Are labels informative in semi-supervised learning? Estimating and leveraging the missing-data mechanism

Aude Sportisse [1] **Hugo Schmutz** [1 2] **Olivier Humbert** [2] **Charles Bouveyron** [1] **Pierre-Alexandre Mattei** [1]

Abstract

Semi-supervised learning is a powerful technique for leveraging unlabeled data to improve machine learning models, but it can be affected by the presence of "informative" labels, which occur when some classes are more likely to be labeled than others. In the missing data literature, such labels are called missing not at random. In this paper, we propose a novel approach to address this issue by estimating the missing-data mechanism and using inverse propensity weighting to debias any SSL algorithm, including those using data augmentation. We also propose a likelihood ratio test to assess whether or not labels are indeed informative. Finally, we demonstrate the performance of the proposed methods on different datasets, in particular on two medical datasets for which we design pseudo-realistic missing data scenarios.

1. Introduction

Technological advancements have enabled the collection and storage of vast amounts of data, offering real hope for better prediction of phenomena. Unfortunately, this also leads to dirty data, and more specifically, missing data. In this paper, we focus on the scenario where a large amount of data is available, but labeling the data is costly, time-consuming, or even risky (for instance, medical data collection requires invasive tests for patients). Semi-supervised learning (Scudder, 1965) has emerged as a crucial problem to leverage both labeled and unlabeled data in predictive model (Chapelle et al., 2006, Chapter 1). The unlabeled data are treated as observations with missing labels, as previously done in various studies (Grandvalet & Bengio, 2004; Ahfock & McLachlan, 2019; Hu et al., 2022; Schmutz et al., 2023). Recently, SSL algorithms have been extended to deep learning techniques demonstrating remarkable empirical successes, particularly through the systematic use of data augmentation (Berthelot et al., 2019a; Xie et al., 2020; Sohn et al., 2020; Rizve et al., 2021).

One of the challenges in semi-supervised learning is that the distribution of labels in the unlabeled dataset is unknown. For instance, it is uncertain whether a class that is well-represented in the labeled images is also well-represented in the unlabeled images. The traditional approach is to assume that the label distributions are identical in the labeled and unlabeled datasets. This assumption implies that people label classes in equal proportions, regardless of the class nature or the quality of the images. However, it disregards the potential unbalance of popularity among classes. For example, in a medical context, doctors may prioritize labeling the class of sick patients or leave unlabeled the data with an ambiguous diagnosis. When the label distribution differs in the labeled and unlabeled datasets, the missing labels are said to be informative or Missing Not At Random (MNAR). The missingness of a label must be taken into account to obtain results from the available data that can be generalized to the entire population (Rubin, 1976). This is usually modeled by the missing-data mechanism, i.e. the probability of a sample to be observed (depending on the values of the label itself). Recently, it has been shown that classical SSL algorithms indeed fail to provide accurate results for the less observed classes in presence of informative labels. As there is a selection bias in the sample, MNAR data also raise the issue that some models can lead to non-identifiable parameters (Baker & Laird, 1988; Miao et al., 2016). A major challenge is that testing whether the data is indeed MNAR is difficult (d'Haultfoeuille, 2010), but it is necessary to provide a guideline for choosing which algorithm to apply. The main objective of this paper is to address these issues by estimating the missing-data mechanism.

Beyond the scope of deep learning, in the missing-data literature, significant works have considered the case of MNAR responses (Tang & Ju, 2018). Ibrahim & Lipsitz (1996) and Ibrahim et al. (2001) estimate the parameters of both the model and the missing-data mechanism using the Expectation-Maximization (EM) algorithm in binomial regression and generalized linear models. They also propose

[1]Université Côte d'Azur, Inria, Maasai, LJAD, CNRS [2]Université Côte d'Azur, TIRO-MATOS, UMR CEA E4320. Correspondence to: Aude Sportisse <aude.sportisse@inria.fr>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

a likelihood ratio test statistic for selecting the variables related to the missingness but leave the identifiability of the parameters in perspective. It is in the semi-parametric setting that a lot of work has been done to obtain identification results, most often using a *shadow* variable (Miao et al., 2019; Miao & Tchetgen Tchetgen, 2016) that adds auxiliary information (Molenberghs et al., 2008). Some works (Shao & Wang, 2016; Morikawa et al., 2017) also propose to debias classical estimators by using inverse probability weighting (IPW) techniques, weighting each sample by the inverse of its probability of being observed as determined by the missing-data mechanism. However, only the recent work of Hu et al. (2022) proposes an extension to deep learning, debiasing the risk estimator with a propensity score, but they do not directly model the missing-data mechanism, which is the main focus of our study (see Section 3.2 for a comprehensive comparison).

Our key contributions are summarized as follows:

- We consider a general self-masked MNAR model and prove its identifiability, showing in the process of identifiability of the model of Hu et al. (2022).
- We propose two estimates of the missing-data mechanism and show their consistency.
- Based on these estimates, we propose an algorithm using IPW techniques able to debias any SSL algorithm in presence of informative labels.
- We provide a heuristic procedure to test whether the labels are indeed MNAR.
- We first demonstrate the efficiency of our methods on classical datasets. Furthermore, we propose two new pseudo-realistic MNAR scenarios using medMNIST datasets (Yang et al., 2023). These contrast with the toy missing-data scenarios often used in existing works, even when the method is designed to handle informative labels. Our code is available here[1].

2. Informative labels

2.1. Missing labels typology

In this paper, we study a dataset of n samples, denoted as $D = (x_i, y_i)_{i=1}^n$, where x_i represents the features and y_i represents the labels, which are drawn from the distribution $p(x, y) = p(x)p(y|x)$. Some of the labels are supposed to be missing, and thus the dataset is split into two subsets: a labeled dataset $D_\ell = (x_i, y_i)_{i=1}^{n_\ell}$ of size n_ℓ and an unlabeled dataset $D_u = (x_i)_{i=n_\ell+1}^n$ of size $n_u = n - n_\ell$. The distribution of the labeled dataset is denoted as $p^\ell(.)$ (resp. $p^u(.)$ for the unlabeled dataset). In the following, we consider a discrete set of labels denoted as $\mathcal{C} = \{1, \dots, K\}$,

[1] https://github.com/AudeSportisse/SSL_MNAR

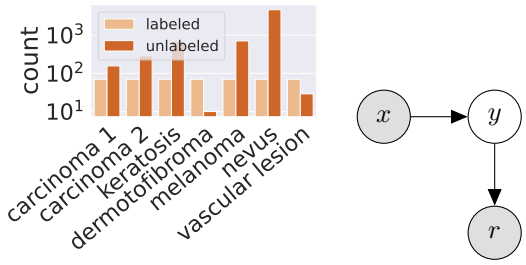

Figure 1. Left panel: MNAR labels for dermaMNIST (log count of labeled and unlabeled images per class). Right panel: Structural causal graph of the self-masked mechanism (Assumption **A2.**). The nodes in grey represent fully observed variables and the edges from x to y means that x causes y.

with K the number of classes.

Most of the semi-supervised learning methods make the following assumption:

A1. The marginal distributions of the features and of the labels are identical in the labeled and unlabeled dataset, i.e. $p^\ell(x) = p^u(x)$, $\forall x$ and $p^\ell(y) = p^u(y)$, $\forall y$.

Assumption **A1.** means that people label classes in equal proportions, regardless of their nature (label) or the quality of the data (features). Modeling two separate distributions, $p^\ell(.)$ and $p^u(.)$, is not always convenient, so we instead use a notation commonly used in missing-data studies. We introduce an additional random variable called missing-data indicator, $r \in \{0, 1\}$, where $r = 1$ if y is observed and $r = 0$ if y is missing. For example, this notation implies $p^\ell(x) = p(x|r = 1)$ and $p^u(x) = p(x|r = 0)$. According to Rubin's (1976) typology, labels can be: (i) Missing Completely At Random (MCAR) if the cause of the missingness is completely independent from the data values, i.e. $r \perp\!\!\!\perp x, y$ (equivalent to Assumption **A1.**), (ii) Missing At Random (MAR) if the cause of the missingness can be explained by the features, $r \perp\!\!\!\perp y|x$ and (iii) Missing Not At Random (MNAR) in all other cases. For example, labels will be MAR if medical doctors are less likely to label analyses that are of poor quality but they will be MNAR if they prefer to label the class of sick patients first. This last situation creates an "unbalanced class popularity" for which Assumption **A1.** is no longer valid.

In this paper, we focus on the MNAR case, specifically, the labels are assumed to be "self-masked MNAR", meaning their unavailability only depends on their own values. This assumption allows us to model the unbalanced class popularity situation (see Figure 1) and is widely used in the missing-data literature (Mohan, 2018; Sportisse et al., 2020). Our assumption is formalized as follows:

A2. The labels are self-masked MNAR, i.e. $r \perp\!\!\!\perp x|y$.

Assumption **A2.** is weaker than Assumption **A1.**, since the equality of marginal distributions, either for features or labels, in both the labeled and unlabeled datasets, is relaxed. It only requires that the conditional distribution of the features given the class is the same in the labeled and unlabeled datasets (i.e. $p^\ell(x|y) = p^u(x|y), : \forall x, y$). For example, it does not cover the case where the radiography of sick patients does not have the same resolution whether it is labeled or not.

Remark 2.1 (More general assumptions). The notation introduced here does not allow to consider different sets for the labels present in the labeled dataset and in the unlabeled dataset. Label distribution mismatch has already been considered, such as when new classes appear in the unlabeled dataset (Guo et al., 2020; Cao et al., 2021) or when none of the classes present in the labeled dataset are present in the unlabeled dataset (Chen et al., 2020; Huang et al., 2021). However, these works are beyond the context of our work, as they do not allow to directly account for the informative nature of the labels, which is the main focus of this paper.

2.2. Non-ignorable missing-data mechanism

This typology of missing-data mechanism is important in determining the appropriate method to use: statistical inference can be performed on $p(x, y)$ for MCAR or MAR labels, but it should be performed on $p(x, y, r)$ for MNAR labels (Little & Rubin, 2019). We denote the parameter of interest, $\theta \in \Theta$, of $p(y|x; \theta)$. In most cases, this parameter corresponds to the weights of a neural network, or in simpler cases, a logistic regression. The parameter ϕ of the missing data mechanism $p(r|x, y; \phi)$ lives in $\Phi = [0, 1]^K$. In the following, we assume that the parameters are distinct, meaning that the joint parameter space is equal to $\Theta \times \Phi$. Following the common notation introduced by Le Morvan et al. (2020), the observed label vector is $(y \odot r)$, where $\odot$ represents the term-by-term product, such that $(y \odot r)_i = y_i$ if $r_i = 1$ and $(y \odot r)_i = \text{NA}$ if $r_i = 0$.

The traditional method for estimating θ is to minimize the negative observed log-likelihood:

$$\begin{aligned}
\ell(\theta, \phi) &= -\sum_{i=1}^{n} \log p(x_i, y_i \odot r_i, r_i; \theta, \phi) \\
&= -\sum_{i=1}^{n} \log p(r_i|x_i, y_i \odot r_i; \phi) p(y_i \odot r_i|x_i; \theta) p(x_i) \\
&= -\sum_{i=1}^{n} \begin{cases} \log p(r_i|x_i, y_i; \phi) p(y_i|x_i; \theta) & \text{if } r_i = 1 \\ \log \sum_{\tilde{y} \in \mathcal{C}} p(r_i|x_i, \tilde{y}; \phi) p(\tilde{y}|x_i; \theta) & \text{if } r_i = 0 \end{cases} + C \qquad (1) \\
&= -\sum_{i=1}^{n} r_i \log p(y_i|x_i; \theta) + C \ \text{ under M(C)AR assumption,}
\end{aligned}$$

where C is a constant independent of θ. For M(C)AR labels, we use in the last step that $p(r_i|x_i, y; \phi)$ does not depend on y and that $\sum_{\tilde{y} \in \mathcal{C}} p(\tilde{y}|x_i; \theta) = 1$; the result follows if ϕ is considered as a nuisance term. This simple calculation is a common technique in the missing data literature (Little & Rubin, 2019). It implies that for MCAR or MAR labels, it is not necessary to estimate the missing-data mechanism and minimizing the complete likelihood (on labeled data only) is sufficient. However, for MNAR labels, the missing-data mechanism cannot be ignored and must be taken into account.

2.3. Identification of the joint distribution

To fix ideas, Figure 1 (right panel) shows the causal relationships between the variables x, y and the missing-data indicator r through a structural causal graph (Neuberg, 2003). Assumption **A2.** allows us to get the nonparametric identification of the joint distribution $p(y, x, r)$, i.e. it can be expressed with quantities involving only observed data. Specifically, in the self-masked setting, the features act as shadow variables, providing enough auxiliary observed information to achieve the identifiability of the parameters.

Proposition 2.2 (Identification of the joint distribution)**.** *Under Assumptions **A2.** (self-masked MNAR), the joint distribution $p(y, x, r)$ is identified.*

This result is a corollary of Theorem 1 in (Miao et al., 2019) and is proved in Appendix A. It is worth noting that this result also demonstrates the identification in (Hu et al., 2022).

3. Debiasing classical SSL algorithms

3.1. Complete-case: learning with labeled data

In classical supervised learning, the aim is to learn a predictive model $p(y|x; \theta)$, parametrized by $\theta \in \Theta$, by minimizing the theoretical risk:

$$\theta^\star = \text{argmin}_{\theta \in \Theta} \quad \mathcal{R}(\theta) := \mathbb{E}_{(x,y) \sim p(x,y)}[\ell_\ell(\theta; x, y)],$$

where ℓ_ℓ is typically the negative log-likelihood function $\ell_\ell(\theta; x_i, y_i) = -\log p(y_i|x_i; \theta)$ but can be any loss function. The theoretical risk is never observed, as it requires the knowledge of the true distribution $p(x, y)$. A typical learning procedure is then the minimization of the empirical risk, which is an unbiased estimate of the theoretical risk:

$$\hat{\theta} = \text{argmin}_{\theta \in \Theta} \quad \hat{\mathcal{R}}(\theta) := \frac{1}{n} \sum_{i=1}^{n} \ell_\ell(\theta; x_i, y_i).$$

This quantity is known in the supervised learning setting, but is still unobserved in the presence of missing labels. For MCAR labels, the natural estimator for $\mathcal{R}(\theta)$ is the complete-case empirical risk, computed for the labeled data only: $\hat{\mathcal{R}}^{CC}(\theta) := \frac{1}{n_\ell} \sum_{i=1}^{n} r_i \ell_\ell(\theta; x_i, y_i)$. It is unbiased for MCAR labels but not in the other cases. For MAR

labels, Liu & Goldberg (2020) propose to use the IPW estimator defined as $\tilde{\mathcal{R}}^{\text{IPW,MAR}}(\theta) := \frac{1}{n}\sum_{i=1}^{n} \frac{r_i \ell_\ell(\theta; x_i, y_i)}{\pi^{\text{MAR}}(x_i)}$, where $\pi^{\text{MAR}}(x) = \mathbb{P}(r = 1|x)$. Similarly, for self-masked MNAR labels, we propose the following IPW estimator:

$$\hat{\mathcal{R}}_\phi(\theta) := \frac{1}{n}\sum_{i=1}^{n} \frac{r_i \ell_\ell(\theta; x_i, y_i)}{\phi_{y_i}}, \tag{2}$$

where $\phi = (\phi_0, ..., \phi_K) \in \Phi = [0, 1]^K$, and $\forall k \in \mathcal{C}$, $\phi_k := \mathbb{P}(r = 1|y = k)$.

The idea behind the IPW technique is that one labeled sample $(x_i, y_i, r_i = 1)$ should not be counted only once but should take into account that there are unlabeled samples $(x_j, y_j, r_j = 0), j \neq i$ that belong to the same class $(y_j = y_i)$. As a result, it is then counted $1/\phi_{y_i}$ times. For example, if the probability of being observed in a class is one third, an observed sample from that class will be counted three times.

Proposition 3.1 (Unbiasedness of the IPW estimator)**.** *The IPW estimator proposed in* (2) *is unbiased, if the mechanism is well specified, i.e.* $\mathbb{E}[r|y] = \phi_y$.

Proof.

$$\begin{aligned}\mathbb{E}\left[\frac{r}{\phi_y}\ell_\ell(\theta; x, y)\right] &= \mathbb{E}\left[\mathbb{E}\left[\frac{r}{\phi_y}\ell_\ell(\theta; x, y)|y\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\frac{r}{\phi_y}|y\right]\mathbb{E}\left[\ell_\ell(\theta; x, y)|y\right]\right] \\ &= \mathbb{E}\left[\frac{\mathbb{E}[r|y]}{\phi_y}\mathbb{E}\left[\ell_\ell(\theta; x, y)|y\right]\right] = \mathcal{R}(\theta),\end{aligned}$$

using $r \perp\!\!\!\perp x|y$ in the second equality. □

3.2. Incorporating the unlabeled data

A major drawback of the classical IPW estimator in (2) is that it only uses labeled data and not all available data. To address this, traditional SSL algorithms for MCAR labels add a regularization term to the classical supervised objective:

$$\hat{\mathcal{R}}^{\text{SSL}}(\theta) := \frac{1}{n}\sum_{i=1}^{n} r_i \frac{\ell_\ell(\theta; x_i, y_i)}{n_\ell/n} + \frac{\lambda}{n}\sum_{i=1}^{n}(1 - r_i)\frac{\ell_u(\theta; x_i)}{n_u/n}, \tag{3}$$

where $\lambda > 0$ is a regularization term. The function ℓ_u is a loss function which does not depend on the labels; Schmutz et al. (2023) note that ℓ_u can be viewed in many cases as a surrogate of ℓ_ℓ.

For example, Grandvalet & Bengio (2004) use the Shannon entropy. Another popular approach is to use "pseudo-labels" (Rizve et al., 2021) for unlabeled data by selecting the class with the highest posterior probability.Only the pseudo-labels that have a predicted probability higher than a predefined threshold τ are used as targets. Both methods encourage the model to have a high level of confidence when imputing unlabeled data, but pseudo-label methods only use data points that have already been predicted with high confidence. Recently, state-of-the-art methods such as (Sohn et al., 2020; Berthelot et al., 2019a) have also been developed to make the model more robust to data augmentation of the features.

For MNAR labels, the standard estimator in (3) is biased, and we propose the following estimator, which is unbiased if the mechanism is correctly specified, using the same argument as Proposition 3.1:

$$\hat{\mathcal{R}}_\phi^{\text{SSL}}(\theta) := \frac{1}{n}\sum_{i=1}^{n} \frac{r_i \ell_\ell(\theta; x_i, y_i)}{\phi_{y_i}} - \frac{\lambda}{n}\sum_{i=1}^{n} \frac{r_i - \phi_{y_i}}{\phi_{y_i}}\ell_u(\theta; x_i). \tag{4}$$

This estimator has the significant advantage of being able to debias any SSL algorithm, including methods using data augmentation, with the knowledge of the weights ϕ_{y_i}. This is the MNAR counterpart of the estimator proposed by Schmutz et al. (2023) for MCAR data and of the one suggested by Liu & Goldberg (2020) for MAR data with $\lambda = 1$. The only difference is the form of the mechanism: for MCAR data, $\phi_{y_i} = n_\ell/n, \forall i$ and for MAR data, they use $\pi^{\text{MAR}}(x)$ defined in Section 3.1 instead.

Comparison with the work of Hu et al. (2022) Our estimator also shares similarities with the "doubly-robust" estimate suggested by Hu et al. (2022), for debiasing the classical SSL estimators of the risk for self-masked MNAR labels (Assumption **A2.**). They build their estimator upon a very interesting strategy also used in the missing-data literature (Rabe-Hesketh & Skrondal, 2023) and leverage from it to indirectly account for the informative nature of the missing labels. The form of the risk estimator is as (4), but they target a composite likelihood (which does not encompass the cross entropy) by starting from $\hat{\theta} \in \text{argmin}_\theta - \log p(x|y; \theta)$ instead of $\hat{\theta} \in \text{argmin}_\theta - \log p(y|x; \theta)$. The biggest advantage of their strategy is that it allows not to directly model the missing-data mechanism, which can be tedious in some missing data scenarios. Besides, the fundamental difference between our work and theirs is their weight does not involve the missing- data mechanism, but only the class proportions $p(y)$. Their method thus encourages the model to be accurate for the least frequent classes (when $p(y)$ is small) but will not detect or favour the least labeled classes (when ϕ_y is small). On the contrary, our method will benefit from the estimation of the missing-data mechanism, which can be obtained at no extra computational cost (see Section 4.5).

4. Estimating the missing-data mechanism

In Section 3, we proposed two unbiased estimates of the risk when the labels are informative. However, both require the knowledge of the missing-data mechanism. In practice, we will use the estimators given in (2) and (4) by plugging-in an estimation $\hat{\phi}$ of the mechanism, resulting in $\hat{\mathcal{R}}_{\hat{\phi}}$ and $\hat{\mathcal{R}}_{\hat{\phi}}^{\text{SSL}}(\theta)$. In this section, we provide two estimators of the missing-data mechanism by using either the method of moments or the method of maximum likelihood.

4.1. Moment estimator

A possible estimator of the missing-data mechanism is obtained by the method of moments applied to $p(r=1, y=y) = \mathbb{E}\left[\mathbb{1}_{\{r=1,y\}}\right] = \phi_y p(y)$. It implies

$$\phi_y^M = \frac{\sum_{i=1}^n \mathbb{1}_{\{r=1,y_i=y\}}}{n} \frac{1}{p(y)}, \forall y \in \mathcal{C}. \tag{5}$$

This estimator allows us to leverage the information we have on the labeled data, because $p(r = 1, y)$ is estimated by counting the number of labeled data in each class ($\mathbb{1}_{r=1,y_i=y}$). The challenge now is to estimate the class distribution $p(y)$. This allows for two simple cases where the mechanism can be calculated directly: (i) when we know that the entire dataset is balanced (use $p(y) = 1/K$) and (ii) when we have prior information on the class proportions (use $p(y) = p_{\text{prior}}(y)$). This last case can happen when we have data from the general population (e.g. we know the prevalence rate of a disease).

When the class proportions are unknown, we propose to estimate $p(y)$ as follows:

$$\hat{p}(y;\theta) = \frac{1}{n}\sum_{i=1}^n p(y_i|x_i;\theta), \forall \theta \in \Theta, \tag{6}$$

which is a consistent estimator of $p(y)$ noting that $p(y) = \int p(y|x;\theta)p(x)dx$. The estimator of the missing-data mechanism is thus:

$$(\hat{\phi}_y^M)_\theta = \frac{\sum_{i=1}^n \mathbb{1}_{\{r=1,y_i=y\}}}{n} \frac{1}{\hat{p}(y,\theta)}, \forall y \in \mathcal{C}. \tag{7}$$

Remark 4.1 (Computation of $p_\theta(y)$). A SGD algorithm is commonly employed to estimate $\theta^\star$. However, since the estimator of class proportions, $\hat{p}(y,\theta)$, depends on θ, it is crucial to be mindful of the gradient propagation through it, as it may not always be necessary. We propose two ways to compute it within a mini-batch:

- use a moving averaging strategy inspired by (Hu et al., 2022), by using a buffer $\hat{p}_{\text{buffer}}(y)$ updated at each iteration with $\hat{p}(y, \theta = \theta_b)$ (6), where θ_b is the parameter of the current mini-batch b[2]:

$$\hat{p}_{\text{buffer}}(y) = \mu\hat{p}_{\text{buffer}}(y) + (1-\mu)\hat{p}(y,\theta_b). \tag{8}$$

- propagate the gradients through $(\hat{\phi}_y^M)_\theta$.

4.2. Maximum likelihood estimator

The second estimator of the missing-data mechanism relies on the method of maximum likelihood, already carried out by Ibrahim & Lipsitz (1996) and Ibrahim et al. (2001) outside the scope of deep learning. It is obtained by minimizing the negative observed log-likelihood (1):

$$\hat{\theta}^L, \hat{\phi}^L = \text{argmin}_{\theta\in\Theta,\phi\in\Phi} \quad \ell(\theta,\phi;x,y\odot r,r). \tag{9}$$

We highlight the following points:

- (Two-steps algorithm for SSL). Even if (9) gives an estimator $\hat{\theta}^L$ of θ, the latter can be really improved by incorporating the unlabeled data as in $\hat{\mathcal{R}}_{\hat{\phi}}^{\text{SSL}}(\theta)$ in a second step (see Algorithm 1 in Section 4.5).
- (MCAR setting). For not informative labels, the unlabeled data are not used for the estimation, as noted in Section 2.2. Besides, as expected, the minimum of the function is attained for a mechanism equal to the proportion of the labeled data. Indeed, we have $\frac{\partial\ell(\theta,\phi_0)}{\partial\phi_0} = -\frac{n_l}{\phi_0} + \frac{n-n_l}{1-\phi_0}$ and $\frac{\partial\ell(\theta,\phi_0)}{\partial\phi_0} = 0 \Leftrightarrow \phi_0 = \frac{n_\ell}{n}$.
- (Convexity). The negative observed log-likelihood (1) is convex in ϕ, for a fixed $\theta \in \Theta$ (Appendix B.2).

Remark 4.2 (Solving (9) in practice). To our knowledge, there is no closed form for the minimization problem. In practice, we propose to calculate the gradients by using the automatic differentiation package in PyTorch (Paszke et al., 2019). To comply with the constraint $\phi \in \Phi$, we consider $\sigma(\phi_k) = \frac{1}{1+\exp(-\phi_k)}, \forall k \in \mathcal{C}$ instead of ϕ_k. In addition, we suggest solving (9) subject to the constraint of $\sum_y \frac{\sum_{i=1}^n \mathbb{1}_{\{r=1,y_i=y\}}}{n}\frac{1}{\hat{\phi}_y^L} = 1$, to comply with $\sum_y p(y) = 1$ (see (5)), by using the mdmm package.

4.3. Theoretical results

In this section, we provide theoretical results that validate the relevance of the chosen estimators. The proofs are detailed in Appendix B. We first demonstrate the consistency of the moment estimator for a fixed $\theta \in \Theta$ by applying general results such as the law of large numbers and Slutsky's theorem.

Proposition 4.3 (Consistency of $\hat{\phi}^M$)**.** *The moment estimator defined by* (7) *is consistent for a fixed* $\theta \in \Theta$.

[2] In this method, while $(\hat{\phi}_y^M)_\theta$ depends on θ, we do not propagate the gradients through θ.

Additionally, the consistency and asymptotic normality of the maximum likelihood estimator are obtained by applying Theorem 5.7 and Theorem 5.23 of (Van der Vaart, 1998) (stated in the more general case of M-estimators). We consider the negative observed log-likelihood for a fixed θ, denoted as $\ell_\theta : \phi \mapsto \ell(\theta, \phi; x, y \odot r, r)$. In Appendix, we prove that the associated statistical model $\mathcal{P}_\phi = \{p(r|y \odot r; \phi) : \phi \in \Phi\}$ is identifiable and that under interchangeability of differentiation w.r.t. ϕ and integration over (x, y, r) (Assumption **A3.**), the Fisher information evaluated at the oracle estimate is invertible. Besides, we assume that ϕ is in the interior of the set $\Phi = [0, 1]^K$, i.e. it cannot be on its boundary (Assumption **A4.**).

Proposition 4.4 (Consistency, asymptotic normality of $\hat{\phi}^L$)**.** *Under Assumptions **A3.** and **A4.**, the estimator $\hat{\phi}^L$ is consistent and asymptotically normal.*

Finally, the consistency of the estimator of the missing-data mechanism directly implies the consistency of the risks that we minimize in our SSL algorithms.

Proposition 4.5 (Consistency of the risk)**.** *If $\hat{\phi}$ is a consistent estimator of ϕ and if the mechanism is well specified, the risks $\hat{\mathcal{R}}_{\hat{\phi}}$ and $\hat{\mathcal{R}}^{\mathrm{SSL}}_{\hat{\phi}}(\theta)$ are consistent estimators of the theoretical risk $\mathcal{R}(\theta) = \mathbb{E}[\ell_\ell(\theta; x, y)]$.*

Remark 4.6 (Consistency of $\hat{\mathcal{R}}_{\hat{\phi}}$ using unlabeled data). As a consequence of the ignorability of the MCAR mechanism (see Section 2.2), the estimator of the theoretical risk using only labeled data is consistent in presence of MCAR labels. Proposition 4.5 shows that the IPW estimator $\hat{\mathcal{R}}_{\hat{\phi}}$ is consistent for MNAR labels. It is worth noting that its expression refers only to labeled data but involves an estimator of the missing-data mechanism, computed on both labeled and unlabeled data (see Equations (7) and (9)). This underlines the relevance of unlabeled data in SSL with MNAR labels.

Remark 4.7 (Double-robustness of the SSL risk $\hat{\mathcal{R}}^{\mathrm{SSL}}_{\hat{\phi}}$). An interesting property is double-robustness, meaning that the estimator is consistent even if either the missing-data mechanism estimation or imputation is inaccurate. Hu et al. (2022) prove that double-robustness of their debiased risk (see Section 3.2) holds, if they assume that under inaccurate propensity estimation, the imputation is perfect (in the sense that the model always predicts the right class). This is a strong assumption. In our work, double-robustness (in that sense) of $\hat{\mathcal{R}}^{\mathrm{SSL}}_{\hat{\phi}}$ is directly implied by Proposition 4.5 and by the unbiasedness of the risk. The strong assumption above could also be relaxed, applying Theorem 2 of Miao et al. (2019) to our case: this is left as a perpective of our work.

4.4. Testing the assumption on the mechanism

We present here a heuristics for estimating the missing-data mechanism to test if the labels are MCAR or not in the case of semi-supervised learning. The aim of such a test is to encourage the use of a specific method if the labels are not MCAR, or to support the selection of a traditional method if they are.

We want to test

$$H_0 : \phi \in \Phi^{\mathrm{MCAR}} \text{ against } H_1 : \phi \notin \Phi^{\mathrm{MCAR}},$$

where $\Phi^{\mathrm{MCAR}} = \{\phi \in [0, 1]^K, \forall k, k', \phi_k = \phi_{k'}\}$. For the maximum likelihood estimator given in (9), we consider the following test statistic:

$$-2 \left(\inf_{\theta,\phi} \ell(\theta, \phi) - \inf_{\theta} \ell(\theta, \phi^{\mathrm{MCAR}}) \right). \tag{10}$$

Under the same assumptions of Proposition 4.4, we know that the test stastic $2(\ell_\theta(\phi) - \ell_\theta(\phi^{\mathrm{MCAR}})$, for a fixed $\theta \in \Theta$, converges in distribution to a chi-squared random variable χ^2_d (Theorem 16.7 of Van der Vaart (1998)), where d is the difference in degrees of freedom between the null hypothesis (H_0) and the alternative hypothesis (H_1), i.e. $d = K - 1$ for K classes. We conjecture that an extension of Proposition 4.4 developed for a fixed $\theta \in \Theta$ can be obtained by considering the profile likelihood $\phi \mapsto \inf_{\theta \in \Theta} \ell(\theta, \phi)$, and by applying the asymptotic results on it (Murphy & Van der Vaart, 2000). This implies that the test statistic (10) also converges in distribution to χ^2_d. Based on this asymptotic distribution, it is possible to calculate a p-value from (10) and to easily test if the MCAR assumption is rejected.

4.5. Algorithms

To ensure clarity, we explain how the proposed estimators can be applied to any SSL algorithm. As previously mentioned, the goal is to plug-in the estimation of the mechanism into the estimators of the theoretical risk.

The moment estimator presented in (7) is continuously updated throughout the SSL algorithm (Algorithm 2), and thus the estimation of the mechanism does not add any additional computational cost when using the moment estimator, as the estimation of the mechanism and the model are performed in a single step. On the other hand, when using the maximum likelihood method, the estimation process is divided into two steps: (i) the estimation of the mechanism by optimizing (1) (Algorithm 1) and (ii) the estimation of the model θ using the SSL algorithm (Algorithm 2 using the estimator (i) as input for $\hat{\phi}$). In both algorithms, the hyperparameters are classical: the sizes of the mini-batch ($N_{\mathcal{B}}$,$N'_{\mathcal{B}}$), the learning rates (γ_ϕ,γ_θ,γ'_θ) and the number of epochs (N, N').

Remark 4.8 (Adaptive threshold). SSL algorithms (Rizve et al., 2021; Sohn et al., 2020) that utilize pseudo-label techniques often employ a fixed threshold to select relevant imputations from unlabeled data. However, several recent studies (Hu et al., 2022; Wei et al., 2021; Berthelot et al., 2019a) have noted that an adaptive threshold can improve the performance of the classifier on the rarest classes, particularly in unbalanced semi-supervised learning. For instance,

Algorithm 1 Maximum likelihood estimator for ϕ

Input: labeled data D_ℓ, unlabeled data D_u
Initialize θ_0 (at random), ϕ_0 (MCAR case: n_ℓ/n).
for $k = 0$ **to** N **iteratively do**
  Sample a Mini-Batch $\mathcal{B}$ of size $N_\mathcal{B}$ from D_ℓ and from D_u.
  $\phi_{k+1} = \phi_k - \gamma_\phi \partial_\phi \frac{1}{N_\mathcal{B}} \sum_{i\in\mathcal{B}} \ell(\theta_k, \phi_k)$
  Sample a Mini-Batch $\mathcal{B}$ of size $N_\mathcal{B}$ from D_ℓ and from D_u.
  $\theta_{k+1} = \theta_k - \gamma_\theta \partial_\theta \frac{1}{N_\mathcal{B}} \sum_{i\in\mathcal{B}} \ell(\theta_k, \phi_k)$
end for
Output: $\phi_N = \hat{\phi}^L, \theta_N$

Algorithm 2 Debiased SSL algorithm for informative labels

Input: labeled data D_ℓ, unlabeled data D_u, $\hat{\phi}$ (if available)
Initialize θ_0 (at random)
for $k = 1$ **to** N' **do**
  Sample a Mini-Batch $\mathcal{B}$ of size $N'_\mathcal{B}$ from D_ℓ and from D_u.
  if $\hat{\phi}$ is not provided **then**
    Compute $\hat{\phi}_y, \forall y \in \mathcal{C}$ by the method of moments (7).
  end if
  $\theta_{k+1} = \theta_k - \gamma'_\theta \partial_\theta \frac{1}{N'_\mathcal{B}} \sum_{i\in\mathcal{B}} \hat{\mathcal{R}}^{\text{SSL}}_{\hat{\phi}}(\theta_k)$
end for
Output: $\theta_{N'}$

Hu et al. (2022) propose to use an adaptive threshold based on class proportions, setting higher requirements for popular classes and lower requirements for rare classes. Another possible adaptive threshold, suggested by our estimation of the missing-data mechanism, would depend on the missingness proportion of a class and set the highest requirement for the most observed class:

$$\forall k \in \mathcal{C}, \tau(y_k) = \tau_0 \left(\frac{\mathbb{P}(r=1|y=k)}{\max_y \mathbb{P}(r=1|y)} \right)^\beta = \tau_0 \left(\frac{\phi_y}{\max_y \phi_y} \right)^\beta,$$

with τ_0 the classical threshold and β the hyper-parameter that determines how adaptive the cutoff is.

5. Numerical experiments

In this study, we evaluate the effectiveness of our proposed estimates of the missing-data mechanism using the benchmark dataset MNIST (LeCun & Cortes, 2010). Additionally, our debiased approach (Algorithm 2) of the classical SSL method using pseudo-labels (Scudder, 1965; Lee, 2013; Rizve et al., 2021) is compared in both its original implementation (**PseudoLabel**) and its debiased version for MCAR labels (**DePseudoLabel**) (Schmutz et al., 2023), using both the MNIST dataset and two datasets of MedMNIST (Yang et al., 2023). Furthermore, we compare our debiased version of Fixmatch (Sohn et al., 2020), designed to handle informative labels, with its original counterpart (**Fixmatch**) and its debiased version for MCAR labels (**DeFixmatch**) (Schmutz et al., 2023) on the CIFAR-10 dataset (Krizhevsky, 2009).

To evaluate the accuracy of our proposed estimates of the missing-data mechanism, we calculate the normalized Mean Squared Error (MSE) using the formula $\|\hat{\phi} - \phi^\star\|_2 / \|\phi^\star\|^2$. This provides a measure of how well our estimate of the missing-data mechanism ($\hat{\phi}$) approximates the true mechanism ($\phi^\star$). We consider four different estimators of the missing-data mechanism.

- **MLE**: the maximum likelihood estimator derived from Algorithm 1. As highlighted in Section 4.2, we use the estimation of θ given by Algorithm 2 when assessing the model's performance.
- **ME**: the moment estimator derived from Algorithm 2 by using a moving averaging strategy (8) for the class distribution.
- **MEg**: the moment estimator derived from Algorithm 2 by propagating the gradients through θ.
- **CADR**: the estimator derived from Hu et al. (2022). Although the authors did not propose an estimation of the missing-data mechanism, we are able to derive it directly from their estimation of the class proportions (see (7)).

5.1. MNIST and CIFAR-10 for toy mechanisms

5.1.1. MNIST

The MNIST dataset is an advantageous choice for SSL as the classes are well-separated, allowing us to verify the effectiveness of our method in simple cases. In order to randomly select the labeled and unlabeled data per class according to a specific distribution, we follow the method proposed by Hu et al. (2022). The number of labeled data (or unlabeled data) in each class k is determined by $n_k = n_1 \gamma^{-\frac{k-1}{K-1}}, \forall k \in \mathcal{C}$, where γ controls the degree of imbalance among the classes, with $\gamma = 1$ resulting in a balanced distribution of labeled data among classes. Additionally, n_1 represents the maximum (or minimum) number of labeled data among all the classes. In particular, we consider two cases (see Figure 3, Appendix C):

S1. when the dataset is balanced, we randomly select labeled data in each class with $n_1 = 400$ and $\gamma = 10$, and the remaining data is considered as unlabeled.

S2. when the dataset is unbalanced, we randomly select labeled data (resp. unlabeled data) with $n_1 = 400$ and $\gamma = 10$ for (resp. $\gamma = 0.1$).

S1. (resp. **S2.**) leads to a percentage of observed labels of 3% (resp. 9%). We trained a 3-layer CNN for both Algorithm 1 and 2. In terms of estimation of the missing-data mechanism, all methods have comparable and low MSE values in the balanced setting **S1.**, as reported in Appendix C (Table 6). In the unbalanced case **S2.**, the estimation of the missing-data mechanism **CADR** underestimates the observed proportions in the four rarest classes (i.e. classes 0 to 3), as seen in Figure 2, which leads to a highest MSE (see Table 1). For model estimation (see Figure 4 in Appendix C,

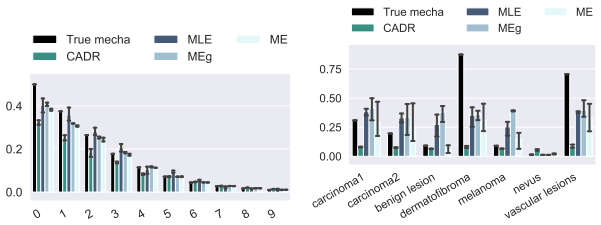

Figure 2. Estimation of the mechanism (coordinates of ϕ for each class) on MNIST (Setting **S2.**) and on dermaMNIST.

while in the balanced case **S1.** the methods have comparable results, there is an improvement in both test accuracy and test loss with our methods that include mechanism estimation (**MLE**, **MEg** and **ME**), especially for the less observed classes (i.e. classes 5 to 9). Note also that in both cases the method of Hu et al. (2022) has the highest test loss, which can be explained by the fact that the objective function that they minimize is quite different as explained in Section 3.2.

Table 1. Test accuracy and test loss on MNIST, Setting **S2.**.

METHOD	LOSS	ACCURACY	MSE ϕ
PSEUDOLABEL	0.141 ± 0.018	92.95 ± 0.55	0.594
DEPSEUDOLABEL	0.138 ± 0.015	93.18 ± 0.71	0.594
CADR	0.160 ± 0.029	89.15 ± 0.99	0.106 ± 0.012
MLE (OURS)	0.116 ± 0.021	94.29 ± 0.11	0.027 ± 0.012
MEG (OURS)	0.103 ± 0.009	94.83 ± 0.38	0.022 ± 0.004
ME (OURS)	0.111 ± 0.005	94.59 ± 0.28	0.037 ± 0.002

Compatibility with MCAR setting. When the labels are MCAR, we verify that our methods do not deteriorate compared to conventional methods specifically designed for the MCAR scenario (**PseudoLabel** and **DePseudoLabel**). In this case, all classes maintain an identical ratio of labeled to unlabeled data: we consider 164 labeled data and 1635 unlabeled data in each class. In Table 2, all the methods have similar results.

Table 2. Test accuracy and test loss on MNIST, MCAR labels.

METHOD	LOSS	ACCURACY
PSEUDOLABEL	0.21 ± 0.024	95.03 ± 0.53
DEPSEUDOLABEL	0.203 ± 0.029	94.85 ± 0.67
CADR	0.214 ± 0.03	95.04 ± 0.67
MLE (OURS)	0.206 ± 0.029	95.3 ± 0.76
MEG (OURS)	0.208 ± 0.024	94.84 ± 0.37
ME (OURS)	0.204 ± 0.023	94.88 ± 0.49

Our algorithm is not specifically designed for MAR labels, as the risk estimator given in (4) will be biased. However, the algorithm might perform well, particularly in scenarios where the classification problem is easy (e.g., when there is a strong correlation between x and y). In such cases, the MAR and MNAR scenarios are close, as the missing-data mechanism is either a function of r given x or given y.

5.1.2. CIFAR-10

In the CIFAR-10 dataset and considering Setting **S2.**, we compare the original version of Fixmatch (Sohn et al., 2020) with its debiased versions for MCAR data (Schmutz et al., 2023) and for MNAR data[3]. In Table 3[4], our method **ME** demonstrates improved performance with higher overall accuracy. Again, we observe in Figure 5 (Appendix C) that the classes with less observations (classes 5 to 10, particularly "dog" and "ship") are more accurately predicted when the missing-data mechanism is taken into account.

Table 3. Test accuracy and test loss for CIFAR10, Setting **S2.**.

METHOD	LOSS	ACCURACY
FIXMATCH	0.426 ± 0.017	90.91 ± 0.12
DEFIXMATCH	0.536 ± 0.020	89.71 ± 0.16
FIXMATCH + CADR	0.452 ± 0.006	91.14 ± 0.30
FIXMATCH + ME (OURS)	0.321 ± 0.016	91.88 ± 0.24

In Appendix C, Table 8 presents a comparison of the methods, considering various overall proportions of labeled data (1%, 4%, and 10%). Our method **ME** outperforms the other approaches, across all cases.

5.2. medMNIST with pseudo-realistic mechanismns

We first consider the dermaMNIST dataset (Codella et al., 2019), which consists of 10,015 dermatoscopic images (including 2000 for test data) categorized into 7 different categories of skin diseases, both benign and malignant. This dataset is unbalanced, its the most frequent class (71% of the images) being benign naevi. To simulate a more realistic MNAR scenario, we assume that a medical doctor would like to classify the conditions equally and select 70 images per class for labeling, resulting in 7% of observed labels (see Figure 1, left panel). Note that despite this selection, the dataset remains unbalanced due to the original distribution of the classes. Our three estimators (**MLE**, **ME**, **MEg**) of the missing-data mechanism detect that the class of naevi is very little observed compared to other classes (see Figure 2), whereas **CADR** gives a mechanism where all classes are

[3]Only the results for the moment estimator using averaging strategy (**ME**) are reported, as Algorithm 1 (**MLE**) has not yet been implemented with data augmentation and **MEg** proved difficult to calibrate. It can be challenging to find the right balance between too much initialization using $\hat{\phi} = n_\ell/n$ (leading to deviation from the optimal solution) and too little (leading to numerical problems). Therefore, we recommend considering the ME estimator when using data augmentation in practice.

[4]Table 7 (Appendix C) shows the results in the complete case, i.e. without taking unlabeled data into account.

equally observed. In Table 4, only our methods determine if a lesion is a nevus or not with a high accuracy, which can be used as a pre-processing step before the images are reviewed by a medical doctor. Note that, even if the class proportions are known, **CADR** fails to give accurate results for the nevus class (see Table 9 in Appendix C), which shows the relevance of taking into account the missing-data mechanism in this case. Finally, we use **MLE** together with the test presented in Section 4.4 to assess whether or not labels are informative. If we generate MCAR labels, the likelihood ratio test is rigthfully unable to reject the MCAR hypothesis (p-value of 0.68 ± 0.2 over 10 runs). But if we give it images with MNAR labels, the test rejects the MCAR hypothesis with very high confidence (p-value$< 10^{-4}$ for all 10 runs).

Table 4. Test accuracies on dermaMNIST and noduleMNIST3D.

METHOD	DERMAMNIST	NODULEMNIST
PSEUDOLABEL	57.72 ± 1.95	84.91
CADR	49.36 ± 1.91	80.32
MLE (OURS)	66.4 ± 0.81	85.8
MEG (OURS)	66.65 ± 1.76	82.26
ME (OURS)	65.8 ± 0.78	85.16

We now consider the noduleMNIST3D dataset (Armato III et al., 2011) on images from thoracic CT scans, which is of particular interest to simulate the MNAR labels. The dataset was originally designed for a 5-level malignancy classification task, but as described by Yang et al. (2023), the task is made binary into two levels: malignant and benign tumors. We have also access to the subtlety score s, which describes from 1 (extremely subtle) to 5 (obvious) the difficulty of nodule detection. According to these scores, we simulate the missing-data mechanism using $p(r|y) = \sum_{s=1}^{5} p(r|s)p(s|y)$, with $p(s|y)$ computed on the data. The only quantities to choose are the probability of being observed given the subtlety score, we fix a low probability when the detection was difficult ($p(r|s \in \{1, 2, 3\}) = 0.1$) and a high probability when the detection was easy ($p(r|s \in \{4, 5\}) = 0.9$). At the end, the class of benign nodules has a missing-data proportion (43%) higher than the class of malignant nodules (8%). On the contrary to the missing-data setting chosen for dermaMNIST, the more observed class is also the less frequent (Figure 7). **MLE** performs better in terms of accuracy (Table 4) and all the methods designed for informative labels has a highest specificity than the classical one **PseudoLabel** for the class of malignant nodules (Table 10 in Appendix C).

5.3. Discussion

In our experiments, we noted that our methods **MLE**, **ME** and **MEg** handle well the case of poorly observed classes, with a small amount of labeled data but a significant amount of unlabeled data (refer to classes 5 to 10 for CIFAR10 in Figure 5 and classes 5 to 9 for MNIST in Figure 4).

On the other hand, our methods may be unfavorable to rare and fully observed classes (i.e. containing little data, but only labeled data), as can occur when a rare disease is always diagnosed. In the MNIST dataset, we consider Setting **S2.**, with the "0" digit class that is rare in the labeled set (approximately 1%) and absent from the unlabeled data. In Table 5, our methods **ME** and **MEg** perform less effectively compared to other SSL methods with respect to accuracy for the rare class, but outperform them for the least observed class. However, **PseudoLabel** and **CADR** clearly outperform the supervised baseline (i.e. the oracle case where all the data are labeled) for the rarest class, which may indicate that they are subject to a bias that favors this class.

Table 5. Test accuracy for MNIST, Setting **S2.** with a rare class.

METHOD	RARE CLASS	LEAST OBSERVED CLASS
SUPERVISED BASELINE	82 ± 7.4	97.6 ± 0.9
PSEUDOLABEL	92 ± 1.9	91.2 ± 1.8
CADR	94.6 ± 1.1	94.2 ± 0.8
MEG (OURS)	84 ± 1.6	95 ± 1
ME (OURS)	85.3 ± 3.1	94.2 ± 0.8

Finally, it is important to note that all our experiments were conducted on datasets containing a maximum of 10 classes. We cannot therefore determine whether the estimation of the missing-data mechanism will perform as well in classification problems involving thousands of classes. However, in such scenarios, interpretability, such as the precise estimation of the probabilities of observed data per class, may carry less significance.

6. Conclusion

For future works, we would be eager to (i) provide a more realistic theoretical grounding without freezing θ and (ii) propose another statistical test for the moment estimator, for example using bootstrap strategies. Note that the latter perspective is quite challenging, because of the sample bias in the informative data.

7. Acknowledgments

The authors thank Xinting Hu for her help with her code and article, and Lucas Boiteau for fruitful discussions about the work. This work has been supported by the French government, through the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002. The authors are also grateful to the OPAL infrastructure from Université Côte d'Azur for providing resources and support.

References

Ahfock, D. and McLachlan, G. J. On missing label patterns in semi-supervised learning. *arXiv preprint arXiv:1904.02883*, 2019.

Armato III, S. G., McLennan, G., Bidaut, L., McNitt-Gray, M. F., Meyer, C. R., Reeves, A. P., Zhao, B., Aberle, D. R., Henschke, C. I., Hoffman, E. A., et al. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical physics*, 38(2):915–931, 2011.

Baker, S. G. and Laird, N. M. Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *Journal of the American Statistical Association*, 83(401):62–69, 1988.

Berthelot, D., Carlini, N., Cubuk, E. D., Kurakin, A., Sohn, K., Zhang, H., and Raffel, C. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *International Conference on Learning Representations*, 2019a.

Cao, K., Brbic, M., and Leskovec, J. Open-world semi-supervised learning. *arXiv preprint arXiv:2102.03526*, 2021.

Chapelle, O., Schölkopf, B., and Zien, A. (eds.). *Semi-Supervised Learning*. The MIT Press, 2006.

Chen, Y., Zhu, X., Li, W., and Gong, S. Semi-supervised learning under class distribution mismatch. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 3569–3576, 2020.

Codella, N., Rotemberg, V., Tschandl, P., Celebi, M. E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019.

d'Haultfoeuille, X. A new instrumental method for dealing with endogenous selection. *Journal of Econometrics*, 154 (1):1–15, 2010.

Grandvalet, Y. and Bengio, Y. Semi-supervised learning by entropy minimization. *Advances in neural information processing systems*, 17, 2004.

Guo, L.-Z., Zhang, Z.-Y., Jiang, Y., Li, Y.-F., and Zhou, Z.-H. Safe deep semi-supervised learning for unseen-class unlabeled data. In *International Conference on Machine Learning*, pp. 3897–3906. PMLR, 2020.

Hu, X., Niu, Y., Miao, C., Hua, X.-S., and Zhang, H. On non-random missing labels in semi-supervised learning. In *International Conference on Learning Representations*, 2022.

Huang, Z., Xue, C., Han, B., Yang, J., and Gong, C. Universal semi-supervised learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 26714–26725, 2021.

Ibrahim, J. G. and Lipsitz, S. R. Parameter estimation from incomplete data in binomial regression when the missing data mechanism is nonignorable. *Biometrics*, 52(3):1071–1078, 1996.

Ibrahim, J. G., Chen, M.-H., and Lipsitz, S. R. Missing responses in generalised linear mixed models when the missing data mechanism is nonignorable. *Biometrika*, 88 (2):551–564, 2001.

Krizhevsky, A. Learning multiple layers of features from tiny images. *MSc thesis*, 2009.

Le Morvan, M., Prost, N., Josse, J., Scornet, E., and Varoquaux, G. Linear predictor on linearly-generated data with missing values: non consistency and solutions. In *International Conference on Artificial Intelligence and Statistics*, pp. 3165–3174. PMLR, 2020.

LeCun, Y. and Cortes, C. MNIST handwritten digit database. 2010. URL `http://yann.lecun.com/exdb/mnist/`.

Lee, D.-H. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML workshop on Challenges in Representation Learning (WREPL)*, 2013.

Little, R. J. and Rubin, D. B. *Statistical analysis with missing data*. John Wiley & Sons, 2019.

Liu, T. and Goldberg, Y. Kernel machines with missing responses. *Electronic Journal of Statistics*, 14(2):3766–3820, 2020.

Miao, W. and Tchetgen Tchetgen, E. J. On varieties of doubly robust estimators under missingness not at random with a shadow variable. *Biometrika*, 103(2):475–482, 2016.

Miao, W., Ding, P., and Geng, Z. Identifiability of normal and normal mixture models with nonignorable missing data. *Journal of the American Statistical Association*, 111 (516):1673–1683, 2016.

Miao, W., Liu, L., Tchetgen, E. T., and Geng, Z. Identification, doubly robust estimation, and semiparametric efficiency theory of nonignorable missing data with a shadow variable. *arXiv preprint arXiv:1509.02556*, 2019.

Mohan, K. On handling self-masking and other hard missing data problems. In *AAAI Symposium 2018*, 2018.

Molenberghs, G., Beunckens, C., Sotto, C., and Kenward, M. G. Every missingness not at random model has a missingness at random counterpart with equal fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(2):371–388, 2008.

Morikawa, K., Kim, J. K., and Kano, Y. Semiparametric maximum likelihood estimation with data missing not at random. *Canadian Journal of Statistics*, 45(4):393–409, 2017.

Murphy, S. A. and Van der Vaart, A. W. On profile likelihood. *Journal of the American Statistical Association*, 95 (450):449–465, 2000.

Neuberg, L. G. Causality: models, reasoning, and inference, by judea pearl, cambridge university press, 2000. *Econometric Theory*, 19(4):675–685, 2003.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. 2019.

Rabe-Hesketh, S. and Skrondal, A. Ignoring non-ignorable missingness. *Psychometrika*, 88:31–50, 2023.

Rizve, M. N., Duarte, K., Rawat, Y. S., and Shah, M. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. *International Conference on Learning Representations*, 2021.

Rubin, D. B. Inference and missing data. *Biometrika*, 63(3): 581–592, 1976.

Schmutz, H., Humbert, O., and Mattei, P.-A. Don't fear the unlabelled: safe deep semi-supervised learning via simple debiasing. *International Conference on Learning Representations*, 2023.

Scudder, H. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, 11(3):363–371, 1965.

Shao, J. and Wang, L. Semiparametric inverse propensity weighting for nonignorable missing data. *Biometrika*, 103(1):175–187, 2016.

Sohn, K., Berthelot, D., Li, C.-L., Zhang, Z., Carlini, N., Cubuk, E. D., Kurakin, A., Zhang, H., and Raffel, C. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems*, 33:596–608, 2020.

Sportisse, A., Boyer, C., and Josse, J. Estimation and imputation in probabilistic principal component analysis with missing not at random data. *Advances in Neural Information Processing Systems*, 33:7067–7077, 2020.

Tang, N. and Ju, Y. Statistical inference for nonignorable missing-data problems: a selective review. *Statistical Theory and Related Fields*, 2(2):105–133, 2018.

Van der Vaart, A. W. *Asymptotic statistics*. Cambridge university press, 1998.

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge university press, 2019.

Wei, C., Sohn, K., Mellina, C., Yuille, A., and Yang, F. Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10857–10866, 2021.

Xie, Q., Dai, Z., Hovy, E., Luong, T., and Le, Q. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33: 6256–6268, 2020.

Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., and Ni, B. MedMNIST v2: A large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Scientific Data*, 10(1):41, 2023.

A. Proof of identification

Proposition A.1 (Identification of the joint distribution)**.** *Under Assumptions **A2.** (self-masked MNAR), the joint distribution $p(y, x, r)$ is identified.*

Proof. This proof is a direct application of Theorem 1 of (Miao et al., 2019) (stated in a more general case). For clarity, we give it again here in our setting for clarity by proving the intermediate results.

Let us consider the odds ratio $OR(y) = \frac{p(y|r=0)p(y=1|r=1)}{p(y|r=1)p(y=1|r=0)}$ where $y = 1$ is used as a reference value. The goal is to determine $p(y|r = 0, x)$ by only quantities involving observed data, and it will imply that the joint distribution $p(y, x, r)$ is identified. Proposition 2 of (Miao et al., 2019) gives the two following equalities:

$$p(y|r = 0, x) = \frac{OR(y)p(y|r = 1, x)}{\mathbb{E}[OR(y)|r = 1, x]} \tag{11}$$

$$\mathbb{E}[\tilde{OR}(y)|r = 1, x] = \frac{p(x|r = 0)}{p(x|r = 1)} \text{ with } \tilde{OR}(y) = \frac{OR(y)}{\mathbb{E}[OR(y)|r=1]} \tag{12}$$

The first equation (11) indicates that the identification of the odds ratio function involves the identification of $p(y|r = 0, x)$. Note that in (12), as $p(x|r = 0), p(x|r = 1)$ and $p(y|r = 1, x)$ are obtained from observed data, so is $\mathbb{E}[\tilde{OR}(y)|r = 1, x]$. We just have to prove that (11) has a unique solution. Let us consider $\tilde{OR}^\star(y) \neq \tilde{OR}^\star(y)$, we have

$$\mathbb{E}[\tilde{OR}^\star(y)|r = 1, x] = \frac{p(x|r = 0)}{p(x|r = 1)},$$

which implies

$$\mathbb{E}[\tilde{OR}^\star(y) - \tilde{OR}(y)|r = 1, x] = 0 \Leftrightarrow \tilde{OR}^\star(y) = \tilde{OR}(y),$$

This is obtained by using Condition 1 of (Miao et al., 2019). In our case, it amounts to assuming that for any image, there is a non-zero probability that it has any label. $\tilde{OR}(y)$ is identified and so is $OR(y)$, noting that $OR(y) = \frac{\tilde{OR}(y)}{\tilde{OR}(y=0)}$.

□

B. Proofs of the theoretical results of Section 4

B.1. Moment estimator

The moment estimator has the following form, for a fixed $\theta \in \Theta$:

$$\hat{\phi}_y^M = \frac{\sum_{i=1}^n \mathbb{1}_{\{r=1, y_i=y\}}}{n} \frac{1}{\hat{p}(y, \theta)}, \forall y \in \mathcal{C}$$

with

$$\hat{p}(y, \theta) := \hat{p}_\theta(y) = \frac{1}{n} \sum_{i=1}^n p(y_i|x_i; \theta)$$

In this section, we prove the consistency of this estimator.

Proposition B.1 (Consistency of $\hat{\phi}^M$)**.** *The moment estimator defined by* (7) *is consistent for a fixed $\theta \in \Theta$.*

Proof. We have by the law of large numbers,

$$\frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{r_i=1, y_i=y\}} \xrightarrow[n \to +\infty]{\mathbb{P}} \mathbb{E}[\mathbb{1}_{\{r_i=1, y_i=y\}}] = p(r = 1, y) \tag{13}$$

$$\frac{1}{n} \sum_{i=1}^n p(y_i|x_i; \theta) \xrightarrow[n \to +\infty]{\mathbb{P}} \mathbb{E}[p(y|x)] = p(y) \tag{14}$$

We can apply the continuous mapping theorem to $f(x) = 1/x, x \in]0,1]$ (assuming that the probability of each class is greater than 0) to get

$$\frac{1}{\frac{1}{n}\sum_{i=1}^{n} p(y_i|x_i)} \xrightarrow[n\to+\infty]{\mathbb{P}} \frac{1}{p(y)} \tag{15}$$

In (13) and (15), we have the convergence in probability to a constant, which implies by Slutsky's theorem that the product converges in probability to the product of the constants (Slutsky's theorem gives the property in distribution but convergence in probability and law is equivalent if the limit is a constant), which gives:

$$\forall y \in \mathcal{C}, \hat{\phi}_y^M \xrightarrow[n\to+\infty]{\mathbb{P}} \phi_y = \frac{p(r=1,y)}{p(y)}$$

□

B.2. Maximum likelihood estimator

In this section, we study the negative observed log-likelihood $\phi \mapsto \ell(\theta, \phi; x, y \odot r, r)$ and derive the consistency and the asymptotic normality of the maximum likelihood estimator ϕ^L given in (9). All the results are obtained for a fixed $\theta \in \Theta$. Let us recall:

$$\ell(\theta, \phi; x, y \odot r, r) = -\sum_{i=1}^{n} r_i \log p(y_i|x_i;\theta)\phi_{y_i} - \sum_{i=1}^{n}(1-r_i)\log \sum_{\tilde{y}\in\mathcal{C}} p(\tilde{y}|x_i;\theta)(1-\phi_{\tilde{y}}) \tag{16}$$

For simplicity, as θ is fixed in this section, let us define the function $\ell_\theta : \mathbb{R}^{\mathbb{K}} \mapsto \mathbb{R}$ such that $\ell_\theta(\phi) = \ell(\theta, \phi, x, y \odot r, r)$.

B.2.1. Convexity

Proposition B.2 (Convexity of $\ell_\theta(\phi; x, y \odot r, r)$)**.** *For a fixed $\theta \in \Theta$, the negative observed log-likelihood $\phi \mapsto \ell(\theta, \phi; x, y \odot r, r)$ is convex.*

Proof. Let us remark that:

$$\begin{aligned}
\ell_\theta(\phi) &= -\sum_{i=1}^{n} r_i \log p(y_i|x_i)\phi_{y_i} - \sum_{i=1}^{n}(1-r_i)\log \sum_{\tilde{y}\in\mathcal{C}} p(\tilde{y}|x_i)(1-\phi_{\tilde{y}}) \\
&= -\sum_{i=1}^{n} r_i \log p(y_i|x_i)\phi_{y_i} - \sum_{i=1}^{n}(1-r_i)\log(-\phi_k B_{i,k} + C_{i,k}),
\end{aligned}$$

where $B_{i,k} = p(y_i = k|x_i;\theta)$ and $C_{i,k} = B_{i,k} + \sum_{\tilde{y}\in\mathcal{C}\setminus\{k\}} p(\tilde{y}|x_i;\theta)(1-\phi_{\tilde{y}})$. Let us compute the Hessian H of F, such that

$$H(\phi) := \left(\frac{\partial^2 \ell_\theta}{\partial\phi_\ell \partial\phi_k}(\phi)\right)_{k,\ell\in\mathcal{C}}$$

We have:

$$\forall k \in \mathcal{C}, \frac{\partial \ell_\theta}{\partial \phi_k}(\phi) = -\left(\sum_{i=1}^{n} r_i \frac{1}{\phi_k}\mathbb{1}_{y_i=k} - \sum_{i=1}^{n}(1-r_i)\frac{B_{i,k}}{-\phi_k B_{i,k} + C_{i,k}}\right).$$

Thus,

$$\begin{aligned}
\forall k \in \mathcal{C}, \frac{\partial^2 \ell_\theta}{\partial\phi_k\partial\phi_k}(\phi) &= \sum_{i=1}^{n} r_i \frac{1}{\phi_k^2}\mathbb{1}_{y_i=k} + \sum_{i=1}^{n}(1-r_i)\frac{(B_{i,k})^2}{(-\phi_k B_{i,k} + C_{i,k})^2} \\
\forall \ell \in \mathcal{C}, \ell \neq k, \frac{\partial^2 \ell_\theta}{\partial\phi_\ell\partial\phi_k}(\phi) &= \sum_{i=1}^{n}(1-r_i)\frac{B_{i,k}B_{i,\ell}}{(-\phi_k B_{i,k} + C_{i,k})^2}
\end{aligned}$$

In the following, let us denote $A_i = -\phi_k B_{i,k} + C_{i,k} = \sum_{\tilde{y}\in\mathcal{C}} p(\tilde{y}|x_i;\theta)(1-\phi_{\tilde{y}})$ We have now to prove that the Hessian is positive-definite . As it is a symmetric matrix, we can show that $\forall v \in \mathbb{R}^k, v \neq \vec{0}, v^T H v > 0$.

$$v^T H v = \sum_{k=1}^{K} \frac{v_k^2}{\phi_k^2} \sum_{i=1}^{n} r_i \mathbb{1}_{y_i=k} + \underbrace{\sum_{k=1}^{K} v_k^2 \sum_{i=1}^{n} (1-r_i) \frac{(B_{i,k})^2}{A_i^2} + 2 \sum_{1 \leq k < \ell \leq K} v_k v_\ell \sum_{i=1}^{n} (1-r_i) \frac{B_{i,k} B_{i,\ell}}{A_i^2}}_{=\mathrm{T}} > 0$$

The first term is trivially greater or equal to 0. Moreover, it is never equal to 0, if at least one sample is observed ($n_\ell > 0$). For the last two terms, note that:

$$T = \sum_{i=1}^{n} \left(\sum_{k=1}^{K} v_k B_{i,k} \sqrt{\frac{(1-r_i)}{A_i^2}} \right)^2 \geq 0$$

□

Remark B.3 (Domain of definition of $\ell_\theta(\phi; x, y \odot r, r)$). We look at the natural domain of the negative observed log-likelihood $\phi \mapsto \ell(\theta, \phi; x, y \odot r, r)$, for a fixed $\theta \in \Theta$. The goal is to know if we can minimize the function without constraint on ϕ (if its domain of definition is included in $[0, 1]$)

In (16), the first term implies $\forall y_i, p(y_i|x_i;\theta)\phi_{y_i} > 0$ i.e. $\phi_k > 0$, $\forall k \in \{1, \dots, K\}$. The second term requires $\forall i \in \{n_\ell + 1; n\}$,

$$\begin{aligned}
&\sum_{\tilde{y} \in \mathcal{C}} p(\tilde{y}|x_i;\theta)(1-\phi_{\tilde{y}}) > 0 \\
\Leftrightarrow &\forall k \in \{1, \dots, K\}, (1-\phi_k) p(\tilde{y} = k|x_i;\theta) + \sum_{\tilde{y} \in \mathcal{C} \setminus \{k\}} p(\tilde{y}|x_i;\theta)(1-\phi_{\tilde{y}}) > 0 \\
\Leftrightarrow &\forall k \in \{1, \dots, K\}, \phi_k < 1 + \frac{1}{p(\tilde{y} = k|x_i;\theta)} \sum_{\tilde{y} \in \mathcal{C} \setminus \{k\}} p(\tilde{y}|x_i;\theta)(1-\phi_{\tilde{y}})
\end{aligned}$$

As $\frac{1}{p(\tilde{y}=k|x_i;\theta)} \sum_{\tilde{y} \in \mathcal{C} \setminus \{k\}} p(\tilde{y}|x_i;\theta)(1-\phi_{\tilde{y}}) > 0$, this last inequality does not necessarily implies $\phi_k \leq 1$. Therefore, a reparametrization trick or constrained optimization is essential. The domain of definition for $\phi_k, k \in \{1, \dots, K\}$ of the negative log-likelihood ℓ is then

$$D_{\phi_k} = \left] 0; 1 + \min_{i \in \{1, \dots, n\}} \frac{1}{p(\tilde{y} = k|x_i;\theta)} \sum_{\tilde{y} \in \mathcal{C} \setminus \{k\}} p(\tilde{y}|x_i;\theta)(1-\phi_{\tilde{y}}) \right[.$$

B.2.2. Consistency and asymptotic normality

The consistency and asymptotic normality of the maximum likelihood estimator is obtained by applying Theorem 5.7 and Theorem 5.23 of (Van der Vaart, 1998). Let us assume the following:

A3. We can interchange differentiation with respect to ϕ and integration over (x, y, r).

A4. $\forall k \in \mathcal{C}$, there exists a compact interval U_k such that $\phi_k \in U_k \subset]0, 1[$.

To get the results, we need first to show the identifiability of the statistical model $\mathcal{P}_\phi = \{p(r|y \odot r; \phi) : \phi \in \Phi\}$ and the nonsingularity of the Fisher information $I_{\phi^\star}$, with $\phi^\star$ the oracle point (i.e. $\phi^\star = \mathrm{argmin}_{\phi \in \Phi} \ell_\theta(\phi)$).

Lemma B.4 (Identifiability of $\mathcal{P}_\phi$)**.** *The model $\mathcal{P}_\phi$ is identifiable.*

Proof. Let us consider $\phi, \phi' \in \Phi$ such that $p(r|y \odot r; \phi) = p(r|y \odot r; \phi')$.

$$\begin{aligned}
&p(r|y \odot r; \phi) = p(r|y \odot r; \phi'), \forall y, r \\
&\Leftrightarrow r\phi_y + \sum_{\tilde{y}} (1-r)(1-\phi_{\tilde{y}}) = r\phi'_y + \sum_{\tilde{y}} (1-r)(1-\phi'_{\tilde{y}}),\ \forall y, r \\
&\Leftrightarrow r(\phi_y - \phi'_y) + \sum_{\tilde{y}} (1-r)(\phi'_y - \phi_{\tilde{y}}) = 0,\ \forall y, r
\end{aligned}$$

The case $r = 1$ leads to $\phi_y = \phi'_y, \forall y$. □

Lemma B.5 (Nonsingularity of the Fisher information $I_{\phi^\star}$)**.** *Under Assumption **A3.**, the Fisher information at the oracle point $\phi^\star$ is nonsingular.*

Proof. Assumption **A3.** implies that $\forall \ell, k \in \mathcal{C}, (I_{\phi^\star})_{(\ell,k)} = -\mathbb{E}_{x,y,r}\left[\frac{\partial^2 \log \ell_\theta}{\partial\phi_\ell \partial\phi_k}(\phi^\star)\right] = -\frac{1}{n}\mathbb{E}_{x,y,r}[(H(\phi))_{(\ell,k)}]$. We can simply use the strict convexity of the function $\phi \mapsto \ell_\theta(\phi)$ proved in Proposition B.2. □

Proposition B.6 (Consistency of $\hat{\phi}^L$)**.** *Under Assumption **A4.**, the maximum likelihood estimator $\hat{\phi}^L$ defined in* (9) *is consistent, for a fixed $\theta \in \Theta$.*

Proof. As said in Section 5.5 of Van der Vaart (1998) for the application to the maximum likelihood estimators, the method consists of applying Theorem 5.7 of Van der Vaart (1998) by noting that the MLE is a M-estimator: $\hat{\phi}^L \in \operatorname{argmin}_{\phi\in\Phi} M_n(\phi)$, with $M_n(\phi) = \frac{1}{n}\sum_{i=1}^n \log \frac{p(x_i, y_i\odot r_i, r_i;\phi^\star)}{p(x_i, y_i\odot r_i, r_i;\phi)}$ and $M(\phi) = \mathbb{E}\left[\log \frac{p(x, y\odot r, r;\phi^\star)}{p(x, y\odot r, r;\phi)}\right]$.

- We have the identifiability of $\mathcal{P}_\phi$ by Lemma B.4. We also have the strong identifiability, which is equivalent to the identifiability, since the restricted set of Φ, $\otimes_{k\in\mathcal{C}} U_k$, is compact (see Assumption **A4.**).
- We have to show that the uniform weak law of large numbers hold for the function $\phi \mapsto \log \frac{p(x_i, y_i\odot r_i, r_i;\phi^\star)}{p(x_i, y_i\odot r_i, r_i;\phi)}$, i.e.

$$\sup_{\phi\in\Phi}|M_n(\phi) - M(\phi)| \xrightarrow[n\to+\infty]{\mathbb{P}} 0.$$

  Following Theorem 4.2 of Wainwright (2019), a sufficient condition is that there exists $M > 0$ sucht that, $\left|\log \frac{p(x, y\odot r, r;\phi^\star)}{p(x, y\odot r, r;\phi)}\right| < M$, $\forall \phi, x, y, r$, i.e. the uniform boundedness. We will prove that there exists $M_1, M_2 > 0$ such that $M_1 \leq \frac{p(x, y\odot r, r;\phi^\star)}{p(x, y\odot r, r;\phi)} \leq M_2$.

  We have

$$\frac{p(x, y\odot r, r;\phi^\star)}{p(x, y\odot r, r;\phi)} = \frac{r\phi^\star_y + (1-r)\sum_{\tilde{y}}(1-\phi^\star_{\tilde{y}})}{r\phi_y + (1-r)\sum_{\tilde{y}}(1-\phi_{\tilde{y}})}$$

  Using Assumption **A4.**, we get $M_1 = \min\left(\frac{\min_k a_k}{\max_k b_k}, \frac{1-\max_k b_k}{1-\min_k a_k}\right)$ and $M_2 = \max\left(\frac{1-\min_k a_k}{1-\max_k b_k}, \frac{\max_k b_k}{\min_k a_k}\right)$
- By identifiability of the statistical model $\mathcal{P}_\phi$, we have: $M(\phi) = M(\phi^\star)$ implies $\phi = \phi^\star$. Therefore $\forall \phi \neq \phi^\star, M(\phi) > M(\phi^\star)$ and the function $\phi \mapsto M(\phi)$ admits a strict minimum in ϕ_0.

□

Proposition B.7 (Asymptotic normality of $\hat{\phi}^L$)**.** *Under Assumption **A3.** and **A4.**, the maximum likelihood estimator $\hat{\phi}^L$ defined in* (9) *is asymptotically normal, for a fixed $\theta \in \Theta$.*

Proof. The proof directly follows from Theorem 5.39 of Van der Vaart (1998) (Corollary of Theorem 5.23 for the maximum likelihood estimators). To apply the theorem, we check the following conditions:

- The statistical model $\mathcal{P}_\phi = \{p(r|y\odot r;\phi) : \phi \in \Phi\}$ is differentiable in quadratic mean at $\phi^\star$, because $p(r|y\odot r;\phi) = r\phi_y + \sum_{\tilde{y}}(1-r)(1-\phi_{\tilde{y}})$ is trivially twice differentiable.
- The score function $S(\phi; r, y\odot r)$ is uniformly bounded in y, r and ϕ, ranging over a compact and continuous in ϕ forall y, r, i.e. we want to show that there exists a real number M,

$$\|S(\phi; r, y)\|_1 = \sum_{k=1}^K \left|\frac{\partial \log p(r|y\odot r;\phi)}{\partial\phi_k}\right| \leq M, \forall \phi, y, r$$

  The score function for its coordinate k is: $S(\phi_k; r, y\odot r) = \frac{\partial \log p(r|y\odot r;\phi)}{\partial\phi_k} = \frac{r\mathbb{1}_{y=k} - (1-r)}{r\phi_y + (1-r)\sum_{\tilde{y}}(1-\phi_{\tilde{y}})}$. If $r = 1$, this amounts to bound $1/\phi_k$ and if $r = 0$, this amounts to bound $\frac{1}{\sum_{\tilde{y}}(1-\phi_{\tilde{y}})} \leq \frac{1}{K(1-\max_k \phi_k)}$. Using Assumption **A4.** is sufficient to get the bound. Let us denote $U_k = [a_k, b_k], \forall k \in \mathcal{C}$, one has $a_k \leq \phi_k \leq b_k \forall k \in \mathcal{C}$ and we can choose for the bound $M = K\max\left(\frac{1}{\min_k a_k}, \frac{1}{K(1-\max_k b_k)}\right)$.

- By Lemma B.5, the Fisher information $I_{\phi^\star}$ is nonsingular.

- By Proposition B.6, the estimator $\hat{\phi}^L$ is consistent.

□

B.3. Consistency of the SSL risk

Proposition B.8 (Consistency of the risk)**.** *If $\hat{\phi}$ is a consistent estimator of ϕ and if the mechanism is well specified, the risks $\hat{\mathcal{R}}_{\hat{\phi}}$ and $\hat{\mathcal{R}}^{\mathrm{SSL}}_{\hat{\phi}}(\theta)$ are consistent estimators of the theoretical risk $\mathcal{R}(\theta) = \mathbb{E}[\ell_\ell(\theta; x, y)]$.*

Proof. We prove the results for the IPW estimator $\hat{\mathcal{R}}_{\hat{\phi}}$ (the proof is similar for $\hat{\mathcal{R}}^{\mathrm{SSL}}_{\hat{\phi}}(\theta)$).

It is a simple application of the law of large numbers, using the unbiasedness of the estimator (by Proposition 3.1).

We have:

$$\frac{1}{n}\sum_{i=1}^{n} r_i \frac{\ell_\ell(\theta; x_i, y_i)}{\hat{\phi}_{y_i}} = \frac{1}{n}\sum_{i=1}^{n} r_i \frac{\ell_\ell(\theta; x_i, y_i)}{\phi_{y_i}} + o_{\mathbb{P}}(1) \xrightarrow[n\to+\infty]{\mathbb{P}} \mathbb{E}\left[r\frac{\ell_\ell(\theta; x, y)}{\phi_y}\right] = \mathbb{E}\left[\ell_\ell(\theta; x, y)\right]$$

□

C. Additional numerical experiments

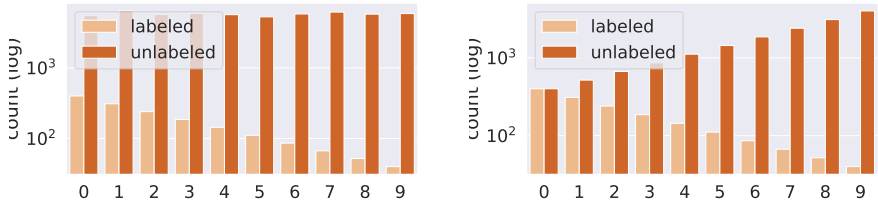

Figure 3. MNAR labels on MNIST (Settings **S1.** and **S2.**)

Table 6. Test accuracy and test loss on MNIST, Setting **S1.**.

METHOD	LOSS	ACCURACY	MSE ϕ
PSEUDOLABEL	0.259 ± 0.034	95.48 ± 0.16	0.318
DEPSEUDOLABEL	0.237 ± 0.045	95.69 ± 0.06	0.318
CADR	0.272 ± 0.046	95.40 ± 0.33	0.014 ± 0.004
MLE (OURS)	0.249 ± 0.050	95.59 ± 0.40	0.031 ± 0.009
MEG (OURS)	0.240 ± 0.029	95.69 ± 0.80	0.004 ± 0.001
ME (OURS)	0.240 ± 0.027	95.34 ± 0.26	0.013 ± 0.001

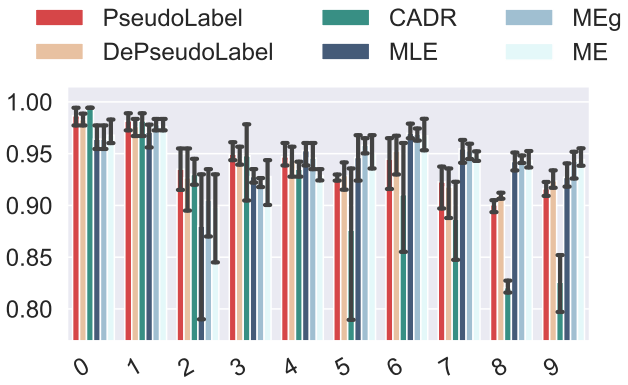

Figure 4. Accuracy per class on MNIST, Setting **S2.**.

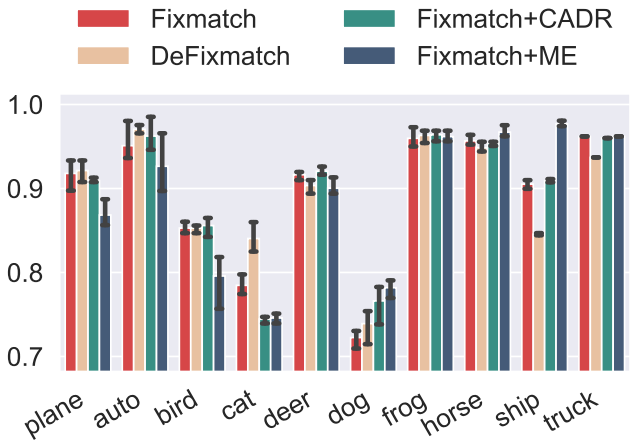

Figure 5. Accuracy per class on CIFAR-10, Setting **S2.**.

Table 7. Test accuracy and test loss for CIFAR10, Setting **S2.**.

METHOD	LOSS	ACCURACY
COMPLETE CASE FIXMATCH	1.647 ± 0.025	68.26 ± 0.56
FIXMATCH	0.426 ± 0.017	90.91 ± 0.12
DEFIXMATCH	0.536 ± 0.020	89.71 ± 0.16
FIXMATCH + CADR	0.452 ± 0.006	91.14 ± 0.30
FIXMATCH + ME (OURS)	0.321 ± 0.016	91.88 ± 0.24

Table 8. Test accuracy and test loss for CIFAR10 varying the proportion of labeled data. There is the same number of unlabeled data per class (4000); only the number of labeled data per class varies, following zthe formula in Section 5 with $\gamma = 10$ and $n_1 = 100, 400, 1000$.

PROPORTION LABELED	10%		4%		1%	
METHOD	LOSS	ACCURACY	LOSS	ACCURACY	LOSS	ACCURACY
PSEUDOLABEL	0.296	93.1	0.347	92.47	0.474	90.55
DEPSEUDOLABEL	0.3052	92.95	0.3152	93.24	0.4036	92.29
CADR	0.299	93.22	0.341	92.87	0.472	89.78
ME (OURS)	0.249	93.99	0.259	93.83	0.342	92.51

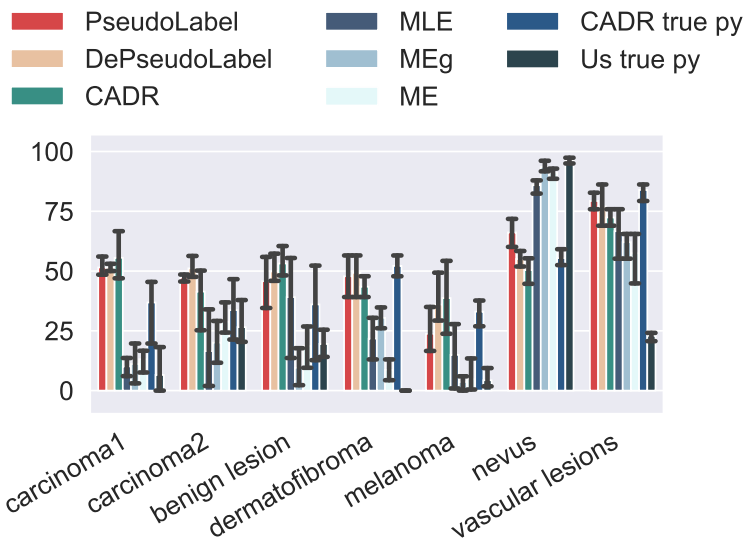

Figure 6. Accuracy per class on dermaMNIST.

Table 9. Accuracy and Loss in the test set on dermaMNIST.

METHOD	LOSS	ACCURACY	ACCURACY NAEVI	MSE ϕ_{NEVI}
PSEUDOLABEL	1.34 ± 0.16	57.72 ± 1.95	66.14 ± 5.86	0.80
CADR	1.42 ± 0.060	49.36 ± 1.91	50.41 ± 5.38	0.77 ± 0.02
MLE (OURS)	0.993 ± 0.020	66.4 ± 0.81	91.16 ± 2.26	0.34 ± 0.03
MEG (OURS)	1.19 ± 0.148	66.65 ± 1.76	93.54 ± 2.30	0.42 ± 0.08
ME (OURS)	1.24 ± 0.087	65.8 ± 0.78	85.91 ± 3.05	0.38 ± 0.15
CADR ($p(y)$ KNOWN)	1.57 ± 0.12	49.44 ± 3.27	55.40 ± 3.39	
ALGORITHM 2 ($p(y)$ KNOWN)	0.943 ± 0.029	68.83 ± 0.26	96.12 ± 1.20	

Table 10. Accuracy and Loss in the test set on noduleMNIST3D

METHOD	LOSS	SPECIFICITY (BEGNIN)	SPECIFICITY (MALIGN)	MSE ϕ
PSEUDOLABEL	0.389	91.06	54.69	0.0627
CADR	0.48	80.08	81.25	0.0143
MLE (OURS)	0.359	87.80	71.88	0.0001
MEG (OURS)	0.353	83.74	76.56	0.0199
ME (OURS)	0.355	86.99	78.13	0.0002

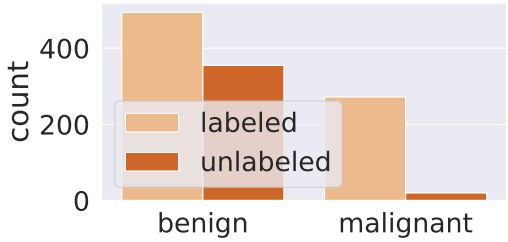

Figure 7. MNAR labels on noduleMNIST (see Section 5.2 for an explanation of the missing-data scenario)