# Naive imputation implicitly regularizes high-dimensional linear models

Alexis Ayme, Claire Boyer, Aymeric Dieuleveut, Erwan Scornet

# Naive imputation implicitly regularizes high-dimensional linear models.

Alexis Ayme, Claire Boyer, Aymeric Dieuleveut
& Erwan Scornet

**Abstract**

Two different approaches exist to handle missing values for prediction: either imputation, prior to fitting any predictive algorithms, or dedicated methods able to natively incorporate missing values. While imputation is widely (and easily) use, it is unfortunately biased when low-capacity predictors (such as linear models) are applied afterward. However, in practice, naive imputation exhibits good predictive performance. In this paper, we study the impact of imputation in a high-dimensional linear model with MCAR missing data. We prove that zero imputation performs an implicit regularization closely related to the ridge method, often used in high-dimensional problems. Leveraging on this connection, we establish that the imputation bias is controlled by a ridge bias, which vanishes in high dimension. As a predictor, we argue in favor of the averaged SGD strategy, applied to zero-imputed data. We establish an upper bound on its generalization error, highlighting that imputation is benign in the $d \gg \sqrt{n}$ regime. Experiments illustrate our findings.

## 1 Introduction

Missing data has become an inherent problem in modern data science. Indeed, most real-world data sets contain missing entries due to a variety of reasons: merging different data sources, sensor failures, difficulty to collect/access data in sensitive fields (e.g., health), just to name a few. The simple, yet quite extreme, solution of throwing partial observations away can drastically reduce the data set size and thereby hinder further statistical analysis. Specific methods should be therefore developed to handle missing values. Most of them are dedicated to model estimation, aiming at inferring the underlying model parameters despite missing values (see, e.g., Rubin, 1976). In this paper, we take a different route and consider a supervised machine learning (ML) problem with missing values in the training and test inputs, for which our aim is to build a prediction function (*and not* to estimate accurately the true model parameters).

**Prediction with NA** A common practice to perform supervised learning with missing data is to simply impute the data set first, and then train any predictor on the completed/imputed data set. The imputation technique can be simple (e.g., using mean imputation) or more

elaborate (Van Buuren and Groothuis-Oudshoorn, 2011; Yoon et al., 2018; Muzellec et al., 2020; Ipsen et al., 2022). While such widely-used two-step strategies lack deep theoretical foundations, they have been shown to be consistent, provided that the approximation capacity of the chosen predictor is large enough (see Josse et al., 2019; Le Morvan et al., 2021). When considering low-capacity predictors, such as linear models, other theoretically sound strategies consist of decomposing the prediction task with respect to all possible missing patterns (see Le Morvan et al., 2020b; Ayme et al., 2022) or by automatically detecting relevant patterns to predict, thus breaking the combinatorics of such pattern-by-pattern predictors (see the specific NeuMiss architecture in Le Morvan et al., 2020a). Proved to be nearly optimal (Ayme et al., 2022), such approaches are likely to be robust to very pessimistic missing data scenarios. Inherently, they do not scale with high-dimensional data sets, as the variety of missing patterns explodes. Another direction is advocated in (Agarwal et al., 2019) relying on principal component regression (PCR) in order to train linear models with missing inputs. However, out-of-sample prediction in such a case requires to retrain the predictor on the training and test sets (to perform a global PC analysis), which strongly departs from classical ML algorithms massively used in practice.

In this paper, we focus on the high-dimensional regime of linear predictors, which will appear to be more favorable to handling missing values via simple and cheap imputation methods, in particular in the missing completely at random (MCAR) case.

**High-dimensional linear models** In supervised learning with complete inputs, when training a parametric method (such as a linear model) in a high-dimensional framework, one often resorts to an $\ell^2$ or ridge regularization technique. On the one hand, such regularization fastens the optimization procedure (via its convergence rate) (Dieuleveut et al., 2017); on the other hand, it also improves the generalization capabilities of the trained predictor (Caponnetto and De Vito, 2007; Hsu et al., 2012). In general, this second point holds for explicit $\ell^2$-regularization, but some works also emphasize the ability of optimization algorithms to induce an implicit regularization, e.g., via early stopping (Yao et al., 2007) and more recently via gradient strategies in interpolation regimes (Bartlett et al., 2020; Chizat and Bach, 2020; Pesme et al., 2021).

**Contributions** For supervised learning purposes, we consider a zero-imputation strategy consisting in replacing input missing entries by zero, and we formalize the induced bias on a regression task (Section 2). When the missing values are said Missing Completely At Random (MCAR), we prove that zero imputation, used prior to training a linear model, introduces an implicit regularization closely related to that of ridge regression (Section 3). This bias is exemplified to be negligible in settings commonly encountered in high-dimensional regimes, e.g., when the inputs admit a low-rank covariance matrix. We then advocate for the choice of an averaged stochastic gradient algorithm (SGD) applied on zero-imputed data (Section 4). Indeed, such a predictor, being computationally efficient, remains particularly relevant for high-dimensional learning. For such a strategy, we establish a generalization bound valid for all $d, n$, in which the impact of imputation on MCAR data is soothed when $d \gg \sqrt{n}$.

These theoretical results legitimate the widespread imputation approach, adopted by most practitioners, and are corroborated by numerical experiments in Section 5. All proofs are to be found in the Appendix.

# 2 Background and motivation

## 2.1 General setting and notations

In the context of supervised learning, consider $n \in \mathbb{N}$ input/output observations $((X_i, Y_i))_{i\in[n]}$, i.i.d. copies of a generic pair $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$. By some abuse of notation, we always use $X_i$ with $i \in [n]$ to denote the $i$-th observation living in $\mathbb{R}^d$, and $X_j$ (or $X_k$) with $j \in [d]$ (or $k \in [d]$) to denote the $j$-th (or $k$-th) coordinate of the generic input $X$ (see Section A for notations).

**Missing values** In real data sets, the input covariates $(X_i)_{i\in[n]}$ are often only partially observed. To code for this missing information, we introduce the random vector $P \in \{0, 1\}^d$, referred to as mask or missing pattern, and such that $P_j = 0$ if the $j$-th coordinate of $X$, $X_j$, is missing and $P_j = 1$ otherwise. The random vectors $P_1, \ldots, P_n$ are assumed to be i.i.d. copies of a generic random variable $P \in \{0, 1\}^d$ and the missing patterns of $X_1, \ldots, X_n$. Note that we assume that the output is always observed and only entries of the input vectors can be missing. Missing data are usually classified into 3 types, initially introduced by (Rubin, 1976). In this paper, we focus on the MCAR assumption where missing patterns and (underlying) inputs are independent.

**Assumption 1** (Missing Completely At Random - MCAR)**.** The pair $(X, Y)$ and the missing pattern $P$ associated to $X$ are independent.

For $j \in [d]$, we define $\rho_j := \mathbb{P}(P_j = 1)$, i.e., $1 - \rho_j$ is the expected proportion of missing values on the $j$-th feature. A particular case of MCAR data requires, not only the independence of the mask and the data, but also the independence between all mask components, as follows.

**Assumption 1'** (Ho-MCAR: MCAR pattern with independent homogeneous components)**.** The pair $(X, Y)$ and the missing pattern $P$ associated to $X$ are independent, and the distribution of $P$ satisfies $P \sim \mathcal{B}(\rho)^{\otimes d}$ for $0 < \rho \leq 1$, with $1 - \rho$ the expected proportion of missing values, and $\mathcal{B}$ the Bernoulli distribution.

**Naive imputation of covariates** A common way to handle missing values for any learning task is to first impute missing data, to obtain a complete dataset, to which standard ML algorithms can then be applied. In particular, constant imputation (using the empirical mean or an oracle constant provided by experts) is very common among practitioners. In this paper, we consider, even for noncentered distributions, the naive imputation by zero, so that the imputed-by-0 observation $(X_{\text{imp}})_i$, for $i \in [n]$, is given by

$$(X_{\text{imp}})_i = P_i \odot X_i. \tag{1}$$

**Risk** Let $f : \mathbb{R}^d \to \mathbb{R}$ be a measurable prediction function, based on a complete $d$-dimensional input. Its predictive performance can be measured through its quadratic risk,

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

Accordingly, we let $f^\star(X) = \mathbb{E}[Y|X]$ be the Bayes predictor for the complete case and $R^\star$ the associated risk.

In the presence of missing data, one can still use the predictor function $f$, applied to the imputed-by-0 input $X_{\text{imp}}$, resulting in the prediction $f(X_{\text{imp}})$. In such a setting, the risk of $f$, acting on the imputed data, is defined by

$$R_{\text{imp}}(f) := \mathbb{E}\left[(Y - f(X_{\text{imp}}))^2\right]. \tag{3}$$

For the class $\mathcal{F}$ of linear prediction functions from $\mathbb{R}^d$ to $\mathbb{R}$, we respectively define

$$R^\star(\mathcal{F}) = \inf_{f \in \mathcal{F}} R(f), \tag{4}$$

and

$$R^\star_{\text{imp}}(\mathcal{F}) = \inf_{f \in \mathcal{F}} R_{\text{imp}}(f), \tag{5}$$

as the infimum over the class $\mathcal{F}$ with respectively complete and imputed-by-0 input data.

For any linear prediction function defined by $f_\theta(x) = \theta^\top x$ for any $x \in \mathbb{R}^d$ and a fixed $\theta \in \mathbb{R}^d$, as $f_\theta$ is completely determined by the parameter $\theta$, we make the abuse of notation of $R(\theta)$ to designate $R(f_\theta)$ (and $R_{\text{imp}}(\theta)$ for $R_{\text{imp}}(f_\theta)$). We also let $\theta^\star \in \mathbb{R}^d$ (resp. $\theta^\star_{\text{imp}}$) be a parameter achieving the best risk on the class of linear functions, i.e., such that $R^\star(\mathcal{F}) = R(\theta^\star)$ (resp. $R^\star_{\text{imp}}(\mathcal{F}) = R_{\text{imp}}(\theta^\star_{\text{imp}})$).

**Imputation bias** Even if the prepocessing step consisting of imputing the missing data by 0 is often used in practice, this imputation technique can introduce a bias in the prediction. We formalize this *imputation bias* as

$$B_{\text{imp}}(\mathcal{F}) := R^\star_{\text{imp}}(\mathcal{F}) - R^\star(\mathcal{F}). \tag{6}$$

This quantity represents the difference in predictive performance between the best predictor on complete data and that on imputed-by-0 inputs. In particular, if this quantity is small, the risk of the best predictor on imputed data is close to that of the best predictor when all data are available. Note that, in presence of missing values, one might be interested in the Bayes predictor

$$f^\star_{\text{mis}}(X_{\text{imp}}, P) = \mathbb{E}[Y|X_{\text{imp}}, P]. \tag{7}$$

and its associated risk $R^\star_{\text{mis}}$.

**Lemma 2.1.** *Assume that regression model $Y = f^\star(X) + \epsilon$ is such that $\epsilon$ and $P$ are independent, then $R^\star \leq R^\star_{\text{mis}}$.*

Intuitively, under the classical assumption $\varepsilon \perp\!\!\!\perp P$ (see Josse et al., 2019), which is a verified under Assumption 1, missing data ineluctably deteriorates the original prediction problem. As a direct consequence, for a well-specified linear model on the complete case $f^\star \in \mathcal{F}$,

$$R_{\text{imp}}(\mathcal{F}) - R^\star_{\text{mis}} \leq B_{\text{imp}}(\mathcal{F}). \tag{8}$$

Consequently, in this paper, we focus our analysis on the bias (and excess risk) associated to impute-then-regress strategies with respect to the complete-case problem (right-hand side term of (8)) thus controlling the excess risk of imputation with respect to the missing data scenario (left-hand side term of (8)).

In a nutshell, the quantity $B_{\text{imp}}(\mathcal{F})$ thus represents how missing values, handled with zero imputation, increase the difficulty of the learning problem. This effect can be tempered in a high-dimensional regime, as rigorously studied in Section 3. To give some intuition, let us now study the following toy example.

*Example* 2.2. Assume an extremely redundant setting in which all covariates are equal, that is, for all $j \in [d]$, $X_j = X_1$ with $\mathbb{E}\left[X_1^2\right] = 1$. Also assume that the output is such that $Y = X_1$ and that Assumption 1' holds with $\rho = 1/2$. In this scenario, due to the input redundancy, all $\theta$ satisfying $\sum_{j=1}^d \theta_j = 1$ minimize $\theta \mapsto R(\theta)$. Letting, for example, $\theta_1 = (1, 0, ..., 0)^\top$, we have $R^\star = R(\theta_1) = 0$ but

$$R_{\text{imp}}(\theta_1) = \mathbb{E}\left[(X_1 - P_1 X_1)^2\right] = \frac{1}{2}.$$

This choice of $\theta_1$ introduces an irreducible discrepancy between the risk computed on the imputed data and the Bayes risk $R^\star = 0$. Another choice of parameter could actually help to close this gap. Indeed, by exploiting the redundancy in covariates, the parameter $\theta_2 = (2/d, 2/d, ..., 2/d)^\top$ (which is not a minimizer of the initial risk anymore) gives

$$R_{\text{imp}}(\theta_2) = \mathbb{E}\left[\left(X_1 - \frac{2}{d}\sum_{j=1}^d P_j X_j\right)^2\right] = \frac{1}{d},$$

so that the imputation bias $B_{\text{imp}}(\mathcal{F})$ is bounded by $1/d$, tending to zero as the dimension increases. Two other important observations on this example follow. First, this bound is still valid if $\mathbb{E}X_1 \neq 0$, thus the imputation by 0 is still relevant even for non-centered data. Second, we remark that $\|\theta_2\|_2^2 = 4/d$, thus good candidates to predict with imputation seem to be of small norm in high dimension. This will be proved for more general settings, in Section 4.

The purpose of this paper is to generalize the phenomenon described in Example 2.2 to less stringent settings. In light of this example, we focus our analysis on scenarios for which some information is shared across input variables: for linear models, correlation plays such a role.

**Covariance matrix** For a generic complete input $X \in \mathbb{R}^d$, call $\Sigma := \mathbb{E}\left[XX^\top\right]$ the associated covariance matrix, admitting the following singular value decomposition

$$\Sigma = \sum_{j=1}^{d} \lambda_j v_j v_j^\top, \tag{9}$$

where $\lambda_j$ (resp. $v_j$) are singular values (resp. singular vectors) of $\Sigma$ and such that $\lambda_1 \geq ... \geq \lambda_d$. The associated pseudo-norm is given by, for all $\theta \in \mathbb{R}^d$,

$$\|\theta\|_\Sigma^2 := \theta^\top \Sigma \theta = \sum_{j=1}^{d} \lambda_j (v_j^\top \theta)^2.$$

For the best linear prediction, $Y = X^\top \theta^\star + \epsilon$, and the noise satisfies $\mathbb{E}[\epsilon X] = 0$ (first order condition). Denoting $\mathbb{E}[\epsilon^2] = \sigma^2$, we have

$$\mathbb{E}Y^2 = \|\theta^\star\|_\Sigma^2 + \sigma^2 = \sum_{j=1}^{d} \lambda_j (v_j^\top \theta^\star)^2 + \sigma^2. \tag{10}$$

The quantity $\lambda_j (v_j^\top \theta^\star)^2$ can be therefore interpreted as the part of the variance explained by the singular direction $v_j$.

*Remark* 2.3. Note that, in the setting of Example 2.2, $\Sigma$ has a unique positive singular values $\lambda_1 = d$, that is to say, all of the variance is concentrated on the first singular direction. Actually, our analysis will stress out that a proper decay of singular values leads to low imputation biases.

Furthermore, for the rest of our analysis, we need the following assumptions on the second-order moments of $X$.

**Assumption 2.** $\exists L < \infty$ such that, $\forall j \in [d]$, $\mathbb{E}[X_j^2] \leq L^2$.

**Assumption 3.** $\exists \ell > 0$ such that, $\forall j \in [d]$, $\mathbb{E}[X_j^2] \geq \ell^2$.

For example, Assumption 2 and 3 hold with $L^2 = \ell^2 = 1$ with normalized data.

# 3 Imputation bias for linear models

## 3.1 Implicit regularization of imputation

Ridge regression, widely used in high-dimensional settings, and notably for its computational purposes, amounts to form an $\ell_2$-penalized version of the least square estimator:

$$\hat{\theta}_\lambda \in \underset{\theta \in \mathbb{R}^d}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - f_\theta(X_i)\right)^2 + \lambda \|\theta\|_2^2 \right\},$$

where $\lambda > 0$ is the penalization parameter. The associated generalization risk can be written as

$$R_\lambda(\theta) := R(\theta) + \lambda \|\theta\|_2^2 .$$

Proposition 3.1 establishes a link between imputation and ridge penalization.

**Proposition 3.1.** *Under Assumption 1, let $V$ be the covariance matrix of $P$ ($V_{ij} = \mathrm{Cov}(P_i, P_j)$) and $H = \mathrm{diag}(\rho_1, \ldots, \rho_d)$, with $\rho_j = \mathbb{P}(P_j = 1)$. Then, for all $\theta$,*

$$R_{\mathrm{imp}}(\theta) = R\left(H\theta\right) + \|\theta\|^2_{V \odot \Sigma} .$$

*In particular, under Assumptions 1', 2 and 3 when $L^2 = \ell^2$,*

$$R_{\mathrm{imp}}(\theta) = R\left(\rho\theta\right) + L^2\rho(1-\rho)\|\theta\|_2^2 . \tag{11}$$

This result highlights the implicit $\ell^2$-regularization at work: performing standard regression on zero-imputed ho-MCAR data can be seen as performing a ridge regression on complete data, whose strength $\lambda$ depends on the missing values proportion. More precisely, using Equation (11), the optimal predictor $\theta^\star_{\mathrm{imp}}$ working with imputed samples verifies

$$\theta^\star_{\mathrm{imp}} = \frac{1}{L^2\rho} \operatorname*{arg\,min}_{\theta \in \mathbb{R}^d} \left\{ R\left(\theta\right) + \lambda_{\mathrm{imp}} \|\theta\|_2^2 \right\},$$

with $\lambda_{\mathrm{imp}} := L^2\left(\frac{1-\rho}{\rho}\right)$. We exploit this correspondence in Section 3.2 and 3.3 to control the imputation bias.

### 3.2 Imputation bias for linear models with ho-MCAR missing inputs

When the inputs admit ho-MCAR missing patterns (Assumption 1'), the zero-imputation bias $B_{\mathrm{imp}}(\mathcal{F})$ induced in the linear model is controlled by a particular instance of the ridge regression bias (see, e.g., Hsu et al., 2012; Dieuleveut et al., 2017; Mourtada, 2019), defined in general by

$$B_{\mathrm{ridge},\lambda}(\mathcal{F}) := \inf_{\theta \in \mathbb{R}^d} \left\{ R_\lambda(\theta) - R^\star(\mathcal{F}) \right\} \tag{12}$$

$$= \lambda \|\theta^\star\|^2_{\Sigma(\Sigma+\lambda I)^{-1}} . \tag{13}$$

**Theorem 3.2.** *Under Assumption 1', 2, and 3, one has*

$$B_{\mathrm{ridge},\lambda'_{\mathrm{imp}}}(\mathcal{F}) \leq B_{\mathrm{imp}}(\mathcal{F}) \leq B_{\mathrm{ridge},\lambda_{\mathrm{imp}}}(\mathcal{F}),$$

*with $\lambda'_{\mathrm{imp}} := \ell^2\left(\frac{1-\rho}{\rho}\right)$ and $\lambda_{\mathrm{imp}} = L^2\left(\frac{1-\rho}{\rho}\right)$.*

As could be expected from Proposition 3.1, the zero-imputation bias is lower and upper-bounded by the ridge bias, with a penalization constant depending on the fraction of missing values. In the specific case where $\ell^2 = L^2$ (same second-order moment), the imputation bias exactly equals a ridge bias with a constant $L^2(1-\rho)/\rho$. Besides, in the extreme case where there is no missing data ($\rho = 1$) then $\lambda_{\text{imp}} = 0$, and the bias vanishes. On the contrary, if there is a large percentage of missing values ($\rho \to 0$) then $\lambda'_{\text{imp}} \to +\infty$ and the imputation bias amounts to the excess risk of the naive predictor, i.e., $B_{\text{imp}}(\mathcal{F}) = R(0_{\mathbb{R}^d}) - R^\star(\mathcal{F})$. For the intermediate case where half of the data is likely to be missing ($\rho = 1/2$), we obtain $\lambda_{\text{imp}} = L^2$.

Thus, in terms of statistical guarantees, performing linear regression on imputed inputs suffers from a bias comparable to that of a ridge penalization, but with a fixed hyperparameter $\lambda_{\text{imp}}$. Note that, when performing standard ridge regression in a high-dimensional setting, the best theoretical choice of the penalization parameter usually scales as $d/n$ (see Sridharan et al., 2008; Hsu et al., 2012; Mourtada and Rosasco, 2022, for details). If $\rho \gtrsim L^2 \frac{n}{d+n}$ (which is equivalent to $\lambda_{\text{imp}} \lesssim \frac{d}{n}$), the imputation bias remains smaller than that of the ridge regression with the optimal hyperparameter $\lambda = d/n$ (which is commonly accepted in applications). In this context, performing zero-imputation prior to applying a ridge regression allows handling easily missing data without drastically increasing the overall bias.

In turns out that the bias of the ridge regression in random designs, and thus the imputation bias, can be controlled, under classical assumptions about low-rank covariance structures (Caponnetto and De Vito, 2007; Hsu et al., 2012; Dieuleveut et al., 2017). In all following examples, we consider that $\operatorname{Tr}(\Sigma) = d$, which holds in particular for normalized data.

*Example* 3.3 (Low-rank covariance matrix with equal singular values). Consider a covariance matrix with a low rank $r \ll d$ and constant eigenvalues ($\lambda_1 = \cdots = \lambda_r = \frac{d}{r}$). Then $\Sigma(\Sigma + \lambda_{\text{imp}} I)^{-1} \preceq \lambda_r^{-1}\Sigma = \frac{r}{d}\Sigma$ and Theorem 3.2 leads to

$$B_{\text{imp}}(\mathcal{F}) \leq \lambda_{\text{imp}} \frac{r}{d} \|\theta^\star\|_\Sigma^2 .$$

Hence, the imputation bias is small when $r \ll d$ (low-rank setting). Indeed, for a fixed dimension, when the covariance is low-rank, there is a lot of redundancy across variables, which helps counterbalancing missing information in the input variables, thereby reducing the prediction bias.

Note that Example 3.3 ($r \ll d$) is a generalization of Example 2.2 (in which $r = 1$), and is rotation-invariant contrary to the latter.

*Remark* 3.4. A first order condition (see equation (29)) implies that $\|\theta^\star\|_\Sigma^2 + \sigma^2 = \mathbb{E}Y^2 = R(0_{\mathbb{R}^d})$, which is independent of the dimension $d$. Thus, in all our upper bounds, $\|\theta^\star\|_\Sigma^2$ can be replaced by $\mathbb{E}Y^2$, which is dimension-free. Consequently, we can interpret Example 3.3 (and the following examples) upper bound as follows: if $r \ll d$, then the risk of the naive predictor is divided by $d/r \gg 1$. As a consequence, $B_{\text{imp}}$ tends to zero when the dimension increases and the rank is fixed.

*Example* 3.5 (Low-rank covariance matrix compatible with $\theta^\star$ ). Consider a covariance matrix with a low rank $r \ll d$ and assume that $\langle \theta^\star, v_1 \rangle^2 \geq \cdots \geq \langle \theta^\star, v_d \rangle^2$ (meaning that $\theta^\star$ is well represented with the first eigendirections of $\Sigma$), Theorem 3.2 leads to

$$B_{\text{imp}}(\mathcal{F}) \lesssim \lambda_{\text{imp}} \frac{r(\log(r)+1)}{d} \|\theta^\star\|_\Sigma^2 .$$

This result is similar to Example 3.3 (up to a log factor), except that assumptions on the eigenvalues of $\Sigma$ have been replaced by a condition on the compatibility between the covariance structure and $\theta^\star$. If $\theta^\star$ is well explained by the largest eigenvalues then the imputation bias remains low. This underlines that imputation bias does not only depend on the spectral structure of $\Sigma$ but also on $\theta^\star$.

*Example* 3.6 (Spiked model, Johnstone (2001)). In this model, the covariance matrix can be decomposed as $\Sigma = \Sigma_{\leq r} + \Sigma_{>r}$ where $\Sigma_{\leq r}$ corresponds to the low-rank part of the data with large eigenvalues and $\Sigma_{>r}$ to the residual high-dimensional data. Suppose that $\Sigma_{>r} \preceq \eta I$ (small operator norm) and that all non-zero eigenvalues of $\Sigma_{\leq r}$ are equal, then Theorem 3.2 gives

$$B_{\text{imp}}(\mathcal{F}) \leq \frac{\lambda_{\text{imp}}}{1-\eta} \frac{r}{d} \|\theta^\star\|_\Sigma^2 + \eta \|\theta^\star_{>r}\|_2^2 ,$$

where $\theta^\star_{>r}$ is the projection of $\theta^\star$ on the range of $\Sigma_{>r}$. Contrary to Example 3.3, $\Sigma$ is only *approximately* low rank, and one can refer to $r$ as the "effective rank" of $\Sigma$ (see Bartlett et al., 2020). The above upper bound admits a term in $O(r/d)$ (as in Example 3.3), but also suffers from a non-compressible part $\eta \|\theta^\star_{>r}\|_2^2$, due to the presence of residual (potentially noisy) high-dimensional data. Note that, if $\theta^\star_{>r} = 0$ (only the low-dimensional part of the data is informative) then we retrieve the same rate as in Example 3.3.

### 3.3 Imputation bias for linear models and general MCAR settings

Theorem 3.2 holds only for Ho-MCAR settings, which excludes the case of dependence between mask components. To cover the case of dependent variables $P_1, \ldots, P_d$ under Assumption 1, recall $\rho_j := \mathbb{P}(P_j = 1)$ the probability that the component $j$ is not missing, and define the matrix $C \in \mathbb{R}^{d \times d}$ associated to $P$, given by:

$$C_{kj} := \frac{V_{k,j}}{\rho_k \rho_j}, \quad (k,j) \in [d] \times [d]. \tag{14}$$

Furthermore, under Assumption 2, define

$$\Lambda_{\text{imp}} := L^2 \lambda_{\max}(C). \tag{15}$$

The following result establishes an upper bound on the imputation bias for general MCAR settings.

**Proposition 3.7.** *Under Assumption 1 and 2, we have*

$$B_{\text{imp}}(\mathcal{F}) \leq B_{\text{ridge},\Lambda_{\text{imp}}}(\mathcal{F}).$$

The bound on the bias is similar to the one of Theorem 3.2 but appeals to $\lambda = \Lambda_{\text{imp}}$ which takes into account the correlations between the components of missing patterns. Remark that, under Assumption 1', there are no correlation and $\Lambda_{\text{imp}} = L^2 \frac{1-\rho}{\rho}$, thus matching the result in Theorem 3.2. The following examples highlight generic scenarios in which an explicit control on $\Lambda_{\text{imp}}$ is obtained.

*Example* 3.8 (Limited number of correlations). If each missing pattern component is correlated with at most $k-1$ other components then $\Lambda_{\text{imp}} \leq L^2 k \max_{j \in [d]} \left\{ \frac{1-\rho_j}{\rho_j} \right\}$.

*Example* 3.9 (Sampling without replacement). Missing pattern components are sampled as $k$ components without replacement in $[d]$, then $\Lambda_{\text{imp}} = L^2 \frac{k+1}{d-k}$. In particular, if one half of data is missing ($k = \frac{d}{2}$) then $\Lambda_{\text{imp}} \leq 3L^2$.

In conclusion, we proved that the imputation bias is controlled by the ridge bias, with a penalization constant $\Lambda_{\text{imp}}$, under any MCAR settings. More precisely, all examples of the previous section (Examples 3.3, 3.5 and 3.6), relying on a specific structure of the covariance matrix $\Sigma$ and the best predictor $\theta^\star$, are still valid, replacing $\lambda_{\text{imp}}$ by $\Lambda_{\text{imp}}$. Additionally, specifying the missing data generation (as in Examples 3.8 and 3.9) allows us to control the imputation bias, which is then proved to be small in high dimension, for all the above examples.

# 4 SGD on zero-imputed data

Since the imputation bias is only a part of the story, we need to propose a proper estimation strategy for $\theta^\star_{\text{imp}}$. To this aim, we choose to train a linear predictor on imputed samples, using an averaged stochastic gradient algorithm (Polyak and Juditsky, 1992), described below. We then establish generalization bounds on the excess risk of this estimation strategy.

## 4.1 Algorithm

Given an initialization $\theta_0 \in \mathbb{R}^d$ and a constant learning rate $\gamma > 0$, the iterates of the averaged SGD algorithm are given at iteration $t$ by

$$\theta_{\text{imp},t} = \left[ I - \gamma X_{\text{imp},t} X_{\text{imp},t}^\top \right] \theta_{\text{imp},t-1} + \gamma Y_t X_{\text{imp},t}, \tag{16}$$

so that after one pass over the data (early stopping), the final estimator $\bar{\theta}_{\text{imp},n}$ is given by the Polyak-Ruppert average $\bar{\theta}_{\text{imp},n} = \frac{1}{n+1} \sum_{t=1}^{n} \theta_{\text{imp},t}$. Such recursive procedures are suitable for high-dimensional settings, and indicated for model miss-specification (induced here by missing entries), as studied in Bach and Moulines (2013). Besides, they are very competitive for large-scale datasets, since one pass over the data requires $O(dn)$ operations.

## 4.2 Generalization bound

Our aim is to derive a generalization bound on the predictive performance of the above algorithm, trained on zero-imputed data. To do this, we require the following extra assumptions on the complete data.

**Assumption 4.** There exist $\sigma > 0$ and $\kappa > 0$ such that $\mathbb{E}[XX^\top \|X\|_2^2] \preceq \kappa \mathrm{Tr}(\Sigma)\Sigma$ and $\mathbb{E}[\epsilon^2 \|X\|_2^2] \leq \sigma^2 \kappa \mathrm{Tr}(\Sigma)$, where $\epsilon = Y - X^\top \theta^\star$.

Assumption 4 is a classical fourth-moment assumption in stochastic optimization (see Bach and Moulines, 2013; Dieuleveut et al., 2017, for details). Indeed, the first statement in Assumption 4 holds, for example, if $X$ is a Gaussian vector (with $\kappa = 3$) or when $X$ satisfies $\|X\|_2 \leq \kappa \mathrm{Tr}(\Sigma)$ almost surely. The second statement in Assumption 4 holds, for example, if the model is well specified or when the noise $\varepsilon$ is almost surely bounded. Note that if the first part holds then the second part holds with $\sigma^2 \leq 2\mathbb{E}[Y^2] + 2\mathbb{E}[Y^4]^{1/2}$.

Our main result, establishing an upper bound on the risk of SGD applied to zero-imputed data, follows.

**Theorem 4.1.** *Under Assumption 4, choosing a constant learning rate* $\gamma = \frac{1}{\kappa \mathrm{Tr}(\Sigma)\sqrt{n}}$ *leads to*

$$\mathbb{E}\left[R_{\mathrm{imp}}\left(\bar{\theta}_{\mathrm{imp},n}\right)\right] - R^\star(\mathcal{F}) \lesssim \frac{\kappa \mathrm{Tr}(\Sigma)}{\sqrt{n}} \left\|\theta^\star_{\mathrm{imp}} - \theta_0\right\|_2^2 + \frac{\sigma^2 + \|\theta^\star\|_\Sigma^2}{\sqrt{n}} + B_{\mathrm{imp}}(\mathcal{F}),$$

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

Theorem 4.1 gives an upper bound on the difference between the averaged risk $\mathbb{E}[R_{\mathrm{imp}}\left(\bar{\theta}_{\mathrm{imp},n}\right)]$ of the estimated linear predictor with imputed missing values (in both train and test samples) and $R^\star(\mathcal{F})$, the risk of the best linear predictor on the complete case. Interestingly, by Lemma 2.1 and under a well-specified linear model, the latter also holds for $\mathbb{E}\left[R_{\mathrm{imp}}\left(\bar{\theta}_{\mathrm{imp},n}\right)\right] - R^\star_{\mathrm{mis}}$. The generalization bound in Theorem 4.1 takes into account the statistical error of the method as well as the optimization error. More precisely, the upper bound can be decomposed into $(i)$ a bias associated to the initial condition, $(ii)$ a variance term of the considered method, and $(iii)$ the aforementioned imputation bias.

The variance term $(ii)$ depends on the second moment of $Y$ (as $\|\theta^\star\|_\Sigma^2 \leq \mathbb{E}Y^2$) and decreases with a slow rate $1/\sqrt{n}$. As seen in Section 3, the imputation bias is upper-bounded by the ridge bias with penalization parameter $\lambda_{\mathrm{imp}}$, which is controlled in high dimension for low-rank data (see examples in Section 3.2).

The bias $(i)$ due to the initial condition is the most critical. Indeed, $\mathrm{Tr}(\Sigma) = \mathbb{E}[\|X\|_2^2]$ is likely to increase with $d$, e.g., under Assumption 2, $\mathrm{Tr}(\Sigma) \leq dL^2$. Besides, the starting point $\theta_0$ may be far from $\theta^\star_{\mathrm{imp}}$. Fortunately, Lemma 4.2 establishes some properties of $\theta^\star_{\mathrm{imp}}$.

**Lemma 4.2.** *Under Assumptions 1 and 3, let* $V$ *be the covariance matrix of* $P$ *defined in Proposition 3.1. If* $V$ *is invertible, then*

$$\left\|\theta^\star_{\mathrm{imp}}\right\|_2^2 \leq \frac{B_{\mathrm{imp}}(\mathcal{F})}{\ell^2 \lambda_{\min}(V)}. \tag{17}$$

*In particular, under Assumption 1',*

$$\left\|\theta^\star_{\mathrm{imp}}\right\|_2^2 \leq \frac{B_{\mathrm{imp}}(\mathcal{F})}{\ell^2 \rho(1-\rho)}. \tag{18}$$

Lemma 4.2 controls the norm of the optimal predictor $\theta^\star_{\text{imp}}$ by the imputation bias: if the imputation bias is small, then the optimal predictor on zero-imputed data is of low norm. According to Section 3, this holds in particular for high-dimensional settings. Thus, choosing $\theta_0 = 0$ permits us to exploit the upper bound provided by Lemma 4.2 in Theorem 4.1. With such an initialization, the bias due to this initial condition is upper bounded by $\frac{\kappa \text{Tr}(\Sigma)}{\sqrt{n}} \|\theta^\star_{\text{imp}}\|_2^2$. Intuitively, as $\theta^\star_{\text{imp}}$ is in an $\ell^2$-ball of small radius, choosing $\theta_0$ within that ball, e.g. $\theta_0 = 0$ is a good choice.

Taking into account Lemma 4.2, Proposition 4.3 establishes our final upper bound on SGD on zero-imputed data.

**Proposition 4.3.** *Under Assumptions 1', 2, 3 and 4, the predictor $\bar{\theta}_{\text{imp},n}$ resulting from the SGD strategy, defined in Section 4.1, with starting point $\theta_0 = 0$ and learning rate $\gamma = \frac{1}{d\kappa L^2 \sqrt{n}}$, satisfies*

$$\mathbb{E}\left[R_{\text{imp}}\left(\bar{\theta}_{\text{imp},n}\right)\right] - R^\star(\mathcal{F}) \lesssim \left(\frac{L^2}{\ell^2} \frac{\kappa d}{\rho(1-\rho)\sqrt{n}} + 1\right) B_{\text{imp}}(\mathcal{F}) + \frac{\sigma^2 + \|\theta^\star\|_\Sigma^2}{\sqrt{n}}.$$

In this upper bound, the first term encapsulates the imputation bias and the one due to the initial condition, whilst the second one corresponds to the variance of the training procedure. As soon as $d \gg \frac{\ell^2}{L^2} \frac{\rho(1-\rho)\sqrt{n}}{\kappa}$ then the imputation bias is negligible compared to that of the initial condition.

### 4.3 Examples

According to Examples 3.3 and 3.6, $B_{\text{imp}}(\mathcal{F})$ decreases with the dimension, provided that $\Sigma$ or $\beta$ are structured. Strikingly, Corollary 4.4 highlights cases where the upper bound of Proposition 4.3 is actually dimension-free.

**Corollary 4.4.** *Suppose that assumptions of Proposition 4.3 hold. Recall that $\lambda_1 \geq \ldots \geq \lambda_d$ are the eigenvalues of $\Sigma$ associated with the eigenvectors $v_1, \ldots, v_d$.*

*(i)* (Example 3.3 - Low-rank $\Sigma$). *If $\Sigma$ has a low rank $r \ll d$ and equal non-zero singular values, then*

$$\mathbb{E}\left[R_{\text{imp}}\left(\bar{\theta}_{\text{imp},n}\right)\right] - R^\star(\mathcal{F}) \lesssim \frac{L^2}{\ell^2}\left(\frac{L^2}{\ell^2}\frac{\kappa}{\rho\sqrt{n}} + \frac{1-\rho}{d}\right)\frac{r\,\|\theta^\star\|_\Sigma^2}{\rho} + \frac{\sigma^2}{\sqrt{n}}.$$

*(ii)* (Example 3.6 - Spiked model). *If $\Sigma = \Sigma_{\leq r} + \Sigma_{>r}$ with $\Sigma_{>r} \preceq \ell^2 \eta I$, $\Sigma_{\leq r}$ has a low rank $r \ll d$ with equal non-zero singular values, and the projection of $\theta^\star$ on the range of $\Sigma_{>r}$ satisfies $\theta^\star_{>r} = 0$, then*

$$\mathbb{E}\left[R_{\text{imp}}\left(\bar{\theta}_{\text{imp},n}\right)\right] - R^\star(\mathcal{F}) \lesssim \frac{L^2}{\ell^2}\left(\frac{L^2}{\ell^2}\frac{\kappa}{\rho\sqrt{n}} + \frac{1-\rho}{d}\right)\frac{r\,\|\theta^\star\|_\Sigma^2}{\rho(1-\eta)} + \frac{\sigma^2}{\sqrt{n}}.$$

Corollary 4.4 establishes upper bounds on the risk of SGD applied on zero-imputed data, for some particular structures on $\Sigma$ and $\theta^\star$. These bounds take into account the statistical error as well as the optimization one, and are expressed as function of $d$ and $n$. Since $\|\theta^\star\|_\Sigma^2$ is upper bounded by $\mathbb{E}Y^2$ (a dimension-free term), the risks in Corollary 4.4 can also be upper bounded by dimension-free quantities, provided $d > \frac{\ell^2}{L^2}\frac{\rho(1-\rho)\sqrt{n}}{\kappa}$.

Besides, Corollary 4.4 shows that, for $d \gg \frac{\ell^2}{L^2}\frac{\rho(1-\rho)\sqrt{n}}{\kappa}$, the imputation bias is negligible with respect to the stochastic error of SGD. Therefore, for structured problems in high-dimensional settings for which $d \gg \frac{\ell^2}{L^2}\frac{\rho(1-\rho)\sqrt{n}}{\kappa}$, the zero-imputation strategy is consistent, with a slow rate of order $1/\sqrt{n}$.

*Remark* 4.5 (Discussion about slow rates). An important limitation of coupling naive imputation with SGD is that fast convergence rates cannot be reached. Indeed, in large dimensions, the classical fast rate is given by $\mathrm{Tr}(\Sigma(\Sigma+\lambda I)^{-1})/n$ with $\lambda$ the penalization hyper-parameter. The quantity $\mathrm{Tr}(\Sigma(\Sigma+\lambda I)^{-1})$, often called degrees of freedom, can be negligible w.r.t. $d$ (for instance when $\Sigma$ has a fast eigenvalue decay). However, when working with an imputed dataset, the covariance matrix of the data is not $\Sigma$ anymore, but $\Sigma_{\mathrm{imp}} = \mathbb{E}X_{\mathrm{imp}}X_{\mathrm{imp}}^\top$. Therefore, in the case of Assumption 1' (Ho-MCAR), all the eigenvalues of $\Sigma_{\mathrm{imp}}$ are larger than $\rho(1-\rho)$ (preventing the eigenvalues decay obtained when working with complete inputs). By concavity of the degrees of freedom (on positive semi-definite matrix), we can show that $\mathrm{Tr}(\Sigma_{\mathrm{imp}}(\Sigma_{\mathrm{imp}}+\lambda I)^{-1}) \geq \frac{d\rho(1-\rho)}{1+\lambda}$, hindering traditional fast rates.

**Link with dropout** Dropout is a classical regularization technique used in deep learning, consisting in randomly discarding some neurons at each SGD iteration (Srivastava et al., 2014). Regularization properties of dropout have attracted a lot of attention (e.g., Gal and Ghahramani, 2016). Interestingly, setting a neuron to 0 on the input layer is equivalent to masking the corresponding feature. Running SGD (as in Section 4) on a stream of zero-imputed data is thus equivalent to training a neural network with no hidden layer, a single output neuron, and dropout on the input layer. Our theoretical analysis describes the implicit regularization impact of dropout in that very particular case. Interestingly, this can also be applied to the fine-tuning of the last layer of any regression network structure.

## 5 Numerical experiments

**Data simulation** We generate $n = 500$ complete input data according to a normal distribution with two different covariance structures. First, in the **low-rank** setting (Ex. 3.3 and 3.5), the output is formed as $Y = \beta^\top Z + \epsilon$, with $\beta \in \mathbb{R}^r$, $Z \sim \mathcal{N}(0, I_r)$ and $\epsilon \sim \mathcal{N}(0, 2)$, and the inputs are given by $X = AZ + \mu$, with a full rank matrix $A \in \mathbb{R}^{d\times r}$ and a mean vector $\mu \in \mathbb{R}^d$. Note that the dimension $d$ varies in the experiments, while $r = 5$ is kept fixed. Besides, the full model can be rewritten as $Y = X^\top\theta^\star + \epsilon$ with $\theta^\star = (A^\dagger)^\top\beta$ where $A^\dagger$ is the Moore-Penrose inverse of $A$. Secondly, in the **spiked model** (Ex. 3.6), the input and the output are decomposed as $X = (X_1, X_2) \in \mathbb{R}^{d/2} \times \mathbb{R}^{d/2}$ and $Y = Y_1 + Y_2$, where $(X_1, Y_1)$

is generated according to the low-rank model above and $(X_2, Y_2)$ is given by a linear model $Y_2 = \theta_2^\top X_2$ and $X_2 \sim \mathcal{N}(0, I_{d/2})$, choosing $\|\theta_2\| = 0.2$.

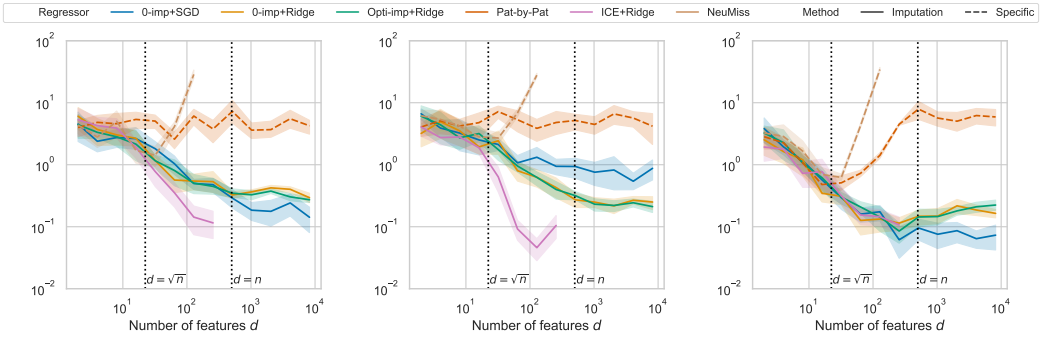
Two missing data scenarios, with a proportion $\rho$ of observed entries equal to 50%, are simulated according to (i) the Ho-MCAR setting (Assumption 1'); and to (ii) the self-masking MNAR setting, which departs significantly from the MCAR case as the presence of missing data depends on the underlying value itself. More precisely, set $\alpha \in \mathbb{R}^d$ such that, for all $j \in [d]$, $\mathbb{P}(P_j = 1|X) = (1 + e^{-\alpha_j X_j})^{-1}$ and $\mathbb{E}[P_j] = 0.5$ (50% of missing data on average per components).

**Regressors** For two-step strategies, different imputers are combined with different regressors. The considered imputers are: the zero imputation method (**0-imp**) complying with the theoretical analysis developed in this paper, the optimal imputation by a constant for each input variable (**Opti-imp**), obtained by training a linear model on the augmented data $(P \odot X, P)$ (see Le Morvan et al., 2020b, Proposition 3.1), and single imputation by chained equations (**ICE**, (Van Buuren and Groothuis-Oudshoorn, 2011))[1]. The subsequent regressors, implemented in scikit-learn (Pedregosa et al., 2011), are either the averaged SGD (**SGD**, package `SGDRegressor`) with $\theta_0 = 0$ and $\gamma = (d\sqrt{n})^{-1}$ (see Proposition 4.3, or the ridge regressor (with a leave-one-out cross-validation, package `ridge`). Two specific methods that do not resort to prior imputation are also assessed: a pattern-by-pattern regressor (Le Morvan et al., 2020b; Ayme et al., 2022) (**Pat-by-Pat**) and a neural network architecture (**NeuMiss**) (Le Morvan et al., 2020a) specifically designed to handle missing data in linear prediction.

**Numerical results** In Figure 1 (a) and (b), we consider Ho-MCAR patterns with Gaussian inputs with resp. a low-rank and spiked covariance matrix. The 2-step strategies perform remarkably well, with the ICE imputer on the top of the podium, highly appropriate to the type of data (MCAR Gaussian) in play. Nonetheless, the naive imputation by zero remains competitive in terms of predictive performance and is computationally efficient, with a complexity of $O(nd)$, especially compared to ICE, whose complexity is of order $n^2d^3$. Regarding Figure 1 (b), we note that ridge regression outperforms SGD for large $d$. Note that, in the regime where $d \geq \sqrt{n}$, the imputation bias is negligible w.r.t. to the method bias, the latter being lower in the case of ridge regression. This highlights the benefit of explicit ridge regularization (with a tuned hyperparameter) over the implicit regularization induced by the imputation.

In practice, missing data are not always of the Ho-MCAR type, we compare therefore the different algorithms on self-masked data. In Figure 1 (c), we note that specific methods remain competitive for larger $d$ compared to MCAR settings. This was to be expected since those methods were designed to handle complex missing not at random (MNAR) data. However, they still suffer from the curse of dimensionality and turns out to be inefficient in large dimension, compared to all two-step strategies.

[1] `IterativeImputer` in scikit-learn (Pedregosa et al., 2011).




(a) Ho-MCAR + Low-rank model

(b) Ho-MCAR + Spiked model

(c) Self-Masked + Low-rank model

Figure 1: Risk w.r.t. the input dimension (evaluated on $10^4$ test samples) when 50% of the input data is missing. The $y$-axis corresponds to $R_{\text{mis}}(f) - R^\star = \mathbb{E}\left[(Y - f(X_{\text{imp}}, P))^2\right] - \sigma^2$. The averaged risk is depicted over 10 repetitions within a 95% confidence interval.

## 6 Discussion and conclusion

In this paper, we study the impact of zero imputation in high-dimensional linear models. We demystify this widespread technique, by exposing its implicit regularization mechanism when dealing with MCAR data. We prove that, in high-dimensional regimes, the induced bias is similar to that of ridge regression, commonly accepted by practitioners. By providing generalization bounds on SGD trained on zero-imputed data, we establish that such two-step procedures are statistically sound, while being computationally appealing.

Theoretical results remain to be established beyond the MCAR case, to properly analyze and compare the different strategies for dealing with missing data in MNAR settings (see Figure 1 (c)). Extending our results to a broader class of functions (escaping linear functions) or even in a classification framework, would be valuable to fully understand the properties of imputation.

## References

Anish Agarwal, Devavrat Shah, Dennis Shen, and Dogyoon Song. On robustness of principal component regression. *Advances in Neural Information Processing Systems*, 32, 2019.

Alexis Ayme, Claire Boyer, Aymeric Dieuleveut, and Erwan Scornet. Near-optimal rate of consistency for linear models with missing values. In *International Conference on Machine Learning*, pages 1211–1243. PMLR, 2022.

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

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

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

Lenaic Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on Learning Theory*, pages 1305–1338. PMLR, 2020.

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

Yarin Gal and Zoubin Ghahramani. A theoretically grounded application of dropout in recurrent neural networks. *Advances in neural information processing systems*, 29, 2016.

Daniel Hsu, Sham M Kakade, and Tong Zhang. Random design analysis of ridge regression. In *Conference on learning theory*, pages 9–1. JMLR Workshop and Conference Proceedings, 2012.

Niels Bruun Ipsen, Pierre-Alexandre Mattei, and Jes Frellsen. How to deal with missing data in supervised deep learning? In *ICLR 2022-10th International Conference on Learning Representations*, 2022.

Iain M. Johnstone. 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.

Julie Josse, Nicolas Prost, Erwan Scornet, and Gaël Varoquaux. On the consistency of supervised learning with missing values. *arXiv preprint arXiv:1902.06931*, 2019.

Marine Le Morvan, Julie Josse, Thomas Moreau, Erwan Scornet, and Gaël Varoquaux. NeuMiss networks: differentiable programming for supervised learning with missing values. In *NeurIPS 2020 - 34th Conference on Neural Information Processing Systems*, Vancouver / Virtual, Canada, December 2020a. URL https://hal.archives-ouvertes.fr/hal-02888867.

Marine Le Morvan, Nicolas Prost, Julie Josse, Erwan Scornet, and Gaël Varoquaux. Linear predictor on linearly-generated data with missing values: non consistency and solutions. In *International Conference on Artificial Intelligence and Statistics*, pages 3165–3174. PMLR, 2020b.

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

Jaouad Mourtada. Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices. *arXiv preprint arXiv:1912.10754*, 2019.

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

Boris Muzellec, Julie Josse, Claire Boyer, and Marco Cuturi. Missing data imputation using optimal transport. In *International Conference on Machine Learning*, pages 7130–7140. PMLR, 2020.

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

Scott Pesme, Loucas Pillaud-Vivien, and Nicolas Flammarion. Implicit bias of sgd for diagonal linear networks: a provable benefit of stochasticity. *Advances in Neural Information Processing Systems*, 34:29218–29230, 2021.

Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

DONALD B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 12 1976. ISSN 0006-3444. doi: 10.1093/biomet/63.3.581. URL https://doi.org/10.1093/biomet/63.3.581.

Karthik Sridharan, Shai Shalev-Shwartz, and Nathan Srebro. Fast rates for regularized objectives. *Advances in neural information processing systems*, 21, 2008.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Stef Van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45:1–67, 2011.

Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

Jinsung Yoon, James Jordon, and Mihaela Schaar. Gain: Missing data imputation using generative adversarial nets. In *International conference on machine learning*, pages 5689–5698. PMLR, 2018.

# A Notations

For two vectors (or matrices) $a, b$, we denote by $a \odot b$ the Hadamard product (or componentwise 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.

Table 1: Notations

| | |
|---|---|
| $P$ | Mask |
| $\mathcal{F}$ | Set of linear functions |
| $B_{\text{imp}}$ | Imputation bias |
| $\Sigma$ | $\mathbb{E}XX^\top$ |
| $\lambda_j$ | eigenvalues of $\Sigma$ |
| $v_j$ | eigendirections of $\Sigma$ |
| $\Sigma_P$ | $\mathbb{E}PP^\top$ |
| $L^2$ | the largest second moments $\max_j \mathbb{E}X_j^2$ (Assumption 2) |
| $\ell^2$ | the smallest second moments $\min_j \mathbb{E}X_j^2$ (Assumption 3) |
| $\theta^\star$ | Best linear predictor on complete data |
| $\theta^\star_{\text{imp}}$ | Best linear predictor on imputed data |
| $r$ | Rank of $\Sigma$ |
| $\rho_j$ | Theoretical proportion of observed entries for the $j$-th variable in a MCAR setting |
| $V$ | Covariance matrix associated to the missing patterns |
| $C$ | Covariance matrix $V$ renormalized by $(\rho_j)_j$ defined in (14) |
| $\kappa$ | Kurtosis of the input $X$ |

# B Proof of the main results

## B.1 Proof of Lemma 2.1

The proof is based on the definition of the conditional expectation, and given that

$$R^\star = \mathbb{E}\left[(Y - \mathbb{E}[Y|X])^2\right].$$

Note that $\mathbb{E}[Y|X,P] = \mathbb{E}[f^\star(X)+\epsilon|X,P] = \mathbb{E}[f^\star(X)|X,P] = f^\star(X)$ (by independence of $\epsilon$ and $P$). Therefore,

$$\begin{aligned} R^\star &= \mathbb{E}\left[(Y - f^\star(X))^2\right] \\ &\leq \mathbb{E}\left[(Y - \mathbb{E}[Y|X,P])^2\right] \\ &\leq \mathbb{E}\left[(Y - \mathbb{E}[Y|X_{\text{imp}},P])^2\right] \\ &\leq R^\star_{\text{mis}}, \end{aligned}$$

using that $\mathbb{E}[Y|X_{\text{imp}},P]$ is a measurable function of $(X,P)$.

## B.2 Preliminary lemmas

**Notation** Let $X_a$ be a random variable of law $\mathcal{L}_a$ (a modified version of the law of the underlying input $X$) on $\mathbb{R}^d$, and for $f \in \mathcal{F}$ define

$$R_a(f) = \mathbb{E}\left[(Y - f(X_a))^2\right],$$

the associate risk. The Bayes risk is given by

$$R_a^\star(\mathcal{F}) = \inf_{f\in\mathcal{F}} \mathbb{E}\left[(Y - f(X_a))^2\right],$$

if the infimum is reached, we denote by $f_a^\star \in \arg\min_{f\in\mathcal{F}} R_a(f)$. The discrepancy between both risks, involving either the modified input $X_a$ or the initial input $X$, can be measured through the following bias:

$$B_a = R_a^\star(\mathcal{F}) - R^\star(\mathcal{F}).$$

**General decomposition** The idea of the next lemma is to compare $R_a(f)$ with the true risk $R(f)$.

**Lemma B.1.** *If $(X_a \perp\!\!\!\perp Y)|X$, then, for all $\theta \in \mathbb{R}^d$,*

$$R_a(f_\theta) = R(g_\theta) + \|\theta\|_\Gamma^2,$$

*where $g_\theta(X) = \theta^\top \mathbb{E}[X_a|X]$ and $\Gamma = \mathbb{E}\left[(X_a - \mathbb{E}[X_a|X])(X_a - \mathbb{E}[X_a|X])^\top\right]$ the integrated conditional covariance matrix. In consequence, if there exists an invertible linear application $H$ such that, $\mathbb{E}[X_a|X] = H^{-1}X$, then*

- *For all $\theta \in \mathbb{R}^d$, $g_\theta$ is a linear function and*

$$R_a^\star(\mathcal{F}) = \inf_{\theta\in\mathbb{R}^d} \left\{R(f_\theta) + \|\theta\|^2_{H^\top \Gamma H}\right\}. \tag{19}$$

- *If* $\lambda_{\max}(H\Gamma H^\top) \leq \Lambda$, *then*

$$B_a(\mathcal{F}) \leq \inf_{\theta \in \mathbb{R}^d} \left\{ R(f_\theta) + \Lambda \|\theta\|_2^2 \right\} = B_{\text{ridge},\Lambda}. \tag{20}$$

- *If* $\lambda_{\min}(\Gamma) \geq \mu > 0$, *then*

$$\|\theta_a^\star\|_2^2 \leq \frac{B_a(\mathcal{F})}{\mu}. \tag{21}$$

*Remark* B.2. Equation (21) is crucial because a bound on the bias $B_a(\mathcal{F})$ actually gives a bound for $\|\theta_a^\star\|_2^2$ too. This will be of particular interest for Theorem 4.1.

*Proof.*

$$\begin{aligned}
R_a\left(f_\theta\right) &= \mathbb{E}\left[\left(Y - \theta^\top X_a\right)^2\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(Y - \mathbb{E}\left[\theta^\top X_a | X\right] + \mathbb{E}\left[\theta^\top X_a | X\right] - \theta^\top X_a\right)^2 \Big| X\right]\right] \\
&= \mathbb{E}\left[\left(Y - \mathbb{E}\left[\theta^\top X_a | X\right]\right)^2\right] + \mathbb{E}\left[\mathbb{E}\left[\left(\mathbb{E}\left[\theta^\top X_a | X\right] - \theta^\top X_a\right)^2 \Big| X\right]\right] \\
&= \mathbb{E}\left[(Y - g_\theta(X))^2\right] + \mathbb{E}\left[\mathbb{E}\left[\left(\mathbb{E}\left[\theta^\top X_a | X\right] - \theta^\top X_a\right)^2 \Big| X\right]\right] \\
&= R(g_\theta) + \mathbb{E}\left[\mathbb{E}\left[\left(\mathbb{E}\left[\theta^\top X_a | X\right] - \theta^\top X_a\right)^2 \Big| X\right]\right].
\end{aligned}$$

since $\mathbb{E}\left[\mathbb{E}\left[\theta^\top X_a | X\right] - \theta^\top X_a | X\right] = 0$. Furthermore,

$$\begin{aligned}
\mathbb{E}\left[\mathbb{E}\left[\left(\mathbb{E}\left[\theta^\top X_a | Z\right] - \theta^\top X_a\right)^2 | X\right]\right] &= \theta^\top \mathbb{E}\left[\left(\mathbb{E}\left[X_a | X\right] - X_a\right)\left(\mathbb{E}\left[X_a | X\right] - X_a\right)^\top\right]\theta \\
&= \mathbb{E}\left[\theta^\top \mathbb{E}\left[\left(\mathbb{E}\left[X_a | X\right] - X_a\right)\left(\mathbb{E}\left[X_a | X\right] - X_a\right)^\top | X\right]\theta\right] \\
&= \mathbb{E}\left[\|\theta\|^2_{\mathbb{E}\left[(\mathbb{E}[X_a|X]-X_a)(\mathbb{E}[X_a|X]-X_a)^\top | X\right]}\right] \\
&= \mathbb{E}\left[\|\theta\|_\Gamma^2\right].
\end{aligned}$$

Finally,

$$R_a\left(f_\theta\right) = R(g_\theta) + \|\theta\|_\Gamma^2.$$

Assume that an invertible matrix $H$ exists such that $g_\theta(X) = \theta^\top H^{-1} X$, thus $g_\theta$ is a linear function. Equation (19) is then obtained by using a change of variable: $\theta' = (H^{-1})^\top \theta = (H^\top)^{-1}\theta$ and $\theta = H^\top \theta'$. Thus, we have $g_{\theta'}(X) = \theta^\top X = f_\theta(X)$ and

$$\begin{aligned}
R_a\left(f_{\theta'}\right) &= R(f_\theta) + \|H^\top \theta'\|_\Gamma^2 \\
&= R(f_\theta) + \|\theta'\|^2_{H\Gamma H^\top}.
\end{aligned}$$

Then using $H\Gamma H^\top \preceq \Lambda I$ proves (20). Note that, without resorting to the previous change of variable, the bias can be written as

$$B_a(\mathcal{F}) = R\left(g_{\theta_a^\star}\right) - R\left(f_{\theta^\star}\right) + \|\theta_a^\star\|_\Gamma^2 . \tag{22}$$

By linearity of $g_{\theta_a^\star}$, $R\left(g_{\theta_a^\star}\right) \geq R\left(f_{\theta^\star}\right) = R^\star(\mathcal{F})$ (because $g_{\theta_a^\star} \in \mathcal{F}$).

Thus, $\|\theta_a^\star\|_\Gamma^2 \leq B_a(\mathcal{F})$. Assuming $\mu I \preceq \Gamma$ gives (21), as

$$\mu \left\|\theta_a^\star\right\|^2 \leq \|\theta_a^\star\|_\Gamma^2 \leq B_a(\mathcal{F}).$$

□

### B.3 Proof of Section 3

We consider the case of imputed-by-0 data, i.e.,

$$X_{\text{imp}} = P \odot X.$$

Under the MCAR setting (Assumption 1),

$$\mathbb{E}\left[X_{\text{imp}}|X\right] = H^{-1}X,$$

with $H = \text{diag}(\rho_1^{-1}, ..., \rho_{\text{d}}^{-1})$ (variables always missing are discarded) and $(\rho_j)_{j\in[d]}$ the observation rates associated to each input variable.

*Proof of Proposition 3.1.* For $i, j \in [d]$,

$$\begin{aligned}
\Gamma_{ij} &= \mathbb{E}\left[\left(\left(X_{\text{imp}}\right)_i - \mathbb{E}\left[\left(X_{\text{imp}}\right)_i|X\right]\right)\left(\left(X_{\text{imp}}\right)_j - \mathbb{E}\left[\left(X_{\text{imp}}\right)_j|X\right]\right)\right] \\
&= \mathbb{E}\left[X_i X_j (P_i - \mathbb{E}P_i)(P_j - \mathbb{E}P_j)\right] \\
&= \mathbb{E}\left[X_i X_j\right] \text{Cov}(P_i, P_j), \\
&= \Sigma_{ij} V_{ij}
\end{aligned} \tag{23}$$

since $P$ and $X$ are independent and with $V$ defined in Proposition 3.1. Therefore, applying Lemma B.1 with $\Gamma = \Sigma \odot V$ proves the first part of Proposition 3.1. Regarding the second part, under the Ho-MCAR assumption, one has $V = \rho(1-\rho)I$, thus $\Gamma = \rho(1-\rho)\text{diag}(\Sigma)$. Furthermore, if $L^2 = \ell^2$, then $\text{diag}(\Sigma) = L^2 I$ which gives $\Gamma = L^2\rho(1-\rho)I$. □

*Proof of Theorem 3.2 and Proposition 3.7.* Under Assumption 1, since $H$ is a diagonal matrix,

$$H^\top \Gamma H = \Sigma \odot C,$$

where $C$ is defined in Equation (14).

- Under Assumption 1', the matrix $C$ satisfies $C = \frac{1-\rho}{\rho} I$. Moreover, under Assumption 2 (resp. Assumption 3), one has $\Sigma \odot C \preceq \frac{1-\rho}{\rho} L^2 I = \lambda_{\text{imp}}$ (resp. $\Sigma \odot C \succeq \frac{1-\rho}{\rho} \ell^2 I = \lambda'_{\text{imp}}$) using (19), we obtain

$$\inf_{\theta \in \mathbb{R}^d} \left\{ R(\theta) + \lambda'_{\text{imp}} \|\theta\|_2^2 \right\} \leq R^\star_{\text{imp}} \leq \inf_{\theta \in \mathbb{R}^d} \left\{ R(\theta) + \lambda_{\text{imp}} \|\theta\|_2^2 \right\}.$$

  Subtracting $R^\star(\mathcal{F})$, one has

$$B_{\text{ridge},\lambda'_{\text{imp}}} \leq B_{\text{imp}} \leq B_{\text{ridge},\lambda_{\text{imp}}},$$

  which concludes the proof of Theorem 3.2.

- Under Assumption 1, we have $H\Gamma H^\top = \Sigma \odot C$. Using Lemma E.2, we obtain for all $\theta$,

$$\|\theta\|^2_{H\Gamma H^\top} = \|\theta\|^2_{\Sigma \odot C} \leq \lambda_{\max}(C) \|\theta\|^2_{\text{diag}(\Sigma)}.$$

  Under Assumption 2, we have $\text{diag}(\Sigma) \preceq L^2 I$, thus

$$\|\theta\|^2_{H\Gamma H^\top} \leq L^2 \lambda_{\max}(C) \|\theta\|_2^2.$$

  This shows that $\lambda_{\max}(H\Gamma H^\top) \leq L^2 \lambda_{\max}(C) = \Lambda_{\text{imp}}$ We conclude on Proposition 3.7 using Equation (19).

□

### B.4 Proof of Lemma 4.2

*Proof.* Using (23), we have $\Gamma = V \odot \Sigma$. Using that $\lambda_{\min}(V) I \preceq V$, by Lemma E.1, we obtain

$$\lambda_{\min}(V) I \odot \Sigma \preceq \Gamma,$$

and equivalently $\lambda_{\min}(V) \odot \text{diag}(\Sigma) \preceq \Gamma$. Under Assumption 3, we have $\ell^2 I \preceq \text{diag}(\Sigma)$, thus

$$\ell^2 \lambda_{\min}(V) I \preceq \Gamma.$$

Therefore, $\lambda_{\min}(\Gamma) \geq \ell^2 \lambda_{\min}(V)$. Thus, using (21), we obtain the first part of Lemma 4.2:

$$\ell^2 \lambda_{\min}(V) \left\|\theta^\star_{\text{imp}}\right\|_2^2 \leq B_{\text{imp}}(\mathcal{F}). \tag{24}$$

Under Assumption 1', $\lambda_{\min}(V) = \rho(1-\rho)$, so that

$$\ell^2 \rho(1-\rho) \left\|\theta^\star_{\text{imp}}\right\|_2^2 \leq B_{\text{imp}}(\mathcal{F}), \tag{25}$$

which proves the second part of Lemma 4.2.

□

# C Stochastic gradient descent

## C.1 Proof of Theorem 4.1

**Lemma C.1.** *Assume $(x_n, \xi_n) \in \mathcal{H} \times \mathcal{H}$ are $\mathcal{F}_n$-measurable for a sequence of increasing $\sigma$-fields $(\mathcal{F}_n)$, $n \geqslant 1$. Assume that $\mathbb{E}[\xi_n \mid \mathcal{F}_{n-1}] = 0, \mathbb{E}\left[\|\xi_n\|^2 \mid \mathcal{F}_{n-1}\right]$ is finite and $\mathbb{E}\left[\left(\|x_n\|^2 x_n \otimes x_n\right) \mid \mathcal{F}_{n-1}\right] \preccurlyeq R^2 H$, with $\mathbb{E}[x_n \otimes x_n \mid \mathcal{F}_{n-1}] = H$ for all $n \geqslant 1$, for some $R > 0$ and invertible operator $H$. Consider the recursion $\alpha_n = (I - \gamma x_n \otimes x_n)\,\alpha_{n-1} + \gamma \xi_n$, with $\gamma R^2 \leqslant 1$. Then:*

$$\left(1 - \gamma R^2\right) \mathbb{E}\left[\langle \bar{\alpha}_{n-1}, H \bar{\alpha}_{n-1} \rangle\right] + \frac{1}{2n\gamma} \mathbb{E}\,\|\alpha_n\|^2 \leqslant \frac{1}{2n\gamma}\,\|\alpha_0\|^2 + \frac{\gamma}{n} \sum_{k=1}^{n} \mathbb{E}\,\|\xi_k\|^2 .$$

*Proof.* The idea is to use Lemma C.1 with

- $x_k = X_{\text{imp},k}$, $y_k = Y_k$
- $H = \Sigma_{\text{imp}} = \mathbb{E}\left[X_{\text{imp},k} X_{\text{imp},k}^\top\right] = \Sigma_P \odot \Sigma$ where $\Sigma_P = \mathbb{E}\left[PP^\top\right]$
- $\alpha_k = \theta_{\text{imp},k} - \theta^\star_{\text{imp}}$
- $\xi_k = X_{\text{imp},k}(Y_k - X_{\text{imp},k}^\top \theta^\star_{\text{imp}})$
- $\gamma = \frac{1}{2R^2\sqrt{n}}$
- $R^2 = \kappa \text{Tr}(\Sigma)$

We can show, with these notations, that recursion (16) leads to recursion $\alpha_n = (I - \gamma x_n \otimes x_n)\,\alpha_{n-1} + \gamma \xi_n$ with $\alpha_0 = \theta_0 - \theta^\star_{\text{imp}}$. Now, let's check the assumption of Lemma C.1.

- Let show that $\mathbb{E}\left[X_{\text{imp}} X_{\text{imp}}^\top \|X_{\text{imp}}\|_2^2\right] \preceq R^2 \Sigma_{\text{imp}}$. Indeed,

$$\mathbb{E}\left[X_{\text{imp}} X_{\text{imp}}^\top \|X_{\text{imp}}\|_2^2\right] \preceq \mathbb{E}\left[X_{\text{imp}} X_{\text{imp}}^\top \|X\|_2^2\right],$$

  using that $\|X_{\text{imp}}\|_2^2 \leq \|X\|_2^2$, and $0 \preccurlyeq X_{\text{imp}} X_{\text{imp}}^\top$. Then,

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

According to Assumption 4, $\mathbb{E}\left[XX^\top \|X\|_2^2\right] \preceq R^2\Sigma$, and Lemma E.1 lead to

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

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

$$\begin{aligned}
\mathbb{E}\left[\|\xi\|_2^2\right] &= \mathbb{E}\left[\epsilon_{\text{imp}}^2 \|X_{\text{imp}}\|_2^2\right] \\
&\leq 3\mathbb{E}\left[\left(\left(X^\top\theta^\star\right)^2 + \epsilon^2 + \left(X_{\text{imp}}^\top\theta_{\text{imp}}^\star\right)^2\right)\|X_{\text{imp}}\|_2^2\right] \\
&\leq 3\left(\mathbb{E}\left[\left(X^\top\theta^\star\right)^2 \|X\|_2^2\right] + \mathbb{E}\left[\epsilon^2\|X\|_2^2\right]\right. \\
&\qquad \left. + \mathbb{E}\left[\left(X_{\text{imp}}^\top\theta_{\text{imp}}^\star\right)^2\|X_{\text{imp}}\|_2^2\right]\right).
\end{aligned}$$

  Let remark that, using Assumption 4

$$\begin{aligned}
\mathbb{E}\left[\left(X^\top\theta^\star\right)^2\|X\|_2^2\right] &= \mathbb{E}\left[\theta^{\star\top}\left(XX^\top\|X\|_2^2\right)\theta^\star\right]\|\theta^\star\|_\Sigma^2 \\
&\leq R^2\theta^{\star\top}\Sigma\theta \\
&= R^2\|\theta^\star\|_\Sigma^2.
\end{aligned}$$

  Using the first point, by the same way, $\mathbb{E}\left[\left(X_{\text{imp}}^\top\theta_{\text{imp}}^\star\right)^2\|X_{\text{imp}}\|_2^2\right] \leq \left\|\theta_{\text{imp}}^\star\right\|_{\Sigma_{\text{imp}}}^2$. By Assumption 4, we have also than $\mathbb{E}\left[\epsilon^2\|X\|_2^2\right] \leq \sigma^2R^2$. Thus,

$$\begin{aligned}
\mathbb{E}\left[\|\xi\|_2^2\right] &\leq 3R^2\left(\sigma^2 + \|\theta^\star\|_\Sigma^2 + \left\|\theta_{\text{imp}}^\star\right\|_{\Sigma_{\text{imp}}}^2\right) \\
&\leq 3R^2\left(\sigma^2 + 2\|\theta^\star\|_\Sigma^2\right),
\end{aligned}$$

  because $\|\theta^\star\|_\Sigma^2 = R\left(\theta^\star\right) \leq R_{\text{imp}}\left(\theta_{\text{imp}}^\star\right) = \left\|\theta_{\text{imp}}^\star\right\|_{\Sigma_{\text{imp}}}^2$.

Consequently we can apply Lemma C.1, to obtain

$$\begin{aligned}
&\left(1 - \frac{1}{2\sqrt{n}}\right)\mathbb{E}\left[\left\langle\bar{\theta}_{\text{imp},n} - \theta_{\text{imp}}^\star, \Sigma_{\text{imp}}(\bar{\theta}_{\text{imp},n} - \theta_{\text{imp}}^\star)\right\rangle\right] + \frac{1}{2n\gamma}\mathbb{E}\left\|\theta_{\text{imp},n} - \theta_{\text{imp}}^\star\right\|^2 \\
&\leqslant \frac{1}{2n\gamma}\left\|\theta_{\text{imp}}^\star - \theta_0\right\|^2 + \frac{\gamma}{n}\sum_{k=1}^{n}\mathbb{E}\left\|\xi_k\right\|^2.
\end{aligned}$$

The choice $\gamma = \frac{1}{2R^2\sqrt{n}}$ leads to

$$\mathbb{E}\left\|\bar{\theta}_{\text{imp},n} - \theta^\star_{\text{imp}}\right\|^2_{\Sigma_{\text{imp}}} \leqslant \frac{2R^2}{\sqrt{n}}\left\|\theta^\star_{\text{imp}} - \theta_0\right\|^2 + 4\frac{\sigma^2 + 2\left\|\theta^\star\right\|^2_\Sigma}{\sqrt{n}}.$$

We conclude on Theorem 4.1 using that,

$$\begin{aligned}\mathbb{E}\left[R_{\text{imp}}\left(\bar{\theta}_{\text{imp}}\right)\right] - R^\star &= \mathbb{E}\left[R_{\text{imp}}\left(\bar{\theta}_{\text{imp}}\right)\right] - R^\star_{\text{imp}} + R^\star_{\text{imp}} - R^\star \\ &= \mathbb{E}\left\|\bar{\theta}_{\text{imp},n} - \theta^\star_{\text{imp}}\right\|^2_{\Sigma_{\text{imp}}} + B_{\text{imp}}.\end{aligned}$$

□

## C.2 Proof of Proposition 4.3 and Corollary 4.4

*Proof of Proposition 4.3.* First, under Assumption 2, $\text{Tr}(\Sigma) \leq dL^2$. Then, initial conditions term with $\theta_0 = 0$,

$$\frac{\kappa\text{Tr}(\Sigma)}{\sqrt{n}}\left\|\theta^\star_{\text{imp}}\right\|^2_2 \leq \frac{\kappa L^2 d}{\sqrt{n}\ell^2\rho(1-\rho)}B_{\text{imp}}(\mathcal{F}), \tag{26}$$

using Lemma 4.2. We obtain Proposition 4.3 using inequality above in Theorem 4.1. □

*proof of Corollary 4.4.* We obtain the upper bounds considered that: according to Theorem 3.2, $B_{\text{imp}} \leq B_{\text{ridge},\lambda_{\text{imp}}}$; under Assumption 3, $\text{Tr}(\Sigma) \geq d\ell^2$. Then, we put together Proposition 4.3 and ridge bias bound (see Appendix D). □

## C.3 Miscellaneous

**Proposition C.2.** *If $X$ statisfies $\mathbb{E}\left[XX^\top\left\|X\right\|^2_2\right] \preceq \kappa\text{Tr}(\Sigma)\Sigma$, then $\mathbb{E}\left[\epsilon^2\left\|X\right\|^2_2\right] \leq \sigma^2\kappa\text{Tr}(\Sigma)$ with $\sigma^2 \leq 2\mathbb{E}[Y^2] + 2\mathbb{E}[Y^4]^{1/2}$.*

*Proof.*

$$\begin{aligned}\mathbb{E}\left[\epsilon^2\left\|X\right\|^2_2\right] &= \mathbb{E}\left[\left(Y - X^\top\theta^\star\right)^2\left\|X\right\|^2_2\right] \\ &\leq 2\mathbb{E}\left[\left(\left(X^\top\theta^\star\right)^2 + Y^2\right)\left\|X\right\|^2_2\right] \\ &\leq 2\mathbb{E}\left[Y^2\left\|X\right\|^2_2\right] + 2\mathbb{E}\left[\left(X^\top\theta^\star\right)^2\left\|X\right\|^2_2\right].\end{aligned}$$

Regarding the first term, by Cauchy Schwarz,

$$\begin{aligned}\mathbb{E}\left[Y^2\left\|X\right\|^2_2\right]^2 &\leq \mathbb{E}\left[Y^4\right]\mathbb{E}\left[\left\|X\right\|^4_2\right] \\ &\leq \mathbb{E}\left[Y^4\right]\mathbb{E}\left[\text{Tr}\left(XX^\top\left\|X\right\|^2_2\right)\right] \\ &\leq \mathbb{E}\left[Y^4\right]\kappa\text{Tr}(\Sigma)^2.\end{aligned}$$

As for the second term,

$$\begin{aligned}\mathbb{E}\left[\left(X^\top\theta^\star\right)^2\|X\|_2^2\right] &= \mathbb{E}\left[(\theta^\star)^\top XX^\top\|X\|_2^2\,\theta^\star\right]\\ &\le \kappa\mathrm{Tr}(\Sigma)\mathbb{E}\left[(\theta^\star)^\top\Sigma\theta^\star\right]\\ &\le \kappa\mathrm{Tr}(\Sigma)\|\theta^\star\|_2^2\,.\end{aligned}$$

$$\mathbb{E}\left[\epsilon^2\|X\|_2^2\right] \le \mathbb{E}\left[Y^4\right]^{\frac{1}{2}}\kappa\mathrm{Tr}(\Sigma) + \kappa\mathrm{Tr}(\Sigma)\|\theta^\star\|_\Sigma^2 \le \sigma^2\kappa\mathrm{Tr}(\Sigma)\|\theta^\star\|_\Sigma^2\,.$$

□

# D Details on examples

Recall that

$$B_{\mathrm{ridge},\lambda}(\mathcal{F}) = \lambda\|\theta^\star\|^2_{\Sigma(\Sigma+\lambda I)^{-1}} \tag{27}$$

$$= \lambda\sum_{j=1}^{d}\frac{\lambda_j}{\lambda_j+\lambda}(v_j^\top\theta^\star)^2. \tag{28}$$

## D.1 Low-rank covariance matrix (Example 3.3)

**Proposition D.1** (Low-rank covariance matrix with equal singular values)**.** *Consider a covariance matrix with a low rank $r \ll d$ and constant eigenvalues ($\lambda_1 = \lambda_2 = ... = \lambda_r$). Then,*

$$B_{\mathrm{ridge},\lambda}(\mathcal{F}) = \lambda\frac{r}{\mathrm{Tr}(\Sigma)}\|\theta^\star\|_\Sigma^2\,.$$

*Proof.* Using that $\lambda_1 = \cdots = \lambda_r$ and $\sum_{j=1}^r\lambda_j = \mathrm{Tr}(\Sigma)$, we have $\lambda_1 = \cdots = \lambda_r = \frac{\mathrm{Tr}(\Sigma)}{r}$. Then $\Sigma(\Sigma+\lambda I)^{-1} \preceq \lambda_r^{-1}\Sigma = \frac{r}{\mathrm{Tr}(\Sigma)}\Sigma$. Thus,

$$B_{\mathrm{ridge},\lambda}(\mathcal{F}) = \lambda\|\theta^\star\|^2_{\Sigma(\Sigma+\lambda I)^{-1}} = \lambda\frac{r}{\mathrm{Tr}(\Sigma)}\|\theta^\star\|_\Sigma^2\,.$$

□

## D.2 Low-rank covariance matrix compatible with $\theta^\star$ (Example 3.5)

**Proposition D.2** (Low-rank covariance matrix compatible with $\theta^\star$)**.** *Consider a covariance matrix with a low rank $r \ll d$ and assume that $\langle\theta^\star, v_1\rangle^2 \ge \cdots \ge \langle\theta^\star, v_d\rangle^2$, then*

$$B_{\mathrm{ridge},\lambda}(\mathcal{F}) \lesssim \lambda\frac{r(\log(r)+1)}{\mathrm{Tr}(\Sigma)}\|\theta^\star\|_\Sigma^2\,.$$

*Proof.* Recall that

$$\|\theta^\star\|_\Sigma^2 = \sum_{j=1}^{d} \lambda_j (v_j^\top \theta^\star)^2. \tag{29}$$

Under the assumptions of Example 3.5, using that $(\lambda_j)_j$ and $\left((v_j^\top \theta^\star)^2\right)_j$ are decreasing, then for all $k \in [r]$,

$$\sum_{j=1}^{k} \lambda_j (v_k^\top \theta^\star)^2 \leq \|\theta^\star\|_\Sigma^2.$$

Thus, for all $k \in [r]$,

$$(v_k^\top \theta^\star)^2 \leq \frac{\|\theta^\star\|_\Sigma^2}{\sum_{j=1}^{k} \lambda_j}.$$

Using that $\sum_{j=1}^{r} \lambda_j = \mathrm{Tr}(\Sigma)$ and that eigenvalues are decreasing, we have $\sum_{j=1}^{k} \lambda_j \geq \frac{k}{r}\mathrm{Tr}(\Sigma)$ using Lemma E.3. Then

$$\begin{aligned}
B_{\mathrm{ridge},\lambda}(\mathcal{F}) &= \lambda \sum_{k=1}^{r} \frac{\lambda_k}{\lambda_k + \lambda} (v_k^\top \theta^\star)^2 \\
&\leq \lambda \sum_{k=1}^{r} (v_k^\top \theta^\star)^2 \\
&\leq \lambda \|\theta^\star\|_\Sigma^2 \sum_{k=1}^{r} \frac{1}{\sum_{j=1}^{k} \lambda_j} \\
&\leq \lambda \sum_{k=1}^{r} \frac{r}{k\mathrm{Tr}(\Sigma)} \\
&\leq \lambda \frac{r}{\mathrm{Tr}(\Sigma)} \sum_{k=1}^{r} \frac{1}{k} \\
&\lesssim \lambda \frac{r}{\mathrm{Tr}(\Sigma)} (\log(r) + 1),
\end{aligned}$$

by upper-bounding the Euler-Maclaurin formula. □

### D.3 Spiked covariance matrix (Example 3.6)

**Proposition D.3** (Spiked model)**.** *Assume that the covariance matrix is decomposed as* $\Sigma = \Sigma_{\leq r} + \Sigma_{>r}$*. Suppose that* $\Sigma_{>r} \preceq \eta I$ *(small operator norm) and that all non-zero eigenvalues of* $\Sigma_{\leq r}$ *are equal, then*

$$B_{\mathrm{ridge},\lambda}(\mathcal{F}) \leq \frac{r}{\mathrm{Tr}(\Sigma) - d\eta} \|\theta^\star\|_\Sigma^2 + \eta \|\theta^\star_{>}\|_2^2 .$$

*where* $\theta^\star_{>r}$ *is the projection of* $\theta^\star$ *on the range of* $\Sigma_{>r}$*.*

*Proof.* One has

$$\begin{aligned}\Sigma(\Sigma+\lambda I)^{-1} &= \Sigma_{\leq}(\Sigma+\lambda I)^{-1} + \Sigma_{>}(\Sigma+\lambda I)^{-1} \\ &\preceq \Sigma_{\leq}(\Sigma_{\leq}+\lambda I)^{-1} + \Sigma_{>}(\Sigma_{>}+\lambda I)^{-1} \\ &\preceq \frac{1}{\mu}\Sigma_{\leq} + \frac{1}{\lambda}\Sigma_{>}\end{aligned}$$

where $\mu$ is the non-zero eigenvalue of $\Sigma_{\leq}$. Thus,

$$\begin{aligned}B_{\mathrm{ridge},\lambda}(\mathcal{F}) &= \|\theta^\star\|^2_{\lambda\Sigma(\Sigma+\lambda I)^{-1}} \\ &\leq \|\theta^\star\|^2_{\frac{\lambda}{\mu}\Sigma_{\leq}+\Sigma_{>}} \\ &\leq \frac{\lambda}{\mu}\|\theta^\star\|^2_{\Sigma} + \|\theta^\star\|^2_{\Sigma_{>}}\,.\end{aligned}$$

Using that $\lambda_{\max}(\Sigma_{>}) \leq \eta$, we have

$$B_{\mathrm{ridge},\lambda}(\mathcal{F}) \leq \frac{\lambda}{\mu}\|\theta^\star\|^2_{\Sigma} + \eta\,\|\theta^\star_{>}\|_2^2\,.$$

Using Weyl's inequality, for all $j \in [d]$, $\lambda_j(\Sigma_{\leq}+\Sigma_{>}) \leq \lambda_j(\Sigma_{\leq})+\eta$. Summing the previous inequalities, we get

$$\mathrm{Tr}(\Sigma) \leq r\mu + d\eta.$$

Thus,

$$\mu \geq \frac{\mathrm{Tr}(\Sigma) - d\eta}{r}.$$

In consequence,

$$B_{\mathrm{ridge},\lambda}(\mathcal{F}) \leq \frac{r}{\mathrm{Tr}(\Sigma) - d\eta}\|\theta^\star\|^2_{\Sigma} + \eta\,\|\theta^\star_{>}\|_2^2\,.$$

□

# E Technical lemmas

**Lemma E.1.** *Let $A, B, V$ be three symmetric non-negative matrix, if $A \preceq B$ then $A \odot V \preceq B \odot V$.*

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

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

□

**Lemma E.2.** *Let $A, B$ be two non-negative symmetric matrices, then $A \odot B$ is non-negative symmetric and, for all $\theta \in \mathbb{R}^d$:*

$$\|\theta\|^2_{A\odot B} \leq \lambda_{\max}(B)\,\|\theta\|^2_{\mathrm{diag(A)}}\,,$$

*where* diag(A) *is the diagonal matrix containing the diagonal terms of $A$.*

*Proof.* Let $X \sim \mathcal{N}(0, A)$, thus $A = \mathbb{E}\left[XX^\top\right]$, then for $\theta \in \mathbb{R}^d$

$$
\begin{aligned}
\|\theta\|_{A\odot B}^2 &= \theta^\top A \odot B\theta \\
&= \theta^\top \left(\left(\mathbb{E}XX^\top\right) \odot B\right)\theta \\
&= \mathbb{E}\left[\theta^\top \left(\left(XX^\top\right) \odot B\right)\theta\right] \\
&= \mathbb{E}\left[\sum_{i,j} \theta_i \left(\left(XX^\top\right) \odot B\right)_{ij} \theta_j\right] \\
&= \mathbb{E}\left[\sum_{i,j} \theta_i X_i X_j B_{ij} \theta_j\right] \\
&= \mathbb{E}\left[\sum_{i,j} \left(\theta_i X_i\right)\left(\theta_j X_j\right) B_{ij}\right] \\
&= \mathbb{E}\left[\left(X \odot \theta\right)^\top B \left(X \odot \theta\right)\right] \\
&\geq 0,
\end{aligned}
$$

using that $B$ is positive. Thus $A \odot B$ is positive. Furthermore,

$$
\begin{aligned}
\|\theta\|_{A\odot B}^2 &= \mathbb{E}\left[\left(X \odot \theta\right)^\top B \left(X \odot \theta\right)\right] \\
&\leq \lambda_{\max}(B)\mathbb{E}\left[\left(X \odot \theta\right)^\top \left(X \odot \theta\right)\right] \\
&= \lambda_{\max}(B)\mathbb{E}\left[\sum_i \theta_i^2 X_i^2\right] \\
&= \lambda_{\max}(B)\sum_i \theta_i^2 \mathbb{E}\left[X_i^2\right] \\
&= \lambda_{\max}(B)\sum_i \theta_i^2 A_{ii} \\
&= \lambda_{\max}(B)\left\|\theta\right\|_{\mathrm{diag}(A)}^2.
\end{aligned}
$$

□

**Lemma E.3.** *Let $(v_j)_{j\in[d]}$ a non-decreasing sequence of positive number, and $S = \sum_{j=1}^d v_j$, for all $k \in [d]$,*

$$
\sum_{j=1}^k v_j \geq \frac{k}{d} S.
$$

*Proof.* We use a absurd m, if $\sum_{j=1}^k v_j < \frac{k}{d}S$. Then, using that $(v_j)_{j\in[d]}$are non-decreasing,

$$
k v_k < \frac{k}{d} S.
$$

Thus $v_{k+1} < \frac{1}{d}S$, summing last elements,

$$\sum_{j=r+1}^{d} v_j < \frac{d-r}{d}S.$$

Then,

$$S = \sum_{j=1}^{k} v_j = \sum_{j=1}^{r} v_j + \sum_{j=r+1}^{d} v_j < \frac{k}{d}S + \frac{d-r}{d}S = S.$$

Thus, this is absurd. □