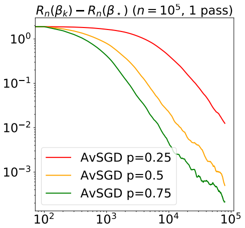
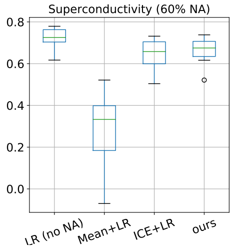
We thank all the reviewers for their insightful comments which help us to improve the article. They acknowledged the theoretical work and the relevance of the method developed for the ML community. In the final version, we will integrate the comments of form given by the reviewers (typos, notations, theorem numbering and English writing).

1 New experimental results (R2)





Increasing missing data proportions on synthetic data. The figure on the top shows the results of our approach on the same data as in Section 5.1 with 25% (green), 50% (orange) and 75% (red) missing values. It shows that the more missing data there are, the more the convergence rate deteriorates. This was expected, as the established theoretical upper bound for the convergence (Th. 4) increases as p gets smaller.

Comparison to other methods. The paper does already include a comparison with the theoretically-grounded competitor for linear regression with missing covariates: the EM algorithm (see Fig. 4), but contrary to our approach, EM requires a distributional assumption on the covariates and prevents from observations in high dimension. Moreover, the proposed strategy results in a practical, efficient and theoretically sound algorithm. For completeness, we ran comparisons on 5 UCI datasets (Boston, Concrete slump, Diabetes, Superconductivity, Wine) to 2-step heuristics: imputation of the covariates (by the mean or ICE[1]) and linear regression (LR) on the completed data, varying the percentage of NA. The coefficient of determination R^2 is plotted on the Figure besides (thus higher is better), for the superconductivity dataset, with 60% of missing values. This is representative of other results where our method greatly outperforms the mean imputation and is better or in the same order of magnitude of ICE (that also does not scale well).

2 Bibliographic addendum

IPW. (R3) We will include the pointed references on IPW, thank you. While the motivation for reweighting may meet our debiasing will, we would like to point out some differences with our work: in the IPW literature, weighting is often used to rebalance samples with missing outcome, while we consider missing values in all the learning task covariates which can be in high dimension. In addition, the expression we use to debiase SGD, and more specifically its gradients (Eq. 4), although involving weights, is more complex than simply weighting the data.

Missing values in deep networks. (R4) A reference to Joonyoung et al., in which they propose a heuristic to debiase zero-imputation in neural networks, will be added in the final version. Due to the high non-linearity of their setting, their debiasing trick significantly departs from ours, and their proposed algorithm comes with no guarantee of convergence.

3 Relevance of the tackled problem and discussion (R1, R3, R4)

Discussion on the bounded feature assumption (R1) As mentioned in the paper, the bounded features assumption is mostly made to ease the readability. It can be actually relaxed into a bound "in average": more precisely only bounds on moments of the random variable can be required, see e.g. Section 6.1. in [1].

Using estimators of $(p_j)_j$. (R3) The available implementation already includes the use of estimated proportions $(\hat{p}_j)_j$ of NA in each column, instead of the oracle ones $(p_j)_j$, and so do all the numerical experiments, always leading to convergent estimators. In addition, we can show that, for the estimator $\hat{\beta}_{k,\hat{p}_j}$ built using our algorithm with $(\hat{p}_j)_j$, *we preserve the optimal* $1/k$ *convergence rate.* More precisely, the supplementary risk w.r.t. the iterate $\bar{\beta}_k$ built with the true $(p_j)_j$ is $\mathbb{E}[R(\hat{\beta}_{k,\hat{p}_j}) - R(\bar{\beta}_k)] = O(1/kp_{\min}^5)$. A remark and the proof of this preliminary[2] result will be added.

Towards a more general MCAR setting. (R3) We indeed considered a specific MCAR setting, in which the missing-pattern random variables were independent. We thank R3 from raising this interesting issue: we can extend the setting to allow coordinates to be dependently missing. To do so, we propose a new way of constructing debiased versions of gradients (Lemma 1), as $\tilde{g}_k(\beta) := (W \odot (\tilde{X}_{k:}\tilde{X}_{k:}^T))\beta - y_k P^{-1} \tilde{X}_{k:}$ with $W \in \mathbb{R}^{d \times d}$, and $W_{ij} := 1/\mathbb{E}[\delta_{ki}\delta_{kj}]$. The noise structure in the SGD iterates (Lemma 2(3.)) becomes even more technical to control. Regarding practical implementation, the matrix W can be estimated, in particular using low-rank strategies on the missing pattern matrix.

Impact of our work and extensions. (R3,R4) Despite the apparent simplicity of the considered setting, we tackle the important issue of performing large-scale Ridge regression with missing covariates (which was not resolved yet), using SGD, thereby handling large dimensionality and online (missing) data. This actually required a lot of technicality: the state-of-the-art proofs cannot be applied directly. An efficient code is also provided. SGD being a keystone block of ML, this paves the way of many developments to handle ubiquitous missing data (e.g., variance reduction, deep learning, etc.). Moreover, theoretical locks are raised, cleared up or resolved: (i) theoretical challenge of multi-pass ERM, (ii) computational optimality with missing data (Th. 4), (iii) information theory optimality (see Sec. 4.4, with an open question to establish a lower bound).

[1] Dieuleveut, Durmus, and Bach. Bridging the gap between constant step size SGD and markov chains. *Ann. Statist.*, 48, 2020.

[1] `sklearn.impute.IterativeImputer` [2] the dependency in $p_{\min}$ may not be optimal and the proof requires the invertibility of the covariates' covariance matrix H and bounded iterates.