Increasing Missingness to Reduce Bias: Richardson-SGD with Missing Data

Ferdinand Genans* **Erwan Scornet**
Sorbonne Université and Université Paris Cité, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, LPSM, F-75005 Paris, France

Abstract

Stochastic gradient methods are central to modern large-scale learning, but their use with incomplete covariates remains delicate since imputation schemes generally introduce systematic gradient biases, as shown for linear models. In this work, we prove that all parametric models exhibit similar gradient bias for various imputation procedures and characterize exactly the dependence on the missingness ratio vector p, with $O(\|p\|)$ as the leading term. We exploit this analysis to propose a simple debiasing procedure for stochastic gradient descent (SGD) with missing values based on Richardson extrapolation, which leverages the exact expression of the gradient bias. The key idea is to *deliberately add missingness*: from an already incomplete observation, we generate a further-thinned version at a higher, controlled missingness level, and combine the two resulting stochastic gradients to cancel the leading bias term. We prove that one Richardson step reduces the gradient bias from $O(\|p\|)$ to $O(\|p\|^2)$ under several missingness scenarios. Our proposed method is computationally efficient, model-agnostic and applies to any parametric loss whose stochastic gradient can be computed after imputation. Furthermore, when missing indicators are independent, the population gradient bias is a multilinear polynomial in p and depends only on population gradient errors induced by declaring a single coordinate missing. In this case, our method generalizes to a multi-step Richardson procedure which recursively cancels higher-order terms. Empirically, Richardson debiasing improves optimization and estimation across several generalized linear models and combines positively with widely used imputation procedures such as MICE. These results suggest that, somewhat counter-intuitively, adding controlled missingness on top of existing missing data can make stochastic learning from incomplete data more accurate.

1 Introduction

Missing data are ubiquitous in modern machine learning. They may arise from database fusion, sensor failure, non-response in surveys, and selective acquisition pipelines, to name only a few. In his seminal paper, Rubin [30] formalized the missing-data framework and introduced the now-standard taxonomy of three missingness regimes: *Missing Completely at Random* (MCAR), in which missingness is independent of the data; *Missing at Random* (MAR), in which missingness depends only on observed entries; and *Missing Not at Random* (MNAR), in which missingness can also depend on the unobserved entries themselves.

Framework. In supervised learning with incomplete covariates, one typically distinguishes two goals: estimating the parameters of a model despite the missing values, and producing a predictor with high test accuracy. The two are aligned when the test set is fully observed—accurate parameter estimation then leads to strong predictive performance—but they decouple when the test set itself contains missing entries, in which case a separate prediction-time strategy is required [see e.g. 14, 35, 13]. We

*Corresponding author: genans.ferdinand@gmail.com

Preprint.

focus on the first objective and assume that missing values appear only in the training set, while the test set is complete. Even in this setting, parameter identifiability is not guaranteed under arbitrary MNAR mechanisms [see, e.g., the examples and discussions in 29, 38, 20]. We therefore restrict our attention to MCAR and a generalization—scalable MAR—in which the conditional missingness probability depends on a known intensity function.

Handling missing data in parametric models. The simplest approach is *complete-case* analysis [26, 16], which discards every sample containing at least one missing entry. This is unbiased under MCAR but throws away samples at a rate that is exponential in the dimension. The next-simplest approach is *imputation*: missing entries are replaced by point estimates, after which any standard learning algorithm can be applied to the completed dataset. Constant imputation (zero or mean) is the most studied [12] and the easiest to analyze, but it injects a systematic bias even under MCAR. Multivariate Imputation by Chained Equations [MICE, 34] and nearest-neighbour or neural-network imputation schemes [33, 19] reduce this bias empirically but offer few formal estimation guarantees. A complementary line of work avoids imputation altogether by working with the joint distribution of inputs and mask. The Expectation–Maximization algorithm of Dempster et al. [9], refined for incomplete data by Ibrahim [10] and extended to logistic regression via Stochastic Approximation EM [11], fits a parametric model to the inputs and the predictor jointly. These algorithms require a known parametric family for the covariates and can be expensive due to the E-step. A first review of estimation procedures for linear regression with missing covariates was given in Little [15].

Related work - SGD with missing data. Gradient descent and its stochastic variants are the workhorse of large-scale learning, but they require a fully observed input to compute a gradient. The natural fix is to impute and then run SGD on the completed dataset; as observed by Jones [12], this leads to a biased gradient. Ayme et al. [2, 3] study the test-time predictive performance of SGD applied to zero-imputed data when data can be missing in both train and test set, relating the imputation bias to a ridge regularization effect and leveraging the implicit bias of SGD [31] towards low-norm solutions to derive convergence rates [see also 7, 36, for missing data in high-dimensional linear models]. Another line of work focuses on parameter estimation or equivalently on test-time performance when the test set is assumed to contain complete data. Loh and Wainwright [17] characterize the exact bias induced by zero imputation in linear models and use this characterization to obtain the first parameter-estimation rates for sparse high-dimensional linear regression under MCAR. Building on their analysis, Needell [23] design a stochastic gradient algorithm that is better suited to large-scale data, and Sportisse et al. [32] establish that averaged debiased SGD attains the optimal one-pass rate for linear regression.

Contributions. We propose and analyze a debiasing procedure for stochastic gradients computed from imputed data, which can be applied to any parametric model and which is valid for a large class of imputation procedures. Our proposed method applies Richardson extrapolation [28] to the missingness scale p, yielding a model-agnostic correction that can be combined with a broad class of imputation rules. While Richardson extrapolation has been used in machine learning to remove leading-order biases in other contexts [5], to the best of our knowledge, this is the first application to the missingness scale of a stochastic gradient. Our contributions are as follows.

Gradient-bias structure. Under several MCAR and MAR settings described below, we establish the exact expression of the population bias of a stochastic gradient computed on imputed data. In doing so, we generalize the expression obtained by Sportisse et al. [32] for linear regression with zero-imputed data and independent MCAR missingness to any parametric model, a large class of imputation procedures, and a broad class of missingness scenarios (non-independent MCAR and MAR). As a consequence, we show that the gradient bias is $O(\|p\|)$, where $p = (p_1, \dots, p_d)$ with p_j being the probability that the jth component is missing. When the mask components are independent, we prove that the order of the remaining terms is $O(\|p\|^2)$. Our bias decomposition holds for generic imputation procedures: better imputation may shrink the constants in $O(\|p\|)$ but cannot generally remove the leading $O(\|p\|)$ bias (Section 3).

Richardson-SGD. We introduce a thinning construction that, from a sample with mask at scale p, generates a further-thinned mask at scale Cp, for some well-chosen $C > 1$, using one extra Bernoulli draw per observed entry. A Richardson combination of the two gradients (computed on an imputed dataset at scales Cp) reduces the bias from $O(\|p\|)$ to $O(\|p\|^2)$ under some MCAR and MAR settings with independent masking components (Section 4). We also introduce a multi-step Richardson-based procedure which cancels higher-order terms, with exact cancellation for d_{miss} steps, where d_{miss} is the number of covariates subject to missingness.

Theory for one-pass SGD. Our bias expansion plugs directly into classical biased-SGD proofs. For one-pass (one-epoch) SGD over n samples, and given a smooth and strongly convex loss, Richardson-SGD attains $\mathbb{E}\|w_n - w^\star\|^2 = O(\|p\|^4) + O(1/n)$, against $O(\|p\|^2) + O(1/n)$ for plain imputed SGD. Multi-step Richardson reduces the missingness term further at the price of increased variance. (Section 5).

Experiments. We validate the theory on synthetic and real datasets. For a variety of generalized linear models, Richardson-SGD improves over plain imputation under several MCAR and MAR mechanisms and combines positively with MICE, Random-Forest MICE, and k-NN imputation (Section 6).

2 Setting

Random covariates and notation. Random variables are written in uppercase (X, Y, M); their realizations are written in the corresponding lowercase (x, y, m). For an integer $d \geq 1$ we set $[d] := \{1, \dots, d\}$. For any $S \subseteq [d]$, we write $S^c := [d] \setminus S$, and for a vector $v \in \mathbb{R}^d$ we let $v^{(S)} \in \mathbb{R}^{|S|}$ denote the subvector indexed by S. Throughout the paper, $\|\cdot\|$ is the Euclidean norm and $\|\cdot\|_\infty$ the supremum norm.

Supervised learning and SGD. We consider a supervised learning setting with random covariates $X \in \mathcal{X} \subseteq \mathbb{R}^d$, response $Y \in \mathcal{Y}$, parameter $w \in W \subseteq \mathbb{R}^q$, and a continuously differentiable loss $\ell : W \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. We aim at minimizing the population risk $L(w) := \mathbb{E}[\ell(w; X, Y)]$. The complete-data single-sample gradient is $g(w; x, y) := \nabla_w \ell(w; x, y)$. Assuming differentiation and expectation commute, we have $\nabla L(w) = \mathbb{E}[g(w; X, Y)]$. Because ∇L has no closed form in general, we use stochastic gradient descent (SGD), whose updates are given by

$$w_{k+1} = w_k - \eta_k \hat{g}_k(w_k), \tag{1}$$

where $\eta_k > 0$ is the step-size and $\hat{g}_k(w_k)$ is a stochastic estimator of $\nabla L(w_k)$ computed at iteration k from a sample or minibatch. When the sample is complete, $g(w; X, Y)$ is unbiased for $\nabla L(w)$.

Missing covariates and imputation. For each training sample, the learner observes a realisation of (X^{obs}, Y, M), where $M \in \{0,1\}^d$ is a missingness mask and, for all $j \in [d]$, $X_j^{\mathrm{obs}} = X_j$ if $M_j = 0$ and $X_j^{\mathrm{obs}} = \mathrm{NA}$ if $M_j = 1$. We write $p_j := \mathbb{P}(M_j = 1)$ for the marginal missingness probability of feature j and $p := (p_1, \dots, p_d)$ for the missingness vector. Missing entries are filled in by an imputation rule $\mathcal{I}$ that produces an imputed covariate vector

$$\tilde{X} := \mathcal{I}\big(X^{\mathrm{obs}}, M, \xi\big), \tag{2}$$

where $\xi \perp M \mid (X, Y)$ collects auxiliary randomness used by $\mathcal{I}$. We focus on *data-independent imputation rules* $\mathcal{I}$ that impute each observation independently of the others. This assumption makes our analysis tractable by enabling a decomposition at the sample level. Standard imputations (mean, iterative) can be slightly modified to fall into this setting by training $\mathcal{I}$ on an auxiliary dataset. The imputed stochastic gradient available to the learner is $\hat{g}(w) := g\big(w; \tilde{X}, Y\big)$, which is, in general, a *biased* estimator of $\nabla L(w)$, with bias

$$\mathcal{B}(w, p) := \mathbb{E}[\hat{g}(w)] - \nabla L(w), \tag{3}$$

the expectation being over (X, Y), the mask M, and the imputation randomness ξ.

Missingness mechanisms. We follow the taxonomy of Rubin [30]: the mask is *Missing Completely at Random* (MCAR) when $M \perp (X, Y)$, and *Missing at Random* (MAR) when, conditionally on the observed entries, M is independent of the missing entries. Throughout the paper, we let $\mathcal{O} \subseteq [d]$ (possibly empty) be the set of indices of variables that are *always-observed*. We let $V := X^{(\mathcal{O})}$ be the vector of *always-observed variables*. To enable a tractable analysis, we focus on two concrete mechanisms, which depend on the probability vector p, assumed to be known.

Heterogeneous MCAR (hMCAR). M is independent of (X, Y) and $\mathbb{P}(M_j = 1) = p_j$.

Scalable MAR (sMAR). $\{M_j\}_{j \in \mathcal{O}^c} \perp (X^{(\mathcal{O}^c)}, Y) \mid V$, and for every $j \in \mathcal{O}^c$,

$$\mathbb{P}\big(M_j = 1 \,\big|\, V\big) = p_j \, q_j(V),$$

for known intensity functions $q_j : \mathbb{R}^{|\mathcal{O}|} \to [0, p_j^{-1}]$ with $\mathbb{E}[q_j(V)] = 1$.

It is known that MAR settings contain scenarios of different difficulties [22], some of which being close to MNAR settings [21], for which identifiability does not always hold [see, e.g., 29, 38, 20]. Thus, we restrict the MAR settings we consider via the sMAR assumption. Note that the condition $\mathbb{E}[q_j(V)] = 1$ in sMAR is necessary to ensure that $p_j = \mathbb{P}(M_j = 1)$. A concrete example of sMAR is a logistic missingness mechanism, as commonly used in simulation studies of missing covariates [e.g. 18, 37]. We say a mask is *independent hMCAR* (resp. *independent sMAR*) if it is hMCAR (resp. sMAR) and the $\{M_j\}_{j \in \mathcal{O}^c}$ are mutually independent (resp. conditionally on V).

Our objective remains the complete-data risk $L(w)$ and its minimizer $w^\star$; missingness and imputation only affect the stochastic gradients used to optimize it. Our goal is to replace the imputed gradient $\hat{g}$ in (1) by a corrected gradient $\hat{g}^{\mathrm{R}}$, computed from the same observation plus a small amount of controlled additional thinning, so as to cancel or shrink the gradient bias (3).

3 First-order structure of the missingness bias

Before designing a debiasing procedure, we describe the structure of the imputation-induced gradient bias as a function of the missingness scale p. The key observation is that, regardless of the loss and the imputation rule, the bias admits a clean expansion whose leading term is *linear* in p and whose coefficients are population gradient gaps that do not depend on p. This expansion will be the structural fact that Richardson extrapolation later exploits.

Proposition 1 (First-order structure of the missingness bias)**.** *Consider any data-independent imputation defined in* (2) *and assume hMCAR or sMAR holds. Then the population gradient bias* (3) *can be decomposed as*

$$\mathcal{B}(w, p) \;=\; \mathcal{A}(w)\, p \;+\; \mathcal{R}(w, p), \tag{4}$$

with $\mathcal{A}(w) \in \mathbb{R}^{q \times d}$ independent of p. Letting $a_j(V) = 1$ for hMCAR and $a_j(V) = q_j(V)$ for sMAR, the j-th column of $\mathcal{A}(w)$ is the population gradient gap obtained by declaring coordinate j missing:

$$\mathcal{A}_{\cdot j}(w) \;=\; \mathbb{E}\left[a_j(V)\left\{G_{\{j\}}(w; X, Y, \xi) - g(w; X, Y)\right\}\right]. \tag{5}$$

The remainder $\mathcal{R}(w, p)$ contains the co-missingness contributions, namely the terms involving simultaneous missingness of two or more coordinates. The exact expression of $\mathcal{R}(w, p)$ is given in Appendix B. The proof is based on a discrete-difference expansion over missingness patterns and separates the contribution of each joint missingness pattern $S \subseteq [d]$.

In full generality, the remainder is at most linear in p, while it is $o(\|p\|)$ in most scenarios. Indeed, strong dependence among mask components can make co-missingness terms contribute at first order. For instance, this may occur when two coordinates are perfectly negatively associated, so that $\mathbb{P}(M_j = 1 \mid M_k = 1) = 0$. The following corollary identifies a key regime motivating Richardson extrapolation.

Corollary 1. *Under the assumptions of Proposition 1, suppose in addition that the missingness indicators $\{M_j\}_{j \in \mathcal{O}^c}$ are conditionally independent given V. Then*

$$\|\mathcal{R}(w, p)\| \;=\; O(\|p\|^2), \qquad \textit{and therefore} \qquad \big\|\mathcal{B}(w, p) - \mathcal{A}(w)\, p\big\| \;=\; O(\|p\|^2). \tag{6}$$

Under independent hMCAR\sMAR, the bias is a *multilinear* in p (see proof of Corollary 1),

$$\mathcal{B}(w, p) = \sum_{\varnothing \neq S \subseteq [d]} \mu_S(w) \Big(\prod_{j \in S} p_j\Big), \quad \mu_S(w) \;:=\; \mathbb{E}\left[\Big(\prod_{j \in S} a_j(V)\Big)\, \Delta_S G_\varnothing(w; X, Y, \xi)\right], \tag{7}$$

where $T_j G_\varnothing := G_{\{j\}}$ declares coordinate j missing and $\Delta_S G_\varnothing := \prod_{j \in S}(T_j - I) G_\varnothing$. The coefficient $\mu_S(w)$ aggregates the effect of $|S|$-fold co-missingness. Equation (7) is the structural fact that drives both first and higher-order Richardson cancellation.

The decomposition has three implications. *(i) The leading bias is linear in p:* under conditional independence, the remainder is $O(\|p\|^2)$, so the first-order behavior is fully captured by $\mathcal{A}(w)\, p$. *(ii) The leading operator is an average gradient gap:* the column $\mathcal{A}_{\cdot j}(w)$ vanishes whenever coordinate j is always observed or is perfectly recovered by $\mathcal{I}$. *(iii) Imputation reduces constants, not the leading*

order: the expansion holds for any data-independent imputation, and a better imputation rule only shrinks the entries of $\mathcal{A}_{\cdot j}(w)$ without changing the order of $\mathcal{B}(w,p)$ in p.

These three points together suggest a clear strategy. Imputation alone cannot remove the leading $O(\|p\|)$ scaling, except if it fully recovers the covariate. Improving the imputation only refines the constants $\mathcal{A}_{\cdot j}(w)$. To eliminate the leading order, we propose to act *on p itself*—that is, evaluate the imputed gradient at two different missingness scales and combine the results so that the linear contribution cancels. This is precisely what Richardson extrapolation achieves, and the construction we develop in the next section turns this idea into a practical SGD update.

4 Richardson-SGD

Richardson extrapolation in a nutshell. Richardson extrapolation [28] cancels the leading term of an asymptotic expansion. If $T(p) = T_0 + pT_1 + p^2T_2 + o(p^2)$ as $p \to 0$ and $C > 1$, the combination

$$T_C^{\mathrm{R}}(p) \;:=\; \tfrac{C\,T(p) - T(Cp)}{C-1} \;=\; T_0 - Cp^2T_2 + o(p^2) \tag{8}$$

eliminates the linear term. With $k+1$ scales $1 = C_0 < C_1 < \cdots < C_k$ and a Vandermonde weight vector, the first k orders are cancelled simultaneously [24].

At first sight, applying (8) to the missingness bias would require evaluating the imputed gradient at *two* missingness scales p and Cp on the same observation. The learner, however, only observes a single mask $M^{(p)}$ at scale p. We resolve this with a single extra Bernoulli draw per observed entry: from a sample at scale p, we *further thin* it to obtain a mask whose conditional law given X is exactly that of an independent draw at scale Cp. No new observation is required. We employ this additional mask to propose a Richardson-corrected gradient, used in lieu of the standard gradient in a SGD procedure.

Further-thinned mask. Fix $C > 1$ such that, for all $j \in \mathcal{O}^c$, $C\,p_j a_j(V) \le 1$ almost surely. Conditional on $(X, M^{(p)})$, draw independent thinning bits r_j with $r_j = 1$ for $j \in \mathcal{O}$ and, for $j \in \mathcal{O}^c$,

$$r_j \,\big|\, (X, M^{(p)}) \;\sim\; \mathrm{Bernoulli}\!\left(\frac{1 - Cp_j a_j(V)}{1 - p_j a_j(V)}\right), \text{and let} \quad M_j^{(Cp)} \;:=\; 1 - (1 - M_j^{(p)})\, r_j. \tag{9}$$

All entries missing under $M^{(p)}$ stay missing under $M^{(Cp)}$; an observed entry is hidden under $M^{(Cp)}$ exactly when $r_j = 0$. A short calculation (Appendix C) gives $\mathbb{P}(M_j^{(Cp)} = 1 \mid V) = Cp_j a_j(V)$, so $M^{(Cp)}$ has the same conditional law as the original mask but at scale Cp.

Richardson-corrected gradient. Equipped with the further-thinned mask, we can apply (8) to the imputed gradient. Crucially, we must *not* impute the same observation at two different missingness levels, since we need common missing values between the two scales to be identical (see Appendix L for further explanation and a numerical illustration). We impute once on the more thinned sample at scale Cp, then *restore* the artificially hidden entries to recover the imputation at scale p:

$$\tilde{X}^{(Cp)} \;:=\; \mathcal{I}\big(X^{\mathrm{obs}}, M^{(Cp)}, \xi\big), \qquad \tilde{X}^{(p)}_{Cp,j} := \begin{cases} X_j, & M_j^{(p)} = 0, \\ \tilde{X}_j^{(Cp)}, & M_j^{(p)} = 1, \end{cases} \quad \text{for all } j \in [d]\,.$$

Set $\hat{g}^{(p)}(w) := g(w; \tilde{X}^{(p)}, Y)$ and $\hat{g}^{(Cp)}(w) := g(w; \tilde{X}^{(Cp)}, Y)$. The *Richardson-corrected gradient* is

$$\hat{g}_C^{\mathrm{R}}(w) \;:=\; \frac{C\,\hat{g}^{(p)}(w) \;-\; \hat{g}^{(Cp)}(w)}{C-1}. \tag{10}$$

Richardson-SGD plugs $\hat{g}_C^{\mathrm{R}}$ into the SGD update (1). For each sampled observation $(x_i^{\mathrm{obs}}, m_i, y_i)$ at iteration k:

1. *Original masked sample.* Read off the mask $m_i^{(p)}$ at scale p.
2. *Further-thinned sample.* Draw r as in (9). For all $j \in [d]$, $m_{ij}^{(Cp)} \leftarrow 1 - (1 - m_{ij}^{(p)})\, r_j$.
3. *One imputation.* Compute $\tilde{x}_i^{(Cp)} \leftarrow \mathcal{I}(x_i^{\mathrm{obs}}, m_i^{(Cp)}, \xi_i)$, then obtain $\tilde{x}_i^{(p)}$ by overwriting the entries hidden by r with their true values from x_i.

4. *Gradient estimates.* Evaluate $\hat{g}_i^{(p)} := g(w_k; \tilde{x}_i^{(p)}, y_i)$ and $\hat{g}_i^{(Cp)} := g(w_k; \tilde{x}_i^{(Cp)}, y_i)$.
5. *Richardson correction & SGD update.* Form $\hat{g}_k^{\mathrm{R}} \leftarrow (C\,\hat{g}_i^{(p)} - \hat{g}_i^{(Cp)})/(C-1)$ and update $w_{k+1} \leftarrow w_k - \eta_k\,\hat{g}_k^{\mathrm{R}}$ (averaged across a minibatch when $b > 1$).

The procedure is a thin wrapper around any imputation-based SGD pipeline: one extra Bernoulli draw per observed entry and one extra gradient evaluation per sample.

5 Theory of Richardson-SGD

We now state the theoretical guarantees of Richardson-SGD. The analysis shows that Richardson corrections successively cancel the terms in the bias expansion, while controlling the associated variance inflation and the error from estimating the missingness mechanism. Combining these bounds with a classical biased-SGD argument yields a convergence rate. Throughout, the result applies to *one-pass (one-epoch) SGD*, as in Sportisse et al. [32] for linear regression: each sample is visited once, and the bias expansion from Section 3 feeds directly into standard biased-SGD arguments. Multi-epoch behavior is outside the scope of the theory and is examined empirically in Section 6.

5.1 First-order bias cancellation

Proposition 2 (First-order debiasing)**.** *Assume independent hMCAR or independent sMAR. Then*

$$\left\| \mathbb{E}\big[\hat{g}_C^{\mathrm{R}}(w)\big] - \nabla L(w) \right\| \;=\; O(\|p\|^2), \qquad \textit{when } \|p\| \to 0. \tag{11}$$

Proposition 2 shows that the debiasing challenge can be met by a deliberately counterintuitive operation: we decrease bias by adding missing values. While the plain imputed gradient has bias of order $\|p\|$, the Richardson-corrected gradient constructed from the original and further-thinned masks cancels this leading term and leaves only an $O(\|p\|^2)$ bias under independent hMCAR or independent sMAR. This gain is uniform in the loss and the imputation rule, and requires only one additional Bernoulli draw and one additional gradient evaluation per sample (proof in Appendix C).

5.2 Higher-order Richardson-SGD under independent masks

When the missing indicators are conditionally independent given V, Section 3 showed that the gradient bias is, in fact, a multilinear polynomial in p. Since Richardson extrapolation is itself linear in the underlying expansion, one can cancel further orders by combining estimators at more than two missingness scales. Iterating the thinning construction with $k+1$ scales $1 = C_0 < C_1 < \cdots < C_k$ (cascaded via (9) with $C \leftarrow C_\ell / C_{\ell-1}$) and Vandermonde weights $\alpha \in \mathbb{R}^{k+1}$ yields the k-th order Richardson estimator $\hat{g}^{[k]}(w) := \sum_{\ell=0}^{k} \alpha_\ell\,\hat{g}^{(C_\ell p)}(w)$.

Corollary 2 (Higher-order cancellation)**.** *Assume $C\,p_j\,a_j(V) \le 1$ for every j. Under independent hMCAR or independent sMAR, $\big\|\mathbb{E}[\hat{g}^{[k]}(w)] - \nabla L(w)\big\| = O(\|p\|^{k+1})$ as $\|p\| \to 0$. Furthermore, with $d_{\mathrm{miss}} := \mathit{Card}(\{j : p_j > 0\})$, the d_{miss}-th order estimator cancels the bias exactly: $\mathbb{E}[\hat{g}^{[d_{\mathrm{miss}}]}(w)] = \nabla L(w)$.*

For linear regression with squared loss, the bias is a polynomial of degree at most 2 in p (Appendix E), so the two-step Richardson-SGD produces an *exact* debiasing under both hMCAR and sMAR. This matches the closed-form correction mechanism of Sportisse et al. [32] as a special case and extends it to sMAR, where no closed form is available. More generally, Corollary 2 suggests that higher-order Richardson-SGD should be most useful when only a few coordinates are subject to missingness (d_{miss} small) so that the corresponding polynomial degree is low, or that the highest polynomial degree in the bias is low, as for linear regression (see Appendix D). Figure 1 illustrates this phenomenon in synthetic linear and logistic regressions.

5.3 Variance inflation

The previous two subsections highlight how Richardson reduces bias. As is standard in extrapolation methods, this comes at a price: the corrected gradient is a difference of two estimators evaluated at different missingness levels, which inflates its variance. Quantifying this inflation is essential, since

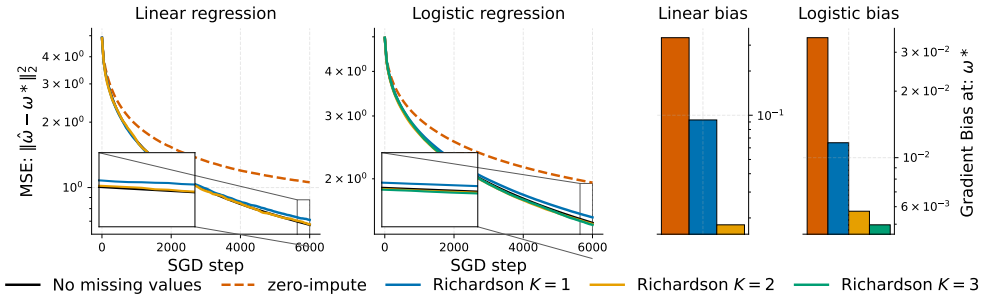

Figure 1: Multi-order Richardson correction in 4-covariate linear and logistic regression under hMCAR with $p = (0.10, 0.15, 0.08, 0.12)$, and bias at $w^\star$, $\|\mathbb{E}[\hat{g}(w^\star)] - \nabla L(w^\star)\|$ As predicted by Corollary 2, each Richardson level removes one further order of bias: first-order Richardson clearly improves over standard SGD on zero-imputed data, second-order matches the complete-data trajectory in linear regression (curves overlap), and third-order yields an additional gain in the logistic case for the bias.

the convergence rate of SGD depends on *both* the bias and the variance of the stochastic gradient. For the first-order estimator,

$$\mathrm{Var}\big[\hat{g}_C^{\mathrm{R}}(w)\big] \;\le\; \frac{2\left(C^2\,\mathrm{Var}\big[\hat{g}^{(p)}(w)\big] \;+\; \mathrm{Var}\big[\hat{g}^{(Cp)}(w)\big]\right)}{(C-1)^2}, \tag{12}$$

with a larger C controlling the multiplicative factor since the function $f : (1, +\infty) \to \mathbb{R}$, $f(x) = x^2/(x-1)^2$, is decreasing. For the k-th order estimator, variance inflates by a factor that grows with k, and requires k missingness upscales, which limits k when some p_j are large. We therefore use first-order Richardson by default and reserve higher-order constructions for small d_{miss} or for losses with low maximum polynomial degree, as linear regression, which is of degree 2 (see Appendix D).

5.4 Richardson-SGD with estimated missingness parameters

So far we have assumed that the quantities (p, q) driving the missing mechanism are known. In practice, p_j is estimated by the empirical missingness frequency on coordinate j, while $q_j(V)$ is fitted by a probabilistic model with input V. We now quantify how the resulting estimation errors propagate into the Richardson bias. Using the identifiability convention $\mathbb{E}[q_j(V)] = 1$, let $\lambda_j(V) := p_j q_j(V)$ and $\hat{\lambda}_j(V) := \hat{p}_j \hat{q}_j(V)$. The plug-in thinning rule replaces (9) by $\tilde{r}_j \sim \mathrm{Bernoulli}((1 - C\hat{\lambda}_j(V))/(1 - \hat{\lambda}_j(V)))$ and yields the plug-in Richardson gradient $\hat{g}^{\mathrm{R}}_{C,\hat{\lambda}}$.

Proposition 3 (Plug-in Richardson)**.** *Assume hMCAR or sMAR with $\lambda_j(V), \hat{\lambda}_j(V) \le \rho < 1$, $C\hat{\lambda}_j(V) \le 1$ for every j, and $\|G_S(w; X, Y, \xi) - \nabla L(w)\|_{L^2} \le G_\star$ for every $S \subseteq [d]$. If $\|\hat{p} - p\|_\infty \le \delta_p$ and $\max_j \sup_v |\hat{q}_j(v) - q_j(v)| \le \delta_q$, then*

$$\left\| \mathbb{E}[\hat{g}^{\mathrm{R}}_{C,\hat{\lambda}}(w)] - \nabla L(w) \right\| \;=\; O\big(\|p\|^2 + \delta_p + \|p\|_\infty \delta_q + \delta_p \delta_q\big)\,. \tag{13}$$

Under hMCAR ($q_j \equiv 1$, $\delta_q = 0$), this collapses to $O(\|p\|^2 + \delta_p)$.

The leading $\mathcal{A}(w)\,p$ contribution is cancelled regardless of plug-in errors, up to an additive $O(\delta_p + \|p\|_\infty \delta_q)$ penalty (proof in Appendix C.4). When δ_p, δ_q shrink fast enough, the $O(\|p\|^2)$ term dominates and the exact-mechanism guarantee is recovered. This behavior further motivates using the first-order Richardson SGD scheme, while higher order might not be conclusive in the plug-in setting. Appendix H reports an empirical sensitivity study.

5.5 One-pass SGD convergence

We have now controlled both the bias of the Richardson-corrected gradient, through Proposition 2 and Corollary 2, and its variance, through (12), including under plug-in mechanisms (Proposition 3). It remains to translate these gradient-level guarantees into a convergence rate for the SGD iterates, which is the quantity of interest. Note that biased SGD schemes have been extensively studied in the literature [see, e.g. 1, 8]. To illustrate the resulting bias improvement of Richardson-SGD compared to plain imputation, we give a result under classic regularity conditions on the loss function.

Corollary 3 (One-pass Richardson-SGD)**.** *Assume L is α-strongly convex and β-smooth, the per-sample stochastic gradients are bounded in L^2. Under independent hMCAR or independent sMAR, after one pass on n i.i.d. samples with $\eta_k = \frac{c}{k+\gamma}, c > \frac{1}{\alpha}, \gamma \geq \frac{6c\beta^2}{\alpha}$, we obtain*

$$\mathbb{E}\left\|w_n - w^\star\right\|^2 \;=\; \begin{cases} O(\|p\|^2) + O(1/n) & \textit{(plain imputed SGD)}, \\ O(\|p\|^4) + O(1/n) & \textit{(Richardson-SGD, first order)}, \end{cases} \tag{14}$$

and the same convergence orders hold for the excess test loss $\mathbb{E}[L(w_n) - L(w^\star)]$. With k-step Richardson-SGD, the missingness floor becomes $O(\|p\|^{2(k+1)})$. Thus, for sufficiently large k, the missingness contribution is dominated by the statistical floor $O(1/n)$.

Two implications of Corollary 3 are worth highlighting. First, when $\|p\| \ll 1$, one-step Richardson-SGD improves the bias floor of plain imputed SGD from $O(\|p\|^2)$ to $O(\|p\|^4)$. Thus, the missingness-induced contribution is reduced by two orders of magnitude in $\|p\|$, while keeping essentially the same per-iteration cost. Second, multi-step Richardson-SGD can, in principle, reduce the missingness term down to the statistical noise level $O(1/n)$. This comes at the price of variance inflation: the bound in (12) compounds across Richardson levels and may become prohibitive when d_{miss} is large or when the loss has heavy stochastic gradients. Consequently, multi-step Richardson-SGD is most appealing when d_{miss} is small, or in settings such as linear regression where order 2 already suffices.

Scope of the theory. We emphasize that Corollary 3 is a one-pass guarantee, in line with the regime studied by Sportisse et al. [32]. The multi-epoch behavior is not covered by our analysis: when iterates revisit the same observations, the gradient noise due to missing values across iterations are no longer independent. The experiments of Section 6 suggest, however, that Richardson-SGD remains effective in multi-epoch training, and we view a formal multi-epoch analysis as an interesting question for future work.

6 Experiments

We empirically study Richardson-SGD on synthetic and real datasets available in `scikit-learn` [25]. Throughout, missing entries are introduced *ex post* into otherwise complete datasets according to the mechanism specified in each subsection, either hMCAR or sMAR, so that the ground truth $w^\star$ is known, or can be estimated by multi-pass training with L-BFGS-B, and the quantities p and q_j are also known. Unless stated otherwise, the average missingness is fixed at $\bar{p} = 0.2$. To keep the main text concise, we report only logistic regression here; analogous experiments for other datasets and models, including linear and Poisson GLMs, together with implementation details, are deferred to Appendix I.

Empirical takeaway. Across datasets, models, missingness mechanisms, and imputation rules, Richardson-SGD behaves as a generic debiasing layer rather than a model-specific correction. It improves imputation-based SGD using only controlled thinning and one additional gradient evaluation, and remains effective when the missingness mechanism is estimated or partially misspecified. In short, the method is simple, fast, model-agnostic, and theoretically grounded, making it a natural add-on for learning with missing covariates.

6.1 Richardson with imputation on logistic regression

This experiment tests the central practical claim of the paper: *Richardson extrapolation can be combined effectively with standard imputation methods*. We run logistic regression under hMCAR missingness, comparing SGD applied on the most standard imputations (namely MICE, MICE with random-forest base learners, and k-nearest-neighbor imputations) used in conjunction with SGD, and the Richardson-SGD counterparts (applied to the same imputation procedures). Across missingness levels and datasets, Richardson consistently acts as a complementary debiasing layer: the imputer reduces the initial missingness bias, while Richardson further reduces the residual gradient bias, with the largest gains obtained when the underlying imputer is already accurate.

Additional experiments. Appendix I extends the numerical study beyond logistic regression to several other Generalized Linear Models (linear and Poisson), datasets, imputation rules, and missingness mechanisms. Across settings, Richardson consistently improves the considered imputations and remains effective beyond the one-pass regime. The gains are largest in the first epoch, matching the

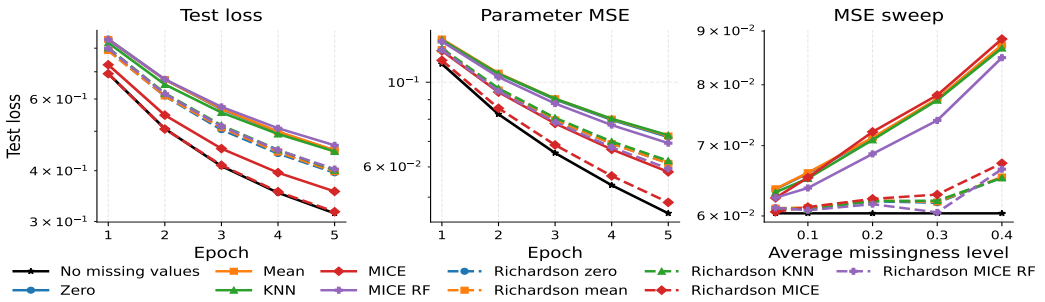

Figure 2: **Covertype missingness sweep.** Test loss and parameter mean-squared error for logistic regression on Covertype [6] under hMCAR missingness, as the average missingness level $\bar{p}$ varies. Richardson improves over each corresponding imputed SGD baseline across a broad range of $\bar{p}$, showing that the correction is not limited to the very small-missingness regime.

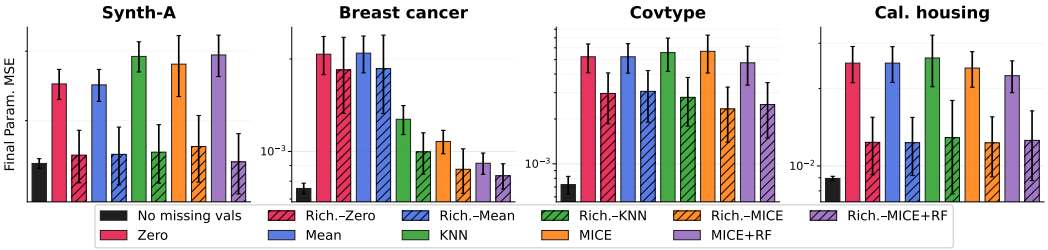

Figure 3: **Cross-dataset comparison.** Final parameter mean-squared error of the last SGD iterate for logistic regression under hMCAR missingness across multiple datasets. Each imputation-based SGD baseline is compared with its first-order Richardson-SGD counterpart. Richardson systematically lowers the final parameter error, with particularly clear gains on the Breast Cancer dataset [39], where stronger imputers lead to substantially smaller Richardson-corrected errors.

theory of Section 5.5. We also show robustness to estimated missingness values by replacing p and q with their estimates in Appendix H, and robustness to misspecification of the missingness mechanism by using Richardson-SGD under an assumed hMCAR mechanism while the true mechanism is sMAR in Appendix K.

7 Conclusion

We introduced Richardson-SGD, a simple debiasing method for stochastic gradient learning computed on imputed data. For arbitrary parametric losses and data-independent imputation rules, we establish that the imputation-induced gradient bias admits a first-order expansion in the missingness vector p. We propose the Richardson-SGD procedure, which turns this structure into an algorithm by deliberately adding controlled missingness. This cancels the leading bias term, reducing gradient bias from $O(\|p\|)$ to $O(\|p\|^2)$ and the one-pass SGD error floor from $O(\|p\|^2)$ to $O(\|p\|^4)$. Our experiments show that one-step Richardson-SGD procedure successfully improves the convergence of SGD for a variety of parametric models and imputation methods. The procedure is lightweight, model-agnostic, and compatible with standard imputation pipelines. Overall, our results show that controlled additional missingness can be more than a nuisance: used carefully, it becomes a practical tool for reducing bias in stochastic learning from incomplete data.

Our debiasing procedure requires generating more missing data with the same distribution as the original sample, but at an increased scale. Doing so is easy for independent hMCAR data, but becomes challenging in the presence of anticorrelation between mask components. In this setting, we are not able to generate more missing data along all coordinates simultaneously while respecting the form of the original missing data distribution. Future research directions are to extend our procedure to such settings. Note however that, in practice, our procedure may be relatively robust to missingness misspecification (Appendix K), which leaves some hope to establish positive results in such settings.

References

[1] A. Ajalloeian and S. U. Stich. On the convergence of sgd with biased gradients. *arXiv preprint arXiv:2008.00051*, 2020.

[2] A. Ayme, C. Boyer, A. Dieuleveut, and E. Scornet. Naive imputation implicitly regularizes high-dimensional linear models. In *International Conference on Machine Learning*, pages 1320–1340. PMLR, 2023.

[3] A. Ayme, C. Boyer, A. Dieuleveut, and E. Scornet. Random features models: a way to study the success of naive imputation. In *Proceedings of the 41st International Conference on Machine Learning*, pages 2108–2134, 2024.

[4] F. Bach. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *The Journal of Machine Learning Research*, 15(1):595–627, 2014.

[5] F. Bach. On the effectiveness of richardson extrapolation in data science. *SIAM Journal on Mathematics of Data Science*, 3(4):1251–1277, 2021.

[6] J. Blackard. Covertype. UCI Machine Learning Repository, 1998. DOI: https://doi.org/10.24432/C50K5N.

[7] K. A. Chandrasekher, A. E. Alaoui, and A. Montanari. Imputation for high-dimensional linear regression. *arXiv preprint arXiv:2001.09180*, 2020.

[8] Y. Demidovich, G. Malinovsky, I. Sokolov, and P. Richtárik. A guide through the zoo of biased sgd. *Advances in Neural Information Processing Systems*, 36:23158–23171, 2023.

[9] A. P. Dempster, N. M. Laird, and D. 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.

[10] J. G. Ibrahim. Incomplete data in generalized linear models. *Journal of the American Statistical Association*, 85(411):765–769, 1990.

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

[12] M. P. Jones. Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American Statistical Association*, 91:222–230, 1996.

[13] J. Josse, J. M. Chen, N. Prost, G. Varoquaux, and E. Scornet. On the consistency of supervised learning with missing values. *Statistical Papers*, 65(9):5447–5479, 2024.

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

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

[16] R. J. Little and D. B. Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2019.

[17] P.-L. Loh and M. J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *Advances in neural information processing systems*, 24, 2011.

[18] A. Marshall, D. G. Altman, P. Royston, and R. L. Holder. Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study. *BMC medical research methodology*, 10(1):7, 2010.

[19] P.-A. Mattei and J. Frellsen. Miwae: Deep generative modelling and imputation of incomplete data sets. In *International conference on machine learning*, pages 4413–4423. PMLR, 2019.

[20] W. Miao, P. Ding, and Z. Geng. Identifiability of normal and normal mixture models with nonignorable missing data. *Journal of the American Statistical Association*, 111(516):1673–1683, 2016.

[21] G. Molenberghs, C. Beunckens, C. Sotto, and M. G. Kenward. Every missingness not at random model has a missingness at random counterpart with equal fit. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(2):371–388, 2008.

[22] J. Näf, E. Scornet, and J. Josse. What is a good imputation under mar missingness? *arXiv preprint arXiv:2403.19196*, 2024.

[23] A. M. Needell. Stochastic gradient descent for linear systems with missing data. *Numerical Mathematics: Theory, Methods and Applications*, 12(1), 2019.

[24] G. Pagès. Multi-step richardson-romberg extrapolation: remarks on variance control and complexity. *Monte Carlo Methods and Applications*, 13, 2007.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[26] T. D. Pigott. A review of methods for missing data. *Educational research and evaluation*, 7(4): 353–383, 2001.

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

[28] L. F. Richardson. The approximate arithmetical solution by finite differences of physical problems involving differential equations, with an application to the stresses in a masonry dam. *Philosophical Transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, 210(459-470):307–357, 1911.

[29] J. M. Robins and Y. Ritov. Toward a curse of dimensionality appropriate (coda) asymptotic theory for semi-parametric models. *Statistics in medicine*, 16(3):285–319, 1997.

[30] D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[31] S. L. Smith, B. Dherin, D. Barrett, and S. De. On the origin of implicit regularization in stochastic gradient descent. In *International Conference on Learning Representations*, 2021.

[32] A. Sportisse, C. Boyer, A. Dieuleveut, and J. Josse. Debiasing averaged stochastic gradient descent to handle missing values. *Advances in Neural Information Processing Systems*, 33: 12957–12967, 2020.

[33] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6): 520–525, 2001.

[34] S. Van Buuren and K. Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45:1–67, 2011.

[35] M. Van Ness, T. M. Bosschieter, R. Halpin-Gregorio, and M. Udell. The missing indicator method: From low to high dimensions. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5004–5015, 2023.

[36] K. A. Verchand and A. Montanari. High-dimensional logistic regression with missing data: Imputation, regularization, and universality. *arXiv preprint arXiv:2410.01093*, 2024.

[37] H. Wang, Z. Lu, and Y. Liu. Score test for missing at random or not under logistic missingness models. *Biometrics*, 79(2):1268–1279, 2023.

[38] S. Wang, J. Shao, and J. K. Kim. An instrumental variable approach for identification and estimation with nonignorable nonresponse. *Statistica Sinica*, pages 1097–1116, 2014.

[39] M. Zwitter and M. Soklic. Breast Cancer. UCI Machine Learning Repository, 1988. DOI: https://doi.org/10.24432/C51P4M.

Contents

A Additional notation and technical preliminaries

This appendix collects notation and elementary identities used throughout the proofs.

Notations. For a mask $m \in \{0,1\}^d$, $S(m) = \{j : m_j = 1\}$. We write $G_S(w; X, Y, \xi) := g(w; \mathcal{I}(X^{\mathrm{NA},S}, M_S, \xi), Y)$ for the gradient when exactly the coordinates in S are declared missing, and $G_\varnothing = g(w; X, Y)$. We use $\Delta_S G_\varnothing := \prod_{j \in S}(T_j - I)G_\varnothing$, where T_j replaces a sample by its version with coordinate j declared missing. We write $a_j(V) \equiv 1$ under hMCAR and $a_j(V) = q_j(V)$ under sMAR.

Inclusion–exclusion identity. For every $S \subseteq [d]$, $\Delta_S G_\varnothing = \sum_{T \subseteq S}(-1)^{|S|-|T|} G_T$, and inversely $G_S = \sum_{T \subseteq S} \Delta_T G_\varnothing$ (Lemma B.1). This is the discrete-difference identity that drives the bias expansion.

B Proofs for the missingness-bias expansion

This appendix proves the structural expansion of the imputation-induced gradient bias.

First, we prove a purely algebraic identity: the gradient obtained after hiding any set of coordinates can be decomposed into a sum of finite-difference effects. These effects isolate what is due to hiding one coordinate, what is due to hiding two coordinates jointly, and so on.

Second, we average this identity over the random missingness mask. This turns the finite-difference effects into a bias expansion whose coefficients are co-missingness probabilities. The first-order terms correspond to single missing coordinates; the remainder contains all simultaneous missingness effects.

Throughout this appendix, fix a parameter value $w \in \mathbb{R}^d$. We suppress the dependence on w whenever this improves readability. By definition, any imputation rule leaves a fully observed sample unchanged:

$$\mathcal{I}(X, \mathbf{0}, \xi) = X.$$

Thus, when no coordinate is declared missing, the imputed gradient equals the complete-data gradient. We assume that all finite differences introduced below are integrable. This is automatic, for instance, if the gradient is continuous, as assumed in the paper.

B.1 Proof of Proposition 1

Gradients indexed by deterministic missingness sets. For a deterministic set $S \subseteq [d]$, let $X^{\mathrm{NA},S}$ be the version of X in which exactly the coordinates in S are replaced by NA. Let $M_S \in \{0,1\}^d$ be the deterministic mask associated with S:

$$(M_S)_j = \mathbf{1}\{j \in S\}.$$

We define

$$G_S := g\big(w; \mathcal{I}(X^{\mathrm{NA},S}, M_S, \xi), Y\big). \tag{15}$$

Thus, G_S is the gradient we would compute if we deliberately declared exactly the coordinates in S missing and then applied the imputation rule. In particular,

$$G_\varnothing = g(w; \mathcal{I}(X, \mathbf{0}, \xi), Y) = g(w; X, Y).$$

Finite missingness differences. The objects G_S describe gradients under different missingness patterns. To separate the effect of one coordinate from the extra effect of several coordinates being missing together, we use finite differences. For every $S \subseteq [d]$, define

$$D_S := \sum_{T \subseteq S} (-1)^{|S|-|T|} G_T, \qquad D_\varnothing := G_\varnothing. \tag{16}$$

The first examples are

$$D_{\{j\}} = G_{\{j\}} - G_\varnothing,$$

and

$$D_{\{j,k\}} = G_{\{j,k\}} - G_{\{j\}} - G_{\{k\}} + G_\varnothing.$$

The interpretation is as follows. The term $D_{\{j\}}$ is the direct effect of hiding coordinate j. The term $D_{\{j,k\}}$ is not the full effect of hiding j and k; it is only the additional interaction left after removing the two separate single-coordinate effects. Higher-order terms D_S have the same meaning: they isolate the part of the missingness effect that appears only when all coordinates in S are hidden together.

The following simple result shows that the full effect of hiding the coordinates in A can be rebuilt by adding all finite-difference effects supported inside A.

Lemma B.1 (Deterministic mask expansion)**.** *For every deterministic set $A \subseteq [d]$,*

$$G_A = \sum_{S \subseteq A} D_S. \tag{17}$$

Proof. Starting from the definition of D_S,

$$\sum_{S \subseteq A} D_S = \sum_{S \subseteq A} \sum_{T \subseteq S} (-1)^{|S|-|T|} G_T.$$

We now group the terms by G_T. A fixed G_T appears only in those sums with $T \subseteq S \subseteq A$, so

$$\sum_{S \subseteq A} D_S = \sum_{T \subseteq A} G_T \sum_{S: T \subseteq S \subseteq A} (-1)^{|S|-|T|}.$$

For fixed $T \subseteq A$, write $S = T \cup R$, where $R \subseteq A \setminus T$. Then the inner sum becomes

$$\sum_{R \subseteq A \setminus T} (-1)^{|R|} = (1-1)^{|A \setminus T|},$$

which results from the binomial expansion of the right-hand side term. This term equals 1 if $T = A$, and 0 otherwise. Therefore, every term cancels except G_A, proving (17). □

From a deterministic mask to a random mask. We now let $M \in \{0,1\}^d$ be the actual random missingness mask. Recall that $S(M) := \{j : M_j = 1\}$ is the set of missing coordinates for the mask M. The corresponding imputed gradient is $G_{S(M)}$. For each $S \subseteq [d]$, we recall that the conditional co-missingness probability is

$$\rho_S := \mathbb{E}\left[\prod_{j \in S} M_j \,\middle|\, X, Y\right] = \mathbb{P}\big(M_j = 1 \text{ for all } j \in S \mid X, Y\big), \qquad \rho_\varnothing := 1. \tag{18}$$

Thus, $\rho_{\{j\}}$ is the conditional probability that coordinate j is missing, while $\rho_{\{j,k\}}$ is the conditional probability that j and k are missing simultaneously.

Lemma B.2 (Random mask expansion)**.** *Consider any data-independent imputation parametrized by ξ, as defined in* (2)*. Let $M \in \{0,1\}^d$ be any missingness mask. Then the corresponding imputed gradient $G_{S(M)}$ satisfies*

$$\mathbb{E}\left[G_{S(M)} \mid X, Y, \xi\right] = G_\varnothing + \sum_{\varnothing \neq S \subseteq [d]} \rho_S D_S. \tag{19}$$

Proof. Apply Lemma B.1 to the random set $A = S(M)$. For a fixed realization of the mask,

$$G_{S(M)} = \sum_{S \subseteq S(M)} D_S.$$

The condition $S \subseteq S(M)$ is equivalent to saying that every coordinate in S is missing, namely $M_j = 1$ for all $j \in S$. Therefore

$$\mathbf{1}\{S \subseteq S(M)\} = \prod_{j \in S} M_j,$$

which yields

$$G_{S(M)} = \sum_{S \subseteq [d]} \left(\prod_{j \in S} M_j\right) D_S.$$

Conditional on (X, Y, ξ), the finite differences D_S are fixed, and only the mask remains random. Moreover, by assumption, we have $\xi \perp M \mid (X, Y)$, which leads to

$$\mathbb{E}\left[\prod_{j\in S} M_j \,\middle|\, X, Y, \xi\right] = \mathbb{E}\left[\prod_{j\in S} M_j \,\middle|\, X, Y\right] = \rho_S.$$

Hence

$$\mathbb{E}\left[G_{S(M)} \mid X, Y, \xi\right] = \sum_{S\subseteq[d]} \rho_S D_S.$$

The term $S = \varnothing$ equals $D_\varnothing = G_\varnothing$, which gives (19). □

Bias expansion. We can now prove the first-order structure of the gradient bias. The random imputed gradient used by the learner is

$$\hat{g}(w) = G_{S(M)}.$$

Since $G_\varnothing = g(w; X, Y)$, we have $\mathbb{E}[G_\varnothing] = \nabla L(w)$. Taking expectations in Lemma B.2 therefore yields

$$\mathcal{B}(w, p) := \mathbb{E}[\hat{g}(w)] - \nabla L(w) = \sum_{\varnothing\neq S\subseteq[d]} \mathbb{E}[\rho_S D_S]. \tag{20}$$

This identity is the key building block. It says that the bias is a sum over all nonempty missingness sets S. Each term has two factors:

- ρ_S, the probability that all coordinates in S are missing;
- D_S, the incremental gradient effect created by hiding exactly the coordinates in S, after lower-order effects have been subtracted.

Thus, singletons $S = \{j\}$ produce the first-order bias, while sets with $|S| \geq 2$ produce the co-missingness remainders.

Proof of Proposition 1. Start from the exact expansion (20). We separate the singleton terms from the terms involving at least two missing coordinates:

$$\mathcal{B}(w, p) = \sum_{j=1}^{d} \mathbb{E}[\rho_{\{j\}} D_{\{j\}}] + \sum_{|S|\geq 2} \mathbb{E}[\rho_S D_S]. \tag{21}$$

We now identify the singleton probabilities under the mechanisms considered in the paper.

Under hMCAR,

$$\rho_{\{j\}} = \mathbb{P}(M_j = 1) = p_j.$$

Under sMAR, with $V = X^{(\mathcal{O})}$,

$$\rho_{\{j\}} = \mathbb{P}(M_j = 1 \mid V) = p_j q_j(V).$$

Both cases can be written as

$$\rho_{\{j\}} = p_j a_j(V), \quad \text{with } a_j(V) = \begin{cases} 1, & \text{hMCAR,} \\ q_j(V), & \text{sMAR.} \end{cases} \tag{22}$$

Substituting (22) into the singleton part of (21) gives

$$\sum_{j=1}^{d} \mathbb{E}[\rho_{\{j\}} D_{\{j\}}] = \sum_{j=1}^{d} p_j \, \mathbb{E}[a_j(V) D_{\{j\}}].$$

Since $D_{\{j\}} = G_{\{j\}} - G_\varnothing$, this is exactly $\mathcal{A}(w)p$, where the j-th column of $\mathcal{A}(w)$ is

$$\mathcal{A}_{\cdot j}(w) = \mathbb{E}\left[a_j(V)\big\{G_{\{j\}}(w; X, Y, \xi) - G_\varnothing(w; X, Y, \xi)\big\}\right].$$

The remaining terms are precisely the co-missingness remainder:

$$\mathcal{R}(w, p) := \sum_{|S|\geq 2} \mathbb{E}[\rho_S D_S]. \tag{23}$$

Combining the singleton part and the remainder proves

$$\mathcal{B}(w, p) = \mathcal{A}(w)p + \mathcal{R}(w, p).$$

□

Remark 1 (What the remainder contains)**.** *The remainder $\mathcal{R}(w,p)$ is the sum of all interaction terms caused by simultaneous missingness. For example, the pair $\{j,k\}$ contributes*

$$\mathbb{E}[\rho_{\{j,k\}} D_{\{j,k\}}].$$

If M_j and M_k are independent and each is missing with probability of order p, then $\rho_{\{j,k\}}$ is of order p^2. If instead the two coordinates are always missing together, then $\rho_{\{j,k\}}$ can be of order p. Thus, without a weak-dependence condition on co-missingness probabilities, $\mathcal{R}(w,p)$ may be a first-order term, proportional to p.

B.2 Proof of Corollary 1

Proof. Assume that the missingness indicators are conditionally independent given the variables driving the missingness mechanism. In hMCAR, this is ordinary independence. In sMAR, this is conditional independence given V.

Then, for every $S \subseteq [d]$, the probability that all coordinates in S are missing factorizes:

$$\rho_S = \prod_{j \in S} \rho_{\{j\}} = \prod_{j \in S} p_j a_j(V).$$

Defining

$$\mu_S(w) := \mathbb{E}\left[\left(\prod_{j \in S} a_j(V) \right) D_S(w) \right], \tag{24}$$

we then have the **multilinear form under independent masks:**

$$\mathcal{B}(w,p) = \sum_{\varnothing \neq S \subseteq [d]} \left(\prod_{j \in S} p_j \right) \mu_S(w), \tag{25}$$

and thus (23),

$$\mathcal{R}(w,p) = \sum_{|S| \geq 2} \left(\prod_{j \in S} p_j \right) \mu_S(w).$$

Every term in this sum contains at least two factors p_j. Since the dimension is fixed and the finite differences are integrable, there exists a finite constant C_w, depending on w but not on p, such that

$$\|\mathcal{R}(w,p)\| \leq C_w \sum_{|S| \geq 2} \prod_{j \in S} p_j.$$

The last sum is $O(\|p\|^2)$ as $\|p\| \to 0$, because each product contains at least two entries of p. Therefore

$$\|\mathcal{R}(w,p)\| = O(\|p\|^2), \qquad \big\|\mathcal{B}(w,p) - \mathcal{A}(w)p\big\| = O(\|p\|^2).$$

□

Remark 2. *The coefficient $\mu_S(w)$ is the average $|S|$-way missingness interaction: it is the effect of declaring all coordinates in S missing, after all lower order effects have been removed by inclusion-exclusion. This formula is the reason Richardson extrapolation applies: the bias is organized by powers of the missingness scale.*

C Proofs for Richardson correction

This appendix collects the proofs of the Richardson-extrapolation results: the joint law of the further-thinned mask, first- and higher-order bias cancellation, the subset-based variant, the plug-in mechanism, and the linear-regression case study. We close with explicit GLM bias formulas.

C.1 Joint law of the further-thinned mask

Fix $C > 1$ and assume $C\,p_j a_j(V) \le 1$ a.s. for every $j \in \mathcal{O}^c$, with a_j as in (22). For $j \in \mathcal{O}^c$, draw r_j as in (9), conditionally independent across j given $(X, M^{(p)})$. Define $M_j^{(Cp)} := 1 - (1 - M_j^{(p)})\, r_j$.

Since the imputation rule is conditionally independent of $M^{(p)}$ given (X, Y), by construction we have $\mathbb{P}(M_j^{(p)} = 0 \mid X, Y) = 1 - p_j a_j(V)$, where $V := X^{(\mathcal{O})}$. Recall that $M_j^{(Cp)} = 0$ iff $M_j^{(p)} = 0$ AND $r_j = 1$. Hence,

$$\begin{aligned}\mathbb{P}\big(M_j^{(Cp)} = 0 \,\big|\, V, Y\big) \;&=\; \mathbb{P}\big(M_j^{(p)} = 0 \,\big|\, V, Y\big)\, \mathbb{P}\big(r_j = 1 \,\big|\, V, Y, M_j^{(p)} = 0\big) \\ &=\; \big(1 - p_j a_j(V)\big)\, \frac{1 - C p_j a_j(V)}{1 - p_j a_j(V)} \\ &=\; 1 - C p_j a_j(V).\end{aligned}$$

Due to the conditional independence of $M^{(p)}$ given V, we obtain the conditional independence of $\{M_j^{(Cp)}\}_{j \in \mathcal{O}^c}$ given V. Hence, $M^{(Cp)}$ has the same conditional law as an independent mask drawn at scale Cp.

C.2 Proof of Proposition 2

Proof. Apply Proposition 1 at scales p and Cp. Both biases admit the decomposition $\mathcal{B}(w, \cdot) = \mathcal{A}(w) \cdot + \mathcal{R}(w, \cdot)$, with the same operator $\mathcal{A}(w)$ (since by (5), $\mathcal{A}$ does not depend on p). Substituting into (10),

$$\begin{aligned}\mathbb{E}\left[\hat{g}_C^{\mathrm{R}}(w)\right] - \nabla L(w) \;&=\; \frac{C\,\mathcal{B}(w, p) - \mathcal{B}(w, Cp)}{C - 1} \\ &=\; \frac{C\,\mathcal{A}(w)\,p - \mathcal{A}(w)\,(Cp)}{C - 1} \;+\; \frac{C\,\mathcal{R}(w, p) - \mathcal{R}(w, Cp)}{C - 1}.\end{aligned}$$

Linearity of $\mathcal{A}$ yields $C\,\mathcal{A}(w)\,p - \mathcal{A}(w)\,(Cp) = 0$, so only the remainder survives:

$$\mathbb{E}\left[\hat{g}_C^{\mathrm{R}}(w)\right] - \nabla L(w) \;=\; \frac{C\,\mathcal{R}(w, p) - \mathcal{R}(w, Cp)}{C - 1}.$$

Under conditional independence of the $\{M_j\}_{j \in \mathcal{O}^c}$ given V, Proposition 1 gives $\|\mathcal{R}(w, p)\| = O(\|p\|^2)$ and, by the same bound applied at scale Cp, $\|\mathcal{R}(w, Cp)\| = O(C^2\|p\|^2) = O(\|p\|^2)$. Combining,

$$\|\,\mathbb{E}[\hat{g}_C^{\mathrm{R}}(w)] - \nabla L(w)\,\| \;=\; O(\|p\|^2),$$

which is (11). □

C.3 Proof of Corollary 2 (higher-order cancellation)

Under independent masks, $\mathcal{B}(w, p) = \sum_{\varnothing \neq S \subseteq [d]} \big(\prod_{j \in S} p_j\big)\mu_S(w)$ from (7). Group terms by $|S|$:

$$\mathcal{B}(w, p) \;=\; \sum_{m=1}^{d} \beta_m(w, p), \qquad \text{with} \qquad \beta_m(w, p) \;:=\; \sum_{|S| = m} \Big(\prod_{j \in S} p_j\Big)\mu_S(w),$$

so that $\beta_m(w, \cdot)$ is homogeneous of degree m, i.e. $\beta_m(w, Cp) = C^m \beta_m(w, p)$. For a sequence of expansion factors $1 = C_0 < C_1 < \cdots < C_k$ with $C_k p_j a_j(V) \le 1$ a.s.,

$$\mathcal{B}(w, C_\ell p) \;=\; \sum_{m=1}^{d} C_\ell^m\, \beta_m(w, p), \qquad \ell = 0, \ldots, k.$$

The Vandermonde system

$$\sum_{\ell=0}^{k} \alpha_\ell \;=\; 1, \qquad \sum_{\ell=0}^{k} \alpha_\ell\, C_\ell^m \;=\; 0, \quad m = 1, \ldots, k,$$

admits a unique solution $\alpha \in \mathbb{R}^{k+1}$ since the matrix $(C_\ell^m)_{\ell,m=0}^k$ is a non-singular Vandermonde. With this choice of α,

$$\begin{aligned}\mathbb{E}\left[\hat{g}^{[k]}(w)\right] - \nabla L(w) &= \sum_{\ell=0}^{k} \alpha_\ell \, \mathcal{B}(w, C_\ell p) \\ &= \sum_{m=1}^{d} \Big(\sum_{\ell=0}^{k} \alpha_\ell C_\ell^m\Big) \beta_m(w,p) \\ &= \sum_{m=k+1}^{d} \Big(\sum_{\ell=0}^{k} \alpha_\ell C_\ell^m\Big) \beta_m(w,p),\end{aligned}$$

where the last equality uses $\sum_\ell \alpha_\ell C_\ell^m = 0$ for $m = 1, \dots, k$ and $\sum_\ell \alpha_\ell C_\ell^0 = 1$ but the $m = 0$ term contributes $\sum_\ell \alpha_\ell \mathcal{B}(w, 0) = 0$ since $\mathcal{B}(w, 0) = 0$. Each $\beta_m(w, p)$ is bounded by $\|p\|_\infty^m \sum_{|S|=m} \|\mu_S(w)\| = O(\|p\|^m)$, hence

$$\big\|\mathbb{E}[\hat{g}^{[k]}(w)] - \nabla L(w)\big\| = O(\|p\|^{k+1}).$$

Finally, let $d_{\text{miss}} = \#\{j : p_j > 0\}$. When $k = d_{\text{miss}}$, every S with $|S| > d_{\text{miss}}$ has at least one coordinate with $p_j = 0$, so $\prod_{j \in S} p_j = 0$ and $\beta_m(w, p) = 0$ for $m > d_{\text{miss}}$. The residual bias vanishes identically: $\mathbb{E}[\hat{g}^{[d_{\text{miss}}]}(w)] = \nabla L(w)$. The argument under sMAR is identical, with p_j replaced by $p_j a_j(V)$ inside the expectation defining $\mu_S(w)$. □

C.4 Proof of Proposition 3 (plug-in mechanism)

Let $\lambda_j(V) := p_j q_j(V)$ and $\hat{\lambda}_j(V) := \hat{p}_j \hat{q}_j(V)$. We write $M^{\tilde{R}}$ for the further-thinned mask produced by the plug-in rule with intensities $\hat{\lambda}_j$, where $\tilde{r}_j \sim \text{Bernoulli}((1 - C\hat{\lambda}_j(V))/(1 - \hat{\lambda}_j(V)))$.

Step 1: Effective intensity of the plug-in further-thinned mask. Conditioning on V and using the conditional independence of $\tilde{r}_j$ and $M^{(p)}$ given V, and the fact that $M_j^{\tilde{R}} = 0$ iff $M_j^{(p)} = 0$ AND $\tilde{r}_j = 1$,

$$\begin{aligned}\mathbb{P}\big(M_j^{\tilde{R}} = 1 \,\big|\, V\big) &= 1 - \mathbb{P}\big(M_j^{(p)} = 0,\, \tilde{r}_j = 1 \,\big|\, V\big) \\ &= 1 - \big(1 - \lambda_j(V)\big)\, \frac{1 - C\hat{\lambda}_j(V)}{1 - \hat{\lambda}_j(V)}.\end{aligned}$$

Define the *effective intensity* after plug-in thinning by

$$\tilde{\lambda}_j(V) := \mathbb{P}(M_j^{\tilde{R}} = 1 \mid V).$$

Expanding the previous display over the common denominator $1 - \hat{\lambda}_j(V)$, we obtain

$$\begin{aligned}\tilde{\lambda}_j(V) &= \frac{1 - \hat{\lambda}_j(V) - \big(1 - \lambda_j(V)\big)\big(1 - C\hat{\lambda}_j(V)\big)}{1 - \hat{\lambda}_j(V)} \\ &= \frac{\lambda_j(V) + (C-1)\hat{\lambda}_j(V) - C\lambda_j(V)\hat{\lambda}_j(V)}{1 - \hat{\lambda}_j(V)} \\ &= \frac{C\lambda_j(V)\big(1 - \hat{\lambda}_j(V)\big) + (C-1)\big(\hat{\lambda}_j(V) - \lambda_j(V)\big)}{1 - \hat{\lambda}_j(V)} \\ &= C\,\lambda_j(V) + (C-1)\,\frac{\hat{\lambda}_j(V) - \lambda_j(V)}{1 - \hat{\lambda}_j(V)}. \qquad (26)\end{aligned}$$

The first term is the desired intensity of a new draw at scale Cp; the second is the plug-in error. Setting

$$e_j^{(C)}(V) := \tilde{\lambda}_j(V) - C\,\lambda_j(V) = (C-1)\,\frac{\hat{\lambda}_j(V) - \lambda_j(V)}{1 - \hat{\lambda}_j(V)},$$

we have $\tilde{\lambda}(V) = C\lambda(V) + e^{(C)}(V)$.

Step 2: Bias of the plug-in Richardson gradient. By the same expansion as in Proposition 1, the singleton part of the bias is obtained by multiplying the singleton gradient gap by the corresponding conditional missingness probability. For the original mask, this probability is

$$\mathbb{P}(M_j = 1 \mid V) = p_j a_j(V),$$

whereas for the plug-in further-thinned mask, Step 1 gives

$$\mathbb{P}(M_j^{\tilde{R}} = 1 \mid V) = \tilde{\lambda}_j(V) = C p_j a_j(V) + e_j^{(C)}(V).$$

Applying the inclusion–exclusion expansion (eq:bias-full-subset-expansion-clear) separately to each of the two stochastic gradients in $\hat{g}^{\mathrm{R}}_{C,\hat{\lambda}}(w) = (C\,\hat{g}^{(p)}(w) - \hat{g}^{(Cp,\hat{\lambda})}(w))/(C-1)$, and using $\rho_{\{j\}} = p_j a_j(V)$ for the original mask and $\mathbb{P}(M_j^{\tilde{R}} = 1 \mid V) = C p_j a_j(V) + e_j^{(C)}(V)$ for the plug-in further-thinned mask, the singleton contributions to the two biases are

$$\sum_{j=1}^{d} p_j\, \mathbb{E}\big[a_j(V) D_{\{j\}}\big] \quad \text{and} \quad \sum_{j=1}^{d} \big(C p_j a_j(V) + e_j^{(C)}(V)\big)\, \mathbb{E}\big[D_{\{j\}} \mid V\big]\,,$$

respectively. In the Richardson combination, the deterministic $C p_j a_j(V)$ contributions cancel exactly, leaving

$$\mathbb{E}[\hat{g}^{\mathrm{R}}_{C,\hat{\lambda}}(w)] - \nabla L(w) = -\frac{1}{C-1} \sum_{j=1}^{d} \mathbb{E}\Big[e_j^{(C)}(V)\big\{G_{\{j\}}(w; X, Y, \xi) - G_{\varnothing}(w; X, Y, \xi)\big\}\Big] + \frac{C\,\mathcal{R}(w,p) - \mathcal{R}^{\tilde{R}}(w)}{C-1},$$

where $\mathcal{R}^{\tilde{R}}(w) := \sum_{|S|\geq 2} \mathbb{E}[\rho_S^{\tilde{R}} D_S]$ is the co-missingness remainder evaluated at the plug-in further-thinned mask. Thus the only remaining first-order contribution is the plug-in intensity error $e_j^{(C)}(V)$. Under the assumed L^2 bound on the singleton gradient gaps, there exists $G_\star$ such that

$$\big\|G_{\{j\}}(w; X, Y, \xi) - G_{\varnothing}(w; X, Y, \xi)\big\|_{L^2} \leq G_\star, \qquad \forall j \in [d].$$

Therefore,

$$\left\| \sum_{j=1}^{d} \mathbb{E}\Big[e_j^{(C)}(V)\big\{G_{\{j\}}(w; X, Y, \xi) - G_{\varnothing}(w; X, Y, \xi)\big\}\Big] \right\| = O\Big(\|e^{(C)}\|_\infty\Big)\,.$$

Moreover, the co-missingness remainder is

$$O\Big(\|p\|^2 + \|p\|\,\|e^{(C)}\|_\infty + \|e^{(C)}\|_\infty^2\Big)\,.$$

Combining gives ,

$$\big\|\mathbb{E}[\hat{g}^{\mathrm{R}}_{C,\hat{\lambda}}(w)] - \nabla L(w)\big\| = O\Big(\|e^{(C)}\|_\infty + \|p\|^2 + \|p\|\,\|e^{(C)}\|_\infty + \|e^{(C)}\|_\infty^2\Big)\,. \tag{27}$$

Step 3: Bound on the plug-in error. Using

$$\begin{aligned} |\hat{p}_j \hat{q}_j(V) - p_j q_j(V)| &= |(\hat{p}_j - p_j) q_j(V) + \hat{p}_j(\hat{q}_j(V) - q_j(V))| \\ &\leq |\hat{p}_j - p_j|\,|q_j(V)| + |\hat{p}_j|\,|\hat{q}_j(V) - q_j(V)| \end{aligned}$$

and the bounds $|q_j(V)| \leq 1$, $|\hat{p}_j| \leq \|p\|_\infty + \delta_p$,

$$|\hat{\lambda}_j(V) - \lambda_j(V)| \;\leq\; \delta_p + (\|p\|_\infty + \delta_p)\,\delta_q.$$

Since $\hat{\lambda}_j(V) \leq \rho < 1$,

$$|e_j^{(C)}(V)| \;\leq\; \frac{C-1}{1-\rho}\,[\delta_p + (\|p\|_\infty + \delta_p)\,\delta_q]\,,$$

hence, defining $\|e^{(C)}\|_\infty := \sup_{j \in \mathcal{O}^c,\, v} |e_j^{(C)}(v)|$,

$$\|e^{(C)}\|_\infty \;=\; O\big(\delta_p + \|p\|_\infty \delta_q + \delta_p \delta_q\big).$$

Step 4: Concluding. Substituting the bound on $\|e^{(C)}\|_\infty$ into (27) and simplifying,

$$\big\|\mathbb{E}[\hat{g}^{\mathrm{R}}_{C,\hat{\lambda}}(w)] - \nabla L(w)\big\| \;=\; O\Big(\|p\|^2 + \delta_p + \|p\|_\infty \delta_q + \delta_p \delta_q\Big),$$

where the implicit constants depend only on C, ρ, and $G_\star$. In the MCAR case $q_j \equiv 1$, $\delta_q = 0$ and the bound collapses to $O(\|p\|^2 + \delta_p)$. □

D Bias formulas for specific generalized linear models

We record explicit expressions for the leading-order population bias $\mathcal{A}(w)\,p$ in three GLMs under heterogeneous MCAR with zero imputation. When the missingness factors are independent, $\mathcal{A}(w)\,p$ is the only leading term in p. In full generality, however, some terms in the remainder may also be linear. In all cases, Richardson-SGD eliminates all linear terms, whether or not additional linear contributions appear in the remainder.

Notably, the bias of linear regression is a polynomial of total degree at most 2 in p, whereas logistic and Poisson regression generally exhibit full-degree bias, up to degree d.

Linear regression (squared loss). For $\ell(w; x, y) = \frac{1}{2}(w^\top x - y)^2$ and zero imputation,

$$\big(\mathcal{A}(w)\,p\big)_j \;=\; -p_j\,\nabla_j L(w) \;-\; \sum_{k \neq j} p_k\, S_{jk} w_k, \qquad S := \mathbb{E}[XX^\top].$$

The detailed derivation, including the exact non-asymptotic version, is reproduced in Appendix E.

Logistic regression. For $\ell(w; x, y) = \log(1 + e^{-y\,w^\top x})$ with $y \in \{-1, +1\}$, the gradient is $g(w; x, y) = -y\,\sigma(-y\,w^\top x)\,x$ where σ is the logistic function. Under zero imputation and heterogeneous MCAR,

$$\big(\mathcal{A}(w)\,p\big)_j \;=\; -p_j \nabla_j L(w) + \sum_{k \neq j} p_k\, \mathbb{E}\left[Y\big(\sigma(-Y\,w^\top X) - \sigma(-Y\,w^\top X^{(-k)})\big) X_j\right],$$

where $X^{(-k)}$ is X with X_k replaced by 0.

Poisson regression. For $\ell(w; x, y) = e^{w^\top x} - y\,w^\top x$, the gradient is $g(w; x, y) = (e^{w^\top x} - y)\,x$, hence

$$\big(\mathcal{A}(w)\,p\big)_j \;=\; -p_j\,\nabla_j L(w) \;+\; \sum_{k \neq j} p_k\, \mathbb{E}\left[(e^{w^\top X^{(-k)}} - e^{w^\top X}) X_j\right].$$

All three expressions are obtained by substituting the corresponding loss into (5). They share the same structural form: a coordinate-wise diagonal contribution $-p_j \nabla_j L(w)$, plus an off-diagonal correction.

E Linear regression: a transparent case study

The goal of this appendix is to show on the simplest GLM that, under heterogeneous MCAR with independent masks and zero imputation, the population gradient bias is a polynomial of degree at most 2 in p. By Corollary 2, second-order Richardson with two factors $C_1 < C_2$ therefore cancels this bias *exactly*, while a single Richardson step already reduces it from $O(\|p\|)$ to $O(\|p\|^2)$.

Setting. We work at the single-observation level with squared loss,

$$\ell(w; x, y) = \tfrac{1}{2}(w^\top x - y)^2, \qquad g(w; x, y) = (w^\top x - y)\,x.$$

The population risk is $L(w) = \frac{1}{2}\mathbb{E}[(w^\top X - Y)^2]$ with $\nabla L(w) = Sw - b$, $S := \mathbb{E}[XX^\top]$, $b := \mathbb{E}[YX]$. We assume heterogeneous MCAR with independent mask coordinates, $\mathbb{P}(M_j = 1) = p_j$, and zero imputation $\tilde{X} = (1 - M) \odot X$. The imputed gradient is $\hat{g}(w) := g(w; \tilde{X}, Y) = (w^\top \tilde{X} - Y)\tilde{X}$.

E.1 Sample-conditional and population biases

Proposition 4 (Sample-conditional bias)**.** *Under heterogeneous MCAR with independent masks, for each $j \in [d]$,*

$$\mathbb{E}_M\left[\hat{g}_j(w;\tilde{X},Y)\,\middle|\,X,Y\right] - g_j(w;X,Y) \;=\; -p_j X_j^2 w_j - \sum_{k\neq j}(p_j + p_k - p_j p_k) X_j X_k w_k + p_j Y X_j, \tag{28}$$

or equivalently

$$\mathbb{E}_M\left[\hat{g}_j(w;\tilde{X},Y)\,\middle|\,X,Y\right] - g_j(w;X,Y) \;=\; -p_j\, g_j(w;X,Y) - (1-p_j)\sum_{k\neq j} p_k\, X_j X_k w_k. \tag{29}$$

Proof. With $\omega_j := 1 - M_j$ and $\tilde{X}_j = \omega_j X_j$, the imputed j-th gradient is

$$\hat{g}_j(w;\tilde{X},Y) = (w^\top \tilde{X} - Y)\tilde{X}_j = \Big(\sum_k w_k \omega_k X_k - Y\Big)\omega_j X_j.$$

Conditioning on (X,Y) and using independence of the mask coordinates ($\mathbb{E}[\omega_j] = \mathbb{E}[\omega_j^2] = 1 - p_j$ and $\mathbb{E}[\omega_j\omega_k] = (1-p_j)(1-p_k)$ for $k \neq j$),

$$\mathbb{E}_M\left[\hat{g}_j(w;\tilde{X},Y)\,\middle|\,X,Y\right] \;=\; (1-p_j)X_j^2 w_j + \sum_{k\neq j}(1-p_j)(1-p_k)X_j X_k w_k - (1-p_j) Y X_j.$$

Subtracting the complete-data gradient $g_j(w;X,Y) = X_j^2 w_j + \sum_{k\neq j} X_j X_k w_k - Y X_j$ yields (28) after expanding $(1-p_j)(1-p_k) - 1 = -(p_j + p_k - p_j p_k)$. To obtain (29), factor $-p_j$ in front of $g_j(w;X,Y)$:

$$\begin{aligned} &- p_j X_j^2 w_j - p_j \sum_{k\neq j} X_j X_k w_k + p_j Y X_j - (1-p_j)\sum_{k\neq j} p_k X_j X_k w_k \\ &\quad = -p_j\, g_j(w;X,Y) - (1-p_j)\sum_{k\neq j} p_k X_j X_k w_k. \end{aligned}$$

□

Corollary 4 (Population bias of zero-imputed linear regression)**.** *Under heterogeneous MCAR with independent masks, with $B_j(w;p) := \mathbb{E}\left[\hat{g}_j(w;\tilde{X},Y)\right] - \nabla_j L(w)$,*

$$B_j(w;p) \;=\; -p_j\,\nabla_j L(w) \;-\; (1-p_j)\sum_{k\neq j} p_k\, S_{jk} w_k, \tag{30}$$

hence $\|B(w;p)\| = O(\|p\|)$.

Proof. Take expectation in (29) and use $\mathbb{E}[g_j(w;X,Y)] = \nabla_j L(w)$ and $\mathbb{E}[X_j X_k] = S_{jk}$ to obtain (30). The norm bound follows from $|B_j(w;p)| \le p_j|\nabla_j L(w)| + \sum_{k\neq j} p_k |S_{jk} w_k| \le \|p\|_\infty\big(|\nabla_j L(w)| + \sum_k |S_{jk} w_k|\big)$. □

E.2 Polynomial structure and exact debiasing in two Richardson steps

We now make explicit that the population gradient bias is a polynomial of degree at most 2 in p, hence is annihilated exactly by second-order Richardson with two factors.

The bias is degree-2 in p. Under heterogeneous MCAR with independent masks, from (30),

$$B_j(w;p) \;=\; \underbrace{\Big(-p_j\,\nabla_j L(w) - \sum_{k\neq j} p_k\, S_{jk} w_k\Big)}_{=:L_j(w;p),\ \text{linear in } p} \;+\; \underbrace{p_j \sum_{k\neq j} p_k\, S_{jk} w_k}_{=:Q_j(w;p),\ \text{quadratic in } p}. \tag{31}$$

The same conclusion follows from the general expansion (7) of Section 3, since for the squared loss the iterated finite differences $\Delta_S G_\varnothing$ vanish identically for $|S| \ge 3$ (the gradient is bilinear in X).

One Richardson step removes the linear part.

Proposition 5 (First-order Richardson cancellation, heterogeneous squared-loss MCAR)**.** *Under heterogeneous MCAR with independent masks, for $C > 1$ with $Cp_j < 1$ for every j, the first-order Richardson gradient* (10) *satisfies*

$$\mathbb{E}\left[\hat{g}_j^{\mathrm{R}}(w)\right] - \nabla_j L(w) \;=\; -C\, p_j \sum_{k \neq j} p_k\, S_{jk} w_k \;=\; -C\, Q_j(w;p).$$

In particular, $\left\|\mathbb{E}[\hat{g}^{\mathrm{R}}(w)] - \nabla L(w)\right\| = O(\|p\|^2)$, while the uncorrected bias is $O(\|p\|)$. If $p_j = 0$ for some j, then $\mathbb{E}[\hat{g}_j^{\mathrm{R}}(w)] = \nabla_j L(w)$, i.e. the Richardson bias vanishes in coordinate j.

Proof. Apply $\mathbb{E}[\hat{g}_j^{\mathrm{R}}] - \nabla_j L = (C\, B_j(w;p) - B_j(w;Cp))/(C-1)$ to (31). Using

$$L_j(w;Cp) = C\, L_j(w;p) \quad \text{and} \quad Q_j(w;Cp) = C^2\, Q_j(w;p),$$

we have

$$\begin{aligned} \frac{C\, B_j(w;p) - B_j(w;Cp)}{C-1} \;&=\; \frac{C\, L_j(w;p) - C\, L_j(w;p)}{C-1} + \frac{C\, Q_j(w;p) - C^2 Q_j(w;p)}{C-1} \\ &=\; -C\, Q_j(w;p), \end{aligned}$$

which gives the claim. □

Two Richardson steps cancel the bias exactly.

Corollary 5 (Exact Richardson debiasing for linear regression)**.** *Suppose heterogeneous MCAR with independent masks. Let $1 = C_0 < C_1 < C_2$ be three expansion factors with $C_2 p_j < 1$ for every j. The unique transposed Vandermonde solution*

$$(\alpha_0, \alpha_1, \alpha_2) \quad \textit{of} \quad \begin{cases} \alpha_0 + \alpha_1 + \alpha_2 = 1, \\ \alpha_0 + \alpha_1 C_1 + \alpha_2 C_2 = 0, \\ \alpha_0 + \alpha_1 C_1^2 + \alpha_2 C_2^2 = 0, \end{cases}$$

gives a second-order Richardson gradient $\hat{g}^{[2]} = \sum_\ell \alpha_\ell \hat{g}^{(C_\ell p)}$ with $\mathbb{E}[\hat{g}^{[2]}(w)] = \nabla L(w)$ exactly. The same statement holds in independent sMAR (conditional on V), after replacing p_j by $p_j a_j(V)$ inside the expectations defining L_j and Q_j.

Proof. The bias (31) is a polynomial of degree ≤ 2 in p with no constant term, so $\mathcal{B}(w, Cp) = C\, L(w;p) + C^2\, Q(w;p)$ for any $C > 0$. Applying $\sum_\ell \alpha_\ell \mathcal{B}(w, C_\ell p)$ and using the Vandermonde conditions, both L and Q contributions vanish. □

This recovers, in our framework, the closed-form debiasing of Sportisse et al. [32] for linear regression with squared loss under independent MCAR, and extends it to the sMAR mechanisms of Section 2, where no closed-form bias is available.

F One-pass biased-SGD convergence consequences

We provide a simple proof for the non-averaged iterates of SGD with biased gradients. The rates stated in the main text are then recovered by applying this result with the bias corresponding to each method. Note that one could also aim for similar guarantees for averaged SGD with a broader class of step sizes, namely $\eta_t \propto t^{-a}$ with $a \in (1/2, 1)$, which notably does not require prior knowledge of the strong convexity constant of the loss [27, 4]. We do not pursue this direction here, as our main focus is bias reduction, and the theorem below already illustrates its practical benefit.

Proposition 6 (One-pass SGD with Bias)**.** *Assume L is α-strongly convex and β-smooth, and let $w^\star$ be its unique minimizer. Run one-pass SGD over n i.i.d. samples,*

$$w_{i+1} = w_i - \eta_i \hat{g}_i(w_i),$$

with step sizes

$$\eta_i = \frac{c}{i+\gamma}, \qquad \alpha c > 1,$$

where γ is large enough that

$$\eta_i \le \frac{\alpha}{6\beta^2} \qquad \textit{for all } i.$$

Suppose that there exists $B(p)$ such that the imputed gradient satisfies, uniformly along the trajectory,

$$\big\|\mathbb{E}[\hat{g}_i(w_i) \mid w_i] - \nabla L(w_i)\big\| \le B(p),$$

and

$$\mathbb{E}\Big[\big\|\hat{g}_i(w_i) - \mathbb{E}[\hat{g}_i(w_i) \mid w_i]\big\|^2 \,\Big|\, w_i\Big] \le \sigma^2 \qquad \textit{a.s.}$$

Then

$$\mathbb{E}\|w_n - w^\star\|^2 = O\big(B(p)^2\big) + O\left(\frac{\sigma^2}{n}\right).$$

In particular, if $\sigma^2 = O(1)$, then

$$\mathbb{E}\|w_n - w^\star\|^2 = O\big(B(p)^2\big) + O(1/n). \tag{32}$$

Proof. Write

$$\mathbb{E}[\hat{g}_i(w_i) \mid w_i] = \nabla L(w_i) + \Delta_i, \qquad \|\Delta_i\| \le B(p).$$

Also write

$$\hat{g}_i(w_i) = \nabla L(w_i) + \Delta_i + \xi_i, \qquad \mathbb{E}[\xi_i \mid w_i] = 0, \qquad \mathbb{E}[\|\xi_i\|^2 \mid w_i] \le \sigma^2.$$

Let

$$\delta_i := \mathbb{E}\|w_i - w^\star\|^2.$$

Since $w^\star$ minimizes L, $\nabla L(w^\star) = 0$. Expanding one SGD step gives

$$\begin{aligned}\mathbb{E}[\|w_{i+1} - w^\star\|^2 \mid w_i] = \|w_i - w^\star\|^2 &- 2\eta_i\langle w_i - w^\star, \nabla L(w_i)\rangle \\ &- 2\eta_i\langle w_i - w^\star, \Delta_i\rangle + \eta_i^2\mathbb{E}[\|\nabla L(w_i) + \Delta_i + \xi_i\|^2 \mid w_i].\end{aligned}$$

By strong convexity,

$$\langle w_i - w^\star, \nabla L(w_i)\rangle \ge \alpha\|w_i - w^\star\|^2.$$

By Young's inequality,

$$2|\langle w_i - w^\star, \Delta_i\rangle| \le \frac{\alpha}{2}\|w_i - w^\star\|^2 + \frac{2}{\alpha}B(p)^2.$$

By smoothness and $\nabla L(w^\star) = 0$,

$$\|\nabla L(w_i)\| \le \beta\|w_i - w^\star\|.$$

Moreover, using $\mathbb{E}[\xi_i \mid w_i] = 0$,

$$\begin{aligned}\mathbb{E}[\|\nabla L(w_i) + \Delta_i + \xi_i\|^2 \mid w_i] &= \|\nabla L(w_i) + \Delta_i\|^2 + \mathbb{E}[\|\xi_i\|^2 \mid w_i] \\ &\le 2\beta^2\|w_i - w^\star\|^2 + 2B(p)^2 + \sigma^2.\end{aligned}$$

Combining these bounds yields

$$\begin{aligned}\mathbb{E}[\|w_{i+1} - w^\star\|^2 \mid w_i] \le \left(1 - \frac{3\alpha\eta_i}{2} + 2\beta^2\eta_i^2\right)\|w_i - w^\star\|^2 \\ + \frac{2\eta_i}{\alpha}B(p)^2 + \eta_i^2\{2B(p)^2 + \sigma^2\}.\end{aligned}$$

Since $\eta_i \le \alpha/(6\beta^2)$, we have

$$2\beta^2\eta_i^2 \le \frac{\alpha\eta_i}{3},$$

and hence

$$1 - \frac{3\alpha\eta_i}{2} + 2\beta^2\eta_i^2 \le 1 - \alpha\eta_i.$$

Taking expectations and absorbing constants gives

$$\delta_{i+1} \le (1 - \alpha\eta_i)\delta_i + C\eta_i B(p)^2 + C\eta_i^2\sigma^2,$$

where $C > 0$ depends only on α, β and the step-size constants.

It remains to solve this recursion. Define

$$\Lambda_i := \delta_i - KB(p)^2,$$

where $K > 0$ is chosen large enough such that, for all i,

$$(1 - \alpha\eta_i)KB(p)^2 + C\eta_i B(p)^2 \le KB(p)^2.$$

Equivalently, it is enough to take $K \ge C/\alpha$. Then

$$\Lambda_{i+1} \le (1 - \alpha\eta_i)\Lambda_i + C\eta_i^2\sigma^2.$$

With $\eta_i = c/(i + \gamma)$ and $\alpha c > 1$, the standard recursion bound gives

$$\Lambda_n = O\big(n^{-\alpha c}\big) + O\left(\frac{\sigma^2}{n}\right).$$

Because $\alpha c > 1$, the initialization term is $O(1/n)$. Therefore,

$$\delta_n = \mathbb{E}\|w_n - w^\star\|^2 = O\big(B(p)^2\big) + O\left(\frac{\sigma^2}{n}\right).$$

If $\sigma^2 = O(1)$, this becomes

$$\mathbb{E}\|w_n - w^\star\|^2 = O\big(B(p)^2\big) + O(1/n).$$

The $O(B(p)^2)$ term is the limiting neighborhood induced by the systematic gradient bias. □

Plug-in for plain imputed SGD. Under independent hMCAR/sMAR, Proposition 1 gives $b(p) = O(\|p\|)$. The variance of $\hat{g}^{(p)}$ is bounded by a constant under the standing L^2 assumptions. Hence

$$\mathbb{E}\|w_n - w^\star\|^2 \;=\; O(\|p\|^2) + O(1/n).$$

Plug-in for Richardson-SGD. Under independent hMCAR/sMAR, Proposition 2 gives $b(p) = O(\|p\|^2)$. Using (12), σ^2 remains $O(1)$ in the moderate-C regime. Substituting in (32) yields

$$\mathbb{E}\|w_n - w^\star\|^2 \;=\; O(\|p\|^4) + O(1/n).$$

The analogous statement for the test loss follows by smoothness of L around $w^\star$. The k-step variant gives $b(p) = O(\|p\|^{k+1})$ and a missingness floor $O(\|p\|^{2(k+1)})$.

Plug-in version with estimated mechanism. Combining Proposition 3 and (32), the one-pass bound becomes

$$\mathbb{E}\|w_n - w^\star\|^2 \;=\; O\big(\|p\|^4 + \delta_p^2 + \|p\|_\infty^2\delta_q^2\big) + O(1/n).$$

Whenever δ_p, δ_q shrink at rate $o(\|p\|^2)$, the exact-mechanism rate is recovered.

Multi-epoch behavior. The above analysis only covers one pass, else the imputed gradients seen on different epochs are not independent. The empirical study of Section 6 indicates that Richardson remains effective in multi-epoch training; a formal multi-epoch analysis is left to future work.

G Implementation details

This appendix describes the experimental protocol used in Section 6 and in Appendix I below. All experiments are run with stochastic gradient descent for 5 epochs, minibatch size 64, average missingness level $\bar{p} = 0.20$, and first-order Richardson scale $C = 2$. Unless stated otherwise, all reported curves are averaged over repeated runs with the same protocol across methods.

Models. We consider three generalized linear models: linear regression with Gaussian noise, logistic regression for binary classification, and Poisson regression for count responses. All models are trained with an ℓ_2 penalty. The regularization parameter is fixed to $\lambda = 10^{-3}$ for every model family and dataset.

Missingness mechanisms. We evaluate three missingness mechanisms. The first is homogeneous MCAR, denoted `mcar`, where each entry is missing independently with the same probability p. The second is heterogeneous MCAR, denoted `hetero_mcar`, where missingness probabilities are generated from row and column multipliers and then calibrated to have average missingness $\bar{p}$. Concretely, the unnormalized missingness scores are sampled uniformly in $[0, 1]$ across covariates and rescaled so that their empirical mean equals 0.20.

The third mechanism is scalable MAR, denoted `smar`. In this case, the oracle missingness intensity is

$$\lambda_j = p_j Q(U_j), \qquad Q(u) = \sigma(1.6u - 0.3), \qquad U = a_{1,j} X_1 + b_{2,j} X_2, \quad a_{1,\cdot}, b_{2,\cdot} \overset{\text{i.i.d}}{\sim} \mathcal{U}(0,1)\,.$$

where σ is the logistic sigmoid. The coordinate-specific constants p_j are calibrated so that the average missingness is $\bar{p} = 0.20$. This is the same scalable MAR mechanism as in Section 2.

Methods compared. We compare the complete-data baseline, plain imputation-based SGD, and Richardson-corrected SGD. The complete-data baseline, denoted `No missing vals`, is trained on the clean unmasked training data. The plain imputation baselines are zero imputation, mean imputation, k-nearest-neighbor imputation, MICE, and MICE with random-forest base learners, denoted respectively by `Zero`, `Mean`, `KNN`, `MICE`, and `MICE+RF`. The corresponding Richardson variants are denoted `Rich.-Zero`, `Rich.-Mean`, `Rich.-KNN`, `Rich.-MICE`, and `Rich.-MICE+RF`. All imputers are taken with default parameters from `scikit-learn`. The experiments are repeated 30 times with different seeds, for the training of SGD methods, and averaged results, along their standard deviations, are displayed.

Metric. The main metric is the parameter mean-squared error

$$\mathrm{MSE}_w(t) = \frac{1}{d_w} \left\| \hat{w}_t - w^\star \right\|_2^2,$$

where d_w is the parameter dimension, $\hat{w}_t$ is the SGD iterate after epoch t, and $w^\star$ is the complete-data reference parameter described below. The metric is reported once per epoch for 5 epochs. For real datasets, $w^\star$ denotes the minimizer of the complete-data ridge 10^{-3} penalized empirical, not a population ground truth (see the paragraph *Reference parameter* below).

Learning-rate calibration. The optimization geometry varies substantially across model families and datasets. To avoid confounding imputation effects with poorly tuned learning rates, we calibrate the initial learning rate η_0 separately for each pair of model family and dataset.

For each pair, we first take the family-level default learning rate $\eta_0 = 10^{-2}$. We then evaluate the geometric grid

$$\eta_0 \in \eta_0^{\mathrm{def}} \cdot \left\{ \frac{1}{4}, \frac{1}{2}, 1, 2, 4 \right\}.$$

For every candidate, we run SGD, without missing data, on the standardized training fold using the same number of epochs, minibatch size, and regularization parameter as in the missing-data experiments. We select the learning rate that minimizes the final iteration parameter MSE,

$$\frac{1}{d_w} \left\| \hat{w}_T - w^\star \right\|_2^2 .$$

This calibration is performed without missingness and without imputation. The selected learning rate is then fixed and reused for all imputation methods, Richardson variants, and missingness mechanisms for that model–dataset pair. Thus, comparisons between MCAR, heterogeneous MCAR, and sMAR within the same row use the same calibrated η_0.

Dataset budget and preprocessing. Each dataset uses $2{,}000$ training samples. Real datasets with fewer observations are bootstrapped to this size when needed. Test sets contain $1{,}000$ samples. Covariates are standardized columnwise on the training fold and the same transformation is applied to the test fold.

The response variable is rescaled depending on the model family. For linear regression on real datasets, the response is z-scored on the loaded sample. This keeps the scale of $w^\star$ comparable across synthetic and real datasets; in particular, using raw elevation in the Covertype regression task

produces parameters much larger than those in the synthetic linear experiments. Synthetic linear responses are left unchanged, since the data-generating process already gives a comparable response scale.

For logistic regression, the response is binary and no rescaling is applied. For Poisson regression, real count responses are rescaled to have mean approximately 2 and then rounded to integer counts, corresponding to a log-mean near 0.7.

Reference parameter. For synthetic datasets, $w^\star$ is the parameter used in the data-generating process. The covariates are generated to be approximately centered and standardized, so the train-fold standardization is nearly idempotent and the generating parameter remains the appropriate reference. For real datasets, there is no closed-form ground-truth parameter. We therefore compute $w^\star$ by L-BFGS-B optimization of the exact ridge penalized empirical loss on the complete, standardized training data, using the same ridge regularization parameter $\lambda = 10^{-3}$ as in the SGD runs. This gives the complete-data regularized minimizer of the observed sample loss and serves as the reference parameter for the reported MSE.

Datasets. The datasets used in the experiments are listed in Table 1. Synthetic datasets are generated with Gaussian covariates. Real datasets are taken from standard `scikit-learn` or OpenML sources and transformed as indicated. For the Bike Sharing Demand dataset, we use the hourly rental count

Table 1: Datasets used in the experiments.

Family	#	Dataset	Source and preprocessing
Linear	1	Synth-A	Synthetic, 10D iid Gaussian
Linear	2	Synth-B	Synthetic, 15D AR-style covariance $\Sigma_{jk} = 0.9^{\lvert j-k \rvert}$
Linear	3	Diabetes	Real, bootstrapped, z-scored response
Linear	4	Covertype-reg	Real, z-scored elevation from 9 continuous features
Logistic	1	Synth-A	Synthetic, 10D iid Gaussian
Logistic	2	Breast cancer	Real, bootstrapped to 2,000 samples
Logistic	3	Covertype	Real, class 1 versus all, 10 continuous features
Logistic	4	California housing	Real; binary response indicating whether the house price exceeds the median
Poisson	1	Synth-A	Synthetic, 10D iid Gaussian
Poisson	2	Synth-B	Synthetic, 8D iid Gaussian
Poisson	3	Bike sharing	Real, hourly rental count from numeric features

as the response and retain eight numeric features: year, month, hour, weekday, temperature, feeling temperature, humidity, and windspeed.

H Robustness to errors in the estimated missingness mechanism

The previous experiments kept p and q known as oracles. In practice, however, these quantities need to be estimated, and Proposition 3 provides an upper bound of the error induced by such estimations. In the following experiment, we test the impact of estimating p and q on Richardson-SGD, on top of imputation by zero, for logistic regression in the hMCAR setting. We perturb the estimated mechanism $(\hat{p}, \hat{q})$ by additive noise with magnitudes δ_p, δ_q and report the parameter MSE as a function of (δ_p, δ_q). Table 2 shows the robustness of Richardson to plug-in estimation. With a reference parameter MSE of 2.647×10^{-2} for no missing data and 5.399×10^{-2} for plain imputation by zero, we see that even under high ratio mismatch, Richardson performs better than simple imputation. The only worse errors occur when $\delta_p = 0.3$ and $\delta_q \geq 0.2$, which we put in italics.

Table 2: MSE across δ_q and δ_p; all entries are $\times 10^{-2}$.

$\delta_q \backslash \delta_p$	0.00	0.05	0.10	0.15	0.20	0.30
0.00	3.27	3.53	3.75	3.97	4.24	4.83
0.05	3.60	3.83	4.14	4.40	4.67	5.25
0.10	3.70	3.89	4.13	4.40	4.57	5.05
0.15	3.96	4.17	4.37	4.66	4.89	5.31
0.20	4.34	4.51	4.73	4.93	5.12	*5.49*
0.30	4.62	4.77	4.89	4.99	5.13	*5.40*

I Additional GLM experiments

We provide additional comparisons of SGD with several imputation rules, with and without Richardson, across linear (Gaussian), logistic, and Poisson regression on synthetic and real datasets. Each figure shows the parameter-MSE trajectory under SGD with and without Richardson on top of several imputation schemes; the bias formulas of App. D predict the per-model behavior. We organize the figures by GLM family $\times$ missingness mechanism. The key empirical observations are as follows.

Key empirical observations.

- Richardson is consistently effective on top of mean, MICE, MICE-RF, and k-NN imputation across the three GLMs and the three mechanisms.
- Although the theory only covers one-pass SGD, Richardson remains robust empirically in multi-epoch training.
- In several settings, the gains are most pronounced in the first epoch, in agreement with the one-pass theory of Section 5.5.

I.1 Linear (Gaussian) regression

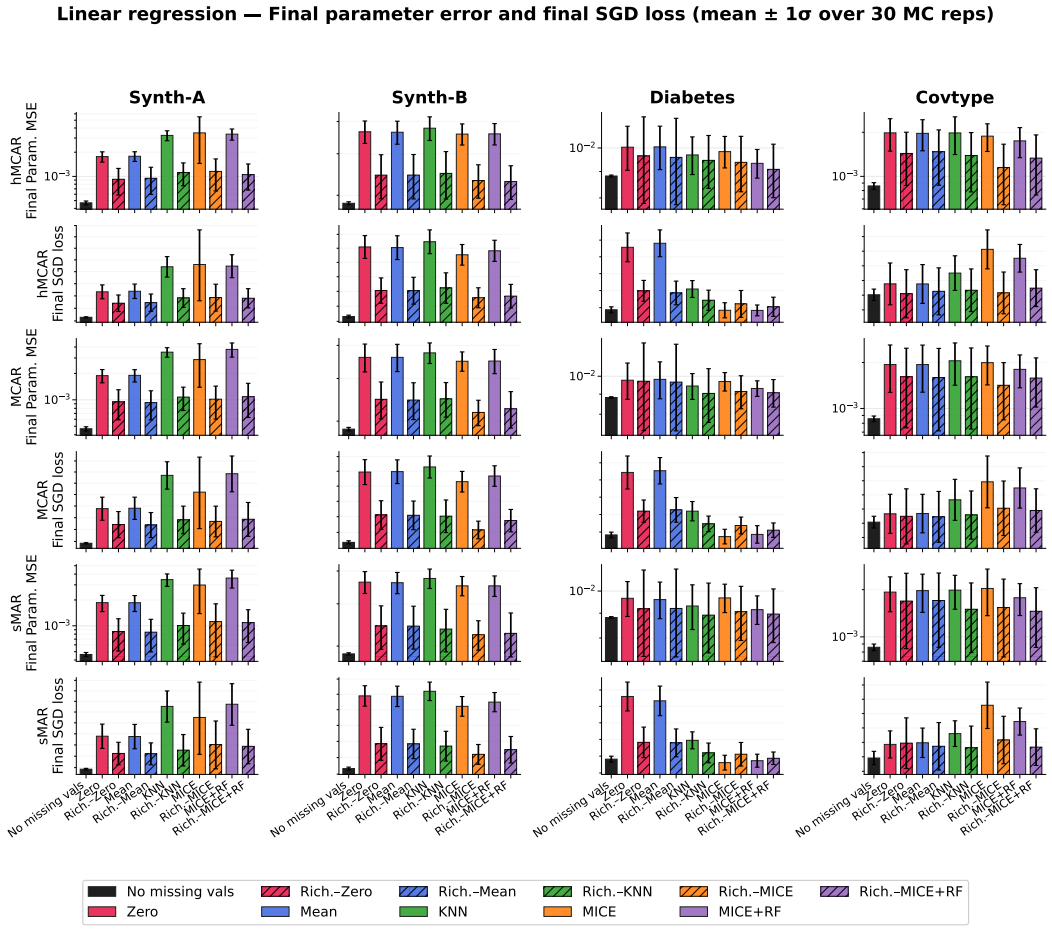

Figure 4: Final parameter MSE and test loss for linear regression, under MCAR, heterogeneous MCAR and sMAR mechanisms, on four different datasets.

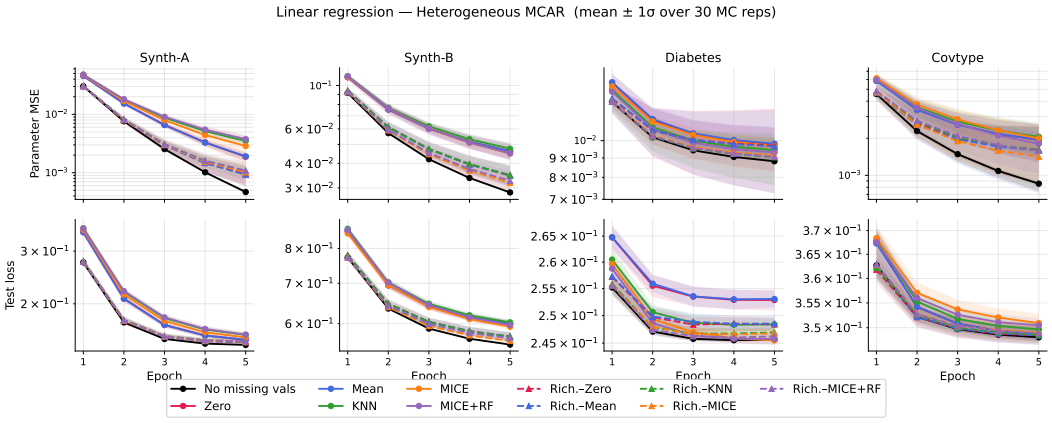

Figure 5: Convergence rate for parameter MSE and test loss for linear regression, under heterogeneous MCAR and sMAR mechanisms, on four different datasets.

I.2 Logistic regression

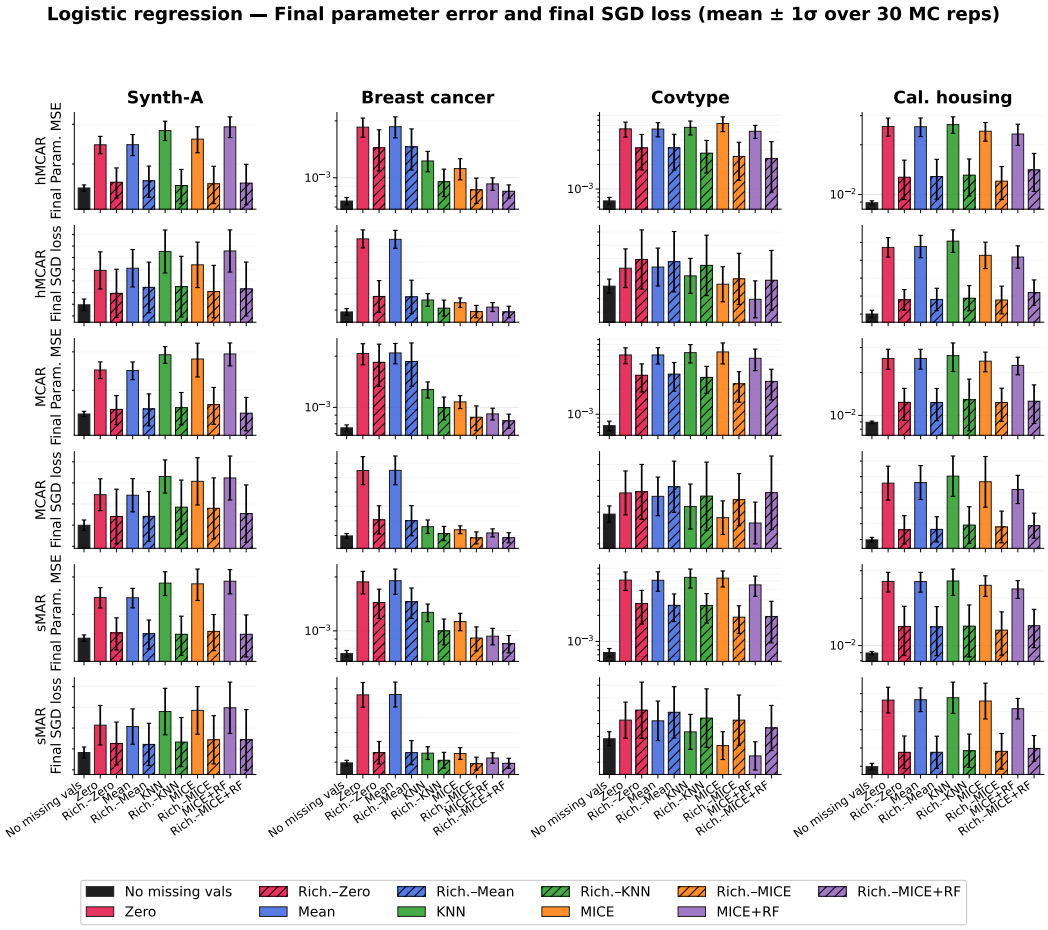

Figure 6: Final parameter MSE and test loss for logistic regression, under MCAR, heterogeneous MCAR and sMAR mechanisms, on four different datasets.

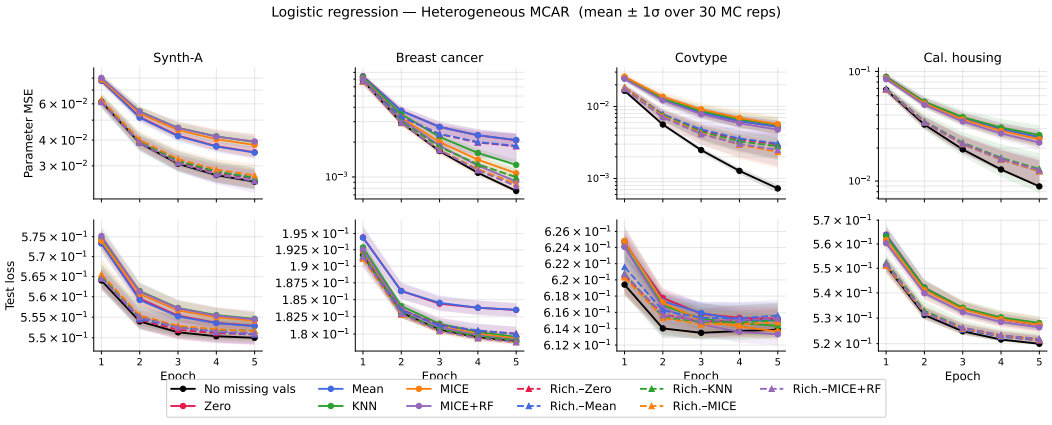

Figure 7: Convergence rate for parameter MSE and test loss for logistic regression, under heterogeneous MCAR and sMAR mechanisms, on four different datasets.

I.3 Poisson regression

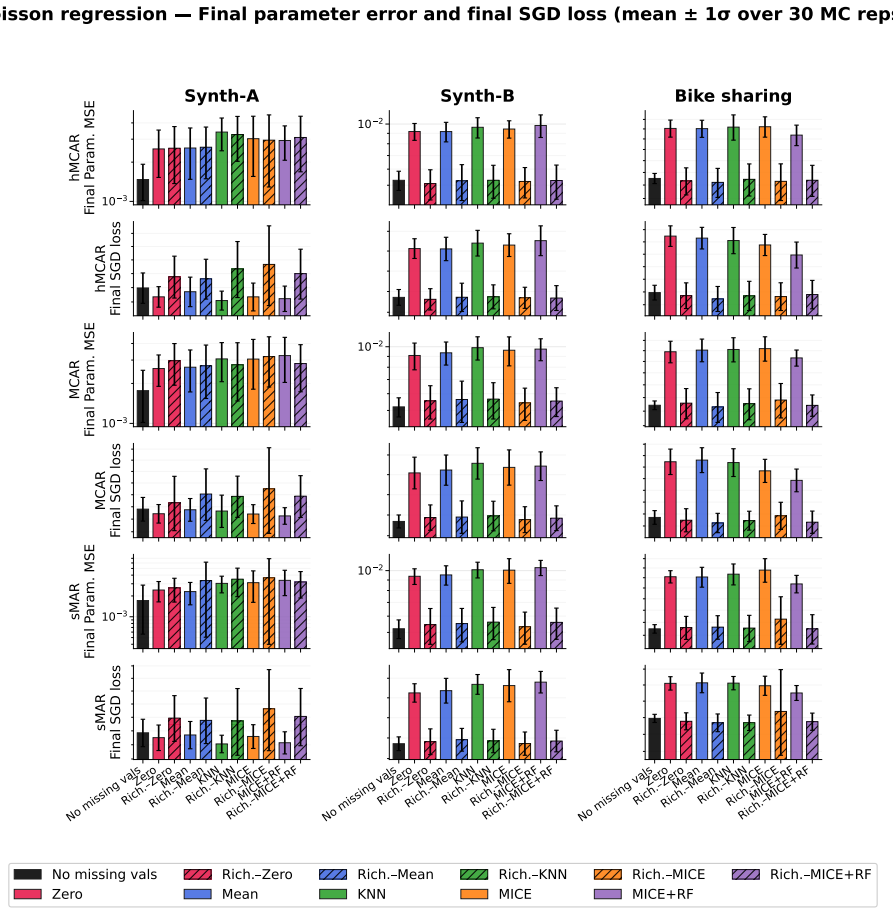

Figure 8: Final parameter MSE and test loss for Poisson regression, under MCAR, heterogeneous MCAR and sMAR mechanisms, on four different datasets.

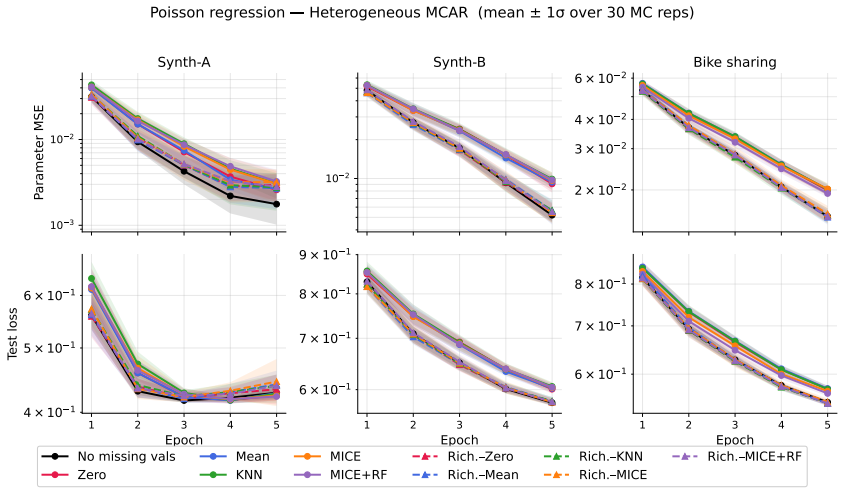

Figure 9: Convergence rate for parameter MSE and test loss for Poisson regression, under heterogeneous MCAR and sMAR mechanisms, on four different datasets.

J Comparison with other schemes

The experiments above show that Richardson-SGD improves performance across the three GLMs considered and across all imputation techniques tested. This is the regime for which the method is primarily intended: when the learner has access to an imputation pipeline, but does not want to impose a model-specific correction or strong structural assumptions on the covariate distribution. In this sense, Richardson-SGD is best viewed as a model-agnostic and distribution-agnostic debiasing layer on top of imputation, rather than as a competitor to specialized estimators designed for a particular statistical model. Consequently, the most direct comparison is with the same imputation pipeline used without Richardson.

For completeness, we nevertheless include a more specialized benchmark in the linear-regression setting. In this case, tailored alternatives are available: debiased SGD under hMCAR [32], and SAEM-type methods under parametric assumptions on the covariate distribution [11]. These methods are designed specifically for this setting, so Richardson-SGD is not expected to dominate them. The point of the comparison is instead to test whether a generic Richardson correction remains competitive even in a regime where model-specific methods have an intrinsic advantage.

The results support this interpretation. Richardson-SGD performs close to debiased SGD and improves over SAEM, especially on non-synthetic datasets where the parametric assumptions underlying SAEM are less well matched to the data. The main failure case occurs for Richardson combined with MICE in the uncorrelated Gaussian-covariate setting. This behavior is expected: when covariates are independent, MICE has essentially no cross-feature signal to exploit and may fit noise, making it a poor base imputer. In such cases, Richardson inherits the limitations of the underlying imputation procedure.

Table 3: Parameter MSE and runtime results. Values are reported as mean $\pm$ standard deviation. SAEM is run only once, with a runtime of 5 s for Synth-A and Synth-B, and 100 s otherwise.

Dataset	Method	PMSE	Time (s)
Synth-A	No missing data (ref)	$5.51\times10^{-5} \pm 1.10\times10^{-6}$	0.019
Synth-A	Zero-Impute	$4.11\times10^{-3} \pm 1.00\times10^{-3}$	0.021
Synth-A	Rich. – Zero	$2.36\times10^{-4} \pm 3.70\times10^{-5}$	0.028
Synth-A	Rich. – MICE	$1.56\times10^{-2} \pm 9.60\times10^{-3}$	0.036
Synth-A	Debiased SGD (Sportisse et al.)	$2.99\times10^{-4} \pm 1.10\times10^{-4}$	0.057
Synth-A	SAEM	$6.00\times10^{-4} \pm 0$	5.00
Synth-B	No missing data (ref)	$2.39\times10^{-3} \pm 1.50\times10^{-5}$	0.029
Synth-B	Zero-Impute	$4.95\times10^{-1} \pm 3.22\times10^{-2}$	0.031
Synth-B	Rich. – Zero	$4.97\times10^{-2} \pm 1.80\times10^{-3}$	0.055
Synth-B	Rich. – MICE	$3.42\times10^{-2} \pm 4.00\times10^{-3}$	0.057
Synth-B	Debiased SGD (Sportisse et al.)	$4.64\times10^{-3} \pm 2.00\times10^{-3}$	0.073
Synth-B	SAEM	$2.33\times10^{-2} \pm 0$	5.00
Covtype	No missing data (ref)	$1.94\times10^{-2} \pm 0$	0.028
Covtype	Zero-Impute	$8.04\times10^{-2} \pm 5.05\times10^{-2}$	0.033
Covtype	Rich. – Zero	$2.60\times10^{-2} \pm 3.55\times10^{-3}$	0.055
Covtype	Rich. – MICE	$2.37\times10^{-2} \pm 1.50\times10^{-3}$	0.057
Covtype	Debiased SGD (Sportisse et al.)	$1.99\times10^{-2} \pm 5.50\times10^{-3}$	0.079
Covtype	SAEM	$2.58\times10^{-1} \pm 0$	100.00
Cal. housing	No missing data (ref)	$2.62\times10^{-2} \pm 5.00\times10^{-5}$	0.026
Cal. housing	Zero-Impute	$1.02\times10^{-1} \pm 9.76\times10^{-3}$	0.021
Cal. housing	Rich. – Zero	$7.15\times10^{-2} \pm 1.23\times10^{-2}$	0.051
Cal. housing	Rich. – MICE	$5.10\times10^{-2} \pm 1.49\times10^{-2}$	0.049
Cal. housing	Debiased SGD (Sportisse et al.)	$3.16\times10^{-2} \pm 9.55\times10^{-3}$	0.065
Cal. housing	SAEM	$2.70\times10^{-1} \pm 0$	100.00

K Robustness to misspecification of the missingness mechanism

This appendix tests Richardson-SGD under misspecification of the thinning mechanism. Missing values are generated under the sMAR mechanism of Appendix G, where missingness depends on X_1 and X_2, but Richardson thinning uses the hMCAR approximation p_j instead of the true conditional probabilities $p_j q_j(V)$. We run linear regression on the four datasets in Table 1, a setting where the zero-imputation gradient bias is a polynomial of degree at most two in the missingness probabilities; see Appendix E. We compare zero imputation and MICE, with and without first-order Richardson correction, over 10 runs. Figure 10 shows the parameter-MSE and test-loss trajectories, and Table 4 reports final values.

Richardson remains robust to this misspecification. With zero imputation, Richardson improves over plain zero imputation on all datasets, often nearly matching the complete-data baseline. With MICE, Richardson improves performance on Synth-A, Synth-B, and Diabetes. The main exception is California Housing, where Richardson–MICE becomes unstable near the last epoch; the plotted variance is capped for readability. This instability is consistent with the variance amplification of Richardson extrapolation discussed in Section 5.3, and may also reflect an imperfect learning-rate choice.

Overall, treating sMAR data as hMCAR does not eliminate the benefit of Richardson-SGD in these experiments, especially with zero imputation. However, the California Housing–MICE case shows that misspecification can interact with the imputation rule and optimization dynamics.

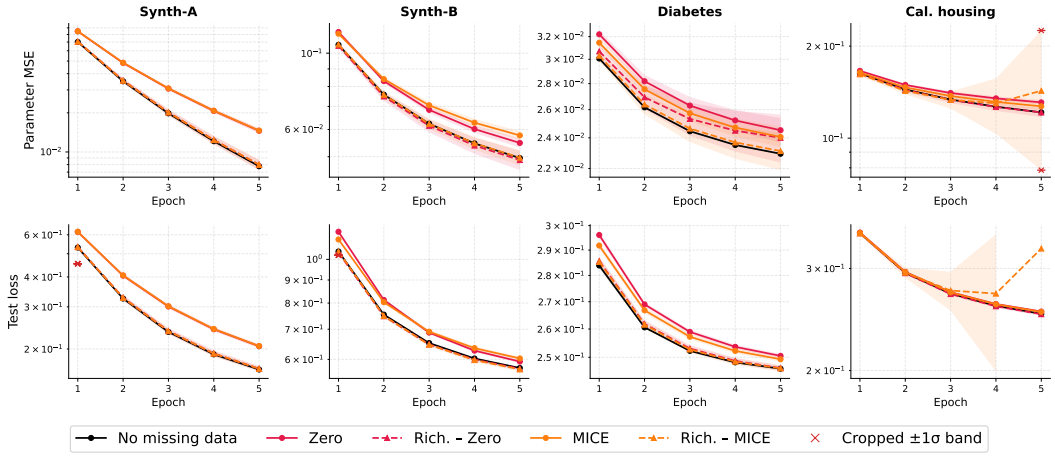

Figure 10: Richardson-SGD under misspecification of the missingness mechanism. Missing values are sMAR, but Richardson-SGD is computed using an hMCAR approximation based only on marginal missingness probabilities.

Table 4: Final parameter MSE and test loss for Richardson-SGD under misspecification of the missingness mechanism. The true mechanism is sMAR, while Richardson thinning uses an hMCAR approximation. Test losses are reported in units of 10^{-1}.

	Synth-A		Synth-B		Diabetes		Cal. housing	
Method	PMSE	Loss	PMSE	Loss	PMSE	Loss	PMSE	Loss
No missing data	$7.77_{\pm 0.045}\,10^{-3}$	1.64	$4.96_{\pm 0.016}\,10^{-2}$	5.74	$2.29_{\pm 0.0037}\,10^{-2}$	2.46	$1.21_{\pm 0.00099}\,10^{-1}$	2.51
Zero	$1.45_{\pm 0.063}\,10^{-2}$	2.05	$5.48_{\pm 0.17}\,10^{-2}$	5.93	$2.45_{\pm 0.086}\,10^{-2}$	2.51	$1.31_{\pm 0.024}\,10^{-1}$	2.53
Rich.–Zero	$7.99_{\pm 0.75}\,10^{-3}$	1.65	$4.89_{\pm 0.32}\,10^{-2}$	5.72	$2.40_{\pm 0.16}\,10^{-2}$	2.46	$1.21_{\pm 0.041}\,10^{-1}$	2.50
MICE	$1.46_{\pm 0.061}\,10^{-2}$	2.06	$5.77_{\pm 0.17}\,10^{-2}$	6.03	$2.41_{\pm 0.070}\,10^{-2}$	2.49	$1.27_{\pm 0.035}\,10^{-1}$	2.52
Rich.–MICE	$7.97_{\pm 0.71}\,10^{-3}$	1.65	$4.97_{\pm 0.25}\,10^{-2}$	5.70	$2.31_{\pm 0.12}\,10^{-2}$	2.46	$1.43_{\pm 1.20}\,10^{-1}$	3.25

L Why the two missingness scales must share the same imputation

Richardson correction compares two gradients evaluated at missingness scales p and Cp. For the linear term to cancel, these two gradients must be generated by the *same imputation operator*. In particular, entries that are missing at both scales must receive the same imputed value. This is why, in Section 4, we impute only once at the higher missingness scale Cp, and then restore the entries that were artificially hidden to obtain the lower-scale covariate.

We formalize this point. Let $\mathcal{I}$ be a data-independent imputation rule and define

$$\hat{g}_{\mathcal{I}}^{(p)}(w) := g\Big(w; \tilde{X}_{\mathcal{I}}^{(p)}, Y\Big), \qquad \mathcal{B}_{\mathcal{I}}(w,p) := \mathbb{E}\Big[\hat{g}_{\mathcal{I}}^{(p)}(w)\Big] - \nabla L(w).$$

By Proposition 1,

$$\mathcal{B}_{\mathcal{I}}(w,p) = \mathcal{A}_{\mathcal{I}}(w)p + \mathcal{R}_{\mathcal{I}}(w,p), \qquad \mathcal{R}_{\mathcal{I}}(w,p) = O(\|p\|^2)$$

under independent hMCAR/sMAR. The first-order operator $\mathcal{A}_{\mathcal{I}}(w)$ depends on the imputation rule, since its j-th column is the expected gradient gap created by declaring coordinate j missing.

If Richardson is applied with two possibly different imputation rules $\mathcal{I}_0$ and $\mathcal{I}_1$ at scales p and Cp, respectively, then

$$\hat{g}^{\mathrm{R}}_{C,\mathcal{I}_0,\mathcal{I}_1}(w) := \frac{C\,\hat{g}_{\mathcal{I}_0}^{(p)}(w) - \hat{g}_{\mathcal{I}_1}^{(Cp)}(w)}{C-1}.$$

Its bias is

$$\begin{aligned}\mathbb{E}\big[\hat{g}^{\mathrm{R}}_{C,\mathcal{I}_0,\mathcal{I}_1}(w)\big] - \nabla L(w) &= \frac{C\,\mathcal{B}_{\mathcal{I}_0}(w,p) - \mathcal{B}_{\mathcal{I}_1}(w,Cp)}{C-1} \\ &= \tfrac{C}{C-1}\big(\mathcal{A}_{\mathcal{I}_0}(w) - \mathcal{A}_{\mathcal{I}_1}(w)\big)p + O(\|p\|^2).\end{aligned}$$

Thus the $O(\|p\|)$ term cancels only if

$$\mathcal{A}_{\mathcal{I}_0}(w) = \mathcal{A}_{\mathcal{I}_1}(w).$$

This condition is automatic when the two gradients are constructed from the same higher-scale imputation, as in Equation (10): common missing entries have identical imputed values at both scales, and the only difference between $\tilde{X}^{(p)}$ and $\tilde{X}^{(Cp)}$ comes from the entries artificially hidden by the thinning step.

By contrast, if one independently runs two stochastic imputers, for example two separate MICE chains at scales p and Cp, then common missing entries may receive different imputations. The corresponding first-order operators need not coincide, so Richardson may leave an even bigger uncancelled $O(\|p\|)$ bias and can amplify stochastic imputation noise through the factor $(C-1)^{-1}$.

Figure 11 illustrates this effect on a linear-regression experiment on California Housing. The linked construction, which imputes once at scale Cp, $C = 1.5$, and restores artificially hidden entries, remains stable and improves over plain MICE. The unlinked construction, which runs independent MICE imputations at the two scales, loses the first-order cancelation and becomes unstable.

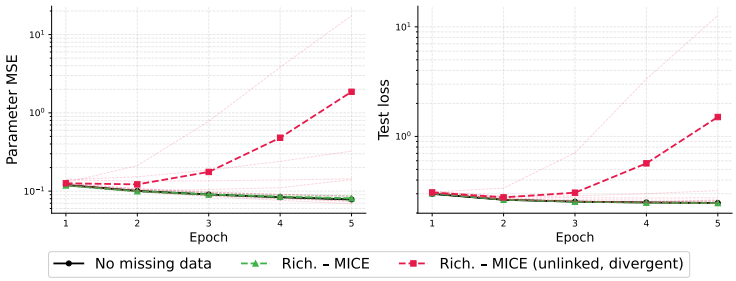

Figure 11: Linked versus unlinked Richardson–MICE on California Housing linear regression. The linked version uses one imputation at scale Cp and restores the artificially hidden entries to obtain the scale-p covariate. The unlinked version runs two separate MICE imputations at scales p and Cp. Only the linked construction preserves the common imputed values required for first-order Richardson cancellation.