatively associated.*

The result above permits to handle the dependence of the terms involved in the summation (4). It is the key argument for proving the following proposition, which extends results for training datasets generated by basic sampling without replacement (*i.e.* in the case of all equal weights: $\pi_i = n/N$ for $i = 1, \ldots, N$), refer to Bardenet and Maillard (2015) (see also Serfling (1974)).

**Proposition 4** *Suppose that the sampling scheme $\boldsymbol{\epsilon}_N$ is rejective with first order inclusion probabilities $\boldsymbol{\pi}_N$ and that the class $\mathcal{G}$ is of finite VC dimension $V < +\infty$. Set $\kappa_N = (n/N)/\min_{i \leq N} \pi_i$. Then, the following assertions hold true.*

*(i) For any $\delta \in (0,1)$, with probability larger than $1 - \delta$, we have: $\forall n \leq N$,*

$$\sup_{g \in \mathcal{G}} \left| \bar{L}_{\boldsymbol{\epsilon}_N}(g) - \hat{L}_N(g) \right| \leq \sqrt{2\kappa_N \frac{\log(\frac{2}{\delta}) + V\log(N+1)}{n}} + 2\kappa_N \frac{\log(\frac{2}{\delta}) + V\log(N+1)}{3n}. \tag{10}$$

*(ii) For any solution $\bar{g}_N$ of the minimization problem $\inf_{g \in \mathcal{G}} \overline{L}_{\boldsymbol{\epsilon}_N}(g)$ is such that, for any $\delta \in (0,1)$, with probability at least $1 - \delta$, we have: $\forall N \geq 1$,*

$$L(\bar{g}_N) - L^* \leq 2\sqrt{2\kappa_N \frac{\log(\frac{4}{\delta}) + V\log(N+1)}{n}} + 4\kappa_N \frac{\log(\frac{4}{\delta}) + V\log(N+1)}{3n}$$
$$+ C\sqrt{\frac{V}{N}} + 2\sqrt{\frac{2\log(\frac{2}{\delta})}{N}} + \inf_{g \in \mathcal{G}} L(g) - L^*.$$

The factor $\kappa_N$ involved in the bounds above reflects the influence of the sampling scheme (notice incidentally that $\kappa_N \geq 1$ since $\sum_{i \leq N} \pi_i = n$). In the SWOR case, *i.e.* when $\pi_i = n/N$ for all $i \in \{1, \ldots, N\}$, it is then minimum, equal to 1. More generally, when $n = o(N)$ as $N \to +\infty$, as soon as the weights cannot vanish faster than $n/N$, the rate achieved by minimizers of the HT risk is of the order $O_{\mathbb{P}}(\sqrt{(\log N)/n})$. Many sampling schemes (*e.g.* Rao-Sampford sampling, Pareto sampling, Srinivasan sampling) of fixed size are actually described by random vectors $\boldsymbol{\epsilon}_N$ with negatively associated components, see Brändén and Jonasson (2012) or Kramer et al. (2011). Hence, Proposition 4's proof shows that the bounds stated above immediately extend to these cases. See Appendix C for more details and references. Before showing how the rate bounds established can be extended to even more general sampling schemes, a few remarks are in order.

**Remark 5** (COMPLEXITY ASSUMPTIONS) *We point out that the results stated can be established, essentially by means of the same argument as that developed in the Appendix, under complexity assumptions of different nature, involving metric entropy conditions for instance (see e.g. van der Vaart and Wellner (1996)). Such straightforward extensions are left to the reader.*

**Remark 6** (MODEL SELECTION) *A slight modification of the argument involved in Proposition 4 straightforwardly leads to bounds on the expected excess risk $\mathbb{E}[L(\bar{g}_{\epsilon_N})] - \inf_{g \in \mathcal{G}} L(g)$. Following the Structural Risk Minimization principle (see Vapnik (2001)), such VC bounds can be next used as complexity regularization terms to penalize additively the HT risk (4) and, for a sequence of model classes $\mathcal{G}_k$ with $k \geq 1$ of finite VC dimension, select the classifier among the minimizers $\{\arg\min_{g \in \mathcal{G}_k} \bar{L}_{\epsilon_N}(g), \ k \geq 1\}$, which has approximately minimal risk. Due to space limitations, details are left to the reader.*

**Remark 7** (BIASED HT RISK) *As recalled in Appendix B, the canonical parameters $\mathbf{p}_N$ are practically used to build a rejective sampling scheme $\epsilon_N$ rather than its vector of first order inclusion probabilities $(\pi_1, \ldots, \pi_N)$, whose explicit computation based on the $p_i$'s is a difficult task, refer to Chen et al. (1994) for dedicated algorithms. For this reason, one could be naturally tempted to minimize the alternative risk estimate $\widetilde{L}_{\epsilon_N}(g) = (1/N) \sum_{i \leq N} (\epsilon_i/p_i) \mathbb{I}\{Y_i \neq g(X_i)\}$. As proved in Appendix B, refinements of Eq. (7)-(8) show that*

$$\sup_{g \in \mathcal{G}} |\widetilde{L}_{\epsilon_N}(g) - \bar{L}_{\epsilon_N}(g)| \leq \frac{1}{N} \sum_{i=1}^{N} \left| \frac{1}{p_i} - \frac{1}{\pi_i} \right| \leq 6N\kappa_N/(nd_N), \tag{11}$$

*one may directly derive a rate bound for solutions of $\inf_{g \in \mathcal{G}} \widetilde{L}_{\epsilon_N}(g)$ from bound (ii) in Proposition 4. In particular, the learning rate achieved by $\bar{g}_N$ is preserved when $1/\sqrt{n} = O(\min_{i \leq N} \pi_i)$ as $N, \ n \to +\infty$.*

### 3.2. Extensions to More General Sampling Schemes

We now extend the rate bound analysis carried out in the previous subsection to more complex sampling schemes (described by a random vector $\epsilon_N^*$ possibly exhibiting a very complex dependence structure). In order to give an insight into the arguments which the extension is based on, additional notations are required. In this section, we consider a general sampling design $R_N^*$ with first order inclusion probabilities $\boldsymbol{\pi}_N^* = (\pi_1^*, \ldots, \pi_N^*)$ described by the vector $\boldsymbol{\epsilon}_N^* = (\epsilon_1^*, \ldots, \epsilon_N^*)$ and investigate the performance of minimizers $\bar{g}_N^*$ of the HT empirical risk $\bar{L}_{\epsilon_N^*}(g) = (1/N) \sum_{i=1}^{N} (\epsilon_i^*/\pi_i^*) \mathbb{I}\{Y_i \neq g(X_i)\}$ over a class $\mathcal{G}$. We also consider a rejective sampling scheme $R_N$ described by the r.v. $\boldsymbol{\epsilon}_N$, with first order inclusion probabilities $\boldsymbol{\pi}_N = (\pi_1, \ldots, \pi_N)$ defined on the same probability space, as well as the following quantity:

$$\check{L}_{\epsilon_N}(g) = \frac{1}{N} \sum_{i=1}^{N} \frac{\epsilon_i}{\pi_i^*} \mathbb{I}\{Y_i \neq g(X_i)\} \tag{12}$$

for any classifier $g$. Observe that (12) differs from the HT empirical risk $\bar{L}_{\epsilon_N}(g)$ related to the rejective sampling scheme $\epsilon_N$ in the weights it involves, the $\pi_i^*$'s instead of the $\pi_i$'s namely. Equipped with this notation, the excess of risk of the HT empirical risk minimizer can be bounded as follows:

$$L(\bar{g}_N^*) - \inf_{g \in \mathcal{G}} L(g) \leq 2 \sup_{g \in \mathcal{G}} \left| L(g) - \widehat{L}_N(g) \right| + 2 \sup_{g \in \mathcal{G}} \left| \widehat{L}_N(g) - \bar{L}_{\epsilon_N}(g) \right|$$

$$+ 2 \sup_{g \in \mathcal{G}} \left| \bar{L}_{\epsilon_N}(g) - \check{L}_{\epsilon_N}(g) \right| + 2 \sup_{g \in \mathcal{G}} \left| \check{L}_{\epsilon_N}(g) - \bar{L}_{\epsilon_N^*}(g) \right|. \tag{13}$$

Whereas the first term on the right hand side of (13) can be classically controlled using Vapnik-Chervonenkis and McDiarmid inequalities (see *e.g.* Vapnik (2001)), assertion $(i)$ of Proposition 4 provides a control of the second term. Following in the footsteps of Hajek (1964), the third term shall be bounded by means of a *coupling* argument, *i.e.* a specific choice of the joint distribution of $(\boldsymbol{\epsilon}_N^*,\ \boldsymbol{\epsilon}_N)$ satisfying the distributional margin constraints, while the second term is controlled by assumptions related to the closeness between the first order inclusion probabilities $\boldsymbol{\pi}_N^*$ and $\boldsymbol{\pi}_N$. More precisely, the assumptions required in the subsequent analysis involve the total variation distance between the sampling plans $R_N$ and $R_N^*$:

$$d_{TV}(R_N,\ R_N^*) \stackrel{def}{=} \frac{1}{2} \sum_{s \in \mathcal{P}(\mathcal{I}_N)} |R_N(s) - R_N^*(s)|.$$

**Theorem 8** *Suppose that Proposition 4's assumptions are fulfilled. Set $\kappa_N^* = (n/N)\min_{i \leq N} \pi_i^*$ and $\kappa_N = (n/N)/\min_{i \leq N} \pi_i$. Then, there exists a universal constant $C < +\infty$ such that we have, $\forall N \geq 1$,*

$$\mathbb{E}\left[L(\bar{g}_N^*) - \inf_{g \in \mathcal{G}} L(g)\right] \leq 2\sqrt{2\kappa_N \frac{\log(\frac{4}{\delta}) + V\log(N+1)}{n}} + 4\kappa_N \frac{\log(\frac{4}{\delta}) + V\log(N+1)}{3n}$$
$$+ C\sqrt{\frac{V}{N}} + 2\sqrt{\frac{2\log(\frac{2}{\delta})}{N}} + 2(\kappa_N^* + \kappa_N)(N/n)d_{TV}(R_N,\ R_N^*), \quad (14)$$

*where the infimum is taken over the set of rejective schemes $R_N$ with first order inclusion probabilities $\boldsymbol{\pi}_N = (\pi_1,\ \ldots,\ \pi_N)$.*

The proof is given in Appendix A. The rate bound obtained depends on the minimum error made when approximating the sampling plan by a rejective sampling plan in terms of total variation distance. In practice, following in the footsteps of Hajek (1964) or Berger (1998), it can be controlled by exhibiting a specific coupling $(\boldsymbol{\epsilon}_N^*,\ \boldsymbol{\epsilon}_N)$. One may refer to Berger (1998) for many coupling results of this nature, in particular when the approximating scheme $\boldsymbol{\epsilon}_N$ is of rejective type.

## 4. Illustrative Numerical Experiments

In this section we display numerical experiments to illustrate the relevance of HT risk minimization. We first consider the case where $g(X) = sign(k(X)^T\theta + b)$, where $k$ is some mapping function, $T$ denotes the transposition operator, $\theta$, $b$ are some parameters. As mentionned in 1, we consider the hinge loss as a convex surrogate of the $0-1$ loss and add some $l_2$ regularization term. This leads to the "Weighted SVM" formulation below:

$$\min_{\theta,b} \frac{1}{N} \sum_{i \in S} \frac{1}{\pi_i} \max(0, 1 - Y_i(k(X_i)^T\theta - b)) + \lambda\|\theta\|^2.$$

We use the gaussian r.b.f kernel and perform cross validation to appropriately choose the value of $\lambda$. We then consider the task of learning classification trees using the CART algorithm. These classifiers are trained using the scikit-learn library Pedregosa et al. (2011) and,

we account for the randomness of our experiments by shuffling our datasets and repeating the experiments 50 times.

We first generate a two class dataset $\mathcal{D}$ in $\mathbb{R}^{10}$ of size 20000 by sampling independent observations from two multivariate normal distribution. A similar dataset $\mathcal{D}_{test}$ of size 2000 is generated to test our classifiers. Denoting by $I_d$ the identity matrix in $\mathbb{R}^d$, the positive class has mean $(0, \ldots, 0)$ and covariance matrix equal to $I_{10}$, the negative class has mean $(1, \ldots, 1)$ and covariance matrix equal to $10 \times I_{10}$. We then build a dataset $\widetilde{\mathcal{D}}$ of size 1100 via a rejective sampling scheme applied to $\mathcal{D}$. Observations from the negative class being more noisy we assign them first order probability equal to 0.1, and assign first order probability equal to 0.01 to observation from the positive class. To allow for a fair comparison, we also build a dataset $\widehat{\mathcal{D}}$ of size 1100 by sampling without replacement within $\mathcal{D}$. We then learn the different classifiers on $\widetilde{\mathcal{D}}$ and $\widehat{\mathcal{D}}$, and display the results in Table 1.

|  | Mean | Standard deviation |
| --- | --- | --- |
| **Weighted SVM on $\widetilde{\mathcal{D}}$** | 0.02 | 0.005 |
| **Unweighted SVM on $\widetilde{\mathcal{D}}$** | 0.18 | 0.02 |
| **SVM on $\widehat{\mathcal{D}}$** | 0.04 | 0.005 |
| **Weighted CART on $\widetilde{\mathcal{D}}$** | 0.06 | 0.01 |
| **Unweighted CART on $\widetilde{\mathcal{D}}$** | 0.11 | 0.03 |
| **CART on $\widehat{\mathcal{D}}$** | 0.08 | 0.01 |

Table 1: *Average over 50 runs of the prediction error on $\mathcal{D}_{test}$ and its standard deviation.*

Overall, taking into account the inclusion probability allows to consider a training set of reduced size and therefore reduce the computationnal complexity of the learning procedure without damaging the quality of the prediction.

The same conclusions can be drawn from the analysis of the following datasets which were obtained via a stratified sampling design. We point out that this sampling scheme involves *negatively associated* (binary) random variables so that the theoretical results obtained in this paper apply to training data sampled by means of this scheme as well.

|  | incaIndiv | GJB | privacy3 | privacy4 |
| --- | --- | --- | --- | --- |
| **N** | 4079 | 2001 | 316 | 301 |
| **Number of features** | 326 | 130 | 95 | 124 |

The dataset *incaIndiv* [2] contain informations on the food consumption of the french population. The dataset *GJB*[3] contains questions about job seeking and the internet, workforce automation, online dating and smartphone use among Americans. The datasets *privacy3*[4] and *privacy4*[5] contain questions about privacy and information sharing. On the

---

2. https://https://www.data.gouv.fr/fr/datasets/

3. http://www.pewinternet.org/datasets/june-10-july-12-2015-gaming-jobs-and-broadband/

4. http://www.pewinternet.org/datasets/nov-26-2014-jan-3-2015-privacy-panel-3/

5. http://www.pewinternet.org/datasets/jan-27-feb-16-2015-privacy-panel-4/

datasets *incaIndiv* and *incaCompl* we try to predict whether or not someone is an adult, on the dataset *GJB* we will try to learn to predict the gender, and on the datasets *privacy3* and *privacy4* we will predict an answer to some questions among 5 possibilities.

We perform our experiments by randomly splitting the datasets *incaIndiv, incaCompl,GJB* into a training set (roughly 70 percent of the initial dataset) and a test set. The size of *privacy3* and *privacy4* being much smaller we perform 10-fold cross-validation.

|  | *incaIndiv* | *GJB* | *privacy3* | *privacy4* |
|---|---|---|---|---|
| **Weighted SVM** | 0.16 | 0.36 | 0.46 | 0.48 |
| **Unweighted SVM** | 0.19 | 0.43 | 0.50 | 0.52 |
| **Weighted CART** | 0.04 | 0.41 | 0.49 | 0.54 |
| **Unweighted CART** | 0.05 | 0.43 | 0.52 | 0.57 |

Table 2: *Average over 50 runs of the prediction error*

## 5. Conclusion

Most theoretical studies providing a statistical explanation for the success of learning algorithms based on the ERM paradigm fully ignore the possible impact of the sampling scheme producing the training data and stipulate that observations are independent replications of a generic r.v. or are uniformly sampled without replacement in a larger dataset. Through the generalizable example of rejective sampling, this paper shows that such studies can be extended to situations where training data are obtained by more general sampling schemes and possibly exhibit a complex dependence structure, provided that related probablity weights are appropriately incorporated in the risk functional.

## References

R Bardenet and O.A. Maillard. Concentration inequalities for sampling without replacement. *Bernoulli*, 21(3):1361–1385, 2015.

P. Bartlett, M. Jordan, and J. McAuliffe. Convexity, classification and risk bounds. *J. of the A.M.S*, 101(473):138–156, 2006.

Y.G. Berger. Rate of convergence to normal distribution for the Horvitz-Thompson estimator. *J. Stat. Plan. Inf*, 67(2):209–226, 1998.

S. N. Bernstein. On a modification of chebyshev's inequality and on the error in laplace formula. *Collected Works, Izd-vo 'Nauka', Moscow (in Russian)*, 4:71–80, 1964.

S. Boucheron, O. Bousquet, and G. Lugosi. Theory of Classification: A Survey of Some Recent Advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.

P. Brändén and J. Jonasson. Negative dependence in sampling. *Scandinavian Journal of Statistics*, 39(4):830–838, 2012.

N.E. Breslow and J.A. Wellner. Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. *Scandinavian Journal of Statistics*, 35:186–192, 2007.

N.E. Breslow, T. Lumley, C. Ballantyne, L. Chambless, and M. Kulich. Improved Horvitz-Thompson estimation of model parameters from two-phase stratified samples: applications in epidemiology. *Stat. Biosc.*, 1:32–49, 2009.

X.H. Chen, Dempster A.P., and J.S. Liu. Weighted finite population sampling to maximize entropy. *Biometrika*, 81(3):457–469, 1994.

J.C. Deville. *Réplications d'échantillons, demi-échantillons, Jackknife, bootstrap dans les sondages*. Economica, Ed. Droesbeke, Tassi, Fichet, 1987.

J. Dupacova. A note on rejective sampling. *Contribution to Statistics (J. Hajek memorial volume) Academia Prague*, pages 71–78, 1979.

L.A. Goodman. On the estimation of the number of classes in a population. *Annals of Mathematical Statistics*, 20:572–579, 1949.

J. Hajek. Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, 35(4):1491–1523, 1964.

D.G. Horvitz and D.J. Thompson. A generalization of sampling without replacement from a finite universe. *JASA*, 47:663–685, 1951.

S. Janson. Large deviation inequalities for sums of indicator variables. 1994.

K. Joag-Dev and F. Proschan. Negative Association of Random Variables with Applications. *The Annals of Statistics*, 11(1):286–295, 1983.

V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization (with discussion). *The Annals of Statistics*, 34:2593–2706, 2006.

J.B. Kramer, J. Cutler, and A.J. Radcliffe. Negative dependence and srinivasan's sampling process. *Combinatorics, Probability and Computing*, 20(3):347–361, 2011.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

P.M. Robinson. On the convergence of the Horvitz-Thompson estimator. *Australian Journal of Statistics*, 24(2):234–238, 1982.

T. Saegusa and J.A. Wellner. Weighted likelihood estimation under two-phase sampling. 2011.

C.E. Särndall and J. Wretman B. Swensson. *Model assisted survey sampling*. Springer-Verlag, NY, 2003.

R.J. Serfling. Probability inequalities for the sum in sampling without replacement. *The Annals of Statistics*, 2:39–48, 1974.

I. Steinwart, D. Hush, and C. Scovel. Learning from dependent observations. *Journal of Multivariate Analysis*, 100(1):175–194, 2009.

Y. Tillé. *Sampling algorithms.* Springer Series in Statistics, 2006.

A.W. van der Vaart and J.A. Wellner. *Weak convergence and empirical processes.* Springer, 1996.

V. Vapnik. *The Nature of Statistical Learning Theory.* Springer-Verlag, New-York, 2001.

## Appendix A. Technical Proofs

### A.1. Proof of Theorem 3

Considering the usual representation of the distribution of $(\epsilon_1, \ldots, \epsilon_N)$ as the conditional distribution of a sample of independent Bernoulli variables $(\epsilon_1^*, \ldots, \epsilon_N^*)$ conditioned upon the event $\sum_{i=1}^{N} \epsilon_i^* = n$ (see subsection 2.2), the result is a consequence of Theorem 2.8 in Joag-Dev and Proschan (1983).

### A.2. Bernstein inequality for sums of negatively associated random variables

For simplicity, we first establish the following tail bound for negatively associated random variables, which extends the usual Bernstein inequality in the i.i.d. setting, see Bernstein (1964). Proofs of Proposition 4 and Theorem 8 are then deduced from Theorem 3 and Theorem 9 (see Appendix A) .

**Theorem 9** *Let $Z_1, \ldots, Z_N$ be negatively associated real valued random variables such that $|Z_i| \leq c < +\infty$ a.s. $\mathbb{E}[Z_i] = 0$ and $\mathbb{E}[Z_i^2] = \sigma_i^2$ for $1 \leq i \leq N$. Then, for all $t > 0$, we have: $\forall N \geq 1$,*

$$\mathbb{P}\left\{\sum_{i=1}^{N} Z_i \geq t\right\} \leq \exp\left(-\frac{t^2}{\frac{2}{3}ct + 2\sum_{i=1}^{N}\sigma_i^2}\right).$$

Before detailing the proof, observe that a similar bound holds true for the tail probability $\mathbb{P}\left(\sum_{i=1}^{N} Z_i \leq -t\right)$ (and for $\mathbb{P}\left(|\sum_{i=1}^{N} Z_i| \geq t\right)$ as well, up to a multiplicative factor 2). Refer also to Theorem 4 in Janson (1994) for a similar result in a more restrictive setting (*i.e.* for tail bounds related to sums of negatively associated r.v.'s).

**Proof** The proof starts off with the usual Chernoff method: for all $\lambda > 0$,

$$\mathbb{P}\left\{\sum_{i=1}^{N} Z_i \geq t\right\} \leq \exp\left(-t\lambda + \log\mathbb{E}\left[e^{t\sum_{i=1}^{N} Z_i}\right]\right). \tag{15}$$

Next, observe that, for all $t > 0$, we have

$$\mathbb{E}\left[e^{t\sum_{i=1}^{n} Z_i}\right] = \mathbb{E}\left[e^{tZ_n} e^{t\sum_{i=1}^{n-1} Z_i}\right] \leq \mathbb{E}\left[e^{tZ_n}\right]\mathbb{E}\left[e^{t\sum_{i=1}^{n-1} Z_i}\right] \leq \prod_{i=1}^{n}\mathbb{E}\left[e^{tZ_i}\right], \tag{16}$$

using the property (9) combined with a descending recurrence on $i$. The proof is finished by plugging (16) into (15), using an adequate control of the log-Laplace transform of the $Z_i$'s and optimizing finally the resulting bound w.r.t. $\lambda > 0$, just like in the proof of the classic Bernstein inequality, see Bernstein (1964). $\blacksquare$

## Appendix B. On Biased HT Risk Minimization

Eq. (11) directly results from the following lemma.

**Lemma 10** *We have, for $p_i's$ such Suppose that $d_N \geq 1$. We have, for all $i \in \{1, \ldots, N\}$,*

$$|1/\pi_i - 1/p_i| \leq \frac{6}{d_N} \times (1 - \pi_i)/\pi_i.$$

**Proof** The proof follows from the representation (5.14) on p1509 in Hajek (1964). Denote by $P_N$ a Poisson sampling distribution on $\mathcal{I}_N$ with inclusion probabilities $p_1, ldots, p_N$, the canonical parameters of $R_N$. For all $i \in \{1, \ldots, N\}$, we have:

$$\frac{\pi_i}{p_i} \frac{1 - p_i}{1 - \pi_i} = \left( \sum_{s \in \mathcal{P}(\mathcal{I}_N): i \in \mathcal{I}_N \setminus \{s\}} P(s) \right)^{-1} \sum_{s \in \mathcal{P}(\mathcal{I}_N): i \in \mathcal{I}_N \setminus \{s\}} P(s) \sum_{h \in s} \frac{1 - p_h}{\sum_{j \in s}(1 - p_j) + (p_h - p_i)}$$

$$= \left( \sum_{s: i \in \mathcal{I}_N \setminus \{s\}} P_N(s) \right)^{-1} \sum_{s: i \in \mathcal{I}_N \setminus \{s\}} P_N(s) \sum_{h \in s} \frac{1 - p_h}{\sum_{j \in s}(1 - p_j) \left( 1 + \frac{(p_h - p_i)}{\sum_{j \in s}(1 - p_j)} \right)}$$

Now recall that for any $x \in ]-1, 1[$, we have:

$$1 - x \leq \frac{1}{1 + x} \leq 1 - x + x^2.$$

It follows that

$$\frac{\pi_i}{p_i} \frac{1 - p_i}{1 - \pi_i} \leq 1 - \left( \sum_{s: i \in \mathcal{I}_N \setminus \{s\}} P(s) \right)^{-1} \sum_{s: i \in \mathcal{I}_N \setminus \{s\}} P(s) \sum_{h \in s} \frac{(1 - p_h)(p_h - p_i)}{\left( \sum_{j \in s}(1 - p_j) \right)^2}$$

$$+ \left( \sum_{s: i \in \mathcal{I}_N \setminus \{s\}} P(s) \right)^{-1} \sum_{s: i \in \mathcal{I}_N \setminus \{s\}} P(s) \sum_{h \in s} \frac{(1 - p_h)(p_h - p_i)^2}{\left( \sum_{j \in s}(1 - p_j) \right)^3}$$

Following now line by line the proof on p. 1510 in Hajek (1964) and noticing that $\sum_{j \in s}(1 - p_j) \geq 1/2 d_N$ (see Lemma 2.2 in Hajek (1964)), we have

$$\left| \sum_{h \in s} \frac{(1 - p_h)(p_h - p_i)}{\left( \sum_{j \in s}(1 - p_j) \right)^2} \right| \leq \frac{1}{\left( \sum_{j \in s}(1 - p_j) \right)} \leq \frac{2}{d_N}$$

155

and similarly

$$\sum_{h \in s} \frac{(1-p_h)(p_h - p_i)^2}{\left(\sum_{j \in s}(1-p_j)\right)^3} \leq \frac{1}{\left(\sum_{j \in s}(1-p_j)\right)^2} \leq \frac{4}{d_N^2}.$$

This yieds: $\forall i \in \{1, \ldots, N\}$,

$$1 - \frac{2}{d_N} \leq \frac{\pi_i}{p_i} \frac{1-p_i}{1-\pi_i} \leq 1 + \frac{2}{d_N} + \frac{4}{d_N^2}$$

and

$$p_i(1-\pi_i)(1 - \frac{2}{d_N}) \leq \pi_i(1-p_i) \leq p_i(1-\pi_i)(1 + \frac{2}{d_N} + \frac{4}{d_N^2}),$$

leading then to

$$-\frac{2}{d_N}(1-\pi_i)p_i \leq \pi_i - p_i \leq p_i(1-\pi_i)(\frac{2}{d_N} + \frac{4}{d_N^2})$$

and finally to

$$-\frac{(1-\pi_i)}{\pi_i}\frac{2}{d_N} \leq \frac{1}{p_i} - \frac{1}{\pi_i} \leq \frac{(1-\pi_i)}{\pi_i}(\frac{2}{d_N} + \frac{4}{d_N^2}).$$

Since $1/d_N^2 \leq 1/d_N$ as soon as $d_N \geq 1$, the lemma is proved. ∎

## Appendix C. Sampling Training Data - Technical Details

### C.1. Further details on the rejective scheme

Let $n \leq N$ and consider a vector $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)$ of first order inclusion probabilities. Further define $\mathcal{S}_n := \{s \in \mathcal{P}(\mathcal{I}_N) : \#s = n\}$, the set of all samples in population $\mathcal{I}_N$ with cardinality $n$. The rejective sampling Hajek (1964); Berger (1998), sometimes called conditional Poisson sampling, exponential design without replacement or maximum entropy design, is the sampling design $R_N$ that selects samples of fixed size $n(s) = n$ so as to maximize the entropy measure $H(R_N) = -\sum_{s \in \mathcal{S}_n} R_N(s) \log R_N(s)$, subject to the constraint that its vector of first order inclusion probabilities coincides with $\boldsymbol{\pi}$. It is easily implemented in two steps:

1. Draw a sample $S$ according to a Poisson plan $P_N$, with properly chosen first order inclusion probabilities $\mathbf{p}_N = (p_1, \ldots, p_N)$. The representation is called canonical if $\sum_{i=1}^{N} p_i = n$. In that case, relationships between each $p_i$ and $\pi_i$, $1 \leq i \leq N$, are established in Hajek (1964).

2. If $n(S) \neq n$, then reject sample $S$ and go back to step one, otherwise stop.

Vector $\mathbf{p}$ must be chosen in a way that the resulting first order inclusion probabilities coincide with $\boldsymbol{\pi}$, by means of a dedicated optimization algorithm Tillé (2006). The corresponding probability distribution is given for all $s \in \mathcal{P}(\mathcal{I}_N)$ by $R_N(s) = \frac{P_N(s)\,\mathbb{I}\{\#s=n\}}{\sum_{s' \in \mathcal{S}_n} P_N(s')} \propto \prod_{i \in s} p_i \prod_{i \notin s}(1-p_i) \times \mathbb{I}\{\#s=n\}$, where $\propto$ denotes the proportionality.

### C.2. Stratified sampling

A stratified sampling design permits to draw a sample $S$ of fixed size $n(S) = n \leq N$ within a population $\mathcal{I}_N$ that can be partitioned into $K \geq 1$ distinct strata $\mathcal{I}_{N_1}, \ldots, \mathcal{I}_{N_K}$ (known a priori) of respective sizes $N_1, \ldots, N_K$ adding up to $N$. Let $n_1, \ldots, n_K$ be non-negative integers such that $n_1 + \cdots + n_K = n$, then the drawing procedure is implemented in $K$ steps: within each stratum $\mathcal{I}_{N_k}$, $k \in \{1, \ldots, K\}$, perform a SWOR of size $n_k \leq N_k$ yielding a sample $S_k$. The final sample is obtained by assembling these sub-samples: $S = \bigcup_{k=1}^K S_k$. The probability of drawing a specific sample $s$ by means of this survey design is $R_N^{\text{str}}(s) = \sum_{k=1}^K \binom{N_k}{n_k}^{-1}$. Naturally, first and second order inclusion probabilities depend on the stratum to which each unit belong: for all $i \neq j$ in $\mathcal{U}_N$, $\pi_i(R_N^{\text{str}}) = \sum_{k=1}^K \frac{n_k}{N_k} \mathbb{I}\{i \in \mathcal{U}_{N_k}\}$ and $\pi_{i,j}(R_N^{\text{str}}) = \sum_{k=1}^K \frac{n_k(n_k-1)}{N_k(N_k-1)} \mathbb{I}\{(i,j) \in \mathcal{U}_{N_k}^2\}$.

### C.3. Rao-Sampford sampling

The Rao-Sampford sampling design generates samples $s \in \mathcal{P}(\mathcal{I}_N)$ of fixed size $n(s) = n$ with respect to some given first order inclusion probabilities $\boldsymbol{\pi}^{RS} := (\pi_1^{RS}, \ldots, \pi_N^{RS})$, fulfilling the condition $\sum_{i=1}^N \pi_i^{RS} = n$, with probability

$$R_N^{RS}(s) = \eta \sum_{i \in s} \pi_i^{RS} \prod_{j \notin s} \frac{\pi_j^{RS}}{1 - \pi_j^{RS}}.$$

Here, $\eta > 0$ is chosen such that $\sum_{s \in \mathcal{P}(\mathcal{I}_N)} R_N^{RS}(s) = 1$. In practice, the following algorithm is often used to implement such a design Berger (1998):

1. select the first unit $i$ with probability $\pi_i^{RS}/n$,

2. select the remaining $n-1$ units $j$ with drawing probabilities proportional to $\pi_j^{RS}/(1 - \pi_j^{RS})$, $j = 1, \ldots, N$,

3. accept the sample if the units drawn are all distinct, otherwise reject it and go back to step one.