Random features models: a way to study the success of naive imputation

Alexis Ayme[1] **Claire Boyer**[1 2] **Aymeric Dieuleveut**[3] **Erwan Scornet**[1]

Abstract

Constant (naive) imputation is still widely used in practice as this is a first easy-to-use technique to deal with missing data. Yet, this simple method could be expected to induce a large bias for prediction purposes, as the imputed input may strongly differ from the true underlying data. However, recent works suggest that this bias is low in the context of high-dimensional linear predictors when data is supposed to be missing completely at random (MCAR). This paper completes the picture for linear predictors by confirming the intuition that the bias is negligible and that surprisingly naive imputation also remains relevant in very low dimension. To this aim, we consider a unique underlying random features model, which offers a rigorous framework for studying predictive performances, whilst the dimension of the observed features varies. Building on these theoretical results, we establish finite-sample bounds on stochastic gradient (SGD) predictors applied to zero-imputed data, a strategy particularly well suited for large-scale learning. If the MCAR assumption appears to be strong, we show that similar favorable behaviors occur for more complex missing data scenarios.

1. Introduction

Missing data appear in most real-world datasets as they arise from merging different data sources, data collecting issues, self-censorship in surveys, just to name a few. Specific handling techniques are required, as most machine learning algorithms do not natively handle missing values, with the notable exception of tree-based methods Stekhoven & Bühlmann (2012); Chen & Guestrin (2016); Perez-Lebel et al. (2022). A common practice consists in imputing missing entries. The resulting complete dataset can then be analyzed using any machine learning algorithm.

While there exists a variety of imputation strategies (single, multiple, conditional, marginal imputation ; see, e.g., Bertsimas et al., 2018, for an overview), mean imputation is definitely one of the most common practices. Such a procedure has been largely criticized in the past as (single) mean imputation distorts data distributions by lowering variances, which can lead to inconsistent parameter estimation. Indeed, a large part of the literature on missing values focuses on inference in parametric models, such as linear (Little, 1992; Jones, 1996) or logistic models (Consentino & Claeskens, 2011; Jiang et al., 2020). From an empirical perspective, benchmarks of imputation techniques (Woźnica & Biecek, 2020) indicate that simple imputation, such as the mean, induces reasonable predictive performances, compared to more complex imputation techniques such as MICE (Perez-Lebel et al., 2022).

On the contrary, a recent line of work (Josse et al., 2024) aims at studying the predictive performances of impute-then-regress strategies that work by first imputing data (possibly with a very simple procedure) and then fitting a learning algorithm on the imputed dataset. Whereas mean imputation leads to inconsistent *parameter estimation*, Josse et al. (2024) and Bertsimas et al. (2021) show that impute-then-regress procedures happen to be consistent if the learning algorithm is universally consistent. Le Morvan et al. (2021) generalize the consistency results of mean-imputation by Josse et al. (2024); Bertsimas et al. (2021) and prove that for any universally consistent regression model, almost all single imputation strategies can lead to consistent predictors. Therefore, the impact of a specific imputation strategy has to be analyzed for specific regression models.

Without dispute, linear models are the most classic regression framework. However, their study becomes challenging in presence of missing values as they can require to build 2^d non-linear regression models (one for each missing data pattern), where d is the number of input variables (Le Morvan et al., 2020; Ayme et al., 2022). In the context of linear models with missing inputs, Le Morvan et al. (2020) establish finite-sample generalization bounds for zero-imputation, showing that this strategy is generally inconsistent. How-

[1] Sorbonne Université, CNRS, Laboratoire de Probabilités, Statistique et Modélisation (LPSM), F-75005 Paris, France [2] Institut Universitaire de France (IUF) [3] CMAP, UMR7641, Ecole Polytechnique, IP Paris, 91128 Palaiseau, France. Correspondence to: Alexis Ayme <alexis.ayme@sorbonne-universite.fr>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

ever, assuming a low-rank structure on the input variables, Ayme et al. (2023) prove that zero-imputation prior to learning is consistent in a high-dimensional setting. Note that the impact of zero-imputation with low-rank inputs has also been analyzed by Agarwal et al. (2019) in the context of Principal Components Regression, where the same type of generalization bounds were established. In this paper, we want to go beyond the low-rank structure by considering a (possibly infinite) latent space, and using the random feature framework.

Related work - Random features First introduced by Rahimi & Recht (2007), random features are used in nonparametric regression to approximate, with a few features, a kernel method where the final predictor belongs to an infinite-dimensional RKHS. Rudi & Rosasco (2017); Carratino et al. (2018) obtain generalization upper bounds for kernel regression learned with a small number of features, leading to computational efficiency. Random features are also used to describe a one-hidden-layer neural network (Bach, 2017).

Related work - high-dimensional linear models Linear models have been widely studied in a fixed design, considering the input variables are fixed (see, e.g., Hastie et al., 2015, for an analysis in the high-dimensional case). Quite notably, few works analyze (high-dimensional) linear models in the random design setting, a necessary framework to assess the predictive performance of linear models on unseen data (Caponnetto & De Vito, 2007; Hsu et al., 2012; Mourtada & Rosasco, 2022). These works mainly focus on the statistical properties of the Empirical Risk Minimizer (ERM) with a ridge regularization using uniform concentration bounds. On the other hand, (Bach & Moulines, 2013; Raskutti et al., 2014; Dieuleveut et al., 2017) directly control the generalization error of predictors resulting of stochastic gradient strategies, while performing a single pass on the dataset. The obtained bounds have therefore the advantage of being dependent on the training algorithm involved.

Contributions In this paper, we analyze the impact of the classic imputation-by-zero procedure on predictive performances, as a function of the input dimension. To this aim, we consider a latent space from which an arbitrary number of input variables are built. The output variable is assumed to result from a linear transformation of the latent variable. Such a framework allows us to analyze how predictive performances vary with the number of input variables, inside a common fixed model. Under this setting, we assume that all entries of input variables are observed with probability $\rho \in (0, 1)$, within a MCAR scenario, and study the performance of a linear model trained on imputed-by-zero data.

- We prove that when the input dimension d is negligible compared to that of the latent space p, the Bayes risk of the zero-imputation strategy is negligible compared to that induced by missing data themselves. Therefore, naive imputation is on par with best strategies.
- When $d \gg p$, both above errors vanish, which highlights that neither the presence of missing data or the naive imputation procedure hinders the predictive performances.
- From a learning perspective, we use Stochastic Gradient Descent to learn parameters on zero-imputed data. We provide finite-sample generalization bounds in different regimes, highlighting that the excess risk vanishes at $1/\sqrt{n}$ for very low dimensions ($d \ll p$) and high dimensions ($d > (1-\rho)\sqrt{n}/\rho$).
- Two different regimes arise from the finite dimension of the latent space. To move beyond this disjunctive scenario, we consider a latent space of infinite dimension and analyze predictors built on d zero-imputed input variables. We prove that the corresponding Bayes excess risk is controlled via the excess risk of a kernel ridge procedure, with a penalization constant depending on ρ and d. A finite-sample generalization bound on the SGD strategy applied on zero-imputed data is established and shows that zero-imputation is consistent in high-dimensional regimes ($d \gg \sqrt{n}$).
- The MCAR assumption considered throughout the paper, and often in the literature, can be attenuated at the cost of weaker theoretical results but which allows to show that naive imputation is relevant in high dimension even for non-trivial Missing Not At Random (MNAR) scenarios.

Notations. For $n \in \mathbb{N}$, we denote $[n] = \{1, \ldots, n\}$. We use $\lesssim$ to denote inequality up to a universal constant. We use i for observations, and j for features.

2. Imputation is adapted for very low and high-dimensional data

2.1. Setting

We adopt the classical regression framework in which we want to predict the value of an output random variable $Y \in \mathbb{R}$ given an input random variable $X \in \mathcal{X} = \mathbb{R}^d$ of dimension d. More precisely, our goal is to build a predictor $\hat{f} : \mathcal{X} \to \mathbb{R}$ that accurately estimates the regression function $f^\star$ (also called Bayes predictor) defined as a minimizer of the quadratic risk

$$R(f) := \mathbb{E}\left[(Y - f(X))^2\right], \tag{1}$$

over the class of measurable functions $f : \mathcal{X} \to \mathbb{R}$. When f is linear, we simply denote $R(\theta)$ the risk of the linear function parameterized by θ, i.e., such that for all $x \in \mathbb{R}^d$, $f(x) = x^\top \theta$.

Random features Real datasets are often characterized by high correlations between variables, or equivalently by a hidden low-rank structure (Johnstone, 2001; Udell & Townsend, 2019). The random feature framework (Rahimi & Recht, 2007) constitutes a general and flexible approach for modeling such datasets. We, therefore, assume that the inputs $(X_i)_{i\in[n]}$, i.i.d. copies of X, actually result from a random feature (RF) model. For pedagogical purposes, we start by restricting ourselves to finite-dimensional latent models.

Assumption 1 (Gaussian random features)**.** *The input variables $(X_i)_{i\in[n]}$ are assumed to be given by*

$$X_{i,j} = Z_i^\top W_j, \qquad \textit{for } i \in [n] \textit{ and } j \in [d] \tag{2}$$

where the p-dimensional latent variables $Z_1, \dots, Z_n$ are i.i.d. copies of $Z \sim \mathcal{N}(0, I_p)$, and where the p-dimensional random weights $W_1, \dots, W_d$ are i.i.d. copies of W uniformly distributed on the sphere $\mathbb{S}^{p-1}$, i.e., $W \sim \mathcal{U}(\mathbb{S}^{p-1})$.

The latent space in Assumption 1 corresponds to $\mathbb{R}^p$. We have only access to n observations $(X_i)_{i\in[n]}$ of dimension d that can be seen as random projections using d directions of the latent features $(Z_i)_{i\in[n]}$. The total amount of information contained in the latent variables $(Z_i)_{i\in[n]}$ cannot be recovered in expectation by that contained in the observations $(X_i)_{i\in[n]}$ if $d < p$. This no longer holds when $d \gg p$, and the observations $(X_i)_{i\in[n]}$ can be therefore regarded as low-rank variables of rank p.

Linear models with RF In the following, we present the model governing the distribution of the output variable Y.

Assumption 2 (Latent linear well specified model)**.** *The target variable Y is assumed to follow a linear model w.r.t. the latent covariate Z, i.e.,*

$$Y = Z^\top \beta^\star + \epsilon, \tag{3}$$

where the model parameter is denoted by $\beta^\star \in \mathbb{R}^p$ and the noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is assumed to be independent of Z.

Considering such a random feature setting is particularly convenient when studying the influence of the input dimension d on learning without modifying the underlying model (indeed, the distribution of Y –and Z– does not depend on the input dimension d). Note that a similar model (with fixed weight $(W_j)_j$) has already been introduced in Hastie et al. (2022).

Missing data Often one does not have access to the full input vector X but rather to a version of X containing missing entries. On the contrary, the output Y is always assumed to be observed. To encode such missing information on X, we introduce, the random variable $P \in \{0,1\}^d$, referred to as the missing pattern (or actually an observation indicator), such that $P_j = 1$ if X_j is observed and $P_j = 0$ otherwise. Assuming that all variables are equally likely to be missing, we define $\rho := \mathbb{P}(P_j = 1)$ for any $j \in [d]$, i.e., $1 - \rho$ is the expected proportion of missing values for any feature. In this paper, we thoroughly analyze the classical Missing Completely At Random (MCAR) setting, where the missing pattern P and the complete observation (X, Y) are independent. Note that we will extend some of our theoretical findings for the MCAR case to relaxed missing scenarios in Section 4.

Assumption 3 (MCAR pattern with independent components)**.** *The complete observation (X, Y) and the missing pattern P are assumed to be independent, i.e., $(X, Y) \perp\!\!\!\perp P$ and such that P follows a Bernoulli distribution $\mathcal{B}(\rho)^{\otimes d}$, i.e., for any $j \in [d]$, $\rho = \mathbb{P}(P_j = 1)$, with $0 < \rho \leq 1$ denoting the expected proportion of observed values.*

Imputation Most machine learning algorithms are not designed to deal directly with missing data. Therefore, we choose to impute the missing values (both in the training and test sets) by zero (or by the mean for non centered inputs). The imputed inputs in the train and test sets are thus denoted, for all i, by

$$\tilde{X}_i = P_i \odot X_i, \tag{4}$$

where $\odot$ represented the component-wise product. The impact of missing data, and their handling by naive imputation, in this supervised learning task can be scrutinized through the evolution of the following key quantities:

- The Bayes risk based on complete random features (without missing entries):

  $$R^\star(d) := \inf_f \mathbb{E}\left[(Y - f(X))^2 | W_1, \dots, W_d\right],$$

  where the infimum is taken over all measurable functions.

- The Bayes risk given $X_{\text{miss}} = (\tilde{X}, P) \in \mathbb{R}^d \times \{0,1\}^d$:

  $$R^\star_{\text{miss}}(d) := \inf_f \mathbb{E}\left[(Y - f(X_{\text{miss}}))^2 | W_1, \dots, W_d\right],$$

  where the infimum is taken over all measurable functions. It has been shown to be attained for a pattern-by-pattern predictor (Le Morvan et al., 2020). The bias or deterministic error due to learning with missing inputs can be characterized as

  $$\Delta_{\text{miss}}(d) := \mathbb{E}\left[R^\star_{\text{miss}}(d) - R^\star(d)\right].$$

- The risk of the best linear predictor relying on zero-imputed inputs:

  $$R^\star_{\text{imp}}(d) := \inf_{\theta \in \mathbb{R}^d} \mathbb{E}\left[(Y - \tilde{X}^\top \theta)^2 | W_1, \dots, W_d\right].$$

The approximation error associated with this specific class of predictors, among those handling missing inputs, is denoted by

$$\Delta_{\mathrm{imp/miss}}(d) := \mathbb{E}\left[R^{\star}_{\mathrm{imp}}(d) - R^{\star}_{\mathrm{miss}}(d)\right].$$

Note that these three risks decrease with the dimension and are ordered as follows,

$$R^{\star}(d) \leq R^{\star}_{\mathrm{miss}}(d) \leq R^{\star}_{\mathrm{imp}}(d). \tag{5}$$

In what follows, we give a precise evaluation of $R^{\star}(d)$ and $R^{\star}_{\mathrm{mis}}(d)$ and provide bounds for $R^{\star}_{\mathrm{imp}}(d)$ as well.

2.2. Theoretical analysis

Our goal is to dissect the systematic errors introduced by either the occurrence of missing inputs or their handling via naive zero imputation. To do so, we start by characterizing the optimal risk over the class of linear predictors when working with complete inputs.

Proposition 2.1. *Under Assumptions 1 and 2, the Bayes risk for linear predictors based on complete random features is*

$$\mathbb{E}\left[R^{\star}(d)\right] = \begin{cases} \sigma^2 + \frac{p-d}{p}\|\beta^{\star}\|_2^2, & \textit{when } d < p, \\ \sigma^2 & \textit{when } d \geq p, \end{cases}$$

where the expectation is taken over $(W_j)_{j\in[d]}$.

Proposition 2.1 highlights that learning with a number d of random features larger than the latent dimension p is equivalent to learning directly with the latent covariate Z. Besides, when $d < p$, the Bayes predictor suffers from an increased risk, as learning is flawed by a lack of information in the (fully observed) inputs. This can be compensated by increasing the number d of random features, as the explained variance of Y, i.e., $\mathbb{E}Y^2 - \mathbb{E}R^{\star}(d) = \frac{d}{p}\|\beta^{\star}\|_2^2$, increases with d for $d \leq p$.

Proposition 2.2. *Under Assumptions 1 to 3, the Bayes risk for predictors working with missing data is given by*

$$\mathbb{E}\left[R^{\star}_{\mathrm{miss}}(d)\right] = \begin{cases} \sigma^2 + \frac{p-\rho d}{p}\|\beta^{\star}\|_2^2 & \textit{when } d < p, \\ \sigma^2 + \frac{\mathbb{E}[(p-B)\mathbb{1}_{B\leq p}]}{p}\|\beta^{\star}\|_2^2 & \textit{when } d \geq p, \end{cases}$$

where the expectation is taken over the random weights $(W_j)_{j\in[d]}$ and $B \sim \mathcal{B}(d,\rho)$ (Binomial law of parameters d and ρ) . Therefore,

$$\Delta_{\mathrm{miss}}(d) = \begin{cases} (1-\rho)\frac{d}{p}\|\beta^{\star}\|_2^2 & \textit{when } d < p, \\ \frac{\mathbb{E}[(p-B)\mathbb{1}_{B\leq p}]}{p}\|\beta^{\star}\|_2^2, & \textit{when } d \geq p. \end{cases}$$

To our knowledge, this result is the first one to precisely evaluate the error induced by missing inputs when learning a linear latent model. More specifically, two regimes are identified. In the first regime $d < p$, i.e., when working with random features of lower dimension than that of the latent model, $R^{\star}_{\mathrm{miss}}$ takes the same form as $R^{\star}$, where the input dimension d is replaced by ρd. This can be interpreted as the cost of learning with ρd observed features in expectation instead of the d initial features.

In the second regime, when $d \geq p$, the error due to missing data becomes more and more negligible as d increases, as the redundancy of the random feature model is sufficient to retrieve the information contained in the latent covariate of lower dimension p. Furthermore, if $d \geq (p+1)\frac{(1-\rho)e}{\rho}$, we can bound $\Delta_{\mathrm{miss}}(d)$ from above and below,

$$\begin{aligned} \frac{\rho}{2e}(1-\rho)^{d-1}\|\beta^{\star}\|_2^2 &\leq \Delta_{\mathrm{miss}}(d) \\ &\leq p\left(\frac{d\rho}{p(1-\rho)}\right)^p (1-\rho)^d\|\beta^{\star}\|_2^2, \end{aligned} \tag{6}$$

showing that $\Delta_{\mathrm{miss}}(d)$ decays exponentially fast with d in the high-dimensional regime: the impact of missing data on learning is therefore completely mitigated in high dimension.

Theorem 2.3. *Under Assumptions 1 to 3, the Bayes risk for predictors based on zero-imputed random features satisfies*

$$\begin{aligned} &\mathbb{E}\left[R^{\star}_{\mathrm{imp}}(d)\right] - \sigma^2 \\ &\leq \begin{cases} \inf\limits_{k\leq d}\left\{\frac{p-\rho k}{p} + \frac{(1-\rho)\rho(k-1)}{p-\rho(k-1)-2}\frac{k}{p}\right\}\|\beta^{\star}\|_2^2 & \textit{if } d < p, \\ \frac{p}{\rho d+(1-\rho)p}\|\beta^{\star}\|_2^2 & \textit{if } d \geq p. \end{cases} \end{aligned} \tag{7}$$

Thus, when $d < p$,

$$\begin{aligned} \Delta_{\mathrm{imp/miss}}(d) &\leq \frac{(1-\rho)\rho(d-1)}{p-\rho(d-1)-2}\frac{d}{p}\|\beta^{\star}\|_2^2 && (8) \\ &= \frac{\rho(d-1)}{p-\rho(d-1)-2}\Delta_{\mathrm{miss}}(d). && (9) \end{aligned}$$

And, when $d \geq p$,

$$\begin{aligned} \frac{(1-\rho)p}{\rho d+(1-\rho)p}\|\beta^{\star}\|_2^2 &\leq \Delta_{\mathrm{imp/miss}}(d) + \Delta_{\mathrm{miss}}(d) && (10) \\ &\leq \frac{p}{\rho d+(1-\rho)p}\|\beta^{\star}\|_2^2. && (11) \end{aligned}$$

Theorem 2.3 is the first result to provide a complete view of the impact of naive imputation on learning linear latent model. In particular, it sheds light on the following low-dimensional behavior. When $\rho d \ll p$, the error due to naive imputation appears to be negligible in comparison to the error $\Delta_{\mathrm{miss}}(d)$ due to missing data. Low-dimensional (missing) random features are unlikely to be strongly correlated, thus making imputation before training competitive (compared to the best predictor based on missing values). This is all the more true as the expected number of observed entries ρd is negligible compared to p.

In high dimensions where $d \gg p$, both errors $\Delta_{\text{imp/miss}}(d)$ and $\Delta_{\text{miss}}(d)$ vanish: neither the occurrence of missing data, nor their naive handling through imputation, hinder the learning task. This provides a refined and more rigorous analysis of this favorable behavior already identified in Ayme et al. (2023). Remark that, contrary to $\Delta_{\text{miss}}(d)$ (Equation (6)), $\Delta_{\text{imp/miss}}(d)$ does not seem to decrease exponentially as d increases, but only as $1/d$. Not that, according to (10), this rate is optimal.

Illustration We illustrate the bounds obtained in Proposition 2.1, Proposition 2.2 and Theorem 2.3 in Figure 1. In particular, we remark that the upper bounds (7) represented in (a) decrease with the number of features d but is loose for d close to and smaller than p. Indeed, for this regime, features are not enough isotropic to say that imputation by the mean (here by 0) is relevant, and not enough correlated to exploit shared information between features. Figure (b) illustrates the shift point in p for R_{mis} and $R_{\text{imp/mis}}$ and the difficulties to learn with missing values (imputed or not) around $d = p$.

2.3. Learning from imputed data with SGD

We leverage on our analysis of the Bayes risk of impute-then-regress strategy to propose a learning algorithm based on an SGD strategy. Our algorithm is computationally efficient, as it requires only one pass over the dataset, and shown to have theoretical guarantees. Note that due to missing data, the model becomes miss-specified (see Ayme et al., 2023), a challenging study case that can still be handled by SGD procedures.

SGD estimator. An averaged stochastic gradient descent (SGD) on the imputed dataset $(\tilde{X}_i)_{i\in[n]}$ is performed to directly minimize the theoretical risk $\theta \longmapsto R_{\text{imp}}(\theta)$ over $\mathbb{R}^d$. The algorithm starts from $\theta_0 = 0$ with a step-size $\gamma > 0$, then follows the recursion

$$\theta_t = \left[I - \gamma \tilde{X}_t \tilde{X}_t^\top\right] \theta_{t-1} + \gamma Y_t \tilde{X}_t, \tag{12}$$

and outputs after n iterations the Polyak-Ruppert averaging $\bar{\theta} = \frac{1}{n+1}\sum_{t=1}^{n} \theta_t$, used to estimate $\theta^\star_{\text{imp}}$. This algorithm (performing one pass over the training data) is optimal in terms of computational complexity. Note that the choice of the step size should depend intimately on the input dimension, with a slight variation, according to whether the setting is low or high-dimensional (see, Dieuleveut & Bach, 2016, for example).

Theorem 2.4. *Under Assumptions 1 to 3, for $d < p - 1$, the SGD recursion with Polyak-Ruppert averaging and step size $\gamma = \frac{1}{d}$ satisfy*

$$\mathbb{E}[R_{\text{imp}}(\bar{\theta}) - R^\star_{\text{miss}}(d)] \lesssim \frac{\rho(1-\rho)d(d-1)}{p(p-\rho(d-1)-2)}\|\beta^\star\|_2^2 + \frac{d}{\rho n}\frac{d}{(p-\rho(d-1)-2)}\|\beta^\star\|_2^2 + \rho\frac{d}{n}(\sigma^2 + \|\beta^\star\|_2^2). \tag{13}$$

For $d \geq p$, the choice $\gamma = \frac{1}{d\sqrt{n}}$ leads to

$$\mathbb{E}[R_{\text{imp}}(\bar{\theta}) - R^\star(d)] \lesssim \frac{1}{\rho}\frac{p}{d}\|\beta^\star\|_2^2 + \frac{p}{\rho^2(1-\rho)\sqrt{n}}\|\beta^\star\|_2^2 + \frac{\sigma^2 + \|\beta^\star\|_2^2}{\sqrt{n}}. \tag{14}$$

Regardless of the regime, these generalization upper bounds are composed of three terms. The first one encapsulates the (deterministic) error due to the imputation of missing values. The last two are stochastic errors inherent to learning, corresponding respectively to the initial condition and the variance of the SGD recursion.

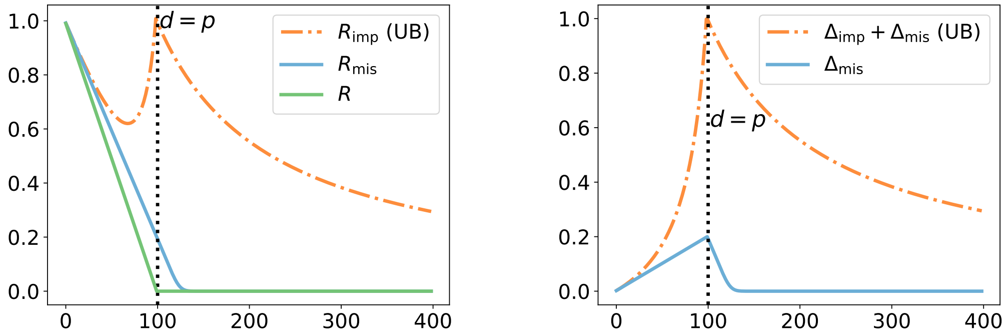
When $d < p - 1$, the learning error

$$\frac{d}{\rho n}\frac{d}{(p-\rho(d-1)-2)}\|\beta^\star\|_2^2 + \rho\frac{d}{n}(\sigma^2 + \|\beta^\star\|_2^2)$$

decreases fast with the number n of observation. However, the error due to the imputation of missing data $\rho(1-\rho)\frac{d(d-1)}{p(p-(d-1)-2)}\|\beta^\star\|_2^2$ becomes negligible only for extremely low-dimensional regimes ($d \ll p$) and remains significant when $d \lesssim p$. Therefore, for such a regime, even with a lot of observations, the imputation produces a large bias, and we recommend using other methods natively capable of handling missing values. In particular, tree-based methods (Stekhoven & Bühlmann, 2012; Chen & Guestrin, 2016) have demonstrated their effectiveness in addressing this specific regime.

In the regime $d \gg p$, the error $\frac{1}{\rho}\frac{p}{d}\|\beta^\star\|_2^2$ due to missing data and the zero-imputation procedure is low. Besides, the learning error $\frac{p}{\rho^2(1-\rho)\sqrt{n}}\|\beta^\star\|_2^2 + \frac{\sigma^2+\|\beta^\star\|_2^2}{\sqrt{n}}$ decreases at a (slow) rate $1/\sqrt{n}$. This slow rate of learning error is due to the fact that the covariance matrix of the imputed data $\Sigma_{\text{imp}} = \mathbb{E}[\tilde{X}\tilde{X}^\top]$ has a rank equal to d and eigenvalues lower bounded by $\rho(1-\rho)$. Hence imputed data are not of low rank, even for $d \gg p$. However, the upper bound (14) becomes dimension-free for the regime $d > \rho(1-\rho)\sqrt{n}$. In this case, the bias due to missing data and zero imputation is negligible compared to the learning error. This gives a clear practical recommendation: if the observed rate is such that $\rho(1-\rho) < d/\sqrt{n}$, then zero-imputation does not deteriorate the learning procedure for d large enough.

Overall, we have fully characterized the inherent error due to missing values in learning linear latent models, and proposed an efficient predictor based on SGD strategies. This



(a)

(b)

Figure 1. Evolution of different risks w.r.t. the number d of random features, with $p = 100$, $\|\beta^\star\| = 1$, $\rho = 0, 8$ and $\sigma = 0$.

section outlines that naive imputation thus remains a competitive and relevant technique not only for high-dimensional models but also for extremely low-dimensional ones, a favorable regime that was not identified so far in the literature. Note in passing that for the latter, the error bound (13) of the SGD predictor enjoys both a fast rate and a negligible approximation error, which can only be marginally improved. However, the analysis conducted so far relies on strong assumptions (finite-dimensional latent model with Gaussian random features and uniform weights, linear model). In the next section, we propose an extension of our high-dimensional results to a more general framework.

3. Extension to infinite-dimensional latent space

In this section, we analyze the influence of missing data in learning, when the random feature model involves an infinite-dimensional latent space.

3.1. The extended random feature framework

Consider a latent space $\mathcal{Z}$ (taking the place of $\mathbb{R}^p$), possibly of infinite dimension. We denote by $(Z_i)_{i\in[n]}$ i.i.d. latent variables, distributed as a generic random variable $Z \in \mathcal{Z}$. We only observe the variables $(X_i)_{i\in[n]}$, i.i.d. copies of $X \in \mathcal{X} = \mathbb{R}^d$, resulting from the following transformation of the latent variables.

Assumption 4 (General random features)**.** *The input variables are assumed to be given by*

$$X_{ij} = \psi(Z_i, W_j), \quad \textit{for all } i \in [n] \textit{ and } j \in [d], \tag{15}$$

where the weights $W_1, \dots, W_d \in \mathcal{W}$ are i.i.d. drawn according to a distribution ν, and where $\psi : \mathcal{Z} \times \mathcal{W} \to \mathbb{R}$. Furthermore, we assume that $\psi(Z_i, .) \in L^2(\nu)$, and there exists $L > 0$ such that $\mathbb{E}[\psi(Z, W)^2|W] \leq L^2$ almost surely.

Assumption 4 is an extension of Assumption 1 considering $\mathcal{Z} = \mathcal{W} = \mathbb{R}^p$, $\psi(Z_i, W_j) = Z_i^\top W_j$, ν the uniform distribution on $\mathbb{S}^{p-1}$ and $L^2 = 1$. But, such a setting of general random features encompasses many more scenarios, and has been extensively studied (Rahimi & Recht, 2007).

In this framework, we aim to study the linear prediction of an output Y given X, i.e., to build a prediction function of the form $g(X) = X^\top \theta$ with $\theta \in \mathbb{R}^d$. Note that this type of prediction can be also obtained as a function of the (latent) variable Z, indeed,

$$g(X) = \sum_{j=1}^{d} \theta_j \psi(Z, W_j) =: f(Z). \tag{16}$$

We can therefore define the corresponding class of functions with input space $\mathcal{Z}$ as

$$\mathcal{F}_\nu^{(d)} := \left\{ f : \mathcal{Z} \to \mathbb{R}, f(z) = \sum_{j=1}^{d} \theta_j \psi(z, W_j), \theta \in \mathbb{R}^d \right\}$$

Note that the class $\mathcal{F}_\nu^{(d)}$ is random because the weights $(W_j)_{j\in[m]}$ themselves are random. When the number d of random features tends to infinity, we can define the set $\mathcal{F}_\nu^{(\infty)}$ of functions which take the form:

$$f(Z) = \int \alpha_f(w) \psi(Z, w) d\nu(w), \tag{17}$$

for any $\alpha_f \in L^2(\nu)$. The associated norm is given by

$$\|f\|_\nu^2 := \inf_{\alpha \in \mathbb{L}_2(\nu)} \int |\alpha(w)|^2 \, d\nu(w)$$
$$\text{s.t.} \quad \forall z \in \mathcal{Z}, f(z) = \int \alpha(w) \psi(z, w) d\nu(w). \tag{18}$$

This norm corresponds to an RKHS norm, we refer the interested reader to Bach (2017) for further details. We denote by $R^\star(\infty)$, the risk of the best predictor belonging to the class $\mathcal{F}_\nu^{(\infty)}$, i.e., $R^\star(\infty) = \inf_{f\in\mathcal{F}_\nu^{(\infty)}} \mathbb{E}\left[(Y - f(Z))^2\right]$.

The setting considered in this section includes, for instance, Fourier random features (Rahimi & Recht, 2007; Rudi & Rosasco, 2017).

Example 3.1 (Fourier random features). Consider $Z \in \mathcal{Z}$ where $\mathcal{Z}$ is a compact subset of $\mathbb{R}^p$, $W = (A, B, C) \in \mathcal{W} = \mathbb{R}^p \times \mathbb{R} \times \{-1, 1\}$ and fix

$$\psi(Z, W) = \cos(A^\top Z + B) + 2C,$$

with $A \sim \mathcal{N}(0, I)$, $B \sim \mathcal{U}([0, 2\pi])$ and $C \sim \mathcal{U}(\{-1, 1\})$. Note that Assumption 4 holds here with $L^2 = 3$. The resulting function class $\mathcal{F}_\nu^{(\infty)}$ described by these random features is dense for $\|\cdot\|_\infty$ in the space of continuous functions.

As shown in this example, with a proper choice of ν and ψ, the class $\mathcal{F}_\nu^{(\infty)}$ can approximate any function that makes the following assumption feasible.

Assumption 5. *The Bayes predictor $f^\star(z) = \mathbb{E}[Y|Z = z]$ belongs to $\mathcal{F}_\nu^{(\infty)}$.*

Under Assumption 5, the model $Y = f^\star(Z) + \epsilon$, is well defined, i.e., $\mathbb{E}[\epsilon|Z] = 0$. Remark that Assumption 5 can be seen as a natural extension of linear model Assumption 2. Indeed, under Assumptions 1 and 2, $\mathcal{F}_\nu^\infty$ is the set of linear functions of Z.

3.2. Impact of missing data and imputation in the RF framework

The general random features $(X_i)_{i\in[n]}$ are assumed to be corrupted by MCAR entries, whether during training and test phases. Our goal is to study the quantity $R^\star_{\text{imp}}(d)$. Note that (5) can be rewritten here as

$$R^\star(\infty) \le R^\star(d) \le R^\star_{\text{miss}}(d) \le R^\star_{\text{imp}}(d).$$

Thus, introducing the quantity

$$\begin{aligned}\Delta^{(\infty)}_{\text{imp}}(d) &:= \mathbb{E}R^\star_{\text{imp}}(d) - R^\star(\infty)\\ &= \Delta_{\text{miss}}(d) + \Delta_{\text{imp/miss}}(d) + \mathbb{E}R^\star(d) - R^\star(\infty), \quad (19)\end{aligned}$$

we encapsulate (i) the error $\mathbb{E}R^\star(d) - R^\star(\infty)$ due to learning from a finite number of random features, (ii) the error $\Delta_{\text{miss}}(d)$ due to learning with missing inputs and, (iii) the approximation error $\Delta_{\text{imp/miss}}(d)$ due to the imputation by zero.

Theorem 3.2. *Under Assumptions 3 and 4,*

$$\Delta^{(\infty)}_{\text{imp}}(d) \le \inf_{f\in\mathcal{F}_\nu^{(\infty)}} \left\{R(f) - R^\star(\infty) + \frac{\lambda_{\text{imp}}}{d}\|f\|_\nu^2\right\},$$

with $\lambda_{\text{imp}} = \frac{L^2}{\rho}$. In particular, under Assumption 5,

$$\Delta^{(\infty)}_{\text{imp}}(d) \le \frac{\lambda_{\text{imp}}}{d}\|f^\star\|_\nu^2.$$

Theorem 3.2 provides an upper bound on $\Delta^{(\infty)}_{\text{imp}}(d)$ for general random feature models. In particular, the latter can be compared to a ridge bias when performing a kernel ridge regression in $\mathcal{F}_\nu^{(\infty)}$, and choosing the penalization strength of the order of λ_{imp}/d. Furthermore, under Assumption 5 (well-specified model), this bias converges to zero with a rate of $\|f^\star\|_\nu^2/(\rho d)$. By applying this result to a finite-dimensional latent model under Assumptions 1 and 2, and remarking that $\|f^\star\|_\nu^2 = p\|\beta^\star\|_2^2$, we recover the same rate $p/(\rho d)\|\beta^\star\|_2^2$ exhibited in Theorem 2.3. According to Theorem 2.3, this rate cannot be improved in general. More globally, missing data in RF models become harmless when learning with a large number of random features. It should be noted that when Assumption 5 does not hold anymore, one can still conclude that the bias $\Delta^{(\infty)}_{\text{imp}}$ tends to zero but at an arbitrarily slow rate. Regarding Assumption 4, it remains a mild requirement; in particular, it does not require centered inputs. This underlines that there is no need to impute by the mean to obtain $\Delta^{(\infty)}_{\text{imp}}$ converging to 0 in high-dimensional regimes.

3.3. SGD generalization upper bound

In this section, we assess the generalization performance of the SGD iterates when working with an underlying general random feature model.

Assumption 6. *There exists $\ell > 0$ such that, almost surely*

$$\mathbb{E}[\psi(Z, W)^2|W] \ge \ell^2.$$

This assumption holds when features are renormalized (i.e., when $\ell = L = 1$) or in the case of random Fourier features (see Example 3.1) with $\ell = 1$.

Assumption 7. *Assume that almost surely,*

$$|\psi(Z, W)|^2 \le \kappa L^2.$$

This assumption is satisfied in Example 3.1 with $\kappa = 2$ and $L^2 = 3$.

Theorem 3.3. *Under the general framework covered by Assumptions 4 to 7 with MCAR data (Assumption 3), the SGD recursion with Polyak-Ruppert averaging and step size $\gamma = 1/(\kappa d\sqrt{n})$ satisfy*

$$\begin{aligned}\mathbb{E}[R_{\text{imp}}(\bar\theta) - R^\star(\infty)] &\lesssim \frac{L^2}{\rho d}\|f^\star\|_\nu^2\\ &+ \frac{L^2}{\ell^2}\frac{L^2}{\rho^2(1-\rho)\sqrt{n}}\|f^\star\|_\nu^2 + \frac{\kappa L^2\mathbb{E}Y^2}{\sqrt{n}}. \quad (20)\end{aligned}$$

Theorem 3.3 outlines that, even for very general random features model (with possibly a latent space of infinite dimension), the impact of (i) the finite number of features, (ii) the missing data, and (iii) the imputation by 0, represented by the quantity $\frac{L^2}{\rho d}\|f^\star\|_\nu^2$, remains negligible in high dimension. Similarly to (14), the learning error $\frac{L^2}{\ell^2}\frac{L^2}{\rho^2(1-\rho)\sqrt{n}}\|f^\star\|_\nu^2 + \frac{\kappa L^2 \mathbb{E}Y^2}{\sqrt{n}}$ decreases with a slow rate. More precisely, when $d \gg \frac{L^2}{\ell^2}\rho(1-\rho)\sqrt{n}$, the upper bound is dimension free and the bias $\Delta_{\text{imp}}^{(\infty)}$ due to imputation becomes completely negligible. Note that, for renormalized features ($L^2 = l^2 = 1$), the transition from a low-dimensional regime to a high-dimensional one is given by $d = \rho(1-\rho)\sqrt{n}$ (as for Theorem 2.4), which is very easy to evaluate.

4. Beyond the MCAR assumption

To go beyond the MCAR missing data framework used in the previous section, we now consider missing not at random (MNAR) data, in which the missingness indicator of any variable can depend on the (possibly missing) value of the variable. In particular, we assume that the missing patterns $(P_i)_i$ depend on the latent features $(Z_i)_i$, which results in a MNAR scenario.

Assumption 8. *Suppose that P and Y are independent. Furthermore, consider that there exists an i.i.d. sequence $(\underline{W}'_j)_{j\in[d]}$ i.i.d. drawn according to some distribution μ supported on a set $\mathcal{W}'$ and assume that $P_1,\dots,P_d|Z,W'_1,\dots,W'_d$ are independent. We assume that the sequences $(\underline{W}'_j)_{j\in[d]}$ and $(W_j)_{j\in[d]}$ are independent and that*

$$\mathbb{P}\left(P_j|Z,\underline{W}'_j\right) = \phi(Z,\underline{W}'_j), \quad \textit{for all } j\in[d],$$

where $\phi : \mathcal{Z}\times\mathcal{W}' \to (0,1]$ is a continuous function.

The following result shows that the asymptotic property of $R^\star_{\text{imp}}$ in a MCAR setting (Theorem 2.3) remains valid in the MNAR setting of Assumption 8.

Theorem 4.1. *Under Assumption 8, consider one of the following settings:*

(i) *(finite-dimensional latent space) Under Assumptions 1 and 2, assume the distribution of the missing mechanism to be given for $\underline{W}' = (W'_0, W') \in \mathbb{R}\times\mathbb{R}^d$, by $\phi(Z,\underline{W}') = \Phi(Z^\top W' + W'_0)$ with Φ a Lipschitz function. Additionally, 0 is required to belong to the support of W'.*

(ii) *(general latent space) Under Assumptions 4 and 5, assume in addition that $\mathcal{Z}$ is compact, $f^\star$ continuous and $\mathcal{F}^{(\nu)}$ dense in the space of continuous functions equipped with the norm $\|\cdot\|_\infty$.*

Then, almost surely,

$$\lim_{d\to+\infty} R^\star_{\text{imp}}(d) = R^\star(\infty).$$

As a consequence,

$$\lim_d \Delta_{\text{imp}}^{(\infty)}(d) = \lim_d \Delta_{\text{miss}}(d) = \lim_d \Delta_{\text{miss/imp}}(d) = 0.$$

This result shows that the benign impact of missing data and imputation on predictive performances in high dimension holds true outside the MCAR assumption, even for missing scenarios (MNAR) often considered as more challenging. Let us consider two non-trivial examples.

Example 4.2 (Gaussian random features with logistic model). Consider the finite-dimensional latent model of Assumptions 1 and 2, where $\mathcal{Z} = \mathbb{R}^p$ and $\underline{W}'_j = (W'_{0j}, W'_j) \in \mathbb{R}\times\mathbb{R}^d$, and assume that the conditional distribution of the missing patterns P_j is given by

$$\mathbb{P}\left(P_j|Z,\underline{W}'_j\right) = \Phi(W'_{0j} + W_j'^\top Z) = \frac{1}{1+e^{W'_{0j}+W_j'^\top Z}}.$$

In this example, the features X_j are assumed to be missing according to a logistic model on the latent variables Z. In this setting, we can show that Theorem 4.1 (i) applies, since in particular, 0 belongs the support of W'_j. Note that, if $W'_j = 0$ almost surely then this model corresponds to a MCAR scenarios but with different proportion of missing values for each feature. The model is no longer MCAR as soon as random variable W'_j is not exactly equal to 0.

Example 4.3 (Fourier random features for any function ϕ). Let us consider the framework of Example 3.1 with a continuous function $f^\star$. Then Theorem 4.1 (ii) applies for any continuous function ϕ (in particular, we can consider the logistic model of Example 4.2 without any condition on W').

For these two MNAR examples,

$$\lim_d \Delta_{\text{imp}}^{(\infty)}(d) = \lim_d \Delta_{\text{miss}}(d) = \lim_d \Delta_{\text{miss/imp}}(d) = 0,$$

which means that missing values and imputations vanishes with the dimension.

5. Discussion and conclusion

Discussion on assumptions The random features model offers a way to capture independence in low dimensions and correlation in high dimensions. Specifically, mean imputation is the most effective approach when data are uncorrelated, which is typically more prevalent in low-dimensional settings than in high-dimensional ones. Such an independence assumption can be easily validated by examining the empirical covariance matrix. On the contrary, random feature models are more likely to hold in high-dimensional datasets, since strong correlations between input variables are often encountered in these settings. Verifying this assumption about dependence is made possible by observing the empirical covariance matrix. In very large dimensions, Principal Component Regression (PCR) can be employed as an automatic method to demonstrate that datasets are approximately of low ranks.

Insights for practitionners This article aims to show that employing a simple imputation technique (in this case, zero imputation) can result in a highly effective predictor in various regimes, as discussed in Section 2. By combining zero imputation with stochastic gradient descent (SGD) recursion, we obtain an optimal algorithm both in terms of generalization capacity and complexity, requiring just a single pass through the data. Therefore, we strongly recommend to evaluate this approach before resorting to more costly methods. This technique is not only computationally efficient, but also readily available in major software packages.

Summary of theoretical results Thanks to the rigorous framework of random features models, we prove that naive imputation is relevant both in high- and low-dimensional regimes. In particular, the bias induced by imputation is negligible compared to the one induced by missing data, therefore showing that zero-imputation strategies may lead to near-optimal predictors. Following this analysis, we prove that an SGD procedure trained on zero-imputed data reaches near-optimal rate of consistency in low-dimensional regimes, but still suffer from slow rates in high-dimensional ones. Obtaining fast rates for the latter setting is still an open and interesting question. Whilst our analysis extends beyond the MCAR scenario, rates of consistency for SGD procedures remain to be derived for such settings.

Impact Statement

This paper is mainly theoretical and presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Agarwal, A., Shah, D., Shen, D., and Song, D. On robustness of principal component regression. *Advances in Neural Information Processing Systems*, 32, 2019.

Ayme, A., Boyer, C., Dieuleveut, A., and Scornet, E. Near-optimal rate of consistency for linear models with missing values. In *International Conference on Machine Learning*, pp. 1211–1243. PMLR, 2022.

Ayme, A., Boyer, C., Dieuleveut, A., and Scornet, E. Naive imputation implicitly regularizes high-dimensional linear models. In *International Conference on Machine Learning*, 2023.

Bach, F. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681, 2017.

Bach, F. and Moulines, E. Non-strongly-convex smooth stochastic approximation with convergence rate o (1/n). *Advances in neural information processing systems*, 26, 2013.

Bertsimas, D., Pawlowski, C., and Zhuo, Y. D. From predictive methods to missing data imputation: an optimization approach. *Journal of Machine Learning Research*, 18 (196):1–39, 2018.

Bertsimas, D., Delarue, A., and Pauphilet, J. Beyond impute-then-regress: Adapting prediction to missing data. *arXiv preprint arXiv:2104.03158*, 2021.

Caponnetto, A. and De Vito, E. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

Carlen, E. Trace inequalities and quantum entropy: an introductory course. *Entropy and the quantum*, 529:73–140, 2010.

Carratino, L., Rudi, A., and Rosasco, L. Learning with sgd and random features. *Advances in neural information processing systems*, 31, 2018.

Chen, T. and Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

Consentino, F. and Claeskens, G. Missing covariates in logistic regression, estimation and distribution selection. *Statistical Modelling*, 11(2):159–183, 2011.

Cook, R. D. and Forzani, L. On the mean and variance of the generalized inverse of a singular wishart matrix. 2011.

Dieuleveut, A. and Bach, F. Nonparametric stochastic approximation with large step-sizes. 2016.

Dieuleveut, A., Flammarion, N., and Bach, F. Harder, better, faster, stronger convergence rates for least-squares regression. *The Journal of Machine Learning Research*, 18(1): 3520–3570, 2017.

Giraud, C. *Introduction to high-dimensional statistics*. CRC Press, 2021.

Hastie, T., Tibshirani, R., and Wainwright, M. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022.

Hsu, D., Kakade, S. M., and Zhang, T. Random design analysis of ridge regression. In *Conference on learning theory*, pp. 9–1. JMLR Workshop and Conference Proceedings, 2012.

Jiang, W., Josse, J., Lavielle, M., Group, T., et al. Logistic regression with missing covariates—parameter estimation, model selection and prediction within a joint-modeling framework. *Computational Statistics & Data Analysis*, 145:106907, 2020.

Johnstone, I. M. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, 29(2):295 – 327, 2001. doi: 10.1214/aos/1009210544. URL https://doi.org/10.1214/aos/1009210544.

Jones, M. P. Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American statistical association*, 91(433):222–230, 1996.

Josse, J., Prost, N., Scornet, E., and Varoquaux, G. On the consistency of supervised learning with missing values. *to appear in Statistical Papers*, 2024.

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.

Le Morvan, M., Josse, J., Scornet, E., and Varoquaux, G. What'sa good imputation to predict with missing values? *Advances in Neural Information Processing Systems*, 34:11530–11540, 2021.

Little, R. J. Regression with missing x's: a review. *Journal of the American statistical association*, 87(420):1227–1237, 1992.

Mourtada, J. and Rosasco, L. An elementary analysis of ridge regression with random design. *arXiv preprint arXiv:2203.08564*, 2022.

Page Jr, T. J. Multivariate statistics: A vector space approach. *Journal of Marketing Research*, 21(2):236–236, 1984.

Perez-Lebel, A., Varoquaux, G., Le Morvan, M., Josse, J., and Poline, J.-B. Benchmarking missing-values approaches for predictive models on health databases. *GigaScience*, 11:giac013, 2022.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.

Raskutti, G., Wainwright, M. J., and Yu, B. Early stopping and non-parametric regression: an optimal data-dependent stopping rule. *The Journal of Machine Learning Research*, 15(1):335–366, 2014.

Rudi, A. and Rosasco, L. Generalization properties of learning with random features. *Advances in neural information processing systems*, 30, 2017.

Stekhoven, D. J. and Bühlmann, P. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

Udell, M. and Townsend, A. Why are big data matrices approximately low rank? *SIAM Journal on Mathematics of Data Science*, 1(1):144–160, 2019.

Von Rosen, D. Moments for the inverted wishart distribution. *Scandinavian Journal of Statistics*, pp. 97–109, 1988.

Woźnica, K. and Biecek, P. Does imputation matter? benchmark for predictive models. *arXiv preprint arXiv:2007.02837*, 2020.

A. Notations

For two vectors (or matrices) a, b, we denote by $a \odot b$ the Hadamard product (or component-wise product). $[n] = \{1, 2, ..., n\}$. For two symmetric matrices A and B, $A \preceq B$ means that $B - A$ is positive semi-definite. The symbol $\lesssim$ denotes the inequality up to a universal constant. Table 1 summarizes the notations used throughout the paper and appendix.

Table 1. Notations

$\|u\|_M^2$	$u^\top M u$ the semi-norm induced by a positive matrix M
$\|M\|_{\mathrm{Fr}}^2$	The Frobenius norm of M
$\mathrm{Tr}(M)$	The sum of diagonal elements of M
$M + \lambda$	Abuse of notation for $M + \lambda I_p$
$M^\dagger$	The Moore-Penrose pseudoinverse of M
$\mathbb{S}_p$	The unit sphere of $\mathbb{R}^p$
$\mathrm{Span}(u_j, j \in [k])$	The linear span induced by $(u_j)_{j \in [k]}$
P	The mask
$\mathbf{W}$	$(W_1, \ldots, W_d)^\top$ the matrix of weights
$R(\theta)$	the risk of linear predictor θ on complete data
$R_{\mathrm{imp}}(\theta)$	the risk of linear predictor θ on imputed data
$\theta^\star$	Best linear predictor on complete data
$\theta^\star_{\mathrm{imp}}$	Best linear predictor on imputed data
Σ	$\mathbb{E}[XX^\top \| \mathbf{W}]$
λ_j	eigenvalues of Σ
u_j	eigendirections of Σ
ρ	Theoretical proportion of observed entries
$L^2(\nu)$	The set of two square ν integrable functions

B. Preliminary results - random matrices

We provide here a reminder on singular values decomposition and Moore-Penrose pseudoinverse. We can found these results and more on linear algebra in Giraud (2021, appendix).

Theorem B.1. *Any $n \times p$ real-valued matrix of rank r can be decomposed as*

$$A = \sum_{j=1}^{r} \sigma_j u_j v_j^\top,$$

where

- $\sigma_1 \geq \cdots \geq \sigma_r > 0$,
- $(\sigma_1, \ldots, \sigma_r)$ *are the nonzero eigenvalues of* $A^\top A$ *and* $AA^\top$*, and*
- $(u_1, \ldots, u_r)$ *and* $(v_1, \ldots, v_r)$ *are two orthonormal families of* $\mathbb{R}^n$ *and* $\mathbb{R}^p$*, such that* $AA^\top u_j = \sigma_j^2 u_j$ *and* $A^\top A v_j = \sigma_j^2 v_j$.

Furthermore, the Moore-Penrose pseudo inverse defined as

$$A^\dagger = \sum_{j=1}^{r} \sigma_j^{-1} v_j u_j^\top,$$

satisfied

1. $A^\dagger A$ *is the orthogonal projector on lines of* A*,*
2. $AA^\dagger$ *is the orthogonal projector on columns of* A*,*

3. $(AO)^\dagger = O^\top A^\dagger$ *for any orthogonal matrix* O.

For any function $f : \mathbb{R} \longmapsto \mathbb{R}$, and any positive matrix $A \in \mathbb{R}^{d\times d}$ with the following spectral decomposition $A = \sum_{j=1}^d \lambda_j v_j v_j^\top$, we denote by $f(A)$ the matrix corresponding to the spectral decomposition

$$f(A) := \sum_{j=1}^d f(\lambda_j) v_j v_j^\top.$$

Theorem B.2 (Jensen inequality for random positive matrix, see Theorem 2.10 in (Carlen, 2010))**.** *Let A be a random positive matrix. For all convex functions f, we have*

$$\mathrm{Tr}(f(\mathbb{E}A)) \le \mathbb{E}\mathrm{Tr}(f(A)).$$

Proposition B.3. *Let $A = \sum_{j=1}^d Z_j Z_j^\top$ with $Z_1, \dots, Z_d$ i.i.d. random vector of $\mathbb{R}^p$ with $\|Z_1\|_2^2 \le 1$ almost surely and $\mathbb{E}ZZ^\top = \alpha I_p$, then, for all $\lambda > 0$*

$$\frac{p}{d\alpha + \lambda} \le \mathbb{E}\mathrm{Tr}\left((A + \lambda I_p)^{-1}\right) \le (1 + 1/\lambda)\frac{p}{d\alpha + \lambda}. \tag{21}$$

Proof of Proposition B.3. This result is a direct application of Mourtada & Rosasco (2022, Lemma 2) considering $\hat{\Sigma} = \frac{1}{d}A$ □

Lemma B.4. *Let $M \in \mathbb{R}^{p\times p}$ be a random symmetric matrix, such that for all vectors $u, v \in \mathbb{S}^{p-1}$, $\mathrm{Law}(u^\top M u) = \mathrm{Law}(v^\top M v)$. Then, for all $\beta \in \mathbb{R}^p$,*

$$\mathbb{E}\left[\beta^\top M \beta\right] = \|\beta\|_2^2 \frac{\mathbb{E}\mathrm{Tr}(M)}{p}.$$

This is in particular satisfied if, for any orthogonal matrix O, $OMO^\top$ has the same law as M.

Proof. By assumption, for all $u, v \in \mathbb{S}^{d-1}$, $\mathbb{E}u^\top M u = \mathbb{E}v^\top M v$. Thus, there exists α such that, for all $v \in \mathbb{S}^d$, $v^\top \mathbb{E}M v = \mathbb{E}v^\top M v = \alpha$, which entails that $\mathbb{E}M = \alpha I$ by characterization of symmetric matrices. Therefore, $\mathbb{E}\mathrm{Tr}(M) = \mathrm{Tr}(\mathbb{E}M) = p\alpha$, and $\mathbb{E}M = \frac{\mathbb{E}\mathrm{Tr}(M)}{p} I$. Hence, for all $\beta \in \mathbb{R}^p$

$$\mathbb{E}\left[\beta^\top M \beta\right] = \beta^\top \mathbb{E}M \beta = \|\beta\|_2^2 \frac{\mathbb{E}\mathrm{Tr}(M)}{p}.$$

The last point easily follows, see for example Page Jr (1984, Proposition 2.14) for the case of invariant distributions by orthogonal transforms.

□

The following result is inspired by the result of Cook & Forzani (2011), that is an adaptation of that of Von Rosen (1988).

Lemma B.5. *For all $0 < d < p - 1$, let $\mathbf{W} \in \mathbb{R}^{p\times d}$ such that columns of $\mathbf{W}$ are i.i.d. and uniform over $\mathbb{S}^{p-1}$, then*

$$\mathbb{E}\|\mathbf{W}^\dagger\|_{\mathrm{Fr}}^2 = d\left(1 + \frac{d-1}{p-d-1}\right).$$

Proof. Up to a polar coordinate change of variable, one can show that the distribution of the columns of $\mathbf{W}$ corresponds to that of normalized Gaussian vectors, i.e., for all $j \in [d]$,

$$W_j = \frac{G_j}{\|G_j\|_2},$$

where (G_j) are i.i.d. of law $\mathcal{N}(0, I_p)$. Note that the columns of $\mathbf{W}$ are the rows of $\mathbf{M} = \mathbf{W}^\top$. As $d < p$,

$$\mathbf{M}\mathbf{M}^\dagger = I_d.$$

because the rows of $\mathbf{M}$ are almost surely linearly independent. For all $j \in [d]$, we let l_j be the j-th row of $\mathbf{M}$, and c_j the j-th column of $\mathbf{M}^\dagger$. Therefore, for all $k \neq j$, $l_k^\top c_j = 0$, then $c_j \in \mathrm{Span}(l_k, k \neq j)^\perp$. Note, that $\mathrm{Span}(l_j, j \in [d]) = \mathrm{Span}(c_j, j \in [d])$ by property of Moore-Penrose pseudoinverse (Theorem B.1). Thus, c_j as the form,

$$c_j = \theta_j P_j l_j,$$

where P_j is the orthogonal projection on $\mathrm{Span}(l_k, k \neq j)^\perp$. Besides, $l_j^\top c_j = 1$ gives us that

$$\theta_j = \frac{1}{\|P_j l_j\|_2^2}.$$

Thus,

$$\|M^\dagger\|_{\mathrm{Fr}}^2 = \sum_j \|c_j\|_2^2 = \sum_j \frac{1}{\|P_j l_j\|_2^2} \tag{22}$$

As $l_j = W_j = \frac{G_j}{\|G_j\|_2}$, we can write

$$\frac{1}{\|P_j l_j\|_2^2} = \frac{\|G_j\|_2^2}{\|P_j G_j\|_2^2}$$

Using that $\|G_j\|_2^2 = \|P_j G_j\|_2^2 + \|(I_p - P_j)G_j\|_2^2$, we have

$$\frac{1}{\|P_j l_j\|_2^2} = 1 + \frac{\|(I_p - P_j)G_j\|_2^2}{\|P_j G_j\|_2^2}.$$

Conditioning by (G_k) with $k \neq j$, and using Cochran theorem $(I_p - P_j)G_j | G_k, k \neq j$ and $P_j G_j | G_k, k \neq j$ are two independent standard normal vector of respective dimensions $p - (p - d + 1) = d - 1$ and $p - d + 1$. Thus,

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{\|P_j l_j\|_2^2} | G_k, k \neq j\right] &= 1 + \mathbb{E}\left[\|(I_p - P_j)G_j\|_2^2 | G_k, k \neq j\right] \mathbb{E}\left[\frac{1}{\|P_j G_j\|_2^2} | G_k, k \neq j\right] && (23)\\
&= 1 + \frac{d-1}{p-d-1}, && (24)
\end{aligned}$$

because $\mathbb{E}\left[\|(I_p - P_j)G_j\|_2^2 | G_k, k \neq j\right] = d - 1$ and $\mathbb{E}\left[\frac{1}{\|P_j G_j\|_2^2} | G_k, k \neq j\right] = \frac{1}{p-d-1}$ as the expectation of an inverse-chi-squared of parameter $p - d + 1 > 2$ (with $d < p - 1$). Then, taking the expectation of (22) leads to the result,

$$\mathbb{E}\|M^\dagger\|_{\mathrm{Fr}}^2 = d + \frac{d(d-1)}{p-d-1}.$$

□

Lemma B.6. *Let A, B, V three symetrics non-negative matrix, if $A \preceq B$ then $A \odot V \preceq B \odot V$.*

Proof. Let $X \sim \mathcal{N}(0, V)$ and $\theta \in \mathbb{R}^d$,

$$
\begin{aligned}
\|\theta\|^2_{A\odot V} &= \theta^\top A \odot V \theta \\
&= \theta^\top \left(\left(\mathbb{E} X X^\top \right) \odot A \right) \theta \\
&= \mathbb{E}\left[\theta^\top \left(\left(X X^\top \right) \odot A \right) \theta \right] \\
&= \mathbb{E}\left[\sum_{i,j} \theta_i \left(\left(X X^\top \right) \odot A \right)_{i,j} \theta_j \right] \\
&= \mathbb{E}\left[\sum_{i,j} \theta_i X_i X_j A_{i,j} \theta_j \right] \\
&= \mathbb{E}\left[\sum_{i,j} \left(\theta_i X_i \right) \left(\theta_j X_j \right) A_{i,j} \right] \\
&= \mathbb{E}\left[\|X \odot \theta\|^2_A \right] \\
&\leq \mathbb{E}\left[\|X \odot \theta\|^2_B \right] \\
&= \|\theta\|^2_{B\odot V}
\end{aligned}
$$

□

C. Proof of Section 2

The following result, established by Ayme et al. (2023), is used to derive an expression of $\Delta_{\text{missing}} + \Delta_{\text{imp/miss}}$.

Lemma C.1 (Proposition 3.1 of (Ayme et al., 2023))**.** *For all $\theta \in \mathbb{R}^d$,*

$$
R_{\text{imp}}(\theta) = R(\rho\theta) + \rho(1-\rho)\|\theta\|^2_{\text{diag}(\Sigma)}. \tag{25}
$$

Recalling that

$$
\Delta_{\text{miss}} + \Delta_{\text{imp/miss}} = \mathbb{E}\left[R^\star_{\text{imp}}(d) - R^\star(d) \right], \tag{26}
$$

we deduce from Lemma C.1 that

$$
\Delta_{\text{miss}} + \Delta_{\text{imp/miss}} = \mathbb{E} \inf_{\theta\in\mathbb{R}^d} \left\{ R(\theta) - R^\star(d) + \frac{1-\rho}{\rho} \|\theta\|^2_{\text{diag}(\Sigma)} \right\}. \tag{27}
$$

Additionally, when $\text{diag}(\Sigma) = I_p$ (in particular for model (1)), by optimization, we obtain,

$$
\Delta_{\text{miss}} + \Delta_{\text{imp/miss}} = \lambda \mathbb{E} \|\theta^\star\|^2_{\Sigma(\lambda I + \Sigma)^{-1}}, \tag{28}
$$

with $\lambda = \frac{1-\rho}{\rho}$.

Lemma C.2. *Under Assumption 3,*

$$
\mathbb{E} R^\star_{\text{miss}}(d) = \sum_{k=0}^{d} \mathbb{P}(B = k) \mathbb{E} R^\star(k),
$$

where B is a binomial random variable of parameters d and ρ.

Proof. Using the decomposition of the Bayes predictor from Le Morvan et al. (2020), we have

$$
R^\star_{\text{miss}}(d) = \sum_{m\in\{0,1\}^d} \mathbb{P}(P = m) R^\star_m, \tag{29}
$$

where

$$R_m^\star = \inf_f \mathbb{E}\left[(Y - f(X_{\mathrm{obs}(m)}))^2 | P = m, W_1, \ldots, W_d\right],$$

is the local Bayes risk given $(P = m)$. Using MCAR assumption (Assumption 3), and Gaussian assumption, according to Le Morvan et al. (2020), each local Bayes predictor are linear, thus

$$R_m^\star = \inf_\theta \mathbb{E}\left[(Y - \theta^\top X_{\mathrm{obs}(m)}))^2 | W_j, j \in \mathrm{obs}(m)\right].$$

As (W_j) are i.i.d. (and independent of Y), $R_m^\star$ has the same law as $R^\star(|m|)$ where $|m|$ is the number of observed components of m. Thus,

$$\mathbb{E}R_{\mathrm{miss}}^\star(d) = \sum_{m \in \{0,1\}^d} \mathbb{P}(P = m)\mathbb{E}R^\star(|m|)). \tag{30}$$

Grouping the missing patterns of the same size, we conclude that,

$$\mathbb{E}R_{\mathrm{miss}}^\star(d) = \sum_{k=0}^{d} \mathbb{P}(B = k)\mathbb{E}R^\star(k),$$

where B is a binomial law of parameters d and ρ.

□

C.1. Proof of Proposition 2.1

By definition,

$$\begin{aligned} R^\star(d) &= \mathbb{E}\left[(X^\top \theta^\star - Y)^2 | W_1, \ldots, W_d\right] \\ &= \mathbb{E}\left[(X^\top \theta^\star - Z^\top \beta^\star - \epsilon)^2 | W_1, \ldots, W_d\right] && \text{(using (2))} \\ &= \sigma^2 + \mathbb{E}\left[(X^\top \theta^\star - Z^\top \beta^\star)^2 | W_1, \ldots, W_d\right], \end{aligned}$$

using that ϵ is an independent noise of variance σ^2. We have $X^\top \theta^\star = Z^\top \sum_j \theta_j^\star W_j$. Then,

$$\begin{aligned} R^\star(d) &= \sigma^2 + \mathbb{E}\left[\left(\left(\sum_{j=1}^{d} \theta_j^\star W_j - \beta^\star)\right)^\top Z\right)^2 \middle| W_1, \ldots, W_d\right] \\ &= \sigma^2 + \left\|\beta^\star - \sum_{j=1}^{d} \theta_j^\star W_j\right\|_2^2, \end{aligned}$$

by isotropy of Z ($Z \sim \mathcal{N}(0, I_p)$ and thus $\mathbb{E}ZZ^\top = I$). Using that $\sum_{j=1}^d \theta_j^\star W_j$ belongs to $\mathrm{Span}(W_1, \ldots, W_d)$, we get that $\sum_{j=1}^d \theta_j^\star W_j = P_d \beta^\star$ where P_d is the orthogonal projection on $\mathrm{Span}(W_1, \ldots, W_d)$. Then,

$$R^\star(d) = \sigma^2 + \|(I - P_d)\beta^\star\|_2^2 = \sigma^2 + (\beta^\star)^\top (I - P_d)\beta^\star.$$

Remark that P_d is a random matrix (since $W_1, \ldots, W_d$ are random). Denoting by $\mathbf{W}$ the matrix admitting $W_1, \ldots, W_d$ as rows, the projection matrix can be rewritten as $P_d = \mathbf{W}^\dagger \mathbf{W}$. Thus, for all orthogonal matrix O, $OP_dO^\top = (\mathbf{W}O^\top)^\dagger \mathbf{W}O^\top$. The matrix $\mathbf{W}O^\top$ has rows $O^\top W_1, \ldots, O^\top W_d$, which is an i.i.d. sequence of random vectors on the unit sphere (since $O^\top$ is an orthogonal). Indeed, $\mathbf{W}O^\top$ and $\mathbf{W}$ have the same distribution, in consequence P and OP_dO have the same distribution too. Thus, by Lemma B.4, if $d < p$,

$$\begin{aligned} \mathbb{E}R^\star(d) &= \sigma^2 + \frac{1}{p}\|\beta^\star\|_2^2 \mathbb{E}\mathrm{Tr}(I - P_d) \\ &= \sigma^2 + \frac{p-d}{p}\|\beta^\star\|_2^2, \end{aligned}$$

using that $\mathrm{Tr}(I - P_d) = \mathrm{rank}(I - P_d) = p - d$. Besides, if $d \geq p$, $P_d = I_p$, and $\mathbb{E}R^\star(d) = \sigma^2$.

C.2. Proof of Proposition 2.2

Using Lemma C.2, we have

$$\mathbb{E}R^\star_{\text{miss}}(d) = \sum_{k=0}^{d} \mathbb{P}(B=k)\mathbb{E}R^\star(k),$$

where B is a binomial random variable of parameters d and ρ. Using Proposition 2.1, we have

$$\mathbb{E}\left[R^\star(k)\right] = \begin{cases} \sigma^2 + \frac{p-k}{p}\|\beta^\star\|_2^2, & \text{when } k<p, \\ \sigma^2 & \text{when } k \geq p. \end{cases}$$

Combining the two previous equalities, we obtain that

$$\mathbb{E}R^\star_{\text{miss}}(d) = \sigma^2 + \frac{\mathbb{E}[(p-B)\mathbb{1}_{B\leq p}]}{p}\|\beta^\star\|_2^2.$$

In the case where $d \leq p$, $\mathbb{1}_{B\leq p}=1$ almost surely, and we obtain

$$\mathbb{E}R^\star_{\text{miss}}(d) = \sigma^2 + \frac{p-\rho d}{p}\|\beta^\star\|_2^2.$$

C.3. Proof of Theorem 2.3

C.3.1. Preliminaries

In the rest of the proof, we denote by $\mathbf{W} = (W_1, \ldots, W_d)^\top \in \mathbb{R}^{d\times p}$ the weight matrix that admits the weight vectors $W_j \sim \mathcal{U}(\mathbb{S}^{p-1})$ for rows. We call $\Sigma = \mathbb{E}\left[XX^\top|\mathbf{W}\right]$ the covariance matrix of an input $X \in \mathbb{R}^d$ given the weight matrix $\mathbf{W}$. Recall that the latter, resulting from a random feature model, is such that $X = \mathbf{W}Z$, for $Z \in \mathbb{R}^p$ the corresponding latent vector.

Lemma C.3. *Under assumptions of Theorem 2.3,*

$$\Delta_{\text{imp/miss}} + \Delta_{\text{miss}} = \begin{cases} \frac{\lambda\|\beta^\star\|_2^2}{p}\mathbb{E}\text{Tr}\left((\Sigma+\lambda I_d)^{-1}\right) & \textit{if } d<p \\ \frac{\lambda\|\beta^\star\|_2^2}{p}\mathbb{E}\text{Tr}\left((\mathbf{W}^\top\mathbf{W}+\lambda I_p)^{-1}\right) & \textit{if } d\geq p, \end{cases}$$

with $\lambda = \frac{1-\rho}{\rho}$.

Proof. One has for $\theta \in \mathbb{R}^p$,

$$R(\theta) = \sigma^2 + \mathbb{E}\left[(Z^\top\beta^\star - \theta^\top X)^2|\mathbf{W}\right].$$

Using that $X = \mathbf{W}Z$, we have

$$\begin{aligned} R(\theta) &= \sigma^2 + \mathbb{E}\left[\left(\left(\beta^\star - \mathbf{W}^\top\theta\right)^\top Z\right)^2|\mathbf{W}\right] \\ &= \sigma^2 + \left\|\beta^\star - \mathbf{W}^\top\theta\right\|_2^2, \end{aligned}$$

by isotropy of Z. Since $\theta^\star$ minimizes the risk R, $\theta^\star$ minimizes the least-squares criterion above. Therefore, in the case where $p>d$ (the "design" $\mathbf{W}^\top$ being long), $\theta^\star$ is unique and given by $\theta^\star = (\mathbf{W}\mathbf{W}^\top)^{-1}\mathbf{W}\beta^\star$. In the case where $d<p$ (the design matrix $\mathbf{W}^\top$ being fat), there exists an infinite number of minimizers (all are solutions of the system $\beta^\star = \mathbf{W}^\top\theta$), but one can look at the solution of minimal ℓ^2-norm. Then, KKT conditions provide the particular solution $\theta^\star = \mathbf{W}(\mathbf{W}^\top\mathbf{W})^{-1}\beta^\star$. In both cases, $\theta^\star$ can be written in the following unified way:

$$\theta^\star = \left(\mathbf{W}^\top\right)^\dagger\beta^\star.$$

Futhermore,

$$\Sigma = \mathbb{E}\left[XX^\top|\mathbf{W}\right] = \mathbb{E}\left[\mathbf{W}Z(\mathbf{W}Z)^\top|\mathbf{W}\right] = \mathbb{E}\left[\mathbf{W}ZZ^\top\mathbf{W}^\top|\mathbf{W}\right] = \mathbf{W}\mathbf{W}^\top.$$

Then, using (28),

$$\begin{aligned}\Delta_{\mathrm{imp/miss}} + \Delta_{\mathrm{miss}} &= \lambda \mathbb{E} \left\| \theta^\star \right\|_{\Sigma(\Sigma+\lambda I)^{-1}} \\ &= \lambda \mathbb{E} \left\| \left(\mathbf{W}^\top\right)^\dagger \beta^\star \right\|^2_{\mathbf{W}\mathbf{W}^\top(\mathbf{W}\mathbf{W}^\top+\lambda I)^{-1}} \\ &= \lambda \mathbb{E} \left\| \beta^\star \right\|^2_{\mathbf{W}^\dagger\mathbf{W}\mathbf{W}^\top(\mathbf{W}\mathbf{W}^\top+\lambda I)^{-1}(\mathbf{W}^\top)^\dagger} \\ &= \frac{\lambda \left\| \beta^\star \right\|_2^2}{p} \mathbb{E}\mathrm{Tr}\left(\mathbf{W}^\dagger\mathbf{W}\mathbf{W}^\top(\mathbf{W}\mathbf{W}^\top + \lambda I)^{-1}\left(\mathbf{W}^\top\right)^\dagger\right),\end{aligned}$$

using Lemma B.4 remarking that, for all orthogonal matrix $O \in \mathbb{R}^{p\times p}$

$$\begin{aligned}O\mathbf{W}^\dagger\mathbf{W}\mathbf{W}^\top(\mathbf{W}\mathbf{W}^\top + \lambda I)^{-1}\left(\mathbf{W}^\top\right)^\dagger O^\top &= O\mathbf{W}^\dagger\mathbf{W}O^\top O\mathbf{W}^\top(\mathbf{W}O^\top O\mathbf{W}^\top + \lambda I)^{-1}\left(\mathbf{W}^\top\right)^\dagger O^\top \\ &= (\mathbf{W}O^\top)^\dagger\mathbf{W}O^\top(\mathbf{W}O^\top)^\top(\mathbf{W}O^\top(\mathbf{W}O^\top)^\top + \lambda I)^{-1}\left((\mathbf{W}O^\top)^\top\right)^\dagger,\end{aligned}$$

by orthogonality of O ($O^\top O = I_p$). Then, $O\mathbf{W}^\dagger\mathbf{W}\mathbf{W}^\top(\mathbf{W}\mathbf{W}^\top + \lambda I)^{-1}\left(\mathbf{W}^\top\right)^\dagger O^\top$ has the same distribution as $\mathbf{W}^\dagger\mathbf{W}\mathbf{W}^\top(\mathbf{W}\mathbf{W}^\top + \lambda I)^{-1}\left(\mathbf{W}^\top\right)^\dagger$, since $\mathbf{W}O^\top \overset{dist}{=} \mathbf{W}$.

Consider the singular value decomposition (SVD) of $\mathbf{W}$,

$$\mathbf{W} = \sum_{j=1}^r \sigma_j u_j v_j^\top,$$

where $r = p \wedge d$ is the rank of $\mathbf{W}^\dagger$, (u_j) is an orthonormal basis of $\mathbb{R}^d$, and (v_j) is an orthonormal basis of $\mathbb{R}^p$. The SVD of its pseudo-inverse is therefore

$$\mathbf{W}^\dagger = \sum_{j=1}^r \sigma_j^{-1} v_j u_j^\top.$$

Then, we obtain

$$\mathbf{W}^\dagger\mathbf{W}\mathbf{W}^\top(\mathbf{W}\mathbf{W}^\top + \lambda I)^{-1}\left(\mathbf{W}^\top\right)^\dagger = \sum_{j=1}^r \frac{1}{\lambda + \sigma_j^2} u_j u_j^\top.$$

Thus,

$$\mathrm{Tr}\left(\mathbf{W}^\dagger\mathbf{W}\mathbf{W}^\top(\mathbf{W}\mathbf{W}^\top + \lambda I)^{-1}\left(\mathbf{W}^\top\right)^\dagger\right) = \sum_{j=1}^r \frac{1}{\lambda + \sigma_j^2}.$$

We recognize $(\lambda + \sigma_j^2)_{j\in[r]}$ as the eigenvalues of $\Sigma + \lambda I_d = \mathbf{W}\mathbf{W}^\top + \lambda I_d$ when $d < p$ and $\mathrm{rank}(\mathbf{W}) = d$, or as the eigenvalues of $\mathbf{W}^\top\mathbf{W} + \lambda I_p$ when $d \geq p$ and $\mathrm{rank}(\mathbf{W}) = p$. Hence,

$$\Delta_{\mathrm{imp/miss}} + \Delta_{\mathrm{miss}} = \begin{cases} \frac{\lambda\|\beta^\star\|_2^2}{p}\mathbb{E}\mathrm{Tr}\left((\Sigma + \lambda I_d)^{-1}\right) & \text{if } d < p \\ \frac{\lambda\|\beta^\star\|_2^2}{p}\mathbb{E}\mathrm{Tr}\left((\mathbf{W}^\top\mathbf{W} + \lambda I_p)^{-1}\right) & \text{if } d \geq p.\end{cases}$$

□

C.3.2. PROOF OF (7) ($d < p - 1$)

(First step) Decomposition of $R^\star_{\mathrm{imp}}(d)$. Note that for $x \geq 0$ (to be chosen later), one has

$$R^\star_{\mathrm{imp}}(d) \leq R_{\mathrm{imp}}(x\theta^\star) \leq R(x\rho\theta^\star) + \rho(1-\rho)\|x\theta^\star\|_2^2 = R(x\rho\theta^\star) + \rho(1-\rho)x^2\|\theta^\star\|_2^2,$$

using Lemma C.1. Then,

$$R^\star_{\mathrm{imp}}(d) - R^\star(d) \leq R(x\rho\theta^\star) - R^\star(d) + \rho(1-\rho)x^2\|\theta^\star\|_2^2.$$

Note that,

$$\begin{aligned} R(x\rho\theta^\star) - R^\star(d) &= \|x\rho\theta^\star - \theta^\star\|_\Sigma^2 \\ &= (1-x\rho)^2\|\theta^\star\|_\Sigma^2. \end{aligned}$$

Thus, we have

$$R^\star_{\mathrm{imp}}(d) - R^\star(d) \leq (1-x\rho)^2\|\theta^\star\|_\Sigma^2 + x^2\rho(1-\rho)\|\theta^\star\|_2^2. \tag{31}$$

(Second step) Calculus of $\mathbb{E}\|\theta^\star\|_2^2$. Since $\theta^\star = (\mathbf{W}^\top)^\dagger\beta^\star$,

$$\mathbb{E}\|\theta^\star\|_2^2 = \mathbb{E}\beta^\top(\mathbf{W}^\dagger(\mathbf{W}^\top)^\dagger)\beta.$$

Again, for an orthonormal matrix O, $O\mathbf{W}^\dagger(\mathbf{W}^\top)^\dagger O^\top = (\mathbf{W}O^\top)^\dagger((\mathbf{W}O^\top)^\top)^\dagger$ has the same law as that of $\mathbf{W}^\dagger(\mathbf{W}^\top)^\dagger$ given that $\mathbf{W}O^\top$ has the same law of $\mathbf{W}$. Using Lemma B.4,

$$\begin{aligned} \mathbb{E}\|\theta^\star\|_2^2 &= \|\beta\|_2^2\frac{\mathbb{E}\mathrm{Tr}\left((\mathbf{W}^\dagger)^\top\mathbf{W}^\dagger\right)}{p} \\ &= \|\beta\|_2^2\frac{\mathbb{E}\|\mathbf{W}^\dagger\|_{\mathrm{Fr}}^2}{p}, \end{aligned}$$

by definition of the Frobenius norm. Then, by Lemma B.5, we obtain

$$\mathbb{E}\|\theta^\star\|_2^2 = \frac{d}{p}\left(1 + \frac{d}{p-d-1}\right)\|\beta\|_2^2. \tag{32}$$

(Third step) Calculus of $\mathbb{E}\|\theta^\star\|_\Sigma^2$. By definition, $R^\star(d) = \mathbb{E}\left[(X^\top\theta^\star - Y)^2|W_1,\ldots,W_d\right]$, therefore by optimality of $\theta^\star$, Fermat's rule gives that $\mathbb{E}[(X^\top\theta^\star - Y)X|\mathbf{W}] = 0$, and thus

$$\mathbb{E}Y^2 = \|\theta^\star\|_\Sigma^2 + R^\star(d).$$

Using that $\mathbb{E}Y^2 = \sigma^2 + \|\beta\|_2^2$, and taking the expectation, we obtain

$$\sigma^2 + \|\beta^\star\|_2^2 = \mathbb{E}\|\theta^\star\|_\Sigma^2 + \mathbb{E}R^\star(d).$$

Furthermore, by Proposition 2.1,

$$\mathbb{E}R^\star(d) = \sigma^2 + \frac{p-d}{p}\|\beta^\star\|_2^2.$$

Thus, we obtain

$$\mathbb{E}\|\theta^\star\|_\Sigma^2 = \frac{d}{p}\|\beta^\star\|_2^2. \tag{33}$$

(Fourth step) Conclusion. Putting things together, one gets

$$\begin{aligned} \mathbb{E}\left[R^\star_{\mathrm{imp}}(d) - R^\star(d)\right] &\leq (1-x\rho)^2\mathbb{E}\|\theta^\star\|_\Sigma^2 + x^2\rho(1-\rho)\mathbb{E}\|\theta^\star\|_2^2 \\ &= \frac{d}{p}\|\beta^\star\|_2^2\left((1-x\rho)^2 + x^2(1-\rho)\rho\left(1 + \frac{d-1}{p-d-1}\right)\right). \end{aligned}$$

The bound on the right hand side can be optimized with respect to x. It corresponds to a strongly convex function of the form $f : x \longmapsto (1-ax)^2 + bx^2$. We have $f'(x) = -2a(1-ax) + 2bx$, so that the only critical point is $x^\star = \frac{a}{a^2+b}$, leading to $\min f = f(x^\star) = \frac{b}{a^2+b}$. Therefore,

$$\begin{aligned}
\mathbb{E}\left[R^{\star}_{\text{imp}}(d) - R^{\star}(d)\right] &\leq \frac{d}{p}\|\beta^{\star}\|_2^2 \frac{(1-\rho)\rho\left(1+\frac{d-1}{p-d-1}\right)}{\rho^2 + (1-\rho)\rho\left(1+\frac{d-1}{p-d-1}\right)} \\
&= \frac{d}{p}\|\beta^{\star}\|_2^2 \frac{(1-\rho)\left(1+\frac{d-1}{p-d-1}\right)}{\rho + (1-\rho)\left(1+\frac{d-1}{p-d-1}\right)} \\
&= \frac{d}{p}\|\beta^{\star}\|_2^2 \frac{(1-\rho)\,(p-2)}{\rho(p-d-1) + (1-\rho)\,(p-2)} \\
&= \frac{d}{p}\|\beta^{\star}\|_2^2 \frac{(1-\rho)\,(p-2)}{p-\rho(d-1)-2} \\
&= \frac{d}{p}\|\beta^{\star}\|_2^2 \left((1-\rho) + (1-\rho)\frac{(p-2)-(p-\rho(d-1)-2)}{p-\rho(d-1)-2}\right) \\
&= (1-\rho)\frac{d}{p}\|\beta^{\star}\|_2^2\left(1 + \frac{\rho(d-1)}{p-\rho(d-1)-2}\right),
\end{aligned}$$

which leads to the desired result. We obtain also (8) and (9) using the equality obtained in Proposition 2.2.

C.3.3. Proof of upper and lower bounds (10) (11) ($d \geq p$)

Using Lemma C.3, we have

$$\Delta_{\text{imp/miss}} + \Delta_{\text{miss}} = \frac{\lambda\left\|\beta^{\star}\right\|_2^2}{p}\mathbb{E}\text{Tr}\left((\mathbf{W}^{\top}\mathbf{W} + \lambda I_p)^{-1}\right)$$

Remark that $\mathbf{W}^{\top}\mathbf{W} = \sum_{j=1}^{d} W_j W_j^{\top}$. Furthermore, note that when $W_1 \sim \mathcal{U}(\mathbb{S}^{p-1})$, for any $1 \leq k \leq p$, $W_1 = (W_{11}, \ldots, W_{1k}, \ldots, W_{1p})^{\top}$ has the same distribution as $(W_{11}, \ldots, -W_{1k}, \ldots, W_{1p})^{\top}$. Therefore, for all $1 \leq k \neq k' \leq p$, $\mathbb{E}[W_{1k}W_{1k'}] = -\mathbb{E}[W_{1k}W_{1k'}]$, leading to $\mathbb{E}[W_{1k}W_{1k'}] = 0$. Furthermore $\mathbb{E}[W_{11}^2 + \ldots + W_{1p}^2] = \mathbb{E}[W_{11}^2] + \ldots + \mathbb{E}[W_{1p}^2] = 1$, so that by exchangeability, for all $1 \leq k \leq p$, $\mathbb{E}[W_{1k}^2] = 1/p$ and finally $\mathbb{E}W_1W_1^{\top} = \frac{1}{p}I_p$.

Applying Proposition B.3, we obtain

$$\frac{\lambda\left\|\beta^{\star}\right\|_2^2}{p}\frac{p}{d/p+\lambda} \leq \Delta_{\text{imp/miss}} + \Delta_{\text{miss}} \leq \frac{\lambda\left\|\beta^{\star}\right\|_2^2}{p}(1+1/\lambda)\frac{p}{d/p+\lambda},$$

and

$$\left\|\beta^{\star}\right\|_2^2(1-\rho)\frac{p}{\rho d + (1-\rho)p} \leq \Delta_{\text{imp/miss}} + \Delta_{\text{miss}} \leq \left\|\beta^{\star}\right\|_2^2 \frac{p}{\rho d + (1-\rho)p}.$$

D. Proof of Section 3

D.1. Proof of Theorem 3.2

Start by writing

$$\begin{aligned}
\Delta_{\text{imp}}^{(\infty)} &= \mathbb{E}R^{\star}_{\text{imp}}(d) - R^{\star}(\infty) \\
&= \Delta_{\text{imp/miss}} + \Delta_{\text{miss}} + \mathbb{E}R^{\star}(d) - R^{\star}(\infty) \\
&= \mathbb{E}\inf_{\theta\in\mathbb{R}^d}\left\{R(\theta) - R^{\star}(\infty) + \frac{1-\rho}{\rho}\|\theta\|_{\text{diag}(\Sigma)}^2\right\},
\end{aligned}$$

using (27). Considering Assumption 4, $\text{diag}(\Sigma) \preceq L^2 I_d$, which leads to

$$\Delta_{\text{imp}}^{(\infty)} \leq \mathbb{E}\inf_{\theta\in\mathbb{R}^d}\left\{R(\theta) - R^{\star}(\infty) + L^2\frac{1-\rho}{\rho}\|\theta\|_2^2\right\}.$$

Fixing $\lambda = L^2 \frac{1-\rho}{\rho}$, we aim at providing an upper bound for

$$\Delta_\lambda := \mathbb{E} \inf_{\theta \in \mathbb{R}^d} \left\{ R(\theta) - R^\star(\infty) + \lambda \|\theta\|_2^2 \right\}.$$

Let $f \in \mathcal{F}_\nu^{(\infty)}$ admitting $\alpha \in L^2(\nu)$ as a representer. We set $\theta^{(\alpha)} \in \mathbb{R}^d$ such that for all $j \in [d]$, $\theta_j^{(\alpha)} = \frac{1}{d}\alpha(W_j)$. We have

$$\begin{aligned}
\Delta_\lambda + R^\star(\infty) &= \mathbb{E} \inf_{\theta \in \mathbb{R}^d} \left\{ R\left(\theta\right) + \lambda \mathbb{E}\left[\|\theta\|_2^2 \,|\mathbf{W} \right] \right\} \\
&\leq \mathbb{E}\left[R\left(\theta^{(\alpha)}\right) + \lambda \mathbb{E}\left[\left\|\theta^{(\alpha)}\right\|_2^2 |\mathbf{W} \right]\right] \\
&= \mathbb{E} R\left(\theta^{(\alpha)}\right) + \lambda \mathbb{E}\left[\left\|\theta^{(\alpha)}\right\|_2^2 \right].
\end{aligned}$$

First term. Remark that by definition of random features (15), $X^\top \theta^{(\alpha)} = \sum_{j=1}^d \theta_j^{(\alpha)} \psi(Z, W_j) = \frac{1}{d}\sum_{j=1}^d \alpha(W_j)\psi(Z, W_j)$. In consequence, $\mathbb{E}\left[X^\top \theta^{(\alpha)}|Z\right] = \int \alpha(W)\psi(Z,W)d\nu(W) = f(Z)$. Then,

$$\begin{aligned}
\mathbb{E} R\left(\theta^{(\alpha)}\right) &= \mathbb{E}\left[\mathbb{E}\left[\left(X^\top \theta^{(\alpha)} - Y\right)^2 |\mathbf{W}\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(X^\top \theta^{(\alpha)} - Y\right)^2 |Z\right]\right] && \text{using Fubini's theorem} \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(X^\top \theta^{(\alpha)} - f(Z) + f(Z) - Y\right)^2 |Z\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(X^\top \theta^{(\alpha)} - f(Z)\right)^2 + \left(f(Z) - Y\right)^2 |Z\right]\right] && \text{using } \mathbb{E}\left[X^\top \theta^{(\alpha)}|Z\right] = f(Z) \\
&= \mathbb{E}\left[\mathbb{V}\left[X^\top \theta^{(\alpha)}|Z\right]\right] + R(f).
\end{aligned}$$

Then,

$$\begin{aligned}
\mathbb{V}\left[X^\top \theta^{(\alpha)}|Z\right] &= \mathbb{V}\left[\frac{1}{d}\sum_{j=1}^d \alpha(W_j)\psi(Z, W_j)|Z\right] \\
&= \frac{1}{d}\mathbb{V}_\nu\left[\alpha(W)\psi(Z,W)|Z\right] && (W_j) \text{ being i.i.d} \\
&\leq \frac{1}{d}\mathbb{E}_\nu\left[\alpha(W)^2\psi(Z,W)^2|Z\right] && .
\end{aligned}$$

Then using Fubini's theorem,

$$\mathbb{V}\left[X^\top \theta^{(\alpha)}\right] \leq \frac{1}{d}\mathbb{E}\left[\mathbb{E}\left[\alpha(W)^2\psi(Z,W)^2|W\right]\right] = \frac{1}{d}\mathbb{E}\left[\alpha(W)^2\mathbb{E}\left[\psi(Z,W)^2|W\right]\right].$$

Under Assumption 4,

$$\mathbb{V}\left[X^\top \theta^{(\alpha)}\right] \leq \frac{L^2}{d}\mathbb{E}\left[\alpha(W)^2\right].$$

Thus,

$$\mathbb{E} R\left(\theta^{(\alpha)}\right) \leq \frac{L^2}{d}\mathbb{E}_\nu\left[\alpha(W)^2\right] + R(f).$$

Second term.

$$\begin{aligned}\mathbb{E}\left[\left\|\theta^{(\alpha)}\right\|_2^2\right] &= \mathbb{E}\left[\sum_{j=1}^{d}\theta_j^2\right]\\ &= \frac{1}{d}\mathbb{E}\left[\frac{1}{d}\sum_{j=1}^{d}\alpha(W_j)^2\right]\\ &= \frac{1}{d}\mathbb{E}\left[\alpha(W)^2\right],\end{aligned}$$

using that $(W_j)_j$ are i.i.d..

Conclusion. Combining these two terms, we have

$$\Delta_\lambda + R^\star(\infty) \leq R(f) + \frac{\lambda + L^2}{d}\mathbb{E}_\nu\left[\alpha(W)^2\right].$$

This result is valid for any f and α, thus

$$\Delta_{\text{imp}}^{(\infty)} + R^\star(\infty) \leq \Delta_\lambda + R^\star(\infty) \leq \inf_{f\in\mathcal{F}^{(\nu)}}\left\{R(f) + \frac{\lambda+L^2}{d}\left\|f\right\|_\nu^2\right\}.$$

E. Proof of error bounds for SGD estimators

E.1. Proof of Equation (13) of Theorem 2.4

In this section, we apply results from the SGD literature, in particular, Bach & Moulines (2013, Theorem 1), to our framework.

Theorem E.1. *In the framework of Section 2, for $\gamma = \frac{1}{6d}$, and when $d < n$, we have*

$$R_{\text{imp}}(\bar{\theta}) - R^\star_{\text{imp}}(d) \lesssim \frac{d}{n}\|\theta^\star_{\text{imp}}\|_2^2 + \frac{d}{n}\sigma^2_{\text{imp}}, \tag{34}$$

with $\sigma^2_{\text{imp}} := \sigma^2 + \rho^{-1}(R^\star_{\text{imp}}(d) - R^\star(d) + \|\theta^\star\|_\Sigma^2)$.

Proof. The proof of this theorem consists of verifying that assumptions of Bach & Moulines (2013, Theorem 1) hold in our case. Assumptions (A1-5) are easily satisfied. Let us show that $\mathbb{E}\left[\tilde{X}\tilde{X}^\top\left\|\tilde{X}\right\|_2^2|\mathbf{W}\right] \preceq R^2\Sigma_{\text{imp}}$. Indeed,

$$\mathbb{E}\left[\tilde{X}\tilde{X}^\top\left\|\tilde{X}\right\|_2^2|\mathbf{W}\right] \preceq \mathbb{E}\left[\tilde{X}\tilde{X}^\top\left\|X\right\|_2^2|\mathbf{W}\right],$$

using that $\left\|\tilde{X}\right\|_2^2 \leq \|X\|_2^2$, and $0 \preceq \tilde{X}\tilde{X}^\top$. Then,

$$\begin{aligned}\mathbb{E}\left[\tilde{X}\tilde{X}^\top\left\|X\right\|_2^2|\mathbf{W}\right] &= \mathbb{E}\mathbb{E}\left[\tilde{X}\tilde{X}^\top\left\|X\right\|_2^2|P,\mathbf{W}\right]\\ &= \mathbb{E}\mathbb{E}\left[PP^\top\odot XX^\top\left\|X\right\|_2^2|P,\mathbf{W}\right]\\ &= \mathbb{E}\left[\Sigma_P\odot XX^\top\left\|X\right\|_2^2|\mathbf{W}\right]\\ &= \Sigma_P\odot\left(\mathbb{E}\left[XX^\top\left\|X\right\|_2^2|\mathbf{W}\right]\right).\end{aligned}$$

X is Gaussian vector, thus $\mathbb{E}\left[XX^\top\|X\|_2^2\right] \preceq R^2\Sigma$ with $R^2 = 3d$, and Lemma B.6 lead to

$$\mathbb{E}\left[\tilde{X}\tilde{X}^\top\left\|\tilde{X}\right\|_2^2\right] \preceq R^2\Sigma_P\odot\Sigma = R^2\Sigma_{\text{imp}}.$$

Define $\epsilon_{\text{imp}} = Y - \tilde{X}^\top \theta^\star_{\text{imp}} = X^\top \theta^\star + \epsilon - \tilde{X}^\top \theta^\star_{\text{imp}}$. First, we have $\epsilon^2_{\text{imp}} \leq 3\left(\epsilon^2 + \left(\tilde{X}^\top \theta^\star_{\text{imp}}\right)^2 + \left(X^\top \theta^\star\right)^2\right)$, then

$$\mathbb{E}\left[\epsilon^2_{\text{imp}} \tilde{X}\tilde{X}^\top\right] \preceq 3\mathbb{E}\left[\epsilon^2 \tilde{X}\tilde{X}^\top\right] + 3\mathbb{E}\left[\left(\tilde{X}^\top \theta^\star_{\text{imp}}\right)^2 \tilde{X}\tilde{X}^\top\right] + 3\mathbb{E}\left[\left(\tilde{X}^\top \theta^\star\right)^2 \tilde{X}\tilde{X}^\top\right]. \tag{35}$$

Using that ϵ is an independent noise, $\mathbb{E}\left[\epsilon^2 \tilde{X}\tilde{X}^\top\right] = \sigma^2 \Sigma_{\text{imp}}$. Let u, v in $\mathbb{R}^d$, note that

$$\begin{aligned} v^\top \mathbb{E}\left[\left(u^\top \tilde{X}\right)^2 \tilde{X}\tilde{X}^\top\right] v &= \mathbb{E}\left[\left(u^\top \tilde{X}\right)^2 \left(v^\top \tilde{X}\right)^2\right] \\ &\leq \mathbb{E}\left[\left(u^\top X\right)^2 \left(v^\top \tilde{X}\right)^2\right] \\ &\leq \rho \mathbb{E}\left[\left(u^\top X\right)^2 \left(v^\top X\right)^2\right] \\ &\leq \rho \sqrt{\mathbb{E}\left[(u^\top X)^4\right] \mathbb{E}\left[(v^\top X)^4\right]}, \end{aligned}$$

using the Cauchy-Schwarz inequality. Then, by the kurtosis boundedness of Gaussian vectors, we have $\mathbb{E}\left[\left(u^\top X\right)^4\right] \leq 3\left\|u\right\|^4_\Sigma$ and $\mathbb{E}\left[\left(v^\top X\right)^4\right] \leq 3\left\|v\right\|^4_\Sigma$. Then,

$$\begin{aligned} v^\top \mathbb{E}\left[\left(u^\top \tilde{X}\right)^2 \tilde{X}\tilde{X}^\top\right] v &= 3\rho \left\|u\right\|^2_\Sigma \left\|v\right\|^2_\Sigma \\ &\leq 3\rho^{-1} \left\|u\right\|^2_\Sigma \left\|v\right\|^2_{\Sigma_{\text{imp}}}. \end{aligned}$$

This shows that

$$\mathbb{E}\left[\left(u^\top \tilde{X}\right)^2 \tilde{X}\tilde{X}^\top\right] \preceq 3\rho^{-1} \left\|u\right\|^2_\Sigma \Sigma_{\text{imp}}. \tag{36}$$

Using similar arguments,

$$\mathbb{E}\left[\left(u^\top X\right)^2 \tilde{X}\tilde{X}^\top\right] \preceq 3\rho^{-1} \left\|u\right\|^2_\Sigma \Sigma_{\text{imp}}. \tag{37}$$

These two above equations can be used when u is equal to $\theta^\star_{\text{imp}}$ and $\theta^\star$, to transform (35) into

$$\mathbb{E}\left[\epsilon^2_{\text{imp}} \tilde{X}\tilde{X}^\top\right] \preceq (3\sigma^2 + 9\rho^{-1}\left\|\theta^\star_{\text{imp}}\right\|^2_\Sigma + 9\rho^{-1}\left\|\theta^\star\right\|^2_\Sigma)\Sigma_{\text{imp}}.$$

Remarking, that $\left\|\theta^\star_{\text{imp}}\right\|^2_\Sigma \leq 2\left\|\theta^\star_{\text{imp}} - \theta^\star\right\|^2_\Sigma + 2\left\|\theta^\star\right\|^2_\Sigma = 2(R^\star_{\text{imp}}(d) - R^\star(d) + \left\|\theta^\star\right\|^2_\Sigma)$, we get

$$3\sigma^2 + 9\rho^{-1}\left\|\theta^\star_{\text{imp}}\right\|^2_\Sigma + 9\rho^{-1}\left\|\theta^\star\right\|^2_\Sigma \lesssim \sigma^2 + \rho^{-1}(R^\star_{\text{imp}}(d) - R^\star(d) + \left\|\theta^\star\right\|^2_\Sigma),$$

leading to the desired results. □

Lemma E.2. *Under assumptions of Theorem 2.4. The norm of $\theta^\star_{\text{imp}}$, the best predictor working with imputed by 0 inputs, satisfies*

$$\mathbb{E}\|\theta^\star_{\text{imp}}\|^2 \leq \begin{cases} \frac{d}{p\rho}\frac{p-2}{p-d-1}\|\beta\|^2_2 & \textit{when } d < p-1, \\ \frac{p}{d\rho^2(1-\rho)}\|\beta\|^2_2, & \textit{when } d \geq p-1. \end{cases}$$

Proof. Let's begin by,

$$\rho(1-\rho)\|\theta^\star_{\text{imp}}\|^2_2 \leq R(\rho\theta^\star_{\text{imp}}) - R(\rho\theta^\star) + \rho(1-\rho)\|\theta^\star_{\text{imp}}\|^2_2. \tag{38}$$

because $R(\theta^\star) \leq R(\rho\theta^\star_{\text{imp}})$. Using, Lemma C.1, we obtain

$$\rho(1-\rho)\mathbb{E}\|\theta^\star_{\text{imp}}\|^2_2 \leq \Delta_{\text{miss}} + \Delta_{\text{imp/miss}} \tag{39}$$

First case: $d < p - 1$**.** In this, case

$$\Delta_{\text{miss}} + \Delta_{\text{imp/miss}} \leq (1-\rho)\frac{d}{p}\|\beta\|_2^2 \left(1 + \frac{\rho(d-1)}{p - \rho(d-1) - 2}\right).$$

We obtain, using (39),

$$\mathbb{E}\|\theta^\star_{\text{imp}}\|_2^2 \leq \frac{d}{\rho p}\|\beta\|_2^2 \left(1 + \frac{\rho(d-1)}{p-\rho(d-1)-2}\right) \leq \frac{p-2}{\rho p}\|\beta\|_2^2 \frac{\rho(d-1)}{p-\rho(d-1)-2}.$$

Second case: $d > p - 1$**.** In this case,

$$\Delta_{\text{miss}} + \Delta_{\text{imp/miss}} \leq \frac{p}{\rho d + (1-\rho)p}\|\beta^\star\|_2^2 \leq \frac{p}{\rho d}\|\beta^\star\|_2^2.$$

We obtain, using (39),

$$\mathbb{E}\|\theta^\star_{\text{imp}}\|_2^2 \leq \frac{p}{\rho^2(1-\rho)d}\|\beta^\star\|_2^2.$$

□

Proof of (13).

$$R_{\text{imp}}(\bar\theta) - R^\star(d) = R_{\text{imp}}(\bar\theta) - R^\star_{\text{imp}}(d) + R^\star_{\text{imp}}(d) - R^\star(d)$$

Using Theorem E.1 to bound the first term, we find

$$R_{\text{imp}}(\bar\theta) - R^\star(d) \lesssim \left(1 + \frac{d}{\rho n}\right)(R^\star_{\text{imp}}(d) - R^\star(d)) + \frac{d}{n}\|\theta^\star_{\text{imp}}\|_2^2 + \frac{d}{n}(\sigma^2 + \|\theta^\star\|_\Sigma^2).$$

Thus, taking the expectation,

$$\mathbb{E}[R_{\text{imp}}(\bar\theta) - R^\star(d)] \lesssim \left(1 + \frac{d}{\rho n}\right)(\Delta_{\text{imp/miss}} + \Delta_{\text{miss}}) + \frac{d}{n}\mathbb{E}\|\theta^\star_{\text{imp}}\|_2^2 + \frac{d}{n}(\sigma^2 + \mathbb{E}\,\|\theta^\star\|_\Sigma^2).$$

Note that $\mathbb{E}\|\theta^\star_{\text{imp}}\|_2^2 \leq \frac{1}{(1-\rho)\rho}(\Delta_{\text{imp/miss}} + \Delta_{\text{miss}})$ (using (39)) and $\mathbb{E}\,\|\theta^\star\|_\Sigma^2 = \frac{d}{p}\,\|\beta^\star\|_2^2$ (using Proposition 2.1). Thus,

$$\begin{aligned}\mathbb{E}[R_{\text{imp}}(\bar\theta) - R^\star(d)] &\lesssim \left(1 + \frac{d}{\rho n} + \frac{d}{(1-\rho)\rho n}\right)(\Delta_{\text{imp/miss}} + \Delta_{\text{miss}}) + \frac{d}{n}\left(\sigma^2 + \frac{d}{p}\,\|\beta^\star\|_2^2\right) \\ &\lesssim \left(1 + \frac{d}{(1-\rho)\rho n}\right)(\Delta_{\text{imp/miss}} + \Delta_{\text{miss}}) + \frac{d}{n}\left(\sigma^2 + \frac{d}{p}\,\|\beta^\star\|_2^2\right).\end{aligned}$$

Thus,

$$\begin{aligned}\mathbb{E}[R_{\text{imp}}(\bar\theta) - R^\star_{\text{miss}}(d)] &\lesssim \Delta_{\text{imp/miss}} + \frac{d}{\rho n} + \frac{d}{(1-\rho)\rho n}(\Delta_{\text{imp/miss}} + \Delta_{\text{miss}}) + \frac{d}{n}\left(\sigma^2 + \frac{d}{p}\,\|\beta^\star\|_2^2\right) \\ &\lesssim (1-\rho)\frac{d}{p}\|\beta\|_2^2 \frac{\rho(d-1)}{p-\rho(d-1)-2} + \frac{d}{(1-\rho)\rho n}\frac{d}{p}\|\beta\|_2^2 \frac{(1-\rho)\,(p-2)}{p-\rho(d-1)-2} + \frac{d}{n}\left(\sigma^2 + \frac{d}{p}\,\|\beta^\star\|_2^2\right) \\ &\lesssim (1-\rho)\frac{d}{p}\|\beta\|_2^2 \frac{\rho(d-1)}{p-\rho(d-1)-2} + \frac{d}{(1-\rho)\rho n}\frac{d}{p}\|\beta\|_2^2 \frac{(1-\rho)\,(p-2)}{p-\rho(d-1)-2} + \frac{d}{n}\sigma^2 \\ &\lesssim \frac{d^2\|\beta\|_2^2}{p(p-\rho(d-1)-2)}\left((1-\rho)\rho + \frac{p-2}{\rho n}\right) + \frac{d}{n}\sigma^2.\end{aligned}$$

□

E.2. Proof of Equation (14) in Theorem 2.4 and proof of Theorem 3.3

Let's start with a result of Ayme et al. (2023) for the deterministic case (without random features).

Assumption 9. *There exist $\sigma > 0$ and $R > 0$ such that $\mathbb{E}[XX^\top \|X\|_2^2] \preceq R^2\Sigma$ and $\mathbb{E}[\epsilon^2 \|X\|_2^2] \leq \sigma^2 R^2$, where $\epsilon = Y - X^\top\theta^\star$.*

Theorem E.3. *(Ayme et al., 2023) Under Assumption 9, choosing a constant learning rate $\gamma = \frac{1}{\kappa \mathrm{Tr}(\Sigma)\sqrt{n}}$ leads to*

$$\mathbb{E}\left[R_{\mathrm{imp}}\left(\bar{\theta}_{\mathrm{imp}}\right)\right] - R(\theta^\star) \lesssim \frac{R^2}{\sqrt{n}}\left\|\theta^\star_{\mathrm{imp}}\right\|_2^2 + \frac{\sigma^2 + \|\theta^\star\|_\Sigma^2}{\sqrt{n}} + R^\star_{\mathrm{imp}}(d) - R^\star(d),$$

where $\theta^\star$ (resp. $\theta^\star_{\mathrm{imp}}$) is the best linear predictor for complete (resp. with imputed missing values) case.

Proof of (14) of Theorem 2.4. By using the Gaussianity of X, we obtain that $\mathbb{E}\left[XX^\top \|X\|_2^2\right] \preceq R^2\Sigma$ with $R^2 = 2d$. Furthermore, in the case $p \leq d$, we have $X^\top\theta^\star = Z^\top\beta^\star$ and $\epsilon = Y - X^\top\theta^\star$ thus the noise and X are independent and $\mathbb{E}[\epsilon^2 \|X\|_2^2] \leq \sigma^2 \mathrm{Tr}(\Sigma)$. Then, Assumption 9 is satisfied with $\kappa = 3$. Taking the expectation in Theorem E.3, we found

$$\mathbb{E}\left[R_{\mathrm{imp}}\left(\bar{\theta}_{\mathrm{imp}}\right)\right] - \mathbb{E}R(\theta^\star) \lesssim \frac{\mathrm{Tr}(\Sigma)}{\sqrt{n}}\mathbb{E}\left\|\theta^\star_{\mathrm{imp}}\right\|_2^2 + \frac{\sigma^2 + \mathbb{E}\|\theta^\star\|_\Sigma^2}{\sqrt{n}} + \Delta_{\mathrm{imp/miss}} + \Delta_{\mathrm{miss}}.$$

Recall that $R(\theta^\star) = \sigma^2$ almost-surely, $\|\theta^\star\|_\Sigma^2 = \|\beta\|_2^2$ almost surely and $\mathrm{Tr}(\Sigma) = d$. Then,

$$\mathbb{E}\left[R_{\mathrm{imp}}\left(\bar{\theta}_{\mathrm{imp}}\right)\right] - \sigma^2 \lesssim \frac{d}{\sqrt{n}}\mathbb{E}\left\|\theta^\star_{\mathrm{imp}}\right\|_2^2 + \frac{\sigma^2 + \|\beta\|_2^2}{\sqrt{n}} + \Delta_{\mathrm{imp/miss}} + \Delta_{\mathrm{miss}}.$$

Then applying (39),

$$\mathbb{E}\left[R_{\mathrm{imp}}\left(\bar{\theta}_{\mathrm{imp}}\right)\right] - \sigma^2 \lesssim \left(1 + \frac{d}{\rho(1-\rho)\sqrt{n}}\right)\left(\Delta_{\mathrm{imp/miss}} + \Delta_{\mathrm{miss}}\right) + \frac{\sigma^2 + \|\beta\|_2^2}{\sqrt{n}}.$$

Applying (11), we finally get

$$\begin{aligned}\mathbb{E}\left[R_{\mathrm{imp}}\left(\bar{\theta}_{\mathrm{imp}}\right)\right] - \sigma^2 &\lesssim \left(1 + \frac{d}{\rho(1-\rho)\sqrt{n}}\right)\frac{p\|\beta^\star\|_2^2}{\rho d + (1-\rho)p} + \frac{\sigma^2 + \|\beta\|_2^2}{\sqrt{n}} \\ &\lesssim \left(1 + \frac{d}{\rho(1-\rho)\sqrt{n}}\right)\frac{p\|\beta^\star\|_2^2}{\rho d + (1-\rho)p} + \frac{\sigma^2}{\sqrt{n}}.\end{aligned}$$

□

Proof of Theorem 3.3. Under Assumption 7, $\|X\|_2^2 \leq \kappa L^2 d$ almost surely, then $\mathbb{E}[XX^\top \|X\|_2^2] \preceq \kappa L^2 d\Sigma$. And,

$$\mathbb{E}[\epsilon^2 \|X\|_2^2] \leq \mathbb{E}[\epsilon^2]\kappa L^2 d = R^\star(\infty)\kappa L^2 d.$$

Thus we can applied Theorem E.3, that gives us,

$$\mathbb{E}\left[R_{\mathrm{imp}}\left(\bar{\theta}_{\mathrm{imp}}\right)\right] - R(\theta^\star) \lesssim \frac{\kappa L^2 d}{\sqrt{n}}\left\|\theta^\star_{\mathrm{imp}}\right\|_2^2 + \frac{R^\star(\infty)L\kappa + \|\theta^\star\|_\Sigma^2}{\sqrt{n}} + R^\star_{\mathrm{imp}}(d) - R^\star(d).$$

Note that,

$$\begin{aligned}\Delta_{\mathrm{imp}}^{(\infty)} &= \mathbb{E}R^\star_{\mathrm{imp}}(d) - R^\star(\infty) \\ &= \Delta_{\mathrm{imp/miss}} + \Delta_{\mathrm{miss}} + \mathbb{E}R^\star(d) - R^\star(\infty).\end{aligned}$$

Thus taking the expectation,

$$\mathbb{E}\left[R_{\mathrm{imp}}\left(\bar{\theta}_{\mathrm{imp}}\right)\right] - R^\star(\infty) \lesssim \frac{\kappa L^2 d}{\sqrt{n}}\mathbb{E}\left\|\theta^\star_{\mathrm{imp}}\right\|_2^2 + \frac{R^\star(\infty)\kappa L^2 + \|\theta^\star\|_\Sigma^2}{\sqrt{n}} + \Delta_{\mathrm{imp}}^{(\infty)}.$$

Under Assumption 6, it holds that $\ell^2 I \preceq \mathrm{diag}(\Sigma)$, and

$$\begin{aligned}\ell^2\rho(1-\rho)\mathbb{E}\left\|\theta^\star_{\mathrm{imp}}\right\|_2^2 &\leq \rho(1-\rho)\mathbb{E}\left\|\theta^\star_{\mathrm{imp}}\right\|_{\mathrm{diag}(\Sigma)}^2\\ &\leq \Delta_{\mathrm{imp/miss}} + \Delta_{\mathrm{miss}}\\ &\leq \Delta_{\mathrm{imp}}^{(\infty)}.\end{aligned}$$

Then,

$$\begin{aligned}\mathbb{E}\left[R_{\mathrm{imp}}\left(\bar{\theta}_{\mathrm{imp}}\right)\right] - R^\star(\infty) &\lesssim \frac{\kappa L^2 d}{\ell^2\rho(1-\rho)\sqrt{n}}\Delta_{\mathrm{imp}}^{(\infty)} + \frac{R^\star(\infty)\kappa + \|\theta^\star\|_\Sigma^2}{\sqrt{n}} + \Delta_{\mathrm{imp}}^{(\infty)}\\ &\lesssim \left(1 + \frac{\kappa L^2 d}{\ell^2\rho(1-\rho)\sqrt{n}}\right)\Delta_{\mathrm{imp}}^{(\infty)} + \frac{R^\star(\infty)\kappa L^2 + \|\theta^\star\|_\Sigma^2}{\sqrt{n}}.\end{aligned}$$

Recall that using Theorem 3.2,

$$\Delta_{\mathrm{imp}}^{(\infty)} \leq \frac{\lambda_{\mathrm{imp}}}{d}\left\|f^\star\right\|_\nu^2 = \frac{L^2}{\rho d}\left\|f^\star\right\|_\nu^2,$$

and $\|\theta^\star\|_\Sigma^2 \leq \mathbb{E}Y^2$ almost-surely. Thus,

$$\begin{aligned}\mathbb{E}\left[R_{\mathrm{imp}}\left(\bar{\theta}_{\mathrm{imp}}\right)\right] - R^\star(\infty) &\lesssim \left(1 + \frac{\kappa L^2 d}{\ell^2\rho(1-\rho)\sqrt{n}}\right)\frac{l^2}{\rho d}\left\|f^\star\right\|_\nu^2 + \frac{R^\star(\infty)\kappa + \mathbb{E}Y^2}{\sqrt{n}}\\ &\lesssim \left(1 + \frac{\kappa L^2 d}{\ell^2\rho(1-\rho)\sqrt{n}}\right)\frac{L^2}{\rho d}\left\|f^\star\right\|_\nu^2 + (1+\kappa L^2)\frac{\mathbb{E}Y^2}{\sqrt{n}}.\end{aligned}$$

□

F. Proof of Theorem 4.1 (under MNAR assumption)

First step (bias-variance decomposition) We denote by $\mathbf{W}'$ the matrix of $\underline{W'_j}$ Let $\theta \in \mathbb{R}^p$,

$$\begin{aligned}R_{\mathrm{imp}}(\theta) &= \mathbb{E}_Z\mathbb{E}\left[\left(Y - \tilde{X}^\top\theta\right)^2 |Z, \mathbf{W}, \mathbf{W}'\right]\\ &= \mathbb{E}_Z\mathbb{E}\left[\left(Y - \mathbb{E}\left[\tilde{X}^\top\theta|Z, \mathbf{W}, \mathbf{W}'\right]\right)^2 |Z, \mathbf{W}, \mathbf{W}'\right] + \mathbb{E}_Z\mathbb{V}\left[\tilde{X}^\top\theta|Z, \mathbf{W}, \mathbf{W}'\right],\end{aligned}$$

using a classical bias-variance decomposition. Futhermore,

$$\mathbb{E}\left[\tilde{X}^\top\theta|Z, \mathbf{W}, \mathbf{W}'\right] = \sum_{j=1}^d \theta_j\mathbb{E}\left[\tilde{X}_j|Z, \mathbf{W}, \mathbf{W}'\right] = \sum_{j=1}^d \theta_j\phi(Z, \underline{W'_j})\psi(Z, W_j)$$

and

$$\begin{aligned}\mathbb{V}\left[\tilde{X}^\top\theta|Z, \mathbf{W}, \mathbf{W}'\right] &= \sum_{j=1}^d \theta_j^2\mathbb{V}\left[\tilde{X}_j|Z, W_j, \underline{W'_j}\right]\\ &= \sum_{j=1}^d \theta_j^2\phi(Z, \underline{W'_j})(1-\phi(Z, \underline{W'_j}))\psi(Z, W_j)^2.\end{aligned}$$

Let $\alpha \in L^2(\mu \otimes \nu)$, and define $\theta^{(d)} \in \mathbb{R}^d$ such that $\theta_j^{(d)} = \alpha(W_j, \underline{W'_j})/d$. We have

$$\begin{aligned} R^\star_{\mathrm{imp}}(d) &\leq R_{\mathrm{imp}}(\theta^{(d)}) \\ &= \mathbb{E}_Z \left[\left(Y - \frac{1}{d} \sum_{j=1}^{d} \alpha(W_j, \underline{W'_j}) \phi(Z, \underline{W'_j}) \psi(Z, W_j) \right)^2 \right] \\ &+ \mathbb{E}_Z \left[\frac{1}{d^2} \sum_{j=1}^{d} \alpha(W_j, \underline{W'_j})^2 \phi(Z, \underline{W'_j})(1 - \phi(Z, \underline{W'_j})) \psi(Z, W_j)^2 \right]. \end{aligned}$$

Convergence of the variance term Using that $\phi(Z, \underline{W'_j})(1 - \phi(Z, \underline{W'_j})) \leq 1$ almost-surely, we have

$$\begin{aligned} \mathbb{E}_Z \left[\frac{1}{d^2} \sum_{j=1}^{d} \alpha(W_j, \underline{W'_j}) \phi(Z, \underline{W'_j})(1 - \phi(Z, \underline{W'_j})) \psi(Z, W_j)^2 \right] &\leq \mathbb{E}_Z \left[\frac{1}{d^2} \sum_{j=1}^{d} \alpha(W_j, \underline{W'_j})^2 \psi(Z, W_j)^2 \right] \\ &= \frac{1}{d^2} \sum_{j=1}^{d} \alpha(W_j, \underline{W'_j})^2 \mathbb{E}_Z \psi(Z, W_j)^2 \\ &\leq \frac{1}{d^2} \sum_{j=1}^{d} \alpha(W_j, \underline{W'_j})^2 L^2. \end{aligned}$$

Using that $\left(\alpha(W_j, \underline{W'_j})^2\right)_j$ are an i.i.d. sequences of integrable random variables, we obtain that

$$\lim_{d \to +\infty} \mathbb{E}_Z \left[\frac{1}{d^2} \sum_{j=1}^{d} \alpha(W_j, \underline{W'_j}) \phi(Z, \underline{W'_j})(1 - \phi(Z, \underline{W'_j})) \psi(Z, W_j)^2 \right] = 0,$$

almost-surely.

Convergence of the bias term Note that $\alpha(W_j, \underline{W'_j}) \phi(Z, \underline{W'_j}) \psi(Z, W_j)$ is integrable since $|\alpha(W_j, \underline{W'_j}) \phi(Z, \underline{W'_j})| \leq |\alpha(W_j, \underline{W'_j})|$, and $\psi(z, W_j) \in L^2(\nu)$. Then, using Kolmogorov's law and mapping continuous theorem, we obtain

$$\begin{aligned} &\lim_{d \to +\infty} \mathbb{E}_Z \left[\left(Y - \frac{1}{d} \sum_{j=1}^{d} \alpha(W_j, \underline{W'_j}) \phi(Z, \underline{W'_j}) \psi(Z, W_j) \right)^2 \right] = \\ &\qquad \mathbb{E}_Z \left[\left(Y - \int \alpha(w, w') \phi(Z, w') \psi(Z, w) d\mu \otimes \nu(w, w') \right)^2 \right]. \end{aligned} \tag{40}$$

Thus,

$$\limsup_{d \to +\infty} R^\star_{\mathrm{imp}}(d) \leq \mathbb{E}_Z \left[\left(Y - \int \alpha(w, w') \phi(Z, w') \psi(Z, w) d\mu \otimes \nu(w, w') \right)^2 \right].$$

Denoting by $\mathcal{G}$ the functions of the form

$$g(Z) = \int \alpha(w, w') \phi(Z, w') \psi(Z, w) d\mu \otimes \nu(w, w'),$$

we obtain that

$$\limsup_{d \to +\infty} R^\star_{\mathrm{imp}}(d) \leq \inf_{g \in \mathcal{G}} \mathbb{E}_Z \left[(Y - g(Z))^2 \right].$$

Using Fubini theorem, functions of the form

$$\begin{aligned} g(Z) &= \int \alpha(w)\psi(Z,w)d\nu(w) \int \beta(w')\phi(Z,w')d\mu(w') \\ &= f(Z)h(Z), \end{aligned}$$

are included in $\mathcal{G}$. In the following, we denote by $\mathcal{H}$ the set of functions, of the form

$$h(Z) = \int \beta(w')\phi(Z,w')d\mu(w'),$$

with $\beta \in L^2(\mu)$. We finally obtain the following bound

$$\limsup_{d\to+\infty} R^\star_{\mathrm{imp}}(d) \leq \inf_{(f,h)\in\mathcal{F}_\nu\times\mathcal{G}} \mathbb{E}_Z\left[(Y - f(Z)h(Z))^2\right]. \tag{41}$$

Proof for the case where $\mathcal{Z}$ is compact, $\mathcal{F}$ is dense in the set of continuous functions, and $f^\star$ is continuous First, let's show that the risk is continuous for the set of continuous prediction functions. Let f and g be two continuous functions on $B_\infty(0,B)$ (ball for $\|\|_\infty$), then

$$\begin{aligned} |R(f) - R(g)| &= \left|\mathbb{E}\left[(f(Z) - f^\star(Z))^2 - (g(Z) - f^\star(Z))^2\right]\right| \\ &= \left|\mathbb{E}\left[(f(Z) - g(Z))(g(Z) + f(Z) - f^\star(Z))\right]\right| \\ &\leq \|f-g\|_\infty \left(\|f\|_\infty + \|g\|_\infty + \|f^\star\|_\infty\right) \\ &\leq \|f-g\|_\infty \left(2B + \|f^\star\|_\infty\right). \end{aligned}$$

This shows the continuity of the risk. Then, we consider $\beta = 1$, thus $h(Z) = \int \beta(w')\phi(Z,w')d\mu(w') > 0$ almost-surely because $\phi(Z,w') > 0$ almost surely. Thus considering in (41), $f(Z) = f^\star(Z)/h(Z)$, f is continuous, we conclude using the continuity of risk and that $f \in \bar{\mathcal{F}}$.

Proof for Gaussian RF In this case $\phi(z,(w,w_0')) = \Phi(z^\top w' + w_0')$. Consider $I = \int \Phi(w_0')d\mu$ and introduce $h_\epsilon(z) = \int_{\|w'\|\leq\epsilon} \phi(Z,(w',w_0')d\mu(w',w_0)/\int_{\|w'\|\leq\epsilon} \Phi(w_0')d\mu(w',w_0')$,

$$|I - h_\epsilon(Z)| \leq \frac{1}{\int_{\|w'\|\leq\epsilon} \Phi(w_0')d\mu(w')} \int_{\|w'\|\leq\epsilon} |\phi(0,(0,w_0')) - \phi(Z,(w',w_0))| d\mu(w',w_0').$$

Using that that Φ is C-Lipschitz, one has

$$|I - h_\epsilon(Z)| \leq C\epsilon\|Z\|.$$

We conclude using $h = h_\epsilon$ for a decreasing sequence of ϵ and $f = f^\star/I$ in (41).