Missing but not Missed: On Learnability Under Imputation

Andrea Campagner ✉

IRCCS Ospedale Galeazzi Sant'Ambrogio, Milan, Italy
andrea.campagner@unimib.it

Abstract. Missing data represents one of the most ubiquitous data quality issues, and also one of the most impactful on machine learning (ML) pipelines. Indeed, not only most commonly applied ML methods cannot directly employ incomplete data, but also the techniques employed to manage this issue can impact on the performance and evaluation of ML models. Among such techniques to manage missing data, *imputation*, that is filling in the missing values using information from the observed data, remains among the most popular and effective in practice. Yet, from a theoretical point of view, it is still not clear under which conditions it is possible to learn effectively after imputation. In this article we address this gap by studying learnability under imputation in the framework of statistical learning theory. After giving a general definition of learnability under imputation, we show three main contributions: 1) we introduce a novel stability condition, called *noise risk stability*, which we prove to be both sufficient and, under weak assumptions, necessary for learnability under imputation; 2) we show that a large class of ML models (including linear and kernel methods) satisfies noise risk stability; 3) we characterize the learning-theoretic properties of two common imputation methods (constant and regression imputation). Our results set the stage for a rigorous study of imputation and missing data management in the framework of statistical learning theory, by also describing relevant open questions.

Keywords: Imputation · Missing Data · Learnability · Statistical Learning Theory

1 Introduction

Missing data is one of the most commonly occurring data quality issues in real-world datasets. In fact, in many practical settings, facing missing data is the norm more than the exception, and failure to account for this issue can have profound consequences on the development and evaluation of machine learning (ML) models [1, 34], especially since most commonly used ML algorithms do not have a way to directly use missing data in the training process.

Consequently, a variety of approaches have been developed to manage missing data and enable the development of downstream ML tasks [21]. Among them, imputation (i.e., filling in the missing data with some replacement values) is

one of the most popular approaches, with a variety of techniques ranging from constant imputation (encompassing commonly used approaches such as mean imputation) to regression imputation [43, 44], to multiple imputation [7, 38].

Nonetheless, the conditions under which missing data imputation is expected to work—i.e., it does not negatively impact the learnability of downstream ML tasks—are yet unclear. As such, the exploration of imputation methods is still mostly guided by empirical experimentation and ad-hoc strategies. Indeed, despite extensive empirical validation and comparisons [18, 24, 29, 35], often providing conflicting results, there is no consensus on when and how imputation should be applied [31], and when learning is still possible despite the potential noise introduced by imputation [27]. Most relevantly, from a theoretical point of view, there is still a lack of research studying the impact of missing data imputation on learnability in the statistical learning theory framework, especially in regard to the development of practical finite-sample guarantees that ensure the feasibility of learning after imputation.

Research on the interaction between missing data, imputation, and learnability has been carried out mainly within the algorithmic learning theoretic framework of *learning under partial observability* [14, 13, 22, 33, 39]. While these approaches can provide finite-sample guarantees, most work in the area has focused on model classes (e.g., propositional formulas) and settings (e.g., concept learning) which, while interesting from the point of view of exploring the computational limits to learnability with missing data, are rather distant from the current practice of ML. In contrast, research in the statistical learning framework has been primarily concerned with asymptotic consistency or optimality guarantees [8, 25, 28, 27, 40], especially so for specific classes of ML models such as linear predictors, which despite being closer to the methodology most commonly adopted in modern ML do not provide finite-sample guarantees. More recently, Ayme et al. [4–6, 37] have investigated finite-sample (as well as optimality) guarantees for learning after imputation: however, this line of research has focused only on linear models under specific imputation strategies (constant and pattern-by-pattern imputation), and does not provide general conditions for learnability in this setting that can be applied for general learning models.

In this paper, we start to address this gap by studying conditions under which learning under imputation is feasible. In particular, we provide three main contributions. First, we introduce a condition for learnability under imputation, called *noise stability*, which we show to be sufficient and, under weak assumptions, also necessary. Intuitively, this condition guarantees that a learning algorithm is stable under noisy variations of the real data introduced by an imputation mechanism. Second, we study two common imputation strategies (namely, constant and regression imputation), providing bounds on the noise they may introduce in learning problems. Finally, we show that a large class of ML models are noise stable, and provide finite-sample guarantees for learnability with missing data using these model classes.

2 Background and Mathematical Notation

Let X be the instance space: we assume that X is a (subset of a) d-dimensional real vector space. Given any vector $x \in X$, $x^{(i)}$ denotes the i-th dimension of x. Let Y be the target space. Y can be either discrete (finite or countable), in which case we consider a classification task, or also continuous (that is, uncountable), in which case we consider a regression task. Let $\mathcal{D}$ be a probability measure over $X \times Y$, called *data-generating process*. $\mathcal{D}$ represents the process that generates the complete, possibly unobserved, data samples. We also assume the existence of d probability measures $\mathcal{M}_1, \ldots, \mathcal{M}_d$, where, for each i, $\mathcal{M}_i$ is a probability measure over $X \times Y \times \tilde{X}^{(i)}$, where $\tilde{X}^{(i)} = X^{(i)} \cup \{\perp\}$. In particular, the symbol $\perp$ denotes a missing value. Let $\mathcal{M} = \Pi_i \mathcal{M}_i$ and $\tilde{X} = \Pi_i \tilde{X}^{(i)}$. We will assume that data is generated according to the following process: first, a sample (x, y) is drawn from $\mathcal{D}$; subsequently, an *incomplete* sample $(x', y) \in \tilde{X} \times Y$ is drawn from the conditional $\mathcal{M}(\cdot|(x, y))$. The conditional distributions $\mathcal{M}_i(\perp|x^{(1)}, \ldots, x^{(i)}, \ldots, x^{(d)})$ are of particular interest since, based on the dependency structure of the conditionals, one can distinguish different types of missingness mechanisms: as we will not discuss further this categorization, we refer the interested reader to [30].

We will represent the action of imputation methods by the abstract definition of an *imputation mechanism*. A imputation mechanism is a randomized algorithm Impute : $(\tilde{X} \times Y)^m \times (\tilde{X} \times Y) \to X$ that takes as input a training set S, a (partially observed) instance (x, y) and gives as output an imputed, possibly corrupted, instance x'. Given a missing data mechanism $\mathcal{M}$ and an imputation mechanism Impute, we define a randomized algorithm Corrupt : $(X \times Y)^m \times (X \times Y) \to X$: Corrupt takes as input a dataset S, an instance (x, y), applies the missingness mechanism $\mathcal{M}$ to both S and (x, y), and then returns the result of Impute$(S, (x, y))$. We will use the notation $\text{Corrupt}_S(x, y)$ to denote the action of Corrupt on an instance (x, y) for a given dataset S. We call the probability measure defined by Corrupt a *corruption mechanism*. We say that Corrupt satisfies the *small noise condition*, with noise function $\epsilon_{\text{Corrupt}} : \mathbb{N} \times \mathbb{R} \to \mathbb{R}$, if, for each $m > 0, \delta > 0$, with probability larger than $1 - \delta$ over the sampling of a dataset $S \sim \mathcal{D}^m$ and $T = \{Corrupt_S(x, y) | (x, y) \in A\}$, it holds that:

$$\mathbb{E}[\|Corrupt_S(x, y) - x\|_X] \leq \epsilon_{\text{Corrupt}}(m, \delta), \tag{1}$$

where $\epsilon_{\text{Corrupt}}$ is non-increasing in its arguments and $\|\cdot\|_X$ is a norm on X.

Let $\mathcal{H}$ be a set of models, that are, functions $h : X \to Z$, where Z is a set[1]. A *learning algorithm* is a function $A : \bigcup_{m \in \mathbb{N}^+} (X \times Y)^m \to \mathcal{H}$. A loss function is a map $l : X \times Y \times \mathcal{H} \to \mathbb{R}$. A loss function l is L-Lipschitz if $\mathcal{H}$ has a norm $\|\cdot\|_{\mathcal{H}}$ and $\forall x \in X, y \in Y, h_1, h_2 \in \mathcal{H}, |l(x, y, h_1) - l(x, y, h_2)| \leq L\|h_1 - h_2\|$. l is μ-strongly convex if $\forall h_1, h_2 \in \mathcal{H}, \alpha \in [0, 1]$, $\alpha h_1 + (1 - \alpha)h_2$ exists and is in $\mathcal{H}$ and it holds that $l(x, y, \alpha h_1 + (1-\alpha)h_2) \leq \alpha l(x, y, h_1) + (1-\alpha) l(x, y, h_2) - \frac{\mu}{2}\alpha(1-\alpha)\|h_1 - h_2\|^2$. l is convex if the previous requirement holds only for $\mu = 0$. If l is differentiable,

[1] In general, Z can be different from Y: for example, in the case of binary classification (where $Y = \{-1, 1\}$), Z is often set equal to $\mathbb{R}$.

then l is M-smooth if $\forall x \in X, y \in Y, h_1, h_2 \in \mathcal{H}, \|\nabla l(x,y,h_1) - \nabla l(x,y,h_2)\|_* \leq M\|h_1 - h_2\|$, where $\|\cdot\|_*$ is the dual norm.

Given a data-generating distribution $\mathcal{D}$ and a model $h \in \mathcal{H}$, we define the *true risk* of h as $R(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[l(x,y,h)]$. Given a finite sample $S = \{(x_1,y_1),\ldots,(x_m,y_m)\} \in (X\times Y)^m$, we define the *empirical risk* of h as $\hat{R}_S(h) = \frac{1}{m}\sum_{i=1}^m l(x_i,y_i,h)$. We define empirical risk minimization (ERM) to be any learning algorithm A s.t. $A(S) \in \inf_{h\in\mathcal{H}} \hat{R}_S(h)$. We say that a model class $\mathcal{H}$ is *(agnostic) learnable*[2] if there exists a learning algorithm $A : \bigcup_{m\in\mathbb{N}^*}(X\times Y)^m \to \mathcal{H}$ and a function $\epsilon_L^{\mathcal{H},A} : \mathbb{N}\times\mathbb{R} \to \mathbb{R}$ s.t., for each data-generating mechanism $\mathcal{D}$, $\delta > 0$ and sample size m, with probability larger than $1-\delta$ over the sampling of $S \sim \mathcal{D}^m$ it holds that:

$$|R(A(S)) - \hat{R}_S(A(S))| \leq \epsilon_L^{\mathcal{H},A}(m,\delta), \tag{2}$$

with $\lim_{m\to\infty} \epsilon_L^{\mathcal{H},A}(m,\delta) = 0$. We say that $\mathcal{D}$ is *realizable* (w.r.t. $\mathcal{H}$) if $\exists h \in \mathcal{H}$ s.t. $R(h) = 0$. Finally, $\mathcal{H}$ satisfies *uniform convergence* if there exists a function $\epsilon_{UC}^{\mathcal{H}} : \mathbb{N}\times\mathbb{R} \to \mathbb{R}$ s.t., for all data-generating mechanisms $\mathcal{D}$, $\delta > 0$ and sample sizes $m > 0$, with probability larger than $1-\delta$ over the sampling of $S \sim \mathcal{D}^m$, it holds that $\sup_{h\in\mathcal{H}} |R(h) - \hat{R}_S(h)| \leq \epsilon_{UC}^{\mathcal{H}}(m,\delta)$, with $\lim_{m\to\infty} \epsilon_{UC}^{\mathcal{H}}(m,\delta) = 0$.

3 Learnability with Missing Data

As a first step, we extend the definition of learnability to the case of learning with missing data using imputation.

Definition 1. *Let $\mathcal{H}$ be a class of models. Then, $\mathcal{H}$ is* learnable under imputation *if there exists a learning algorithm $A : \bigcup_{m\in\mathbb{N}}(X\times Y)^m \to \mathcal{H}$ and two functions $\epsilon_1^{\mathcal{H},A} : \mathbb{N}\times\mathbb{R} \to \mathbb{R}$ and $\epsilon_2^{\mathcal{H},A} : \mathbb{N}\times\mathbb{R}^2 \to \mathbb{R}$ s.t. for any $\delta,\epsilon > 0$, $m \in \mathbb{N}^+$, data-generating mechanism $\mathcal{D}$, corruption mechanism Corrupt with $\epsilon_{Corrupt}(m,\delta) \leq \epsilon$, it holds with probability larger than $1-\delta$ (over the sampling of $S \sim \mathcal{D}^m$ and $T = \{Corrupt_S(x,y)|(x,y) \in S\}$) that:*

$$R(A(T)) \leq \hat{R}_T(A(T)) + \epsilon_1^{\mathcal{H},A}(m,\delta) + \epsilon_2^{\mathcal{H},A}(m,\epsilon,\delta), \tag{3}$$

with $\epsilon_1^{\mathcal{H},A}, \epsilon_2^{\mathcal{H},A}$ decreasing in all of their arguments, and $\lim_{m\to\infty} \epsilon_1^{\mathcal{H},A}(m,\delta) = 0$ and $\lim_{m\to\infty,\epsilon\to 0} \epsilon_2^{\mathcal{H},A}(m,\epsilon,\delta) = 0$.

Definition 1 is a natural generalization of the notion of *agnostic learnability* in the classical PAC learning framework. Indeed, for a model class to be learnable under imputation means that the generalization gap can be bounded as

[2] The definition of agnostic learnability we provide is sometimes called *generalizability* in the literature, whereas (strict) learnability (also called, universal consistency) is defined by replacing Eq. (2) with $|R(h^*) - R(h(S))| \leq \epsilon_L^{\mathcal{H},A}(m,\delta)$, with $h^* \in \arg\inf_{h\in\mathcal{H}} R(h)$. Since, under weak assumptions (in particular, under uniform convergence), generalizability is both necessary and sufficient for strict learnability, in this paper we focus on the former notion, and term it learnability.

a quantity that vanishes with the noise introduced by the considered imputation mechanism (and increasing data). Thus, if the imputation is approximately faithful to the real, unobserved data, the empirical risk on the imputed data should be (with high probability) close to the true risk. Specifically, note that Definition 1 requires that if we are able to make the noise due to imputation ϵ arbitrarily small, we recover exactly the definition of agnostic learnability. In contrast, when the noise due to imputation is too large, we cannot expect a model to work as well, as the data used for training will essentially be out-of-distribution compared to the data-generating mechanism $\mathcal{D}$ w.r.t. which the true risk is computed: nonetheless, the penalty in performance can be upper bounded by a quantity that only depends on the noise itself.

As an additional comment, we note that Definition 1 also implies that a good learning algorithm could be obtained by minimizing the right side of Eq. (3). While in standard supervised learning such an algorithm exists (e.g., in both binary classification and regression, ERM), it is not similarly easy to identify such an optimal algorithm in the setting of learning under imputation, as the learning algorithm A affects not only the empirical risk $\hat{R}_T$ but also the term $\epsilon_2^{\mathcal{H},A}$: as a consequence, it is not clear whether ERM minimizes the bound in Eq. 3, even under uniform convergence (in which case, the term $\epsilon_2^{\mathcal{H}}$ is independent of A). This situation is not unexpected, as also in the setting of general learning there exist natural learning tasks for which ERM is not always an optimal strategy [41]. For this reason, in the following, we allow arbitrary learning algorithms A.

We now introduce the central definition in our mathematical development, which we will use to provide a characterization of learnability under imputation.

Definition 2. *Let $\mathcal{H}$ be a set of models, l a loss function and A a learning algorithm for $\mathcal{H}$. We say that A is* noise risk stable *(NRS) if there exists a function $\epsilon_{NRS}^{A} : \mathbb{N} \times \mathbb{R}^2 \to \mathbb{R}$ s.t., for every $\delta, \epsilon > 0, m > 0$ and corruption mechanism Corrupt with $\epsilon_{Corrupt}(m, \delta) \leq \epsilon$, with probability larger than $1 - \delta$ over the sampling of $S \in (X \times Y)^m$ and $T = \{Corrupt_S(x, y) : (x, y) \in S\}$, it holds that:*

$$\hat{R}_S(h) - \hat{R}_T(h) \leq \epsilon_{NRS}^{A}(m, \epsilon, \delta) \tag{4}$$

where $h = A(T)$, and $\limsup_{m\to\infty,\epsilon\to 0} \epsilon_{NRS}^{A}(m, \epsilon, \delta) \leq 0$.

Intuitively, a learning algorithm is NRS if the losses of the models trained on the complete (unobserved) data and the data corrupted by the imputation mechanism are close. Thus, when the imputation does not introduce an excessive amount of noise, a NRS algorithm will not amplify this noise by more than a negligible quantity (that goes to 0 as the noise itself goes to 0). Then, we prove our main result: model classes for which there exists a NRS algorithm are learnable even under corruptions introduced by imputation.

Theorem 1. *Let $\mathcal{H}$ be a model class and l be a loss function bounded in $[0, b]$. Assume that $\mathcal{H}$ is learnable, and define $\mathcal{A}_{\mathcal{H}} = \{A : \bigcup_{m\in\mathbb{N}^+}(X \times Y)^m \to \mathcal{H} | \mathcal{H}$ is learnable using $A\}$. For each $A \in \mathcal{A}_{\mathcal{H}}$, let $\epsilon_L^{\mathcal{H},A}$ be the generalization*

gap as defined in Eq. (3)*. Then, a sufficient condition for learnability under imputation is that there exists $A \in \mathcal{A}_{\mathcal{H}}$ that is NRS. Furthermore, when such an A exists, then, with probability larger than $1-\delta$ over the sampling of $S \sim \mathcal{D}^m$ and $T = \{Corrupt_S(x,y) | (x,y) \in S\}$, it holds that $R(A(T))$ can be upper bounded by $\hat{R}_T(A(T)) + \epsilon_L^{\mathcal{H},A}(m, \frac{\delta}{3}) + \epsilon_{NRS}^{\mathcal{H},A}(m, \epsilon_{Corrupt}(m, \frac{\delta}{3}), \frac{\delta}{3})$.*

Under the assumption that $\mathcal{H}$ satisfies uniform convergence, the existence of a NRS $A \in \mathcal{A}_{\mathcal{H}}$ is also necessary for learnability under imputation.

Theorem 1 provides a tight characterization of the notion of learnability under imputation, as it shows that, under weak assumptions, it is equivalent to the notion of noise risk stability. Thus, if the gap in empirical risk ($\hat{R}_S(A(T)) - \hat{R}_T(A(T))$) introduced by imputation can be controlled, then also the generalization gap $R(A(T)) - \hat{R}_T(A(T))$ can be controlled: in particular, if we can make the noise $\epsilon_{\text{Corrupt}}$ arbitrarily small (as a function of m), then we can make the generalization gap also arbitrarily small. Note that, under uniform convergence and with probability larger than $1-\delta$, this also implies that $R(A(T)) - \inf_{h \in \mathcal{H}} R(h)$ can be made arbitrarily small[3]: this guarantees that, in particular and despite the presence of noise introduced by imputation, it is possible to get a model that (with high probability) will be arbitrarily close to the optimal Bayes predictor $h^* = \inf_{h \in \mathcal{H}} R(h)$. Note that this result is considerably stronger than agnostic learnability, as in the case of learning under imputation the learning algorithm A is not able to observe the real data sampled from $\mathcal{D}$ but only an out-of-distribution sample drawn from Corrupt. Nonetheless, we note that Theorem 1 only provides an upper bound on the generalization gap. In general, a lower bound could easily be obtained by using a lower bound for learnability, however it is not clear whether this bound is tight: that is, there may exist learning problems for which, despite the presence of imputation, we can recover the same generalization gap as if the data were completely observed. We leave the search for lower bounds (as well as to understand whether there exist matching lower and upper bounds) as future work.

Before proceeding further, we make some further comments on Theorem 1 and highlight some potential related open questions. Firstly, a known result from statistical learning theory implies that a class of models $\mathcal{H}$ is (agnostic) learnable if and only if there exists a uniform replace-one (RO) stable[4] asymptotic ERM[5] algorithm A that learns it [41]. Thus, in a sense, stability is already a necessary condition for (conventional) learnability. However, even though both definitions

[3] Indeed, $R(A(T)) - \inf_{h \in \mathcal{H}} R(h) = R(A(T)) - \hat{R}_T(A(T)) + \hat{R}_T(A(T)) - \inf_{h \in \mathcal{H}} R(h)$. If uniform convergence holds, then, without loss of generality, A can be set to be an ERM [41]. Hence, letting $h^* = \inf_{h \in \mathcal{H}} R(h)$, it holds that $R(A(T)) - R(h^*) \leq R(A(T)) - \hat{R}_T(A(T)) + \hat{R}_T(h^*) - R(h^*) \leq \epsilon_{NRS}^{A}(m, \epsilon_{\text{Corrupt}}(m, \frac{\delta}{4}), \frac{\delta}{4}) + \epsilon_L^{\mathcal{H},A}(m, \frac{\delta}{4}) + \epsilon_{UC}^{\mathcal{H}}(m, \frac{\delta}{4}) \to 0$, as $\epsilon_{\text{Corrupt}} \to 0$ and $m \to \infty$.

[4] A is uniform RO stable if there exists $\epsilon_{\text{stable}} : \mathbb{R} \to \mathbb{R}$ s.t. for all possible $S = \{(x_1, y_1), \ldots, (x_m, y_m) \in (X \times Y)^m$ and $(x,y) \in (X \times Y)$ it holds that $\frac{1}{m} \sum_{i=1}^{m} |l(x,y,A(S)) - l(x,y,A(S^i)| \leq \epsilon_{\text{stable}}(m)$, where $S^i = \{(x_1,y_1), \ldots, (x_{i-1}, y_{i-1}), (x,y), (x_{i+1}, y_{i+1}), \ldots, (x_m, y_m)$.

[5] A is an asymptotic ERM if $\lim_{m \to \infty} \mathbb{E}_{S \sim \mathcal{D}^m}[|\hat{R}_S(A(S)) - \inf_{h \in \mathcal{H}} \hat{R}_S(h)|] = 0$.

of stability require that the risk of a learning algorithm does not change too much under small variations of the data, the constraints they impose on a learning algorithm are rather different from an intuitive point of view. Indeed, while uniform RO stability requires the loss being stable under arbitrary changes to a single instance, noise risk stability requires stability under constrained changes to the entire data-generating distribution. We leave the problem of further characterizing the relationships among these two notions of stability (and, in particular, the identification of conditions under which they are equivalent) as future work.

Second, while Theorem 1 shows that noise risk stability is necessary for learnability under imputation, the proof of this fact relies on uniform convergence. Such an assumption is reasonable for many commonly occurring learning tasks: e.g., for both binary classification [12, 45] and bounded regression [3], uniform convergence is equivalent to learnability, and, more generally, uniform convergence is equivalent to learnability whenever the loss function l is bounded, has linear dependence on $h \in \mathcal{H}$ and $\mathcal{H}$ itself is bounded [42]. However, uniform convergence may also fail to hold in many natural scenarios in which learnability is nevertheless possible [19, 41]. In these scenarios, noise risk stability is still sufficient for learnability under imputation, but the proof of necessity in Theorem 1 fails. Therefore, a more general proof technique should be adopted to understand whether noise risk stability characterizes learnability under imputation also in these more general settings: we leave this open question as future work.

Finally, while in this work we focus solely on learnability under imputation, the small noise condition in Eq. (1) only requires that the average distance between instances sampled from $\mathcal{D}$ and instances sampled from a different distribution $\mathcal{L}$ can be bounded from above by a quantity that is monotonically decreasing (in m). While in our setting $\mathcal{L}$ is the distribution derived by the corruption mechanism Corrupt, in principle one could set $\mathcal{L}$ arbitrarily as long as it satisfies the above-mentioned condition. In this sense, learnability under imputation can be seen as a special case of *domain adaptation* [36] and *credal learning* [15, 16]. That is, given a set of data sampled from $\mathcal{L}$, we want to control the out-of-distribution (OOD) risk on data sampled from $\mathcal{D}$, under the constraint that both $\mathcal{D}$ and $\mathcal{L}$ are contained in a set of probability distributions which, by the small noise condition, is not too large: hence information about $\mathcal{L}$ can be used as a proxy for information about $\mathcal{D}$. We leave studying the relationships between learning under imputation and OOD learnability as future work.

4 Linear-in-Parameter Models are Learnable under Imputation

While in the previous section we established that noise risk stability is equivalent to learnability under imputation, we have not yet proved the existence of NRS algorithms. The following Theorem establishes this fact.

Theorem 2. *Let $l : X \times Y \times \mathcal{H} \to \mathbb{R}$ be a loss function that is L-Lipschitz in its first argument, and assume that X is s.t. $\sup_{x \in X} \|x\|_X = B$. Then, any learning*

algorithm $A : \bigcup_{m=1}^{\infty}(X \times Y)^m \to \mathcal{H}$ *is NRS and:*

$$\epsilon^A_{NRS}(m, \epsilon_{Corrupt}, \delta) \leq L\epsilon_{Corrupt}(m, \frac{\delta}{2}) + 2BL\sqrt{\frac{\log(2/\delta)}{2m}}. \tag{5}$$

In particular, assume $l(x, y, h) = g(\langle h, \phi(x)\rangle_K)$*, where* $g : \mathbb{R} \to \mathbb{R}$ *is L-Lipschitz,* K *is a Reproducing Kernel Hilbert Space,* $\langle \cdot, \cdot \rangle_K$ *is an inner product on* K *and* $\| \cdot \|_K$ *the derived norm,* $\mathcal{H} \subseteq K$ *is s.t.* $\sup_{h\in\mathcal{H}} \|h\|_K = H$*,* $\phi : X \to K$*, and* $\phi(X) = \{\phi(x) : x \in X\}$ *is s.t.* $\sup_{k\in\phi(X)} \|k\|_K = B$*. Then, any learning algorithm* A *is NRS and:*

$$\epsilon^A_{NRS}(m, \epsilon_{Corrupt}, \delta) \leq LH\epsilon_{Corrupt}(m, \frac{\delta}{2}) + 2BHL\sqrt{\frac{\log(2/\delta)}{2m}}. \tag{6}$$

Assume that $l(x, y, h) = g(\langle h, \phi(x)\rangle_K) + r(\|h\|^2_K)$*, with* g *as defined above,* g *and* $r : \mathbb{R} \to \mathbb{R}$ *differentiable and* M_g*-smooth,* M_r*-smooth respectively. Assume that* l *is* b*-bounded and* $\mu(m)$*-strongly convex in* $\mathcal{H}$*, with* $\mu : \mathbb{N} \to \mathbb{R}$ *being a function monotone increasing with* m*. Then, assuming* $\inf_{h\in\mathcal{H}} \hat{R}_S(h) \leq \hat{R}_T(A(T))$ *with* S *and* T *defined as above, there exists an algorithm* A *for which it holds, with probability larger than* $1 - \delta$*, that:*

$$\epsilon^A_{NRS}(m, \epsilon_{Corrupt}, \delta) \leq b(1 - \frac{1}{2(B^2M_g + 2M_r)})^T + \frac{\psi^2}{\mu(m)}, \tag{7}$$

where $\psi = HL\epsilon_{Corrupt}(m, \frac{\delta}{2}) + 2BHL\sqrt{\frac{\log(2/\delta)}{2m}} + \delta BHL$*. The algorithm* A *is gradient descent with step size* $\gamma \leq \min\{\frac{1}{(B^2M_g+2M_r)}, \frac{1}{\mu(m)T}\}$ *and executed for* T *iterations. The result above holds, in expectation, also for stochastic gradient descent (with the same step size and the same number of iterations) as long as for all* $h \in \mathcal{H}$ *the noisy gradient estimates* $g(h)$ *satisfy* $\mathbb{E}[g(h)] = \nabla\hat{R}_T(h)$*.*

Theorem 2 shows that when the loss function is Lipschitz w.r.t. X, any algorithm in $\mathcal{A}_{\mathcal{H}}$ (and, in particular if uniform convergence holds, ERM) is NRS. Hence, not only ERM w.r.t. such a loss function provides an algorithm for learning under imputation but, under these assumptions, learnability under imputation is equivalent to learnability. In particular, as a consequence of Eq. (6), the result holds for linear-in-parameter models (which include any kernel method), as long as the models have bounded norm. Nevertheless, we note that this does not imply that learning under imputation is as easy as conventional learning. Indeed, the bound in Eqs. (5) and (6) encompass also the $\epsilon_{\text{Corrupt}}$ term. Thus, fixed a particular class of imputation mechanisms and a class of models $\mathcal{H}$, if the term $\epsilon_{\text{Corrupt}}$ does not converge to 0 as $m \to \infty$, it may be that ϵ_{NRS} (and, thus, the generalization gap $R(A(T)) - \hat{R}_T(A(T))$) does not converge to 0, even though $\mathcal{H}$ is learnable under imputation using A. The reason behind this seemingly paradoxical phenomenon is that the definition of learnability under imputation guarantees the above-mentioned convergence only when we consider a class of imputation mechanisms whose small noise constant also converges to 0 as the sample size grows. In this sense, learnability under imputation is in

general strictly harder than learnability, as the former requires not only a good learning algorithm, but also a good imputation mechanism.

Notably, however, under the stronger assumptions on the learning problem required for Eq. (7), learnability under imputation is always possible, regardless of the specific imputation mechanism adopted. Indeed, even for asymptotically inconsistent imputation mechanisms for which $\lim_{m\to\infty} \epsilon_{\text{Corrupt}}(m, \delta) > 0$, Eq. (7) guarantees that ϵ_{NRS} converges to 0 as m grows, since the learning algorithm (gradient descent) ensures that the noise due to imputation is attenuated at a rate $\frac{1}{\mu(m)}$. Nonetheless, also in this setting, not all imputation mechanisms are equally good: indeed, the guarantee in Eq. (7) is particularly favorable when it is possible to find a class of imputation mechanisms that is consistent (i.e., $\lim_{m\to\infty} \epsilon_{\text{Corrupt}}(m, \delta) > 0$) as, in this case, the generalization gap can decrease exponentially fast in the size of the training set. Also, we note that, though stronger than the assumptions of Eqs. (5) and (6), the assumptions required for Eq. (7) to hold (with the exception of the requirement that $\inf_{h\in\mathcal{H}} \hat{R}_S(h) \leq \hat{R}_T(A(T))$, which is intuitively reasonable[6] but not testable, as S is not available) are naturally satisfied by several natural learning problems, such as (regularized) kernel logistic regression as well as kernel ridge regression. As we discuss in the next section, this result provides a generalization of the main result in [5].

Before proceeding to the next section, we provide some further discussion of the previous result, also highlighting some open questions. First, we note that Theorem 2 requires that the loss function be Lipschitz w.r.t. X: this is different from the usual definition of a Lipschitz loss function, which requires Lipschitzness w.r.t. $\mathcal{H}$. Nonetheless, whenever the loss function is symmetric w.r.t. X and $\mathcal{H}$ (this happens, e.g., for linear-in-parameter models, as in Eq. (6)) the two requirements are equivalent (though with different parameters).

Secondly, Theorem 2 only provides a necessary condition for noise risk stability, and furthermore only provides an upper bound on ϵ_{NRS}. In particular, this implies that even though under the assumptions of Theorem 2 learnability and learnability under imputation are equivalent, there may be learning problems for which this equivalence does not hold. We leave the analysis of noise risk stability for other algorithms, as well as possibly the computation of lower bounds (as well as tighter upper bounds) on ϵ_{NRS}, to future work.

Thirdly, while Eq. (7) provides better guarantees than both Eqs. (5) and (6), it also enforces the selection of a learning algorithm: indeed, while Eqs. (5) and (6) hold for any learning algorithm A, Eq. (7) is guaranteed to hold only for (stochastic) gradient descent. While this is generally not a problem, as descent methods are among the most commonly employed ML algorithms and can be usually implemented efficiently, the previous observation implies that when it is not possible to use gradient descent (e.g., for non-differentiable functions, or also for black-box problems in which only a zero-th order oracle is available), Eq. (7) cannot be applied. As our analysis strictly relies on the possibility to use a gradient descent procedure, it is therefore not clear whether it would be

[6] For example, the assumption holds when the original distribution $\mathcal{D}$ is realizable.

possible to achieve similar convergence guarantees, that hold irrespective of the imputation mechanism and its (in)consistency, also for other learning algorithms.

Finally, our proof of Eq. (7) relies on previous results on the convergence gradient descent with biased oracles [2]. Other papers have investigated the properties of descent-based algorithms with inexact gradient oracles [9, 10, 20, 23]. In particular, Chen et al. [17] studied convergence of (sub)gradient descent under a noise model according to which a biased but asymptotically consistent estimator of the gradient is available: at each iteration of the descent procedure, the estimator is computed on the basis of multiple samples, and the estimator is required to converge (in probability) to the actual gradient. This setting seems to be of particular relevance to learnability under imputation: indeed, there exist imputation mechanisms that allow the construction of multiple filled-in datasets, so-called multiple imputation methods [38], and these alternative imputations could, in turn, be used to obtain the multiple gradient approximations required for the procedure described in [17]. As in this paper we do not discuss multiple imputation methods, we leave the investigation of such a scenario to future work.

5 Characterizing the Small Noise Condition

In this section we provide a characterization of the small noise condition in Eq. (1) for two commonly used imputation strategies, namely constant imputation and regression imputation. These results provide computable certificates for the small noise condition and thus, together with Theorem 2, also provide a way to explicit finite-sample guarantees from the bounds in Theorem 1.

Theorem 3. *Let FV_v, with $v : (\tilde{X} \times Y)^m \to X$, be the imputation mechanism defined coordinate-wise by:*

$$FV_v(S,(x,y))_i = \begin{cases} x_i & x_i \neq \perp \\ v_i(S) & otherwise \end{cases}. \tag{8}$$

Then, for every ℓ_p norm it holds, with probability 1, that: $\mathbb{E}[\|x - Corrupt_S(x,y)\|_X] \leq Tr(\Sigma) + \mathbb{E}[\|v(T) - \mu\|_X] \leq \sup_{x_1,x_2 \in X} \|x_1 - x_2\|_X$, where Σ is the covariance matrix of $\mathcal{D}_X = \int_Y \mathcal{D}$ and $\mu = \mathbb{E}[x]$. Let avg be defined coordinate-wise by $avg_i(S) = \frac{1}{|S_i^C|} \sum_{(x_j,y_j) \in S} x_j^{(i)} \mathbb{1}_{x_j^{(i)} \neq \perp}$. Then, if $\|\cdot\|_X$ is the ℓ_2 norm, then it holds with probability 1, that $c \cdot Tr(\Sigma) \leq \mathbb{E}[\|x - Corrupt_S(x,y)\|_X]$.

Additionally, if $\mathcal{M}$ is s.t. $\forall i, \mathcal{M}_i(\perp|(x,y)) = c$ and $\sup_{x \in X} \|x\|_X = D$, then, with probability larger than $1 - \delta$ over the sampling of $S \sim \mathcal{D}^m$ and $T = \{Corrupt_S(x,y) : (x,y) \in S\}$ it holds that $c \cdot Tr(\Sigma) \leq \mathbb{E}[\|x - Corrupt_S(x,y)\|_X] \leq Tr(\Sigma) + 2D \sum_{i=1}^d \sqrt{\frac{\log(1/\delta)}{2|S_i^C|}}$.

Theorem 4. *Let $\mathcal{R}_i$ be a set of regression models $r : X^{-i} \to \mathbb{R}$, where $X^{-i} = \Pi_{i' \neq i} X^{(i')}$. Let $Fill : \cup_{m=1}^{\infty} \Pi_{i=1}^d \tilde{X}^{-i} \to \cup_{m=1}^{\infty} \Pi_{i=1}^d \tilde{X}^{-i}$ be an arbitrary function satisfying: 1) $|Fill(S)| = |S|$; 2) $Fill(S)_j^{(i)} = S_j^{(i)}$ iff $S_j^{(i)} \neq \perp$. For each feature i,*

assume that l_i is a loss function bounded in $[0, b_i]$ on $X^{(i)}$ and s.t. there exists a $\nu : \mathbb{R} \to \mathbb{R}$ with $\|x_1 - \Pi_{i=1}^d r_i(x_2^{-i})\|_X = \nu(\sum_{i=1}^d l_i(x^{-i}, x^{(i)}, r(x^{-i})))$[7]. Assume that $\mathcal{R}_i$ is learnable over $(\Pi_{i=1}^d X^{-i}) \times X^{(i)}$ through an algorithm A_i with $\epsilon_L^{\mathcal{R}_i, A_i}$. We set Reg to be the imputation mechanism defined coordinate wise by:

$$Reg(S, (x, y))_i = \begin{cases} x^{(i)} & x^{(i)} \neq \perp \\ A_i(T_C^i)(Fill(T^{-i})_j) & otherwise \end{cases}, \tag{9}$$

where $T_C^i = \{(Fill(T^{-i})_j, x_j^{(i)}) : x_j \in T \wedge x_j^{(i)} \neq \perp\}$.

Then, with probability larger than $1 - \delta$ over the sampling of $S \sim \mathcal{D}^m$ and $T = \{(Corrupt_S(x, y) : (x, y) \in S\}$, it holds that $\mathbb{E}[\|x - Corrupt_S(x, y)\|_X]$ can be upper bounded by $\nu(\phi(T, l, \mathcal{R}_1, \ldots, \mathcal{R}_d))$, with $\phi(T, l, \mathcal{R}_1, \ldots, \mathcal{R}_d)$ being:

$$\sum_{i=1}^d 2Rad(l_i \circ \mathcal{R}_i \circ T_C^i) + \frac{1}{S_i^C} \sum_{j=1}^m l_1(x^{-i}, x^{(i)}, A_i(T_C^i)) \mathbb{1}_{x_j^{(i)} \neq \perp} + 4b_i \sqrt{\frac{2\log(4d/\delta)}{|S_i^C|}}, \tag{10}$$

where $T_C^i = \{(Fill(T^{-i})_j, x_j^{(i)}) : x_j \in T \wedge x_j^{(i)} \neq \perp\}$ and Rad is the empirical Rademacher complexity.

Corollary 1. *Assume the same setting as in Theorem 4. Let $\mathcal{L}$ be the distribution over X induced by the composition of $\mathcal{D}$, $\mathcal{M}$ and Fill. Assume, further, that: 1) $\mathcal{L}$ is realizable; 2) $Rad(l_i \circ \mathcal{R}_i \circ T_C^i) \to 0$ as $|T_C^i| \to \infty$[8]; 3) ν is monotone increasing and $\lim_{z \to 0^+} \nu(z) = 0$. Then $\lim_{m \to \infty} \epsilon_{Corrupt}(m, \delta) = 0$.*

Finally, we prove instantiantions of Theorem 4 for two concrete class of regression models, namely linear models (used, for example, by the MICE library [44]) and tree ensembles (used, for example, by the missForest library [43]).

Corollary 2. *Assume the same setting as in Theorem 4. Assume that each of the loss functions l_i is L-Lipschitz and can be written as $g_i(\langle h, \phi(x) \rangle_K)$, where g, $\langle \cdot, \cdot \rangle_K$ and K are defined as in Theorem 2. Assume further that the derived norm $\|\cdot\|_K$ is the ℓ_2 norm over K and $\sup_{h \in \mathcal{H}} \|h\|_K = B$. Then, it holds that Eq. (10) can be upper bounded by $\sum_{i=1}^d \hat{R}_{Fill}(A_i(T_C^i)) + 2\sup_{x \in X} \frac{LB\|\phi(x)\|_K}{\sqrt{|S_i^C|}} + 4b_i\sqrt{\frac{2\log(4d/\delta)}{|S_i^C|}}$, where $\hat{R}_{Fill}(h) = \frac{1}{S_i^C} \sum_{j=1}^m l_1(Fill(T^{-i})_j, x_j^{(i)}, h(T_C^i)) \mathbb{1}_{x_j^{(i)} \neq \perp}$ Similarly, under the same assumptions as above but requiring that the derived norm $\|\cdot\|_K$ is the ℓ_1 norm over K, it holds that Eq. (10) can be upper bounded by $\sum_{i=1}^d \hat{R}_{Fill}(A_i(T_C^i)) + 2LB\sup_{x \in X} \|\phi(x)\|_\infty \sqrt{\frac{2\log(2(d-1))}{|S_i^C|}} + 4b_i\sqrt{\frac{2\log(4d/\delta)}{|S_i^C|}}$.*

[7] For example, if $\|\cdot\|_X^{\ell_2}$, then, setting $l_i(x, y, r) = (y - r(x))^2$ and $\nu(x) = \sqrt{x}$ satisfies the mentioned condition. Indeed, $\|x - \Pi_i r_i(x^{-i})\|_X^{\ell_2} = \sqrt{\sum_i (x^{(i)} - r_i(x^{-i}))^2} = \nu(\sum_i l_i(x_i, y_i, r_i))$. Similarly, if $\|\cdot\|_X^{\ell_1}$ then setting $l_i(x, y, r) = l_1(x, y, r) = |y - r(x)|$ and ν the identity, also satisfies the mentioned condition.

[8] This condition holds, in particular, for both the class of linear models bounded in ℓ_1 or ℓ_2 norm, as well as the class of ensembles of regression trees.

Corollary 3. *Assume the same setting as in Theorem 4. Assume that, for each i, $X^{(i)}$ is bounded in $[-B, B]$. Assume, further, that for each i, $\mathcal{R}_i$ is the class of ensembles composed by at most r regression trees with depth at most k. Then, it holds that Eq. (10) can be upper bounded by* $\sum_{i=1}^{d} \hat{R}_{Fill}(A_i(T_C^i)) + 4b_i\sqrt{\frac{2\log(4d/\delta)}{|S_i^C|}} + O\left(b_i\sqrt{\frac{r2^{k+1}\log(d\cdot 2^{k+1})\log(Br)}{|S_i^c|}}\right)$

Theorems 3 and 4 provide a bound on the noise introduced by constant and regression imputation. Theorem 3 reveals that noise risk stability is not sufficient as a condition for learnability under imputation if we restrict the imputation mechanism to be mean imputation, as this latter does not provide a consistent estimator for the real (unobserved) data, even in the best case. Indeed, in such a situation, one would need the stronger condition $\lim_{m\to\infty} \epsilon_{NRS}(m, \epsilon, \delta) = 0$, irrespective of ϵ, because the small noise constant ϵ will be lower bounded by the trace of the covariance matrix. This fact is a manifestation of the issue we mentioned in the previous section wherein for a fixed class of imputation mechanisms (in this case, mean imputation), it may not be possible to have a vanishing generalization gap, even for a class of models that is learnable under imputation. At the same time, if the conditions for Eq. (7) are satisfied, and using gradient descent as a learning algorithm, learning is possible even while using such an inconsistent imputation mechanism. This allows deriving the following:

Corollary 4. *Let FV_v be defined as in Theorem 3. Let $\mathcal{H}, X \subseteq K$, with K being a Reproducing Kernel Hilbert Space with $\|\cdot\|_K$ being the corresponding norm, that is H-bounded on $\mathcal{H}$ and B-bounded on X. Assume that $l(x, y, h) = g(\langle h, \phi(x)\rangle_K) + \frac{\lambda(m)}{2}\|h\|_K^2$, with g differentiable, M-smooth, b-bounded and satisfying the assumptions required for Eq. (6) to hold. Let $A = GD$ be gradient descent with step-size as defined in Theorem 2 and executed for T steps. Then, with probability larger than $1-\delta$ over the sampling of $S \sim \mathcal{D}^m$ and $T = \{(Corrupt_S(x, y), y) : (x, y) \in S\}$, and requiring $\lambda(m) \leq \sqrt{m}$ and increasing in m, it holds that $R(GD(T)) - \hat{R}_T(GD(T))$ goes to 0 as $m \to \infty$, and is upper bounded by* $\frac{2B\max\{HL, \lambda(m)H^2\}}{\sqrt{\min_i |S_i^C|}} + 4(b + \frac{\lambda(m)}{2}H^2)\sqrt{\frac{2\log(8/\delta)}{\min_i |S_i^C|}} + (b + \frac{\lambda(m)}{2}H^2)\left(1 - \frac{1}{2(B^2M+\lambda(m))}\right)^T + \frac{HL(Tr(\Sigma)+\mathbb{E}[\|v(T)-\mu\|_X])+2BHL\sqrt{\frac{\log(4/\delta)}{2m}}+\frac{\delta}{2}BHL}{\lambda(m)}$.

In particular, the previous bound holds for (kernel) ridge regression (i.e., $g(z) = (y-z)^2$) with $M = 2$, $L = 2BH$, and for (kernel) shrinked logistic regression (i.e., $g(z) = \log(1 + e^{-yz})$) with $M \leq \frac{1}{4}$, $L = 1$.

Corollary 4 is related to and, in some sense, generalizes the main result in [5]. Indeed, while the results in [5] are only applicable to linear ridge regression in the realizable setting, using naive imputation (i.e., constant imputation with 0), and assuming that $\mathcal{M}$ satisfies $\forall i, \mathcal{M}_i(\perp|(x, y)) = c$, in contrast, Corollary 4 can be applied to a larger class of regularized linear-in-parameters models, any form of constant imputation, no assumptions on the missingness mechanism and also in the agnostic setting. On the other hand, the results in [5] provide tighter finite-sample bounds for learnability under imputation of linear ridge

regression. In this sense, the two results provide a complementary perspective on the generalization properties of linear models in learning under imputation: we leave to future work the study of additional conditions under which tighter bounds in the style of [5] for the class of models discussed in Corollary 4.

Even though the previous Corollary illustrates that learnability under imputation is still possible even when considering only inconsistent imputation mechanisms, in general the convergence of the generalization gap could be much slower than optimal. Indeed, as previously shown in [25], one can expect regression imputation to provide better generalization guarantees in general scenarios. For example, as a consequence of Theorem 4, whenever the features are not uncorrelated and the regression models are consistent, the small noise constant $\epsilon_{Corrupt}$ for regression imputation vanishes as the sample size grows to infinity. This holds, in particular, for regression models that are universal approximators (e.g., kernel regression model [32], studied in Corollary 2, or also model ensembles [11]), studied in Corollary 3). We note, however, that the assumptions mentioned in Theorem 4 are rather strong, as they imply that features are not just merely strongly correlated under the data generating distribution $\mathcal{D}$, but are so also under the perturbed distribution determined by the missingness mechanism $\mathcal{M}$ and the Fill function. In practice, Fill may itself depend on the corrupted sampled dataset T [7, 27, 43], making the study of correlation particularly hard. Therefore, precisely characterizing the conditions under which regression imputation provides an asymptotically optimal imputation strategy, remains an open question. We leave to future work the study of more general conditions under which the noise introduced by regression imputation vanishes, as well as the study of other imputation strategies and their properties.

6 Illustrative Experiments

In this section we illustrate our results by means of two simple experiments. In the first example we demonstrate the application of Theorem 2, for both mean and regression imputation, in the simplified (but practically relevant) setting of (regularized) linear regression (as previously studied in [4]). In the second example, by contrast, we demonstrate the application of Theorems 3 and 4.

For the first experiment, we first generated a 10-dimensional vector w. Then, given a sample size m, we generated a dataset X of dimensionality 10, which was subsequently split into a training set T and a validation set V, with sizes, respectively, equal to $0.8m$ and $0.2m$. The target variable was generated as $y = Xw$. Data were generated uniformly at random between 0 and 1, and then instances were normalized[9]. 20% of the entries, randomly selected, in both X and V were set to missing. We then applied, respectively, mean imputation and regression imputation (using Ridge regression as the regression model): by construction, $\epsilon_{Corrupt} \sim \frac{1}{2}$, regardless of the imputation mechanism. Then, we fitted

[9] By construction, he covariance matrix Σ is the identity matrix.

a ridge regression model[10], using regularization coefficient $\lambda = \log(|T|)$, on the imputed training set, reconstructing a regularized linear model w'. In the experiment, we varied the sample size m from 100 to 10000 and assessed the error of the reconstructed model w' in terms of the l_2 loss on the validation set: i.e., $\frac{1}{|V|}\sum_{(x,y)\in V}(y - \langle w', x\rangle)^2$. The empirical error was compared with the theoretical estimates given by Theorem 2, Eq. (7). To account for uncertainty due to randomization, the above simulations were repeated 10 times, and we report the average error and 95% confidence intervals resulting from the experiments.

The results of the first experiment are illustrated in Figure 1. First of all, we note that there were no significant differences between mean and regression imputation. This result is not surprising: indeed, by construction, regression imputation cannot be expected to outperform mean imputation, as no information in the features can reliably be used to predict the missing data[11]. More generally, in both cases, the results provide a confirmation of Theorem 2 since, in all cases, the empirical error is upper bounded by the generalization curve predicted by the proven results. In general, the theoretical bounds become tighter with increasing sample size: indeed, as shown in Figure 1, the theoretical generalization curve rapidly (in fact, exponentially) decreases, approaching the observed empirical error. At the same time, the results of the experiment show that the proven bounds are not the tightest possible (at least in the simple setting of independent features and missing completely at random data, as assumed in the experiments), and hence finding tighter bounds (as well as matching lower bounds) could be of significant practical interest.

As for the second experiment, given a sample size m, we generated a dataset of dimensionality 10, which was subsequently split into a training set T and a validation set V, with sizes, respectively, equal to $0.8m$ and $0.2m$. The dataset was generated according to a random probabilistic graphical model [26]: first, we selected, uniformly at random between 1 and 10, a number of features (root features) to be generated uniformly at random; each subsequent feature f was generated by first randomly selecting a subset of the already generated features (the parents of f) and then defining f as a random linear combination of its parents. The instances in the generated dataset were then normalized. For the root features, 20% of the entries, randomly selected, were set to missing. By contrast, for each other feature f, we generated a random logistic regression model based on the parents of f, setting to missing values for which the logistic regression model predicted a target smaller than 0.5. We then applied, respectively, mean imputation and regression imputation (using ridge regression as the regression model), fitting the model on the training set and then evaluating imputation error on the validation set. These empirical error estimates were compared with the theoretical upper bounds given by Theorems 3 and 4. To account for uncertainty

[10] We note that ridge regression with the above regularization setting defines a $||w'||-$Lipschitz learning problem satisfying the additional assumptions for Eq. (7), with $M_g = M_r = 2$, $b = 1$, $B = 1$, and $\mu(m) = 2\log(m)$.

[11] Indeed, the data-generating distribution adopted in the experiment ensures independence of the features and missing completely at random data [30]

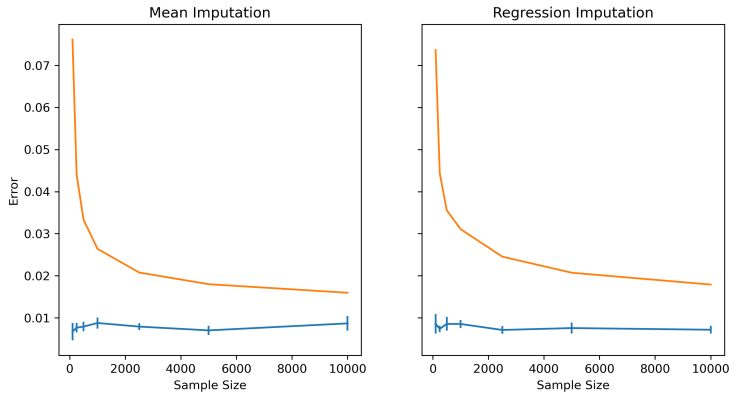

Fig. 1. Results of the experiment illustrating Theorem 2. The blue curve represents the empirical error, the orange one the theoretical bound given by Eq. (7).

due to randomization, the simulations were repeated 10 times, and we report the average error and 95% confidence intervals resulting from the experiments.

The results of the experiment are illustrated in Figure 2. In contrast with the previously described experiment, regression imputation reported on average a smaller imputation error: in several cases, the observed differences were significant. This result shows that, in general, regression imputation can be more effective than mean imputation (more generally, constant imputation) whenever the data is not missing completely at random[12]. More in general, we observe that the theoretical upper bound given by Theorem 3 provides a tight approximation to the expected imputation error of constant imputation, as the predicted value was almost always in the 95% confidence interval for the actual imputation error (only for $m = 2500$ the theoretical guarantee strictly upper bounded the imputation error): this shows that, in general, the given result for constant imputation may be tight. By contrast, the guarantee in Theorem 4 always strictly upper bounded the imputation error of regression imputation: while the deviation between theoretical and actual imputation error rapidly decreased with increasing sample size, the upper approximation is not the tightest possible, showing that the bound in Theorem 4 could be further improved.

7 Conclusion

In this article, we studied the interplay between learnability and imputation within the framework of statistical learning theory. In particular, we provided a necessary and sufficient condition for learnability under imputation, and showed that a large class of ML models satisfies this condition. Finally, we studied the

[12] By construction, the adopted experimental design ensures that the data is missing at random [30].

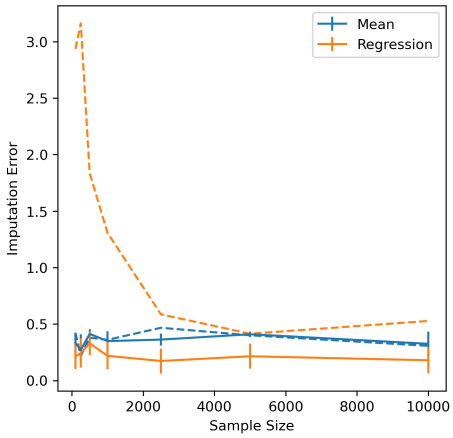

Fig. 2. Results of the experiment illustrating Theorems 3 and 4. For each imputation method, the solid line represents the empirical imputation error, while the dashed one represents the theoretical estimate.

behaviour of two commonly employed imputation strategies. Taken together, our results establish finite-sample guarantees and computable generalization certificates for learning with missing data. Our results set the stage for further formal exploration of missing data: to this aim, we discussed several open questions as well as highlighted relevant connections with other ML settings.

Code and Proof Availability

Proofs of all the results, as well as the code employed in the illustrative experiments, are available on GitHub at the following link: `https://github.com/AndreaCampagner/missing_ecml`.

References

1. Ahmad, A.F., Sayeed, M.S., Alshammari, K., Ahmed, I.: Impact of missing values in machine learning: A comprehensive analysis. arXiv preprint arXiv:2410.08295 (2024)
2. Ajalloeian, A., Stich, S.: Analysis of sgd with biased gradient estimators. arXiv preprint arXiv:2008.00051 (2020)
3. Alon, N., Ben-David, S., Cesa-Bianchi, N., Haussler, D.: Scale-sensitive dimensions, uniform convergence, and learnability. Journal of the ACM **44**(4), 615–631 (1997)
4. Ayme, A., Boyer, C., Dieuleveut, A., Scornet, E.: Near-optimal rate of consistency for linear models with missing values. In: International Conference on Machine Learning. pp. 1211–1243. PMLR (2022)
5. Ayme, A., Boyer, C., Dieuleveut, A., Scornet, E.: Naive imputation implicitly regularizes high-dimensional linear models. In: International Conference on Machine Learning. pp. 1320–1340. PMLR (2023)

6. Ayme, A., Boyer, C., Dieuleveut, A., Scornet, E.: Random features models: a way to study the success of naive imputation. In: International Conference on Machine Learning. pp. 2108–2134. PMLR (2024)
7. Azur, M.J., Stuart, E.A., Frangakis, C., Leaf, P.J.: Multiple imputation by chained equations: what is it and how does it work? International journal of methods in psychiatric research **20**(1), 40–49 (2011)
8. Bertsimas, D., Delarue, A., Pauphilet, J.: Simple imputation rules for prediction with missing data: Theoretical guarantees vs. empirical performance. Transactions on Machine Learning Research (2024)
9. Bhaskara, A., Cutkosky, A., Kumar, R., Purohit, M.: Descent with misaligned gradients and applications to hidden convexity. In: The Thirteenth International Conference on Learning Representations (2025)
10. Bhavsar, N., Prashanth, L.: Nonasymptotic bounds for stochastic optimization with biased noisy gradient oracles. IEEE Transactions on Automatic Control **68**(3), 1628–1641 (2022)
11. Biau, G., Devroye, L., Lugosi, G.: Consistency of random forests and other averaging classifiers. Journal of Machine Learning Research **9**(9) (2008)
12. Blumer, A., Ehrenfeucht, A., Haussler, D., Warmuth, M.: Learnability and the vapnik-chervonenkis dimension. Journal of the ACM **36**(4), 929–965 (1989)
13. Bouthinon, D., Soldano, H.: Learning first order rules from ambiguous examples. In: 2014 IEEE 26th International Conference on Tools with Artificial Intelligence. pp. 39–46. IEEE (2014)
14. Bouthinon, D., Soldano, H., Ventos, V.: Concept learning from (very) ambiguous examples. In: Machine Learning and Data Mining in Pattern Recognition: 6th International Conference, MLDM 2009, Leipzig, Germany, July 23-25, 2009. Proceedings 6. pp. 465–478. Springer (2009)
15. Campagner, A.: Credal learning: Weakly supervised learning from credal sets. In: ECAI 2023, pp. 327–334. IOS Press (2023)
16. Caprio, M., Sultana, M., Elia, E., Cuzzolin, F.: Credal learning theory. arXiv preprint arXiv:2402.00957 (2024)
17. Chen, J., Luss, R.: Stochastic gradient descent with biased but consistent gradient estimators. arXiv preprint arXiv:1807.11880 (2018)
18. Cismondi, F., Fialho, A.S., Vieira, S.M., Reti, S.R., Sousa, J.M., Finkelstein, S.N.: Missing data in medical databases: Impute, delete or classify? Artificial intelligence in medicine **58**(1), 63–72 (2013)
19. Daniely, A., Shalev-Shwartz, S.: Optimal learners for multiclass problems. In: Conference on Learning Theory. pp. 287–316. PMLR (2014)
20. d'Aspremont, A.: Smooth optimization with approximate gradient. SIAM Journal on Optimization **19**(3), 1171–1183 (2008)
21. Enders, C.K.: Applied missing data analysis. Guilford Publications (2022)
22. Goldman, S.A., Kwek, S.S., Scott, S.D.: Learning from examples with unspecified attribute values. Information and Computation **180**(2), 82–100 (2003)
23. Hu, X., Prashanth, L., György, A., Szepesvari, C.: (bandit) convex optimization with biased noisy gradient oracles. In: Artificial Intelligence and Statistics. pp. 819–828. PMLR (2016)
24. Jäger, S., Allhorn, A., Bießmann, F.: A benchmark for data imputation methods. Frontiers in big Data **4**, 693674 (2021)
25. Josse, J., Chen, J.M., Prost, N., Varoquaux, G., Scornet, E.: On the consistency of supervised learning with missing values. Statistical Papers **65**(9), 5447–5479 (2024)

26. Koller, D., Friedman, N.: Probabilistic graphical models: principles and techniques. MIT press (2009)
27. Le Morvan, M., Josse, J., Scornet, E., Varoquaux, G.: What'sa good imputation to predict with missing values? Advances in Neural Information Processing Systems **34**, 11530–11540 (2021)
28. Le Morvan, M., Prost, N., Josse, J., Scornet, E., 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)
29. Lenz, O.U., Peralta, D., Cornelis, C.: No imputation without representation. In: Benelux Conference on Artificial Intelligence. pp. 1–21. Springer (2023)
30. Little, R.J., Rubin, D.B.: Statistical analysis with missing data. John Wiley & Sons (2019)
31. Lodder, P., et al.: To impute or not impute: That's the question. Advising on research methods: Selected topics **2013** (2013)
32. Micchelli, C.A., Xu, Y., Zhang, H.: Universal kernels. Journal of Machine Learning Research **7**(12) (2006)
33. Michael, L.: Partial observability and learnability. Artificial Intelligence **174**(11), 639–669 (2010)
34. Nijman, S.W., Leeuwenberg, A., Beekers, I., Verkouter, I., Jacobs, J., Bots, M., Asselbergs, F., Moons, K.G., Debray, T.P.: Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review. Journal of clinical epidemiology **142**, 218–229 (2022)
35. Perez-Lebel, A., Varoquaux, G., Le Morvan, M., Josse, J., Poline, J.B.: Benchmarking missing-values approaches for predictive models on health databases. GigaScience **11**, giac013 (2022)
36. Redko, I., Morvant, E., Habrard, A., Sebban, M., Bennani, Y.: Advances in domain adaptation theory. Elsevier (2019)
37. Reyero Lobo, A.D., Ayme, A., Boyer, C., Scornet, E.: Harnessing pattern-by-pattern linear classifiers for prediction with missing data. arXiv preprint arXiv:2405-091096 (2024)
38. Rubin, D.B.: Multiple imputation for nonresponse in surveys. Wiley (1987)
39. Schuurmans, D., Greiner, R.: Learning to classify incomplete examples. Computational Learning Theory and Natural Learning Systems: Addressing Real World Tasks pp. 87–105 (1995)
40. Sell, T., Berrett, T.B., Cannings, T.I.: Nonparametric classification with missing data. The Annals of Statistics **52**(3), 1178 – 1200 (2024)
41. Shalev-Shwartz, S., Shamir, O., Srebro, N., Sridharan, K.: Learnability, stability and uniform convergence. The Journal of Machine Learning Research **11**, 2635–2670 (2010)
42. Sridharan, K., Shalev-Shwartz, S., Srebro, N.: Fast rates for regularized objectives. Advances in neural information processing systems **21** (2008)
43. Stekhoven, D.J., Bühlmann, P.: Missforest—non-parametric missing value imputation for mixed-type data. Bioinformatics **28**(1), 112–118 (2012)
44. Van Buuren, S., Groothuis-Oudshoorn, K.: mice: Multivariate imputation by chained equations in r. Journal of statistical software **45**, 1–67 (2011)
45. Vapnik, V., Červonenkis, A.: On the uniform convergence of relative frequencies of events to their probabilities. In: Doklady Akademii Nauk USSR. vol. 181, pp. 781–787 (1968)