



Debiasing Stochastic Gradient Descent to handle missing values

Aude Sportisse, Claire Boyer, Aymeric Dieuleveut, Julie Josse

► To cite this version:

Aude Sportisse, Claire Boyer, Aymeric Dieuleveut, Julie Josse. Debiasing Stochastic Gradient Descent to handle missing values. NeurIPS 2020 - 34th Conference on Neural Information Processing Systems, Dec 2020, Vancouver, Canada. hal-02483651v2

HAL Id: hal-02483651

<https://hal.archives-ouvertes.fr/hal-02483651v2>

Submitted on 4 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Debiasing Averaged Stochastic Gradient Descent to handle missing values

Aude Sportisse⁽¹⁾, Claire Boyer^(1,2), Aymeric Dieuleveut⁽³⁾, Julie Josse^(3,4)
(¹) Sorbonne Université, (²) ENS Paris, (³) Ecole Polytechnique, (⁴) Google France

Abstract

Stochastic gradient algorithm is a key ingredient of many machine learning methods, particularly appropriate for large-scale learning. However, a major caveat of large data is their incompleteness. We propose an averaged stochastic gradient algorithm handling missing values in linear models. This approach has the merit to be free from the need of any data distribution modeling and to account for heterogeneous missing proportion. In both streaming and finite-sample settings, we prove that this algorithm achieves convergence rate of $\mathcal{O}(\frac{1}{n})$ at the iteration n , the same as without missing values. We show the convergence behavior and the relevance of the algorithm not only on synthetic data but also on real data sets, including those collected from medical register.

1 Introduction

Stochastic gradient algorithms (SGD) [20] play a central role in machine learning problems, due to their cheap computational cost and memory per iteration. There is a vast literature on its variants, for example using averaging of the iterates [19], some robust versions of SGD [18, 11] or adaptive gradient algorithms like Adagrad [6]; and on theoretical guarantees of those methods [16, 1, 5, 22, 8, 17]. More globally, averaging strategies have been used to stabilize the algorithm behaviour and reduce the impact of the noise, giving better convergence rates without requiring strong convexity.

The problem of missing values is ubiquitous in large scale data analysis. One of the key challenges in the presence of missing data is to deal with the half-discrete nature of the data which can be seen as a mixed of continuous data (observed values) and categorical data (the missing values). In particular for gradient-based methods, the risk minimization with incomplete data becomes intractable and the usual results cannot be directly applied.

Context. In this paper, we consider a linear regression model, for $i \geq 1$,

$$y_i = X_{i\cdot}^T \beta^* + \epsilon_i, \tag{1}$$

parametrized by $\beta^* \in \mathbb{R}^d$, where $y_i \in \mathbb{R}$, $\epsilon_i \in \mathbb{R}$ is a real-valued centered noise and $X_{i\cdot} \in \mathbb{R}^d$ stands for the real covariates of the i -th observation. The $(X_{i\cdot})$'s are assumed to be only partially known, since some covariates may be missing: our objective is to derive stochastic algorithms for estimating the parameters of the linear model, which handle missing data, and come with strong theoretical guarantees on excess risk.

Related works. There is a rich literature on handling missing values [13] and yet there are still some challenges even for linear regression models. This is all the more true as we consider such models for large sample size or in high dimension. There are very few regularized versions of regression that can deal with missing values. A classical approach to estimating parameters with missing values consists in maximizing the observed likelihood, using for instance an Expectation Maximization algorithm [3]. Even if this approach can be implemented to scale for large datasets

see for instance [2], one of its main drawbacks is to rely on strong parametric assumptions for the covariates distributions. Another popular strategy to fix the missing values issue consists in predicting the missing values to get a completed data and then in applying the desired method. However matrix completion is a different problem from estimating parameters and can lead to uncontrolled bias and undervalued variance of the estimate [13]. In the regression framework, Jones [10] studied the bias induced by naive imputation.

In the settings of the Dantzig selector [21] and LASSO [14], another solution consists in naively imputing by 0 the incomplete matrix and modifying the algorithm used in the complete case to account for the imputation error. Such a strategy has also been studied by Ma and Needell [15] for SGD in the context of linear regression with missing values and with finite samples: the authors used debiased gradients, in the same spirit as the covariance matrix debiasing considered by Loh and Wainwright [14] in a context of sparse linear regression, or by Koltchinskii et al. [12] for matrix completion. This modified version of the SGD algorithm [15] is conjectured to converge in expectation to the ordinary least squares estimator, achieving the rate of $\mathcal{O}(\frac{\log n}{\mu n})$ at iteration n for the excess empirical risk, assumed to be μ -strongly convex in that work. However, their algorithm requires a step choice relying on the knowledge of the strong-convexity constant μ which is often intractable for large-scale settings.

Contributions.

- We develop a debiased averaged SGD to perform (regularized) linear regression either streaming or with finite samples, when covariates are missing. The approach consists in imputing the covariates with a simple imputation and using debiased gradients accordingly.
- Furthermore, the design is allowed to be contaminated by heterogeneous missing values: each covariate may have a different probability to be missing. This encompasses the classical homogeneous Missing Completely At Random (MCAR) case, where the missingness is independent of any covariate value.
- This algorithm comes with theoretical guarantees: we establish convergence in terms of generalization risk at the rate $1/n$ at iteration n . This rate is remarkable as it is (i) optimal w.r.t. n , (ii) free from any bad condition number (no strong convexity constant is required), and (iii) similar to the rate of averaged SGD without any missing value.
- In terms of performance with respect to the missing entries proportion in large dimension, our strategy results in an error provably several orders of magnitude smaller than the best possible algorithm that would only rely on complete observations.
- We show the relevance of the proposed approach and its convergence behavior on numerical applications and its efficiency on real data; including the TraumaBase[®] dataset to assist doctors in making real-time decisions in the management of severely traumatized patients. The code to reproduce all the simulations and numerical experiments is available on <https://github.com/AudeSportisse/SGD-NA>.

2 Problem setting

In this paper, we consider either the streaming setting, i.e. when the data comes in as it goes along, or the finite-sample setting, i.e. when the data size is fixed and form a finite design matrix $X = (X_{1:} | \dots | X_{n:})^T \in \mathbb{R}^{n \times d}$ ($n > d$). We define $\mathcal{D}_n := \sigma((X_{i:}, y_i), i = 1, \dots, n)$ the σ -field generated by n observations. We also denote \preceq the partial order between self-adjoint operators, such that $A \preceq B$ if $B - A$ is positive semi-definite.

Given observations as in (1) and defining $f_i(\beta) := (\langle X_{i:}, \beta \rangle - y_i)^2 / 2$, the (unknown) linear model parameter satisfies:

$$\beta^* = \arg \min_{\beta \in \mathbb{R}^d} \{R(\beta) := \mathbb{E}_{(X_{i:}, y_i)} [f_i(\beta)]\}, \quad (2)$$

where $\mathbb{E}_{(X_{i:}, y_i)}$ denotes the expectation over the distribution of $(X_{i:}, y_i)$ (which is independent of i as the observations are assumed to be i.i.d.).

In this work, the covariates are assumed to contain missing values, so one in fact observes $X_{i:}^{\text{NA}} \in (\mathbb{R} \cup \{\text{NA}\})^d$ instead of $X_{i:}$, as $X_{i:}^{\text{NA}} := X_{i:} \odot D_{i:} + \text{NA}(\mathbf{1}_d - D_{i:})$, where \odot denotes the element-wise product, $\mathbf{1}_d \in \mathbb{R}^d$ is the vector filled with ones and $D_{i:} \in \{0, 1\}^d$ is a binary vector mask coding for

the presence of missing entries in X_i , i.e. $D_{ij} = 0$ if the (i, j) -entry is missing in X_i , and $D_{ij} = 1$ otherwise. We adopt the convention $\text{NA} \times 0 = 0$ and $\text{NA} \times 1 = \text{NA}$. We consider a *heterogeneous* MCAR setting, i.e. D is modeled with a Bernoulli distribution

$$D = (\delta_{ij})_{1 \leq i \leq n, 1 \leq j \leq d} \quad \text{with} \quad \delta_{ij} \sim \mathcal{B}(p_j), \quad (3)$$

with $1 - p_j$ the probability that the j -th covariate is missing.

The considered approach consists in imputing the incomplete covariates by zero in X_i^{NA} , as $\tilde{X}_i = X_i^{\text{NA}} \odot D_i = X_i \odot D_i$, and in accounting for the imputation error in the subsequent algorithm.

3 Averaged SGD with missing values

The proposed method is detailed in Algorithm 1. The impact of the naive imputation by 0 directly translates into a bias in the gradient. Consequently, at each iteration we use a debiased estimate \tilde{g}_k . In order to stabilize the stochastic algorithm, we consider the Polyak-Ruppert [19] averaged iterates $\bar{\beta}_k = \frac{1}{k+1} \sum_{i=0}^k \beta_i$.

Lemma 1. *Let $(\mathcal{F}_k)_{k \geq 0}$ be the following σ -algebra, $\mathcal{F}_k = \sigma(X_{1:}, y_1, D_1: \dots, X_{k:}, y_k, D_{k:})$. The modified gradient $\tilde{g}_k(\beta_{k-1})$ in Equation (4) is \mathcal{F}_k -measurable and a.s.,*

$$\mathbb{E}[\tilde{g}_k(\beta_{k-1}) \mid \mathcal{F}_{k-1}] = \nabla R(\beta_{k-1}).$$

Lemma 1 is proved in Appendix S2.1. Note that in the case of homogeneous MCAR data, i.e. $p_1 = \dots = p_d = p \in (0, 1)$, the chosen direction at iteration k in Equation (4) boils down to $\frac{1}{p} \tilde{X}_k: \left(\frac{1}{p} \tilde{X}_k^T \beta_k - y_k \right) - \frac{1-p}{p^2} \text{diag} \left(\tilde{X}_k: \tilde{X}_k^T \right) \beta_k$. This meets the classical debiasing terms of covariance matrices [14, 15, 12]. Note also that in the presence of complete observations, meaning that $p = 1$, Algorithm 1 matches the standard least squares stochastic algorithm.

Remark 1 (Ridge regularization) Instead of minimizing the theoretical risk as in (2), we can consider a Ridge regularized formulation: $\min_{\beta \in \mathbb{R}^d} R(\beta) + \lambda \|\beta\|^2$, with $\lambda > 0$. Algorithm 1 is trivially extended to this framework: the debiasing term is not modified since the penalization term does not involve the incomplete data \tilde{X}_i . This is useful in practice as no implementation is available for incomplete ridge regression.

4 Theoretical results

In this section, we prove convergence guarantees for Algorithm 1 in terms of theoretical excess risk, in both the streaming and the finite-sample settings. For the rest of this section, assume the following.

- The observations $(X_{k:}, y_k) \in \mathbb{R}^d \times \mathbb{R}$ are independent and identically distributed.
- $\mathbb{E}[\|X_{k:}\|^2]$ and $\mathbb{E}[\|y_k\|^2]$ are finite.
- Let H be an invertible matrix, defined by $H := \mathbb{E}_{(X_{k:}, y_k)}[X_{k:} X_{k:}^T]$.

The main technical challenge to overcome is proving that the noise in play due to missing values is *structured* and still allows to derive convergence results for a debiased version of averaged SGD. This work builds upon the analysis made by Bach and Moulines [1] for standard SGD strategies.

4.1 Technical results

Bach and Moulines [1] proved that for least-squares regression, averaged SGD converges at rate n^{-1} after n iterations. In order to derive similar results, we prove in addition to Lemma 1, Lemmas 2 and 3:

Algorithm 1 Averaged SGD for Heterogeneous Missing Data

Input: data \tilde{X}, y, α (step size)

Initialize $\beta_0 = 0_d$.

Set $P = \text{diag}((p_j)_{j \in \{1, \dots, d\}}) \in \mathbb{R}^{d \times d}$.

for $k = 1$ **to** n **do**

$$\begin{aligned} \tilde{g}_k(\beta_k) &= P^{-1} \tilde{X}_k: \left(\tilde{X}_k^T P^{-1} \beta_k - y_k \right) \\ &\quad - (I - P) P^{-2} \text{diag} \left(\tilde{X}_k: \tilde{X}_k^T \right) \beta_k \end{aligned} \quad (4)$$

$$\beta_k = \beta_{k-1} - \alpha \tilde{g}_k(\beta_{k-1})$$

$$\bar{\beta}_k = \frac{1}{k+1} \sum_{i=0}^k \beta_i = \frac{k}{k+1} \bar{\beta}_{k-1} + \frac{1}{k+1} \beta_k$$

end for

- Lemma 2 shows that the noise induced by the imputation by zeros and the subsequent transformation results in a *structured noise*. This is the most challenging part technically: having a structured noise is fundamental to obtain convergence rates scaling as n^{-1} – in the unstructured case the convergence speed is only $n^{-1/2}$ [5].
- Lemma 3 shows that the adjusted random gradients $\tilde{g}_k(\beta)$ are almost surely *co-coercive* [24] i.e., for any k , there exists a random “primitive” function \tilde{f}_k which is a.s. convex and smooth, and such that $\tilde{g}_k = \nabla \tilde{f}_k$. Proving that \tilde{f}_k is a.s. convex is an important step which was missing in the analysis of Ma and Needell [15].

Lemma 2. *The additive noise process $(\tilde{g}_k(\beta^*))_k$ with β^* defined in (2) is \mathcal{F}_k -measurable and,*

1. $\forall k \geq 0, \mathbb{E}[\tilde{g}_k(\beta^*) \mid \mathcal{F}_{k-1}] = 0$ a.s..
2. $\forall k \geq 0, \mathbb{E}[\|\tilde{g}_k(\beta^*)\|^2 \mid \mathcal{F}_{k-1}]$ is a.s. finite.
3. $\forall k \geq 0, \mathbb{E}[\tilde{g}_k(\beta^*)\tilde{g}_k(\beta^*)^T] \preceq C(\beta^*) = c(\beta^*)H$.

Sketch of proof (Lemma 2). Property 1 easily followed from Lemma 1 and the definition of β^* . Property 2 can be obtained with similar computations as in [15, Lemma 4]. Property 3 cannot be directly derived from Property 2, since $\tilde{g}_k(\beta^*)\tilde{g}_k(\beta^*)^T \preceq \|\tilde{g}_k(\beta^*)\|^2 I$ leads to an insufficient upper bound. Proof relies on decomposing the external product $\tilde{g}_k(\beta^*)\tilde{g}_k(\beta^*)^T$ in several terms and obtaining the control of each, involving technical computations. \square

Lemma 3. *For all $k \geq 0$, given the binary mask D , the adjusted gradient $\tilde{g}_k(\beta)$ is a.s. $L_{k,D}$ -Lipschitz continuous, i.e. for all $u, v \in \mathbb{R}^d$, $\|\tilde{g}_k(u) - \tilde{g}_k(v)\| \leq L_{k,D}\|u - v\|$ a.s.. Set*

$$L := \sup_{k,D} L_{k,D} \leq \frac{1}{p_m^2} \max_k \|X_k\|^2 \text{ a.s.} \quad (5)$$

In addition, for all $k \geq 0$, $\tilde{g}_k(\beta)$ is almost surely co-coercive.

Lemmas 2 and 3 are respectively proved in Appendices S2.2 and S2.3, and can be combined with Theorem 1 in [1] in order to prove the following theoretical guarantees for Algorithm 1.

4.2 Convergence results

The following theorem quantifies the convergence rate of Algorithm 1 in terms of excess risk.

Theorem 4 (Streaming setting). *Assume that for any i , $\|X_i\| \leq \gamma$ almost surely for some $\gamma > 0$. For any constant step-size $\alpha \leq \frac{1}{2L}$, Algorithm 1 ensures that, for any $k \geq 0$:*

$$\mathbb{E} [R(\bar{\beta}_k) - R(\beta^*)] \leq \frac{1}{2k} \left(\frac{\sqrt{c(\beta^*)d}}{1 - \sqrt{\alpha L}} + \frac{\|\beta_0 - \beta^*\|}{\sqrt{\alpha}} \right)^2,$$

with L given in Equation (24), $p_m = \min_{j=1,\dots,d} p_j$ and

$$c(\beta^*) = \frac{\text{Var}(\epsilon_k)}{p_m^2} + \left(\frac{(2 + 5p_m)(1 - p_m)}{p_m^3} \right) \gamma^2 \|\beta^*\|^2. \quad (6)$$

Note that in Theorem 4, the expectation is taken over the randomness of the observations $(X_i, y_i, D_i)_{1 \leq i \leq k}$. The bounded features assumption in Theorem 4 is mostly convenient for the readability, but it can be relaxed at the price of milder but more technical assumptions and proofs (typically bounds on quadratic mean instead of a.s. bounds).

Remark 2 (Finite-sample setting) Similar results as Theorem 4 can be derived in the case of finite-sample setting. For the sake of clarity, they are made explicit hereafter: for any constant step-size $\alpha \leq \frac{1}{2L}$, Algorithm 1 ensures that for any $k \leq n$:

$$\mathbb{E} [R(\bar{\beta}_k) - R(\beta^*) \mid \mathcal{D}_n] \leq \frac{1}{2k} \left(\frac{\sqrt{c(\beta^*)d}}{1 - \sqrt{\alpha L}} + \frac{\|\beta_0 - \beta^*\|}{\sqrt{\alpha}} \right)^2 \text{ with } L \text{ given in Equation (24) and } c(\beta^*) = \frac{\text{Var}(\epsilon_k)}{p_m^2} + \left(\frac{(2 + 5p_m)(1 - p_m)}{p_m^3} \right) \max_{1 \leq i \leq n} \|X_i\|^2 \|\beta^*\|^2.$$

Convergence rates for the iterates. Note that if a Ridge regularization is considered, the regularized function to minimize $R(\beta) + \lambda \|\beta\|^2$ is 2λ -strongly convex. Theorem 4 and Remark 2 then directly provide the following bound on the iterates: $\mathbb{E} \left[\|\bar{\beta}_k - \beta^*\|^2 \right] \leq \frac{1}{2\lambda k} \left(\frac{\sqrt{c(\beta^*)d}}{1-\sqrt{\alpha L}} + \frac{\|\beta_0 - \beta^*\|}{\sqrt{\alpha}} \right)^2$.

Additional comments. We highlight the following points:

- In Theorem 4, the expected excess risk is upper bounded by (a) a *variance* term, that grows with the noise variance and is increased by the missing values, and (b) a *bias* term, that accounts for the importance of the initial distance between the starting point β_0 and the optimal one β^* .
- The optimal convergence rate is achieved for a *constant* learning rate α . One could for example choose $\alpha = \frac{1}{2L}$, that does *not decrease* with the number of iterations. In such a situation, both the *bias* and *variance* terms scale as k^{-1} . Remark that convergence of the averaged SGD with constant step-size only happens for least squares regression, because the un-averaged iterates converge to a limit distribution whose mean is exactly β^* [1, 4].
- The expected risk scales as n^{-1} after n iterations, without strong convexity constant involved.
- For the generalization risk R , this rate of n^{-1} is known to be statistically optimal for least-squares regression: under reasonable assumptions, no algorithm, even more complex than averaged SGD or without missing observations, can have a better dependence in n [23].
- In the complete case, i.e. when $p_1 = \dots = p_d = 1$, Theorem 4 and remark 2 meet the results from Bach and Moulines [1, Theorem 1]. Indeed, in such a case, $c(\beta^*) = \text{Var}(\epsilon_k)$.
- The noise variance coefficient $c(\beta^*)$ includes (i) a first term as a classical noise one, proportional to the model variance, and increased by the missing values occurrence to $\frac{\text{Var}(\epsilon_k)}{p_m^2}$; (ii) the second term is upper-bounded by $\frac{7(1-p_m)}{p_m^3} \cdot \gamma^2 \|\beta^*\|^2$ corresponds to the multiplicative noise induced by the imputation by 0 and gradient debiasing. It naturally increases as the radius γ^2 of the observations increases (so does the imputation error), and vanishes if there are no missing values ($p_m = 1$).

Remark 3 (Only one epoch) It is important to notice that in a finite-sample setting, as covered by Remark 2, given a maximum number of n observations, our convergence rates are only valid for $k \leq n$: the theoretical bound holds only for one pass on the input/output pairs. Indeed, afterwards, we cannot build unbiased gradients of the risk.

4.3 What about empirical risk minimization (ERM)?

Theoretical locks. Note that the translation of the results in Remark 2 in terms of empirical risk convergence is still an open issue. The heart of the problem is that it seems really difficult to obtain a sequence of unbiased gradients of the empirical risk.

- Indeed, to obtain unbiased gradients, the data should be processed *only once* in Algorithm 1: if we consider the gradient of the loss with respect to an observation k , we obviously need the binary mask D_k and the current point β_{k-1} to be independent for the correction relative to the missing entries to make sense. As a consequence, no sample can be used twice - in fact, running multiple passes over a finite sample could result in over-fitting the missing entries.
- Therefore, with a finite sample at hand, the sample used at each iteration should be chosen *without replacement* as the algorithm runs. But even in the complete data case, sampling without replacement induces a bias on the chosen direction [7, 9]. Consequently, Lemma 1 does not hold for the empirical risk instead of the theoretical one. This issue is not addressed in [15], unfortunately making the proof of their result invalid/wrong.

Comparison to Ma and Needell [15]. Leaving aside the last observation, we can still comment on the bounds in [15] for the empirical risk without averaging. As they do not use averaging but only the last iterate, their convergence rate (see Lemma 1 in their paper) is only studied for μ -strongly convex problems and is expected to be larger (i) by a factor μ^{-1} , due to the choice of their decaying learning rate, and (ii) by a $\log n$ factor due to using the last iterate and not the averaged one [22]. Moreover, the strategy of the present paper does not require to access the strong convexity constant, which is generally out of reach, if no explicit regularization is used. More marginally, we provide the proof of the co-coercivity of the adjusted gradients (Lemma 3), which is required to derive the

convergence results, and which was also missing in Ma and Needell [15]. A more detailed discussion on the differences between the two papers is given in Appendix S1.

ERM hindered by NA. It is also interesting to point out that with missing features, *neither* the generalization risk R , *nor* the empirical risk R_n are observed (i.e., only approximations of their values or gradients can be computed). As a consequence, one cannot expect to minimize those functions with unlimited accuracy. This stands in contrast to the *complete observations setting*, in which the empirical risk R_n is known exactly. As a consequence, with missing data, empirical risk loses its main asset - being an observable function that one can minimize with high precision. Overall it is both more natural and easier to focus on the generalization risk.

4.4 On the impact of missing values

Marginal values of incomplete data. An important question in practice is to understand how much information has been lost because of the incompleteness of the observations. In other words, it is better to access 200 input/output pairs with a probability 50% of observing each feature on the inputs, or to observe 100 input/output pairs with complete observations?

Without missing observations, the variance bound in the expected excess risk is given by Theorem 4 with $p_m = 1$: it scales as $O\left(\frac{\text{Var}(\epsilon_k)d}{k}\right)$, while with missing observations it increases to $O\left(\frac{\text{Var}(\epsilon_k)d}{kp_m^2} + \frac{C(X, \beta^*)}{kp_m^3}\right)$. As a consequence, the variance upper bound is larger by a factor p_m^{-1} for the estimator derived from k incomplete observations than for $k \times p_m$ complete observations. This suggests that there is a higher gain to collecting fewer complete observations (e.g., 100) than more incomplete ones (e.g., 200 with $p = 0.5$). However, one should keep in mind that this observation is made by comparing upper bounds thus does not necessarily reflect what would happen in practice.

Keeping only complete observations? Another approach to solve the missing data problem is to discard all observations that have at least one missing feature. The probability that one input is complete, under our missing data model is $\prod_{j=1}^d p_j$. In the homogeneous case, the number of complete observations k_{co} out of a k -sample thus follows a binomial law $k_{co} \sim \mathcal{B}(k, p^d)$. With only those few observations, the statistical lower bound is $\frac{\text{Var}(\epsilon_k)d}{k_{co}}$. In expectation, by Jensen inequality, we get that the lower bound on the risk is larger than $\frac{\text{Var}(\epsilon_k)d}{kp^d}$.

Our strategy thus leads to an *upper-bound* which is typically p^{d-3} times smaller than the *lower bound* on the error of any algorithm relying only on complete observations. For a large dimension or a high percentage of missing values, our strategy is thus provably several orders of magnitude smaller than the best possible algorithm that would only rely on complete observations - e.g., if $p = 0.9$ and $d = 40$, the error of our method is at least 50 times smaller.

Also note that in Theorem 1 and Lemma 1 in Ma and Needell [15], the convergence rate with missing observations suffers from a similar multiplicative factor $O(p^{-2} + \kappa p^{-3})$.

5 Experiments

5.1 Synthetic data

Consider the following simulation setting: the covariates are normally distributed, $X_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \Sigma)$, where Σ is constructed using uniform random orthogonal eigenvectors and decreasing eigenvalues $1/k$, $k = 1, \dots, d$. For a fixed parameter vector β , the outputs y_i are generated according to the linear model (1), with $\epsilon_i \sim \mathcal{N}(0, 1)$. Setting $d = 10$, we introduce 30% of missing values either with a uniform probability p of missingness for any feature, or with probability p_j for covariate j , with $j = 1, \dots, d$. Firstly, the three following algorithms are implemented:

- (1) **AvSGD** described in Algorithm 1 with a constant step size $\alpha = \frac{1}{2L}$, and L given in (24).
- (2) **SGD** from [15] with iterates $\beta_{k+1} = \beta_k - \alpha_k \tilde{g}_{i_k}(\beta_k)$, and decreasing step size $\alpha_k = \frac{1}{\sqrt{k+1}}$.
- (3) **SGD_cst** from [15] with a constant step size $\alpha = \frac{1}{2L}$, where L is given by (24).

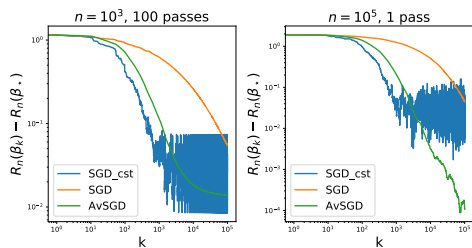


Figure 1: Empirical excess risk $(R_n(\beta_k) - R_n(\beta^*))$. Left: $n = 10^3$ and 100 passes. Right: $n = 10^5$ and 1 pass. $d = 10$, 30% MCAR data. L is assumed to be known in both graphics.

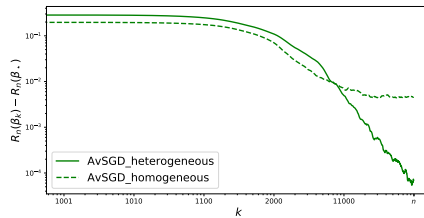


Figure 2: Empirical excess risk $R_n(\beta_k) - R_n(\beta^*)$ for synthetic data where $n = 10^5$, $d = 10$ and with heterogeneous missing values either taking into account the heterogeneity (plain line) in the algorithm or not (dashed line).

Debiased averaged vs. standard SGD. Figure 1 compares the convergence of Algorithms (1), (2) and (3), with either multiple passes or one pass, in terms of excess empirical risk $R_n(\beta) - R(\beta^*)$, with $R_n(\beta) := \frac{1}{n} \sum_{i=1}^n f_i(\beta)$. As expected (see Remark 3 and Subsection 4.3), multiple passes can lead to saturation: after one pass on the observations, AvSGD does not improve anymore (Figure 1, left), while it keeps decreasing in the streaming setting (Figure 1, right). Looking at Figure 1 (right), one may notice that without averaging and with decaying step-size, Algorithm (2) achieves the convergence rate $\mathcal{O}\left(\sqrt{\frac{1}{n}}\right)$, whereas with constant step-size, Algorithm (3) saturates at an excess risk proportional to α after $n = 10^3$ iterations. As theoretically expected, both methods are improved with averaging. Indeed, Algorithm 1 converges pointwise with a rate of $\mathcal{O}\left(\frac{1}{n}\right)$.

About the algorithm hyperparameter. Note that the Lipschitz constant L given in (24) can be either computed from the complete covariates, or estimated from the incomplete data, see discussion and numerical experiments in Appendix S3.

Heterogeneous vs. homogeneous missingness. In Figure 2, the missing values are introduced with different missingness probabilities, i.e. with distinct $(p_j)_{1 \leq j \leq d}$ per feature, as described in Equation (3). When taking into account this heterogeneousness, Algorithm 1 achieves the same convergence rates as in Figure 1. However, ignoring the heterogeneous probabilities in the gradient debiasing leads to stagnation far from the optimum in terms of empirical excess risk.

Polynomial features. Algorithm 1 can be adapted to handle missing polynomial features, see Appendix S4 for a detailed discussion and numerical experiments on synthetic data.

5.2 Real dataset 1: Traumabase dataset

We illustrate our approach on a public health application with the APHP TraumaBase[®] Group (Assistance Publique - Hopitaux de Paris) on the management of traumatized patients. Our aim is to model the level of platelet upon arrival at the hospital from the clinical data of 15785 patients. The platelet is a cellular agent responsible for clot formation and it is essential to control its levels to prevent blood loss and to decide on the most suitable treatment. A better understanding of the impact of the different features is key to trauma management. Explanatory variables for the level of platelet consist in seven quantitative (missing) variables, which have been selected by doctors. In Figure 3, one can see the percentage of missing values in each variable, varying from 0 to 16%, see Appendix S5 for more information on the data.

Model estimation. The model parameter estimation is performed either using the AvSGD Algorithm 1 or an Expectation Maximization (EM) algorithm [3]. Both methods are compared with the ordinary least squares linear regression in the complete case, i.e. keeping the fully-observed rows only (i.e. 9448 rows). The signs of the coefficients for Algorithm 1 are shown in Table 3.

According to the doctors, a negative effect of shock index (SI), vascular filling (VE), blood transfusion (RBC) and lactate ($Lactacte$) was expected, as they all result in low platelet levels and therefore a higher risk of severe bleeding. However, the effects of delta Hemocue ($Delta.Hemocue$) and the heart

Variable	Effect	NA %
Lactate	−	16%
Δ .Hemo	+	16%
VE	−	9%
RBC	−	8%
SI	−	2%
HR	+	1%
Age	−	0%

Figure 3: Percentage of missing features, and effect of the variables on the platelet for the TraumaBase data when the AvSGD algorithm is used. “+” indicates positive effect while “−” negative.

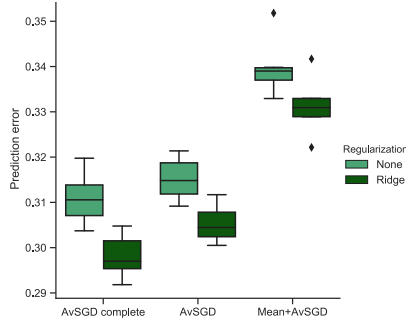


Figure 4: Prediction error boxplots (over 10 replications) for the Superconductivity data. AvSGD complete corresponds to applying the AvSGD on the complete data, AvSGD and Mean+AvSGD use the predictions obtained with the estimated parameters $\hat{\beta}_n^{\text{AvSGD}}$ and $\bar{\beta}_n^{\text{AvSGD}}$ respectively.

rate (*HR*) on platelets are not entirely in agreement with their opinion. Note that using the linear regression in the complete case and the EM algorithm lead to the same sign for the variables effects as presented in Table 3.

5.3 Real dataset 2: Superconductivity dataset

We now consider the Superconductivity dataset (available here), which contains 81 quantitative features from 21263 superconductors. The goal here is to predict the critical temperature of each superconductor. Since the dataset is initially complete, we introduce 30% of missing values with probabilities $(p_j)_{1 \leq j \leq 81}$ for the covariate j , with p_j varying between 0.7 and 1. The results are shown in Figure 4 where a Ridge regularization has been added or not. The regularization parameter λ (see Remark 1) is chosen by cross validation.

Prediction performance. The dataset is divided into training and test sets (random selection of 70 – 30%). The test set does not contain missing values. In order to predict the critical temperature of each superconductor, we compute $\hat{y}_{n+1} = X_{n+1}^T \hat{\beta}$ with $\hat{\beta} = \beta_n^{\text{AvSGD}}$ or β_n^{EM} . We also impute the missing data naively by the mean in the training set, and apply the averaged stochastic gradient without missing data on this imputed dataset, giving a coefficient model $\bar{\beta}_n^{\text{AvSGD}}$. It corresponds to the case where the bias of the imputation has not been corrected. The prediction quality on the test set is compared according to the relative ℓ_2 prediction error, $\|\hat{y} - y\|^2 / \|y\|^2$. The data is scaled, so that the naive prediction by the mean of the outcome variable leads to a prediction error equal to 1. In Figure 4, we observe that the SGD strategies give quite good prediction performances. The EM algorithm is not represented since it is completely out of range (the mean of its prediction error is 0.7), which indicates that it struggles with a large number of covariates. As for the AvSGD Algorithm, it performs well in this setting. Indeed, with or without regularization, the prediction error with missing values is very close to the one obtained from the complete dataset. Note that Algorithm 1 is shown to handle missing polynomial features well even in higher dimensions, see Appendix S4 for a detailed discussion and large-scale experiments on the superconductivity dataset.

6 Discussion

In this work, we thoroughly study the impact of missing values for Stochastic Gradient Descent algorithm for Least Squares Regression. We leverage both the power of averaging and a simple and powerful debiasing approach to derive tight and rigorous convergence guarantees for the generalization risk of the algorithm. The theoretical study directly translates into practical recommendations for the users and a byproduct is the availability of a python implementation of regularized regression with missing values for large scale data, which was not available. Even though we have knocked down some barriers, there are still exciting perspectives to be explored as the robustness of the approach to rarely-occurring covariates, or dealing with more general loss functions as well - for which it is challenging to build a debiased gradient estimator from observations with missing values, or also considering more complex missing-data patterns such as missing-not-at-random mechanisms.

References

- [1] Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $o(1/n)$. In *Advances in neural information processing systems*, pages 773–781, 2013.
- [2] Olivier Cappé and Eric Moulines. On-line expectation–maximization algorithm for latent data models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3): 593–613, 2009.
- [3] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.
- [4] Aymeric Dieuleveut, Alain Durmus, and Francis Bach. Bridging the gap between constant step size stochastic gradient descent and markov chains. *arXiv preprint arXiv:1707.06386*, 2017.
- [5] 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.
- [6] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.
- [7] Mert Gürbüzbalaban, Asu Ozdaglar, and PA Parrilo. Why random reshuffling beats stochastic gradient descent. *Mathematical Programming*, pages 1–36, 2015.
- [8] Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: an optimal algorithm for stochastic strongly-convex optimization. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 421–436, 2011.
- [9] Prateek Jain, Dheeraj Nagaraj, and Praneeth Netrapalli. SGD without Replacement: Sharper Rates for General Smooth Convex Functions. *arXiv e-prints*, art. arXiv:1903.01463, Mar 2019.
- [10] Michael P Jones. Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American statistical association*, 91(433):222–230, 1996.
- [11] Anatoli Juditsky, Arkadi Nemirovski, and Claire Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.
- [12] Vladimir Koltchinskii, Karim Lounici, Alexandre B Tsybakov, et al. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5): 2302–2329, 2011.
- [13] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.
- [14] Po-Ling Loh and Martin J Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. In *Advances in Neural Information Processing Systems*, pages 2726–2734, 2011.
- [15] Anna Ma and Deanna Needell. Stochastic gradient descent for linear systems with missing data. *Numerical Mathematics: Theory, Methods and Applications*, 12(1):1–20, 2018. ISSN 2079-7338. doi: <https://doi.org/10.4208/nmtma.OA-2018-0066>. URL http://global-sci.org/intro/article_detail/nmtma/12689.html.
- [16] Eric Moulines and Francis R Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, pages 451–459, 2011.
- [17] Deanna Needell, Rachel Ward, and Nati Srebro. Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. In *Advances in neural information processing systems*, pages 1017–1025, 2014.

- [18] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4): 1574–1609, 2009.
- [19] Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.
- [20] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.
- [21] Mathieu Rosenbaum, Alexandre B Tsybakov, et al. Sparse recovery under matrix uncertainty. *The Annals of Statistics*, 38(5):2620–2651, 2010.
- [22] Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *International Conference on Machine Learning*, pages 71–79, 2013.
- [23] Alexandre B Tsybakov. Optimal rates of aggregation. In *Learning theory and kernel machines*, pages 303–313. Springer, 2003.
- [24] Dao Li Zhu and Patrice Marcotte. Co-coercivity and its role in the convergence of iterative schemes for solving variational inequalities. *SIAM Journal on Optimization*, 6(3):714–726, 1996.

S1 Discussion on the paper of Ma and Needell [15]

In this section, we make the theoretical issues unlocked in Ma and Needell [15] explicit. For clarity, we directly refer to the lemmas and theorems as numbered in the published version (http://www.global-sci.org/uploads/online_news/NMTMA/201809051633-2442.pdf), the numbering being slightly different than the arXiv version. For readability, we translate their method and results with the notation used in the present paper. In their paper, they consider the finite-sample setting, with at hand $(D_k, \tilde{X}_i)_{1 \leq i \leq n}$, in view of minimizing the empirical risk.

As a preamble, let us remind that the contributions of the present paper go far beyond correcting the approach in [15]: we propose a different algorithm using averaging, that converges faster and in a non-strongly convex regime, with a different proof technique, requiring a more technical proof on the second order moment of the noise, and we allow for heterogeneity in the missing data mechanism.

S1.1 Hurdles to get unbiased gradients of the empirical risk

The stochastic gradients in [15] are not unbiased gradients of the empirical risk (which makes their main result wrong). Indeed, their algorithm uses the debiased direction (4) by sampling uniformly with replacement the $(\tilde{X}_k)_k$'s.

For clarity, we highlight both why the result is not technically correct in their paper, and why it is not intuitively possible to achieve the result they give.

Technically. The proof of the main Theorem 2.2 (Theorem 2.1 being a direct corollary), corresponds to the classical proof in which one upper bounds the expectation of the mean-squared distance from the iterate at iteration $k + 1$ to the optimal point **conditionally to the iterate at iteration k** , or more precisely, conditionally to a σ -algebra making this iterate measurable. This is typically written

$$\mathbb{E} [\|\beta_{k+1} - \beta_*^n\|^2 | \mathcal{F}_k],$$

where β_*^n is the minimizer of the empirical risk R_n and β_k is \mathcal{F}_k -measurable.

The crux of the proof is then to use **unbiased gradients conditionally to \mathcal{F}_k** : the property needed is that

$$\mathbb{E} [g_{i_{k+1}}(\beta_k) | \mathcal{F}_k] = \nabla R_n(\beta_k).$$

In classical ERM (without missing value) it is done by sampling uniformly at iteration $k + 1$ one observation indexed by $i_{k+1} \sim \mathcal{U}[[1; n]]$, **independently from β_k** .

In regression with missing data, one has to deal with another source of randomness, the randomness of the mask D . In Ma and Needell [15], Lemma A.1 states that for a random $i \sim \mathcal{U}[[1; n]]$ and a matrix row A_i , for a random mask D associated to this row,

$$\mathbb{E}_D[\mathbb{E}_i g_i(\beta)] = \nabla R_n(\beta).$$

This lemma is valid. Unfortunately, its usage in the proof of Theorem 2.2 (page 18, line (ii)), is not, as one *does not have*:

$$\mathbb{E}[g_{i_{k+1}}(\beta_k) | \mathcal{F}_k] = \nabla R_n(\beta_k),$$

indeed,

- either the sample i_{k+1} is chosen uniformly at random in $[[1; n]]$ and $D_{i_{k+1}}$ **is not** independent from β_k .
- or the sample i is not chosen uniformly in $[[1; n]]$ (for example without replacement, as we do) and then the gradient is not an unbiased gradient of R_n as the sampling is not uniform anymore.

In other words, the proof would only be valid if **the mask for the missing entries was re-sampled each time the point is used**, which is of course not realistic for a missing data approach (that would mean that the data has in fact been collected without missing entries).

Intuition on why it is hard. A way to understand the impossibility of having a bound for multiple pass on ERM in the context of missing data is to underline that the empirical risk, in the presence of missing data, is an **unknown function**: its value cannot be computed exactly (see Subsection 4.3).

As a consequence we can hardly expect that one could minimize it to unlimited accuracy. This is very similar to the situation for the *generalization risk* in a situation *without missing data*: as the function is not observed, it is impossible to minimize it exactly. Given only n observations, no algorithm can achieve 0-generalization error (and statistical lower bounds [23] prove so).

Conclusion. This highlights how difficult it is to be rigorous when dealing with multiple sources of randomness. Unfortunately, none of these limits are discussed in the current version of [15]. This makes the approach and the main theorem of [15] mathematically invalid. In the present paper, the *generalization risk* is decaying *during the first pass*, and as a consequence, the empirical risk also probably does, but this has not been proved yet.

In the following paragraph, we give details on the missing technical Lemma.

S1.2 Missing key Lemma in the proof.

Proving that (\tilde{f}_k) is a.s. convex is an important step for convergence, which was missing in the analysis of [15]. More precisely, in Lemma A.4. in [15], a condition is missing on $G(x)$: G needs to be smooth *and convex* for its gradient to satisfy the co-coercivity inequality. Note that this condition was also missing in the paper they refer to Needell et al. [17] (Indeed, at the third line of the proof of Lemma A.1. in Needell et al. [17], one needs f to be convex for G to be convex). Co-coercivity of the gradient is indeed a characterization of the fact that the function is smooth and convex, see for example Zhu and Marcotte [24].

S2 Proofs of technical lemmas

Recall that we aim at minimizing the theoretical risk in both streaming and finite-sample settings.

$$\beta^* = \arg \min_{\beta \in \mathbb{R}^d} R(\beta) = \arg \min_{\beta \in \mathbb{R}^d} \mathbb{E}_{(X_i, y_i)} [f_i(\beta)]. \quad (2)$$

In the sequel, one consider the following modified gradient direction

$$\tilde{g}_k(\beta_k) = P^{-1} \tilde{X}_{k:} \left(\tilde{X}_{k:}^T P^{-1} \beta_k - y_k \right) - (I - P) P^{-2} \text{diag} \left(\tilde{X}_{k:}, \tilde{X}_{k:}^T \right) \beta_k. \quad (4)$$

Note that for all k , $D_{k:}$ is independent from $(X_{k:}, y_k)$. In what follows, the proofs are derived considering

$$\mathbb{E} = \mathbb{E}_{(X_{k:}, y_k), D_{k:}} = \mathbb{E}_{(X_{k:}, y_k)} \mathbb{E}_{D_{k:}}$$

where $\mathbb{E}_{(X_{k:}, y_k)}$ and $\mathbb{E}_{D_{k:}}$ denotes the expectation with respect to the distribution of $(X_{k:}, y_k)$ and $D_{k:}$ respectively.

S2.1 Proof of Lemma 1

Lemma S1. Let $(\mathcal{F}_k)_{k \geq 0}$ be the following σ -algebra,

$$\mathcal{F}_k = \sigma(X_{1:}, y_1, D_1, \dots, X_{k:}, y_k, D_{k:}).$$

The modified gradient $\tilde{g}_k(\beta_{k-1})$ in Equation (4) is \mathcal{F}_k -measurable and

$$\mathbb{E}[\tilde{g}_k(\beta_{k-1}) \mid \mathcal{F}_{k-1}] = \nabla R(\beta_{k-1}) \quad a.s.$$

Proof.

$$\begin{aligned} \mathbb{E}_{(X_{k:}, y_k), D_{k:}}[\tilde{g}_k(\beta_{k-1}) \mid \mathcal{F}_{k-1}] &\stackrel{(i)}{=} \mathbb{E}_{(X_{k:}, y_k), D_{k:}} \left[P^{-1} \tilde{X}_{k:} \tilde{X}_{k:}^T P^{-1} \right] \beta_{k-1} - \mathbb{E}_{(X_{k:}, y_k), D_{k:}} \left[P^{-1} \tilde{X}_{k:} y_k \right] \\ &\quad - \mathbb{E}_{(X_{k:}, y_k), D_{k:}} \left[(I - P) P^{-2} \text{diag} \left(\tilde{X}_{k:} \tilde{X}_{k:}^T \right) \right] \beta_{k-1} \\ &\stackrel{(ii)}{=} \mathbb{E}_{(X_{k:}, y_k)} \left[P^{-1} P X_{k:} X_{k:}^T P P^{-1} \beta_{k-1} - P^{-2} (P - P^2) \text{diag} (X_{k:} X_{k:}^T) \beta_{k-1} - P^{-1} P X_{k:} y_k \right] \\ &\quad - \mathbb{E}_{(X_{k:}, y_k)} \left[(I - P) P^{-2} P \text{diag} (X_{k:} X_{k:}^T) \beta_{k-1} \right] \\ &= \nabla R(\beta_{k-1}), \end{aligned}$$

In step (i), we use that β_{k-1} is \mathcal{F}_{k-1} -measurable and (X_k, y_k, D_k) is independent from \mathcal{F}_{k-1} . Step (ii) follows from

$$\begin{cases} \mathbb{E}_{D_{k:}} \left[\tilde{X}_{k:} \tilde{X}_{k:}^T \right] &= P X_{k:} X_{k:}^T P - (P - P^2) \text{diag} (X_{k:} X_{k:}^T), \\ \mathbb{E}_{D_{k:}} \left[\text{diag} (\tilde{X}_{k:} \tilde{X}_{k:}^T) \right] &= P \text{diag} (X_{k:} X_{k:}^T), \\ \mathbb{E}_{D_{k:}} \left[\tilde{X}_{k:} \right] &= P X_{k:}. \end{cases}$$

□

S2.2 Proof of Lemma 2

Lemma S2. The additive noise process $(\tilde{g}_k(\beta^*))_k$ with β^* defined in Equation (2) is \mathcal{F}_k -measurable and has the following properties:

1. $\forall k \geq 0$, $\mathbb{E}[\tilde{g}_k(\beta^*) \mid \mathcal{F}_{k-1}] = 0$ a.s.,
2. $\forall k \geq 0$, $\mathbb{E}[\|\tilde{g}_k(\beta^*)\|^2 \mid \mathcal{F}_{k-1}]$ is a.s. finite,
3. $\forall k \geq 0$, $\mathbb{E}[\tilde{g}_k(\beta^*) \tilde{g}_k(\beta^*)^T] \preceq C(\beta^*) = c(\beta^*)H$, where \preceq denotes the order between self-adjoint operators ($A \preceq B$ if $B - A$ is positive semi-definite).

Proof. 1 The first point is easily verified using Lemma 1 combined with $\nabla R(\beta^*) = 0$ by (2).

2 Let us first remark that by independence $\mathbb{E}[\|\tilde{g}_k(\beta^*)\|^2 \mid \mathcal{F}_{k-1}] = \mathbb{E}[\|\tilde{g}_k(\beta^*)\|^2]$. Then,

$$\mathbb{E}[\|\tilde{g}_k(\beta^*)\|^2] \leq \frac{1}{p_m^2} \mathbb{E} \left[\|X_{k:}\|^2 \left(\tilde{X}_{k:}^T P^{-1} \beta^* - y_k \right)^2 \right] + \frac{(1 - p_m)^2}{p_m^2} \mathbb{E} \left[\|P^{-1} \text{diag} \left(\tilde{X}_{k:} \tilde{X}_{k:}^T \right) \beta^*\|^2 \right].$$

We decompose the computation with respect to \mathbb{E}_{D_k} : first,

$$\begin{aligned}
\mathbb{E}_{D_k} \left[\left(\tilde{X}_{k:}^T P^{-1} \beta^* - y_k \right)^2 \right] &= \mathbb{E}_{D_k} \left[\left(\tilde{X}_{k:}^T P^{-1} \beta^* \right)^2 \right] - 2y_k \mathbb{E}_{D_k} \left[\tilde{X}_{k:}^T P^{-1} \beta^* \right] + y_k^2 \\
&= \mathbb{E}_{D_k} \left[\left(\sum_{j=1}^d \tilde{X}_{kj} p_j^{-1} \beta_j^* \right)^2 \right] - 2y_k \mathbb{E}_{D_k} \left[\sum_{j=1}^d \tilde{X}_{kj} p_j^{-1} \beta_j^* \right] + y_k^2 \\
&= \sum_{j=1}^d \mathbb{E}_{D_k} \left[\tilde{X}_{kj}^2 p_j^{-2} \beta_j^{*2} \right] + 2 \sum_{l < j} \mathbb{E}_{D_k} \left[\tilde{X}_{kj} \tilde{X}_{kl} p_l^{-1} \beta_j^* \beta_l^* \right] - 2y_k \sum_{j=1}^d X_{kj} \beta_j^* + y_k^2 \\
&= \sum_{j=1}^d p_j^{-1} X_{kj}^2 \beta_j^{*2} + 2 \sum_{l < j} X_{kj} X_{kl} \beta_j^* \beta_l^* - 2y_k \sum_{j=1}^d X_{kj} \beta_j^* + y_k^2 \\
&= (X_{k:}^T \beta^* - y_k)^2 + \sum_{j=1}^d (p_j^{-1} - 1) X_{kj}^2 \beta_j^{*2},
\end{aligned}$$

which gives

$$\mathbb{E}_{D_k} \left[\left(\tilde{X}_{k:}^T P^{-1} \beta^* - y_k \right)^2 \right] \leq (X_{k:}^T \beta^* - y_k)^2 + \frac{1 - p_m}{p_m} \beta^{*T} \text{diag}(X_{k:} X_{k:}^T) \beta^*. \quad (8)$$

As for the second term,

$$\begin{aligned}
\mathbb{E}_{D_k} \left[\left\| P^{-1} \text{diag}(\tilde{X}_{k:} \tilde{X}_{k:}^T) \beta^* \right\|^2 \right] &= \mathbb{E}_{D_k} \left[\sum_{j=1}^d \tilde{X}_{kj}^4 p_j^{-2} \beta_j^{*2} \right] \\
&= \sum_{j=1}^d X_{kj}^4 p_j^{-1} \beta_j^{*2} \\
&\leq \frac{1}{p_m} \sum_{j=1}^d X_{ij}^4 \beta_j^{*2} \\
&\leq \frac{1}{p_m} \left(\sum_{j=1}^d X_{kj}^2 \right) \left(\sum_{j=1}^d X_{kj}^2 \beta_j^{*2} \right) \\
&= \frac{1}{p_m} \|X_{k:}\|^2 \beta^{*T} \text{diag}(X_{k:} X_{k:}^T) \beta^*
\end{aligned}$$

Finally, one obtains

$$\mathbb{E}[\|\tilde{g}_k(\beta^*)\|^2 \mid \mathcal{F}_{k-1}] \leq \frac{1}{p_m^2} \mathbb{E}_{(X_{k:}, y_k)} [(\epsilon_k)^2 \|X_{k:}\|^2] + \frac{(1 - p_m) + (1 - p_m)^2}{p_m^3} \mathbb{E}_{(X_{k:}, y_k)} [\|X_{k:}\|^2 \beta^{*T} \text{diag}(X_{k:} X_{k:}^T) \beta^*].$$

3 We aim at proving there exists H such that

$$\mathbb{E}[\tilde{g}_k(\beta^*) \tilde{g}_k(\beta^*)^T] \preceq C = cH.$$

Simple computations lead to:

$$\mathbb{E}[\tilde{g}_k(\beta^*) \tilde{g}_k(\beta^*)^T] = \mathbb{E}[T_1 + T_2 + T_2^T + T_3],$$

with:

$$\begin{aligned}
T_1 &= (\tilde{X}_{k:}^T P^{-1} \beta^* - y_k)^2 P^{-1} \tilde{X}_{k:} \tilde{X}_{k:}^T P^{-1}, \\
T_2 &= -(\tilde{X}_{k:}^T P^{-1} \beta^* - y_k) P^{-1} \tilde{X}_{k:} \beta^{*T} \text{diag}(\tilde{X}_{k:} \tilde{X}_{k:}^T) P^{-2} (I - P), \\
T_3 &= (I - P) P^{-2} \text{diag}(\tilde{X}_{k:} \tilde{X}_{k:}^T) \beta^* \beta^{*T} \text{diag}(\tilde{X}_{k:} \tilde{X}_{k:}^T) P^{-2} (I - P).
\end{aligned}$$

Bound on T_1 . For the first term, we use

$$P^{-1} \tilde{X}_{k:} \tilde{X}_{k:}^T P^{-1} \preceq \frac{1}{p_m^2} \tilde{X}_{k:} \tilde{X}_{k:}^T, \quad (9)$$

since for all vector $v \neq 0$, $v^T \left(\frac{1}{p_m^2} \tilde{X}_{k:} \tilde{X}_{k:}^T - P^{-1} \tilde{X}_{k:} \tilde{X}_{k:}^T P^{-1} \right) v \geq 0$,

$$\begin{aligned} & \sum_{j=1}^d \left(\frac{1}{p_m^2} - \frac{1}{p_j^2} \right) \tilde{X}_{kj}^2 v_j^2 + 2 \sum_{1 \leq j < l \leq d} \left(\frac{1}{p_m^2} - \frac{1}{p_j p_l} \right) \tilde{X}_{kj} \tilde{X}_{kl} v_j v_l \\ & \stackrel{(iii)}{\geq} \sum_{j=1}^d \left(\frac{1}{p_m^2} - \frac{1}{p_j^2} \right) \tilde{X}_{kj}^2 v_j^2 + 2 \sum_{1 \leq j < l \leq d} \sqrt{\left(\frac{1}{p_m^2} - \frac{1}{p_j^2} \right) \left(\frac{1}{p_m^2} - \frac{1}{p_l^2} \right)} \tilde{X}_{kj} \tilde{X}_{kl} v_j v_l \\ & = \left(\sum_{j=1}^d \sqrt{\left(\frac{1}{p_m^2} - \frac{1}{p_j^2} \right)} \tilde{X}_{kj} v_j \right)^2 \geq 0. \end{aligned}$$

Step (iii) uses $\left(\frac{1}{p_m^2} - \frac{1}{p_j p_l} \right) \geq \sqrt{\left(\frac{1}{p_m^2} - \frac{1}{p_j^2} \right) \left(\frac{1}{p_m^2} - \frac{1}{p_l^2} \right)}$. Indeed,

$$\begin{aligned} \left(\frac{1}{p_m^2} - \frac{1}{p_j p_l} \right)^2 & \geq \left(\frac{1}{p_m^2} - \frac{1}{p_j^2} \right) \left(\frac{1}{p_m^2} - \frac{1}{p_l^2} \right) \Leftrightarrow \left(\frac{1}{p_m^4} - 2 \frac{1}{p_j p_l} \frac{1}{p_m^2} + \frac{1}{p_j^2 p_l^2} \right) - \frac{1}{p_m^4} + \frac{1}{p_m^2 p_l^2} + \frac{1}{p_m^2 p_j^2} - \frac{1}{p_j^2 p_l^2} \geq 0 \\ & \Leftrightarrow \left(\frac{1}{p_m p_j} - \frac{1}{p_m p_l} \right)^2 \geq 0. \end{aligned}$$

Let us now prove that

$$\frac{1}{p_m^2} \tilde{X}_{k:} \tilde{X}_{k:}^T \preceq \frac{1}{p_m^2} X_{k:} X_{k:}^T$$

i.e.

$$\tilde{X}_{k:} \tilde{X}_{k:}^T \preceq X_{k:} X_{k:}^T. \quad (10)$$

Indeed, for all vector $v \neq 0$, $v^T (X_{k:} X_{k:}^T - \tilde{X}_{k:} \tilde{X}_{k:}^T) v \geq 0$:

$$\begin{aligned} v^T (X_{k:} X_{k:}^T - \tilde{X}_{k:} \tilde{X}_{k:}^T) v & = \sum_{j=1}^d (1 - \delta_{kj}^2) X_{kj}^2 v_j^2 + 2 \sum_{1 \leq j < l \leq d} (1 - \delta_{kj} \delta_{kl}) X_{kj} X_{kl} v_j v_l \\ & \stackrel{(iv)}{\geq} \sum_{j=1}^d (1 - \delta_{kj}^2) X_{kj}^2 v_j^2 + 2 \sum_{1 \leq j < l \leq d} \sqrt{(1 - \delta_{kj}^2)(1 - \delta_{kl}^2)} X_{kj} X_{kl} v_j v_l \\ & = \left(\sum_{j=1}^d \sqrt{(1 - \delta_{kj}^2)} X_{kj} v_j \right)^2 \geq 0 \end{aligned}$$

Step (iv) is obtained using $(1 - \delta_{kj} \delta_{kl}) \geq \sqrt{(1 - \delta_{kj}^2)(1 - \delta_{kl}^2)}$. Indeed,

$$\begin{aligned} (1 - \delta_{kj} \delta_{kl})^2 & \geq (1 - \delta_{kj}^2)(1 - \delta_{kl}^2) \Leftrightarrow (1 - 2\delta_{kl} \delta_{kj} + \delta_{kj}^2 \delta_{kl}^2) - 1 + \delta_{kj}^2 - \delta_{kj}^2 \delta_{kl}^2 + \delta_{kl}^2 \geq 0 \\ & \Leftrightarrow (\delta_{kj} - \delta_{kl})^2 \geq 0. \end{aligned}$$

Then, by (8) and $(X_{k:}^T \beta^* - y_k)^2 = \epsilon_k^2$,

$$\mathbb{E}_{(X_{k:}, y_k)} [T_1] = \mathbb{E}_{(X_{k:}, y_k)} \left[\frac{1}{p_m^2} \epsilon_k^2 X_{k:} X_{k:}^T \right] + \mathbb{E}_{(X_{k:}, y_k)} \left[\frac{1 - p_m}{p_m^3} (\beta^{*T} \text{diag}(X_{k:} X_{k:}^T) \beta^*) X_{k:} X_{k:}^T \right].$$

Noting that

$$\|\text{diag}(X_{k:}) \beta^*\|^2 \leq \|X_{k:}\|^2 \|\beta^*\|^2, \quad (11)$$

$$\mathbb{E} [T_1] \preceq \frac{1}{p_m^2} \text{Var}(\epsilon_k) H + \frac{1 - p_m}{p_m^3} \|X_{k:}\|^2 \|\beta^*\|^2 H \quad (12)$$

Bound on T_3 . Using the resulting matrix structure of $(I - P)P^{-2}\text{diag}(\tilde{X}_k; \tilde{X}_k^T)\beta^*\beta^{*T}\text{diag}(\tilde{X}_k; \tilde{X}_k^T)P^{-2}(I - P)$,

$$\begin{pmatrix} (\beta_1^*)^2 \delta_{k_1}^4 X_{k_1}^4 & \beta_1^* \beta_2^* \delta_{k_1}^2 \delta_{k_2}^2 X_{k_1}^2 X_{k_2}^2 & & \\ & \ddots & & \\ & & & (\beta_d^*)^2 \delta_{kd}^4 X_{kd}^4 \end{pmatrix},$$

one obtains

$$\begin{aligned} \mathbb{E}_{D_k} [T_3] &= \underbrace{(I - P)P^{-2}P\text{diag}(X_k; X_k^T)\beta^*\beta^{*T}\text{diag}(X_k; X_k^T)PP^{-2}(I - P)}_{=:T_{3a}} \\ &+ \underbrace{(I - P)P^{-2}(P - P^2)\text{diag}(X_k; X_k^T)\text{diag}(\beta^*\beta^{*T})\text{diag}(X_k; X_k^T)P^{-2}(I - P)}_{=:T_{3b}}. \end{aligned} \quad (13)$$

Using similar arguments as in (9), both terms in (13) are bounded as follows

$$\begin{aligned} T_{3a} &\preceq \frac{(1 - p_m)^2}{p_m^2} \text{diag}(X_k; X_k^T)\beta^*\beta^{*T}\text{diag}(X_k; X_k^T) \\ T_{3b} &\preceq \frac{(1 - p_m)^3}{p_m^3} \text{diag}(X_k; X_k^T)\text{diag}(\beta^*\beta^{*T})\text{diag}(X_k; X_k^T) \end{aligned}$$

For T_{3a} , one can go further by using

$$\text{diag}(X_k; X_k^T)\beta^*\beta^{*T}\text{diag}(X_k; X_k^T) \preceq \|\text{diag}(X_k; \beta^*)\|^2 X_k; X_k^T. \quad (14)$$

Let us prove that for all vector $v \neq 0$,

$$v^T (\|\text{diag}(X_k; \beta^*)\|^2 X_k; X_k^T - \text{diag}(X_k; X_k^T)\beta^*\beta^{*T}\text{diag}(X_k; X_k^T))v \geq 0, \text{ i.e.}$$

$$\underbrace{\sum_{j=1}^d \left(\left(\sum_{l=1}^d X_{il}^2 \beta_l^{*2} \right) X_{kj}^2 - X_{kj}^4 \beta_j^{*2} \right) v_j^2 + 2 \sum_{1 \leq j < m \leq d} \left(\left(\sum_{l=1}^d X_{kl}^2 \beta_l^{*2} \right) X_{kj} X_{km} - \beta_j^* \beta_m^* X_{kj}^2 X_{km}^2 \right) v_m v_j}_{=:Q} \geq 0$$

Indeed, $Q \geq \left(\sum_{j=1}^d \sqrt{\left(\sum_{l=1}^d X_{kl}^2 \beta_l^{*2} \right) X_{kj}^2 - X_{kj}^4 \beta_j^{*2}} v_j \right)^2 \geq 0$, since, looking at the term depending only on $v_j v_m$:

$$\begin{aligned} &\left(\left(\sum_{l=1}^d X_{kl}^2 \beta_l^{*2} \right) X_{kj} X_{km} - \beta_j^* \beta_m^* X_{kj}^2 X_{km}^2 \right) \geq \sqrt{\left(\left(\sum_{l=1}^d X_{kl}^2 \beta_l^{*2} \right) X_{kj}^2 - X_{kj}^4 \beta_j^{*2} \right) \left(\left(\sum_{l=1}^d X_{kl}^2 \beta_l^{*2} \right) X_{km}^2 - X_{km}^4 \beta_m^{*2} \right)} \\ &\Leftrightarrow \left(\sum_{l=1}^d X_{kl}^2 \beta_l^{*2} \right) X_{kj}^4 X_{km}^2 \beta_j^{*2} + \left(\sum_{l=1}^d X_{kl}^2 \beta_l^{*2} \right) X_{km}^4 X_{kj}^2 \beta_m^{*2} - 2 \left(\sum_{l=1}^d X_{kl}^2 \beta_l^{*2} \right) X_{kj}^3 X_{km}^3 \beta_j^* \beta_m^* \geq 0 \\ &\Leftrightarrow \left(\sqrt{\left(\sum_{l=1}^d X_{kl}^2 \beta_l^{*2} \right) X_{kj}^2 X_{km}^2} - \sqrt{\left(\sum_{l=1}^d X_{kl}^2 \beta_l^{*2} \right) X_{km}^2 X_{kj}^2} \right)^2 \geq 0 \end{aligned}$$

For T_{3b} , one can also dig deeper noting that

$$\text{diag}(X_k; X_k^T)\text{diag}(\beta^*\beta^{*T})\text{diag}(X_k; X_k^T) \preceq \|\text{diag}(X_k; \beta^*)\|^2 X_k; X_k^T. \quad (15)$$

For all vector $v \neq 0$, we aim at proving

$$\begin{aligned} &v^T (\|\beta^*\beta^{*T}\text{diag}(X_k; \beta^*)\|^2 X_k; X_k^T - \text{diag}(X_k; X_k^T)\text{diag}(\beta^*\beta^{*T})\text{diag}(X_k; X_k^T))v \geq 0 \\ &\Leftrightarrow \underbrace{\sum_{j=1}^d \left(\left(\sum_{l=1}^d X_{kl}^2 \beta_l^{*2} \right) X_{kj}^2 - X_{kj}^4 \beta_j^{*2} \right) v_j^2 + 2 \sum_{1 \leq j < m \leq d} \left(\sum_{l=1}^d X_{kl}^2 \beta_l^{*2} \right) X_{kj} X_{km} v_j v_m}_{=:Q'} \geq 0. \end{aligned}$$

Indeed, $Q' \geq \left(\sum_{j=1}^d \sqrt{\left(\sum_{l=1}^d X_{kl}^2 \beta_l^{*2} \right) X_{kj}^2 - X_{kj}^4 \beta_j^{*2} v_j} \right)^2 \geq 0$ since

$$\begin{aligned} & \left(\left(\sum_{l=1}^d X_{kl}^2 \beta_l^{*2} \right) X_{kj} X_{km} \right) \geq \sqrt{\left(\left(\sum_{l=1}^d X_{kl}^2 \beta_l^{*2} \right) X_{kj}^2 - X_{kj}^4 \beta_j^{*2} \right) \left(\left(\sum_{l=1}^d X_{kl}^2 \beta_l^{*2} \right) X_{km}^2 - X_{km}^4 \beta_m^{*2} \right)} \\ \Leftrightarrow & \left(\sum_{l=1}^d X_{kl}^2 \beta_l^{*2} \right) X_{kj}^4 X_{km}^2 \beta_j^{*2} + \left(\sum_{l=1}^d X_{kl}^2 \beta_l^{*2} \right) X_{km}^4 X_{kj}^2 \beta_m^{*2} - X_{kj}^4 X_{km}^4 \beta_j^{*2} \beta_m^{*2} \geq 0 \end{aligned}$$

Combining (11), (14) and (15) lead to

$$\begin{aligned} \mathbb{E}_{(X_{k:\cdot}, y_k)} [T_{3a}] & \preccurlyeq \frac{(1-p_m)^2}{p_m^2} \|X_{k:\cdot}\|^2 \|\beta^*\|^2 H \\ \mathbb{E}_{(X_{k:\cdot}, y_k)} [T_{3b}] & \preccurlyeq \frac{(1-p_m)^3}{p_m^3} \|X_{k:\cdot}\|^2 \|\beta^*\|^2 H \end{aligned}$$

and to the final bound for T_3 ,

$$\mathbb{E} [T_3] \preccurlyeq \frac{(1-p_m)^2}{p_m^2} \|X_{k:\cdot}\|^2 \|\beta^*\|^2 H + \frac{(1-p_m)^3}{p_m^3} \|X_{k:\cdot}\|^2 \|\beta^*\|^2 H. \quad (16)$$

Bound on $T_2 + T_2^T$. Firstly, focus on T_2 :

$$\begin{aligned} T_2 & = -(\tilde{X}_{k:\cdot}^T P^{-1} \beta^* - y_k) P^{-1} \tilde{X}_{k:\cdot} \beta^{*T} \text{diag}(\tilde{X}_{k:\cdot} \tilde{X}_{k:\cdot}^T) P^{-2} (I - P) \\ & =: -(A - B), \end{aligned}$$

where

$$\begin{aligned} A & = P^{-1} \tilde{X}_{k:\cdot} \tilde{X}_{k:\cdot}^T P^{-1} \beta^* \beta^{*T} \text{diag}(\tilde{X}_{k:\cdot} \tilde{X}_{k:\cdot}^T) P^{-2} (I - P) \\ B & = P^{-1} \tilde{X}_{k:\cdot} y_k \beta^{*T} \text{diag}(\tilde{X}_{k:\cdot} \tilde{X}_{k:\cdot}^T) P^{-2} (I - P). \end{aligned}$$

Computation w.r.t. $\mathbb{E}_{D_{k:\cdot}}$. Term A can be split into three terms,

$$\begin{aligned} A_1 & = P^{-1} \text{diag}(\tilde{X}_{k:\cdot} \tilde{X}_{k:\cdot}^T) P^{-1} \beta^* \beta^{*T} \text{diag}(\tilde{X}_{k:\cdot} \tilde{X}_{k:\cdot}^T) P^{-2} (I - P) \\ A_2 & = P^{-1} (\tilde{X}_{k:\cdot} \tilde{X}_{k:\cdot}^T - \text{diag}(\tilde{X}_{k:\cdot} \tilde{X}_{k:\cdot}^T)) P^{-1} \text{diag}(\beta^* \beta^{*T}) \text{diag}(\tilde{X}_{k:\cdot} \tilde{X}_{k:\cdot}^T) P^{-2} (I - P) \\ A_3 & = P^{-1} (\tilde{X}_{k:\cdot} \tilde{X}_{k:\cdot}^T - \text{diag}(\tilde{X}_{k:\cdot} \tilde{X}_{k:\cdot}^T)) P^{-1} (\beta^* \beta^{*T} - \text{diag}(\beta^* \beta^{*T})) \text{diag}(\tilde{X}_{k:\cdot} \tilde{X}_{k:\cdot}^T) P^{-2} (I - P). \end{aligned}$$

Noting that

$$A_1 = P^{-2} \text{diag}(\tilde{X}_{k:\cdot} \tilde{X}_{k:\cdot}^T) \beta^* \beta^{*T} \text{diag}(\tilde{X}_{k:\cdot} \tilde{X}_{k:\cdot}^T) P^{-2} (I - P),$$

the expectation $\mathbb{E}_{D_{k:\cdot}}$ has already been computed in (13), so

$$\begin{aligned} \mathbb{E}_{D_{k:\cdot}} [A_1] & = P^{-2} P \text{diag}(X_{k:\cdot} X_{k:\cdot}^T) \beta^* \beta^{*T} \text{diag}(X_{k:\cdot} X_{k:\cdot}^T) P P^{-2} (I - P) \\ & \quad + P^{-2} (P - P^2) \text{diag}(X_{k:\cdot} X_{k:\cdot}^T) \text{diag}(\beta^* \beta^{*T}) \text{diag}(X_{k:\cdot} X_{k:\cdot}^T) P^{-2} (I - P). \quad (17) \end{aligned}$$

As for A_2 , making the structure of the term $(\tilde{X}_{k:\cdot} \tilde{X}_{k:\cdot}^T - \text{diag}(\tilde{X}_{k:\cdot} \tilde{X}_{k:\cdot}^T)) P^{-1} \text{diag}(\beta^* \beta^{*T}) \text{diag}(\tilde{X}_{k:\cdot} \tilde{X}_{k:\cdot}^T)$ explicit,

$$A_2 = \begin{pmatrix} 0 & \frac{1}{p_2} \delta_{k1} \delta_{k2}^3 X_{k1} X_{k2}^3 \beta_2^{*2} & \cdots & \frac{1}{p_d} \delta_{k1} \delta_{kd}^3 X_{k1} X_{kd}^3 \beta_d^{*2} \\ \frac{1}{p_1} \delta_{k2} \delta_{k1}^3 X_{k2} X_{k1}^3 \beta_1^{*2} & 0 & & \\ & & \ddots & \\ \frac{1}{p_1} \delta_{kd} \delta_{k1}^3 X_{kd} X_{k1}^3 \beta_1^{*2} & & & 0 \end{pmatrix},$$

one has

$$\mathbb{E}_{D_{k:\cdot}} [A_2] = P^{-1} P (X_{k:\cdot} X_{k:\cdot}^T - \text{diag}(X_{k:\cdot} X_{k:\cdot}^T)) \text{diag}(\beta^* \beta^{*T}) \text{diag}(X_{k:\cdot} X_{k:\cdot}^T) P^{-2} (I - P). \quad (18)$$

As for A_3 , the term $(\tilde{X}_k: \tilde{X}_k^T - \text{diag}(\tilde{X}_k: \tilde{X}_k^T))P^{-1}(\beta^* \beta^{*T} - \text{diag}(\beta^* \beta^{*T}))\text{diag}(\tilde{X}_k: \tilde{X}_k^T)$ can be made explicit as

$$\begin{pmatrix} \sum_{l=2}^d \frac{1}{p_l} \delta_{kl} X_{kl} \beta_l^* \delta_{k1} X_{k1}^3 \beta_1^* & \sum_{l=3}^d \frac{1}{p_l} \delta_{kl} X_{kl} \beta_l^* \delta_{k1} \delta_{k2}^2 X_{k1} X_{k2}^2 \beta_2^* & \cdots & \sum_{l \neq 1, d} \frac{1}{p_l} \delta_{kl} X_{kl} \beta_l^* \delta_{k1} \delta_{kd}^2 X_{k1} X_{kd}^2 \beta_d^* \\ & \ddots & & \\ & & \ddots & \\ & & & \sum_{l=1}^{d-1} \frac{1}{p_l} \delta_{kl} X_{kl} \beta_l^* \delta_{kd} X_{kd}^3 \beta_d^* \end{pmatrix},$$

which gives

$$\begin{aligned} \mathbb{E}_{D_k}: [A_3] &= P^{-1}P(\tilde{X}_k: \tilde{X}_k^T - \text{diag}(\tilde{X}_k: \tilde{X}_k^T))(\beta^* \beta^{*T} - \text{diag}(\beta^* \beta^{*T}))\text{diag}(\tilde{X}_k: \tilde{X}_k^T)PP^{-2}(I-P) \\ &+ P^{-1}(P-P^2)\text{diag}\left((\tilde{X}_k: \tilde{X}_k^T - \text{diag}(\tilde{X}_k: \tilde{X}_k^T))(\beta^* \beta^{*T} - \text{diag}(\beta^* \beta^{*T}))\text{diag}(\tilde{X}_k: \tilde{X}_k^T)\right)P^{-2}(I-P). \end{aligned}$$

Noting the following,

$$\begin{aligned} \text{diag}\left((\tilde{X}_k: \tilde{X}_k^T - \text{diag}(\tilde{X}_k: \tilde{X}_k^T))(\beta^* \beta^{*T} - \text{diag}(\beta^* \beta^{*T}))\text{diag}(\tilde{X}_k: \tilde{X}_k^T)\right) &= \text{diag}\left(\tilde{X}_k: \tilde{X}_k^T \beta^* \beta^{*T} \text{diag}(\tilde{X}_k: \tilde{X}_k^T)\right) \\ &- \text{diag}(\tilde{X}_k: \tilde{X}_k^T)\text{diag}(\beta^* \beta^{*T})\text{diag}(\tilde{X}_k: \tilde{X}_k^T), \end{aligned}$$

one has

$$\begin{aligned} \mathbb{E}_{D_k}: [A_3] &= P^{-1}P(\tilde{X}_k: \tilde{X}_k^T - \text{diag}(\tilde{X}_k: \tilde{X}_k^T))(\beta^* \beta^{*T} - \text{diag}(\beta^* \beta^{*T}))\text{diag}(\tilde{X}_k: \tilde{X}_k^T)PP^{-2}(I-P) \\ &+ P^{-1}(P-P^2)\text{diag}\left(\tilde{X}_k: \tilde{X}_k^T \beta^* \beta^{*T} \text{diag}(\tilde{X}_k: \tilde{X}_k^T)\right)P^{-2}(I-P) \\ &- P^{-1}(P-P^2)\text{diag}(\tilde{X}_k: \tilde{X}_k^T)\text{diag}(\beta^* \beta^{*T})\text{diag}(\tilde{X}_k: \tilde{X}_k^T)P^{-2}(I-P) \quad (19) \end{aligned}$$

Term B can be made explicit as follows

$$\tilde{X}_k: \beta^{*T} \text{diag}(\tilde{X}_k: \tilde{X}_k^T) = \begin{pmatrix} \beta_1^* \delta_{i1}^3 X_{i1}^3 & \beta_1^* \delta_{i1}^2 \delta_{i2} X_{i1}^2 X_{i2} & & \\ \beta_2^* \delta_{i2}^2 X_{i2}^2 \delta_{i1} X_{i1} & \beta_2^* \delta_{i2}^3 X_{i2}^3 & & \\ & & \ddots & \end{pmatrix}$$

which implies

$$\begin{aligned} \mathbb{E}_{D_k}: [B] &= y_k P^{-1}P X_k: \beta^{*T} \text{diag}(X_k: X_k^T)PP^{-2}(I-P) \\ &+ y_k P^{-1}(P-P^2)\text{diag}(X_k: \beta^{*T} \text{diag}(X_k: X_k^T))P^{-2}(I-P). \quad (20) \end{aligned}$$

Putting Equations (17), (18), (19) and (20) together,

$$\mathbb{E}[T_2 + T_2^T] = \mathbb{E}_{(X_k, y_k)} [T_{21} + T_{22} + T_{23} + T_{23}^T + T_{24} + T_{24}^T + T_{25}]$$

$$\begin{aligned} T_{21} &= -2(P^{-1} - I)\text{diag}(X_k: X_k^T)\beta^* \beta^{*T} \text{diag}(X_k: X_k^T)(P^{-1} - I) \\ T_{22} &= -2P^{-3}((I - P)(I - 3P + 2P^2)\text{diag}(X_k: X_k^T)\text{diag}(\beta^* \beta^{*T})\text{diag}(X_k: X_k^T)) \\ T_{23} &= -X_k: X_k^T \text{diag}(\beta^* \beta^{*T})\text{diag}(X_k: X_k^T)(P^{-2}(I - P) - P^{-1}(I - P)) \\ T_{24} &= -(X_k: \beta^* - y_k)X_k: \beta^{*T} \text{diag}(X_k: X_k^T)P^{-1}(I - P) \\ T_{25} &= -2(X_k: \beta^* - y_k)(I - P)\text{diag}(X_k: \beta^{*T} \text{diag}(X_k: X_k^T))P^{-2}(I - P), \end{aligned}$$

Computation w.r.t. $\mathbb{E}_{(X_k, y_k)}$. For T_{21} , it trivially holds that

$$-\text{diag}(X_k: X_k^T)\beta^* \beta^{*T} \text{diag}(X_k: X_k^T) \preceq 0. \quad (21)$$

Indeed, for all vector $v \neq 0$,

$$\sum_{j=1}^d X_{kj}^4 \beta_j^{*2} v_j^2 + 2 \sum_{1 \leq j < m \leq d} \beta_j^* \beta_m^* X_{kj}^2 X_{km}^2 v_j v_m = \left(\sum_{j=1}^d X_{kj}^2 \beta_j^* v_j \right)^2 \geq 0.$$

Denoting the maximum of the coefficients of P as $p_M = \max_j p_j$, one has

$$\begin{aligned} T_{21} &\preceq -2 \frac{(1-p_M)^2}{p_m^2} \text{diag}(X_k: X_k: T) \beta^* \beta^{*T} \text{diag}(X_k: X_k: T) \\ &\preceq 0 \end{aligned} \quad (\text{using (21)}).$$

T_{22} is split into two terms,

$$\begin{aligned} T_{22a} &= -2P^{-3}((I-P)(I+2P^2)) \text{diag}(X_k: X_k: T) \text{diag}(\beta^* \beta^{*T}) \text{diag}(X_k: X_k: T) \\ T_{22b} &= 6P^{-2}(I-P) \text{diag}(X_k: X_k: T) \text{diag}(\beta^* \beta^{*T}) \text{diag}(X_k: X_k: T) \\ T_{22a} &\preceq -2 \frac{(1-p_M)(1+2p_M^2)}{p_m^3} \text{diag}(X_k: X_k: T) \text{diag}(\beta^* \beta^{*T}) \text{diag}(X_k: X_k: T) \preceq 0, \end{aligned}$$

since it is a diagonal matrix with only negative coefficients, and noting that $\frac{(1-p_M)(1+2p_M^2)}{p_m^3} > 0$. Then,

$$T_{22b} \preceq \frac{6(1-p_m)}{p_m^2} \text{diag}(X_k: X_k: T) \text{diag}(\beta^* \beta^{*T}) \text{diag}(X_k: X_k: T)$$

which implies

$$\mathbb{E}_{(X_k, y_k)} [T_{22b}] \preceq \frac{6(1-p_m)}{p_m^2} \|X_k\|^2 \|\beta^*\|^2 H$$

using (14) and (11).

As for $T_{23} + T_{23}^T$, note that

$$T_{23} + T_{23}^T \preceq -2 \frac{(p_M - 1)^2}{p_m^2} (X_k: X_k: T \text{diag}(\beta^* \beta^{*T}) \text{diag}(X_k: X_k: T) + \text{diag}(\beta^* \beta^{*T}) \text{diag}(X_k: X_k: T) X_k: X_k: T)$$

One prove that

$$\begin{aligned} &-(X_k: X_k: T \text{diag}(\beta^* \beta^{*T}) \text{diag}(X_k: X_k: T) + \text{diag}(X_k: X_k: T) \text{diag}(\beta^* \beta^{*T}) X_k: X_k: T) \\ &\preceq -2 \left(\min_{j=1, \dots, d} \beta_j^{*2} X_{kj}^2 \right) X_k: X_k: T \end{aligned} \quad (22)$$

Indeed, denoting $m = \left(\min_{j=1, \dots, d} \beta_j^{*2} X_{kj}^2 \right)$, one has

$$\begin{aligned} &v^T (-2m X_k: X_k: T + (X_k: X_k: T \text{diag}(\beta^* \beta^{*T}) \text{diag}(X_k: X_k: T) + \text{diag}(X_k: X_k: T) \text{diag}(\beta^* \beta^{*T}) X_k: X_k: T)) v \geq 0 \\ \Leftrightarrow &\sum_{j=1}^d (-2m X_{kj}^2 + 2X_{kj}^4 \beta_j^{*2}) v_j^2 + 2 \sum_{1 \leq j < q \leq d} (-2m X_{kj} X_{kq} + X_{kj}^3 X_{kq} \beta_j^{*2} + X_{kq}^3 X_{kj} \beta_q^{*2}) v_j v_q \geq 0 \\ \Leftrightarrow &\sum_{j=1}^d (-2m X_{kj}^2 + 2X_{kj}^4 \beta_j^{*2}) v_j^2 + 2 \sum_{1 \leq j < q \leq d} \sqrt{(-2m X_{kj}^2 + 2X_{kj}^4 \beta_j^{*2}) (-2m X_{kq}^2 + 2X_{kq}^4 \beta_q^{*2})} v_j v_q \geq 0 \\ \Leftrightarrow &\left(\sum_{j=1}^d \sqrt{(-2m X_{kj}^2 + 2X_{kj}^4 \beta_j^{*2})} v_j \right)^2 \geq 0, \end{aligned}$$

using that

$$\begin{aligned} &(-2m X_{kj}^2 + 2X_{kj}^4 \beta_j^{*2}) (-2m X_{kq}^2 + 2X_{kq}^4 \beta_q^{*2}) \geq (-2m X_{kj} X_{kq} + X_{kj}^3 X_{kq} \beta_j^{*2} + X_{kq}^3 X_{kj} \beta_q^{*2})^2 \\ \Leftrightarrow &(X_{kj}^3 X_{kq} \beta_j^{*2} - X_{kq}^3 X_{kj} \beta_q^{*2})^2 \geq 0 \end{aligned}$$

Therefore

$$\mathbb{E}_{(X_k, y_k)} [T_{23} + T_{23}^T] \preceq -2 \frac{(p_M - 1)^2}{p_m^2} \left(\min_{j=1, \dots, d} \beta_j^{*2} X_{kj}^2 \right) H \preceq 0,$$

since H is definite positive.

Finally one uses $(X_{k:}^T \beta^* - y_k) = \epsilon_k$ to conclude by independence that $T_{24} = T_{25} = 0$.

One gets

$$\mathbb{E} [T_2 + T_2^T] \preceq \frac{6(1-p_m)}{p_m^2} \|X_{k:}\|^2 \|\beta^*\|^2 H. \quad (23)$$

Combining (12), (16) and (23) leads to the desired bound. \square

S2.3 Proof of Lemma 3

Lemma S3. For all $k \geq 0$, given the binary mask D , the adjusted gradient $\tilde{g}_k(\beta)$ is a.s. $L_{k,D}$ -Lipschitz continuous, i.e. for all $u, v \in \mathbb{R}^d$,

$$\|\tilde{g}_k(u) - \tilde{g}_k(v)\| \leq L_{k,D} \|u - v\| \text{ a.s.}$$

Set

$$L := \sup_{k,D} L_{k,D} \leq \frac{1}{p_m^2} \max_k \|X_{k:}\|^2 \text{ a.s.} \quad (24)$$

In addition, for all $k \geq 0$, $\tilde{g}_k(\beta)$ is almost surely co-coercive.

Proof. Note that

$$\begin{aligned} \|\tilde{g}_k(u) - \tilde{g}_k(v)\| &= \left\| \left(P^{-1} \tilde{X}_{k:} \tilde{X}_{k:}^T P^{-1} - (I - P) P^{-2} \text{diag}(\tilde{X}_{k:} \tilde{X}_{k:}^T) \right) (u - v) \right\| \\ &\leq \left\| \left(P^{-1} \tilde{X}_{k:} \tilde{X}_{k:}^T P^{-1} - (I - P) P^{-2} \text{diag}(\tilde{X}_{k:} \tilde{X}_{k:}^T) \right) \right\| \|u - v\| \\ &\leq \left\| \frac{1}{p_m^2} \left(\tilde{X}_{k:} \tilde{X}_{k:}^T - (1 - p_m) \text{diag}(\tilde{X}_{k:} \tilde{X}_{k:}^T) \right) \right\| \|u - v\| \\ &\leq \frac{1}{p_m^2} \|\tilde{X}_{k:}\|^2 \|u - v\|, \end{aligned}$$

where we have used the Weyl inequality in the last step.

One can thus choose $L_{k,D} = \frac{1}{p_m^2} \|\tilde{X}_{k:}\|^2$ and

$$L = \sup_{k,D} L_{k,D} \leq \frac{1}{p_m^2} \sup_k \|X_{k:}\|^2 \leq \frac{1}{p_m^2} \max_k \|X_{k:}\|^2$$

Then, let us prove that the primitive of the adjusted gradient \tilde{g}_k is convex. To do this, we check that the derivative of \tilde{g}_k is definite positive:

$$\frac{\partial}{\partial \beta} \tilde{g}_k(\beta) = \frac{1}{p^2} \left(\tilde{X}_{k:} \tilde{X}_{k:}^T - (1 - p) \text{diag} \left(\tilde{X}_{k:} \tilde{X}_{k:}^T \right) \right)$$

since $\left(\tilde{X}_{k:} \tilde{X}_{k:}^T - (1 - p) \text{diag} \left(\tilde{X}_{k:} \tilde{X}_{k:}^T \right) \right)$ is positive semi-definite. Indeed,

$$\begin{aligned} v^T \left(\tilde{X}_{k:} \tilde{X}_{k:}^T - (1 - p) \text{diag} \left(\tilde{X}_{k:} \tilde{X}_{k:}^T \right) \right) v &\geq 0 \\ \Leftrightarrow \sum_{j=1}^d p \tilde{X}_{kj}^2 v_j^2 + 2 \sum_{1 \leq j < l \leq d} \tilde{X}_{kj} \tilde{X}_{kl} v_j v_l &\geq 0 \\ \Leftrightarrow \left(\sum_{j=1}^d \sqrt{p} \tilde{X}_{kj} v_j \right)^2 &\geq 0, \end{aligned}$$

using $p^2 \left(\tilde{X}_{kj} \right)^2 \left(\tilde{X}_{kl} \right)^2 \leq \left(\tilde{X}_{kj} \right)^2 \left(\tilde{X}_{kl} \right)^2$ since $p \leq 1$. \square

S3 Add-on to Section 5: Lipschitz constant computation

The Lipschitz constant L given in (24) is either computed from the complete covariates (oracle estimate) $\hat{L}_n^{\text{OR}} = \frac{1}{\hat{p}_m^2} \max_{1 \leq k \leq n} \|X_{k:}\|^2$, or estimated from the incomplete data matrix, $\hat{L}_n^{\text{NA}} = \frac{1}{\hat{p}_m^2} \max_{1 \leq k \leq n} \frac{\|\tilde{X}_{k:}\|^2 d}{\sum_j D_{kj}}$, with $\hat{p}_m = \min_{1 \leq j \leq d} \hat{p}_j$, and $\hat{p}_j = \frac{\sum_k D_{kj}}{n}$. In \hat{L}_n^{NA} , the squared norm of each row $\|\tilde{X}_{k:}\|^2$ is divided by the proportion of observed values $\frac{d}{\sum_j D_{kj}}$. This way, the value of $\|\tilde{X}_{k:}\|^2$ is renormalized, by taking into account that some rows may contain more missing values than others. Note that theoretically the step size has to satisfy $\alpha \leq \frac{1}{2\hat{L}_n^{\text{NA}}}$, thus \hat{L}_n^{NA} may be overestimated but should not be underestimated at the risk of instability in Algorithm 1. Figure S5 shows that using a slightly overestimated Lipschitz constant estimate does not deteriorate the convergence obtained using the oracle estimate.

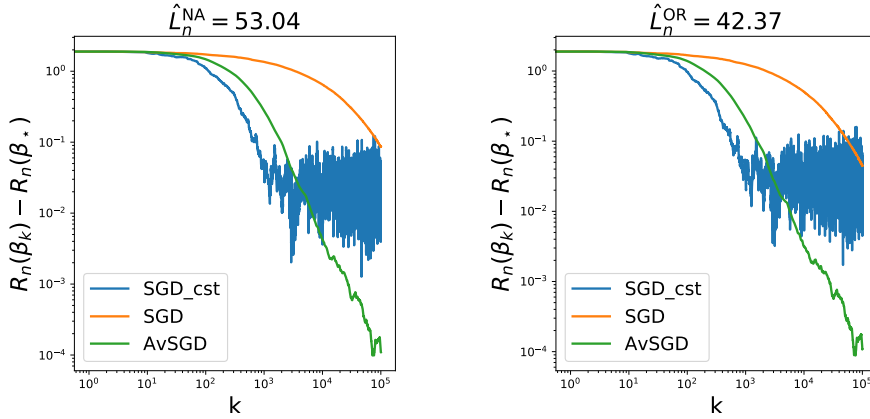


Figure S5: Empirical excess risk ($R_n(\beta_k) - R_n(\beta^*)$) given n for synthetic data ($n = 10^5$, $d = 10$) when there is 30% MCAR data, with 1 pass over the data and estimating the Lipschitz constant.

S4 Add-on to Section 5: Handling polynomial missing features

The debiased averaged SGD algorithm proposed in Section 3 can be further extended to the case of polynomial features by using a different debiasing than in Equation (4).

For example, in dimension $d = 2$, with second-order polynomial features, the interaction effect of $X_{k1}X_{k2}$ and the effects of X_{k1}^2 , X_{k2}^2 are accounted, so the augmented matrix design can be written as

$$(X_{:1}|X_{:2}|X_{:1}X_{:2}|X_{:1}^2|X_{:2}^2)^T.$$

Then, the “descent” direction at iteration k in Equation (4) should be chosen as

$$U^{\odot -1} \odot \tilde{X}_{k:} \tilde{X}_{k:}^T \beta_k - \text{diag}(U)^{\odot -1} \odot \tilde{X}_{k:} y_k$$

where

$$U = \begin{pmatrix} p_1 & p_1 p_2 & p_1 p_2 & p_1 & p_1 p_2 \\ p_1 p_2 & p_2 & p_1 p_2 & p_1 p_2 & p_2 \\ p_1 p_2 & p_1 p_2 & p_1 p_2 & p_1 p_2 & p_1 p_2 \\ p_1 & p_1 p_2 & p_1 p_2 & p_1 & p_1 p_2 \\ p_1 p_2 & p_2 & p_1 p_2 & p_1 p_2 & p_2 \end{pmatrix},$$

and $\text{diag}(U)$ denotes the vector formed by the diagonal coefficients of U and $U^{\odot -1}$ stands for the matrix formed of the inverse coefficients of U .

Synthetic data Considering a second-order model, we simulate data according to $y = (X_{:1}X_{:2}|X_{:1}^2|X_{:2}^2)^T \beta^* + \epsilon$. An additional experiment is given in Figure S6 in Appendix S3, illustrating that Algorithm 1 still achieves a rate of $\mathcal{O}(\frac{1}{n})$ while dealing with polynomial features of degree 2.

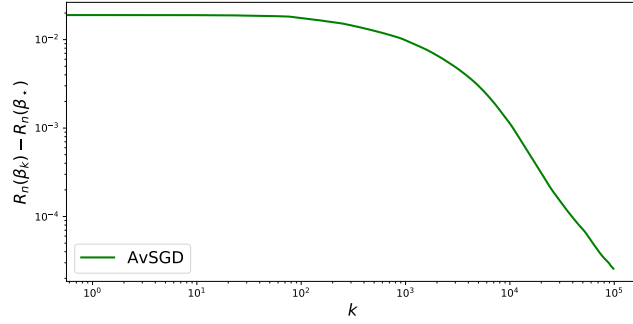


Figure S6: Empirical excess risk ($R_n(\beta_k) - R_n(\beta^*)$) given n for synthetic data ($n = 10^5$, $d = 10$) when the model accounts mixed effects.

Real dataset About large-scale setting there is no computational barrier to apply the proposed method in high dimension, as the computational cost is similar to standard SGD strategies without missing data. These are computationally cheap at each iteration and particularly relevant on large datasets. In this section, we propose to run the proposed algorithm on the superconductivity dataset as in Subsection 5.3. 30% of missing values are uniformly introduced in the initial 81 features, with $n = 21263$. However, here we consider polynomial features of order 2, which increases the initial dimension 81 to 3400.

The empirical proportions of missing values for each variable in the resulting dataset are represented on Figure S7, and the observed convergence rate for one pass on the data is displayed in Figure S8. With the same numerical complexity, Algorithm 1 performs as well as an averaged SGD strategy run on the complete observations, whereas a standard SGD strategy run on imputed-by-0 data saturates far from the optimum.

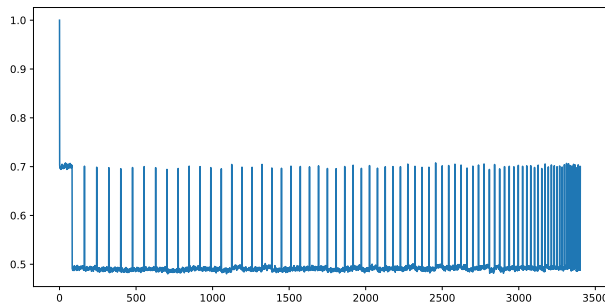


Figure S7: Proportion of missing values for the polynomial features of degree 2 on the superconductivity dataset, when the initial missingness proportion on the raw features is 30%.

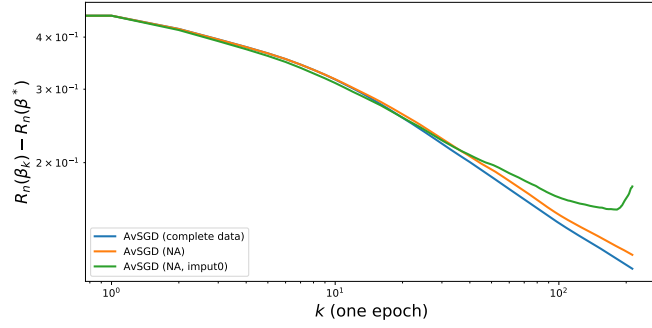


Figure S8: Empirical excess risk ($R_n(\beta_k) - R_n(\beta^*)$) given n for the superconductivity dataset ($n = 21263$) (containing 81 initial features) and $d = 3403$ with polynomial features of degree 2. Three different algorithms are compared: an averaged SGD on complete data (blue), the proposed debiased averaged SGD Algorithm 1 (orange) and an averaged SGD run on imputed-by-0 data without any debiasing (green).

S5 Add-on to Section 5: Description of the TraumaBase data variables

The variables of the TraumaBase dataset which are used in experiments are the following:

- *Lactate*: The conjugate base of lactic acid.
- *Delta.Hemo*: The difference between the hemoglobin on arrival at hospital and that in the ambulance.
- *VE*: A volume expander is a type of intravenous therapy that has the function of providing volume for the circulatory system.
- *RBC*: A binary index which indicates whether the transfusion of Red Blood Cells Concentrates is performed.
- *SI*: Shock index indicates level of occult shock based on heart rate (*HR*) and systolic blood pressure (*SBP*). $SI = \frac{HR}{SBP}$. Evaluated on arrival at hospital.
- *HR*: Heart rate measured on arrival of hospital.
- *Age*: Age.