Debiased stochastic gradient descent to handle missing data in the linear regression case and its perspectives[1]

Laboratoire Heudiasyc, UTC Compiègne

Aude Sportisse [3] Claire Boyer [1] Aymeric Dieuleveut [2]
Julie Josse [3]

[1]Laboratoire de Probabilités Statistique et Modélisation, Sorbonne Université

[2]Centre de Mathématiques Appliquées, Ecole Polytechnique

[3]INRIA Sophia Antipolis

15th February 2021

[1]A. S. et al. "Debiasing Stochastic Gradient Descent to handle missing values". In: *Advances in Neural Information Processing System* (2020).

Motivation: large-scale incomplete data

- **Large-scaling:** large n (number of observations), large d (dimension of the observations)
  - ↪ **Stochastic / online learning algorithms.**
- **Incompleteness** for many reasons: delete observations with NA $\rightarrow$ keep only 5% of the rows.
  - ↪ **Simple algorithmic solution?**

Traumabase: 250 var/ 15 000 patients/ 15 hospitals

Center	Age	Sex	Weight	Height	Heart rate	Lactates
Beaujon	54	m	85	NA	NA	NA
Lille	33	m	80	1.8	180	4.8
Pitie	26	m	NA	NA	NA	3.9
Beaujon	63	m	80	1.8	190	1.66
Pitie	30	w	NA	NA	NA	NA

NA: Not Available.

1. SGD with missing values

2. Theoretical results
   - Without missing values: rates
   - Convergence of our algorithm
   - Adaptation to estimated missing probabilities

3. Experiments

4. Perspectives
   - Finite-sample setting
   - More general loss function
   - More general missing-data mechanisms

Linear regression model

- $(X_{i:}, y_i)_{i \geq 1} \in \mathbb{R}^d \times \mathbb{R}$ i.i.d. observations

$$y_i = X_{i:}^T \beta^\star + \epsilon_i,$$

parametrized by $\beta^\star \in \mathbb{R}^d$, with a noise term $\epsilon_i \in \mathbb{R}$.

- Loss function: $f_i(\beta) = (\langle X_{i:}, \beta \rangle - y_i)^2 / 2$.
- **True risk minimization**:

$$\beta^\star = \underset{\beta \in \mathbb{R}^d}{\arg\min} \left\{ R(\beta) := \mathbb{E}_{(X_{i:}, y_i)} \left[f_i(\beta) \right] \right\}$$

- Stochastic gradient method.
  - At the heart of Machine Learning.
  - Especially useful in high dimension.

Objective

Challenges

- **Large-scaling:** large number of observations, large d.
- **Incomplete data**: missing covariates, $(X_{i:})$'s partially known.
- **Online-setting**: the data come as it goes along.

How to adapt algorithms to the missing data case?

Optimization without missing values

Stochastic gradient descent

- **SGD**: using unbiased estimates of $\nabla R(\beta_{k-1})$.

$$\beta_k = \beta_{k-1} - \alpha g_k(\beta_{k-1})$$

where α is the step-size and $g_k(\beta_{k-1}) = \nabla f_k(\beta_{k-1})$.

$$\mathbb{E}\left[g_k(\beta_{k-1})|\sigma(X_{1:}, y_1, \ldots, X_{k-1:}, y_{k-1})\right] = \nabla R(\beta_{k-1}),$$

- **Averaged SGD**: using the Polyak-Ruppert averaged iterates.

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

✓ It scales with large data

Optimization without missing values

Stochastic gradient descent

- **SGD**: using unbiased estimates of $\nabla R(\beta_{k-1})$.

$$\beta_k = \beta_{k-1} - \alpha g_k(\beta_{k-1})$$

where α is the step-size and $g_k(\beta_{k-1}) = \nabla f_k(\beta_{k-1})$.

$$\mathbb{E}\left[g_k(\beta_{k-1})|\sigma(X_{1:}, y_1, \ldots, X_{k-1:}, y_{k-1})\right] = \nabla R(\beta_{k-1}),$$

- **Averaged SGD**: using the Polyak-Ruppert averaged iterates.

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

✓ It scales with large data

2 challenges

- Obtaining unbiased stochastic gradients with missing data?
- Deriving rates of convergence?

Missing values setting

Formalism

- Missing-data pattern (or mask) D: $D_{i:} \in \{0,1\}^d$, such that

$$D_{ij} = \begin{cases} 0 & \text{if the } (i,j)\text{-entry is missing} \\ 1 & \text{otherwise.} \end{cases}$$

- Access to $X_{i:}^{\mathrm{NA}} \in (\mathbb{R} \cup \{\mathtt{NA}\})^d$ instead of $X_{i:}$

$$X_{i:}^{\mathrm{NA}} := X_{i:} \odot D_{i:} + \mathtt{NA}(1_d - D_{i:}),$$

$\odot$ element-wise product, $1_d = (1 \dots 1)^T \in \mathbb{R}^d$, $\mathtt{NA} \times 0 = 0$, $\mathtt{NA} \times 1 = \mathtt{NA}$.

Missing values setting

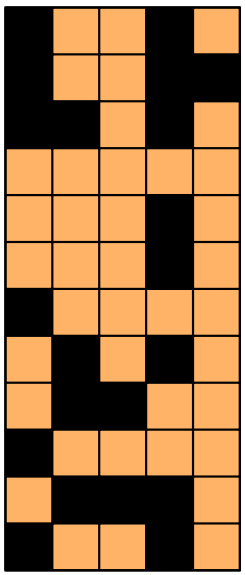
Heterogeneous MCAR

- ***Heterogeneous*** **Missing Completely At Random setting (MCAR)** $\rightarrow$ Bernoulli mask

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

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

✓ different missing probability for each covariate



Heterogeneous case:
$p_1 = 0.5, p_2 = 0.67, p_3 = 0.83, p_4 = 0.33, p_5 = 0.92.$

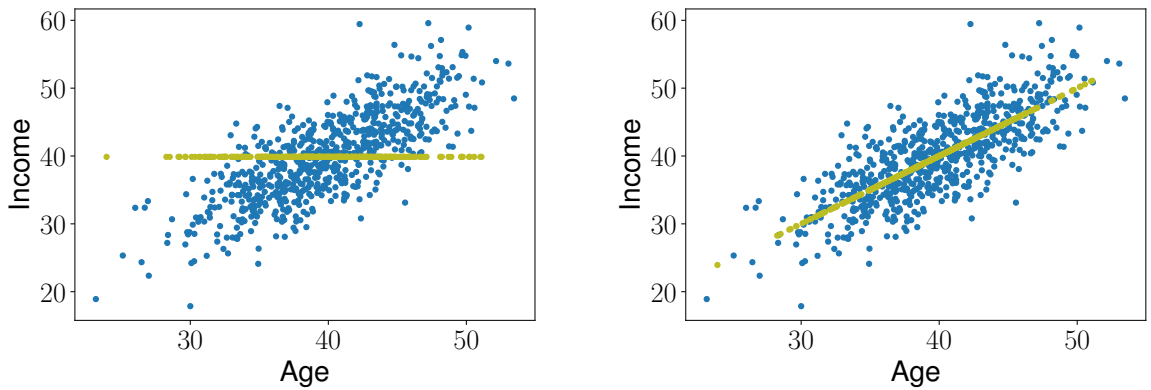
Homogeneous case: $p = 0.65$.

Dealing with missing values

Existing works

Aim: estimate parameters of a linear regression.

- **EM algorithm**[a]: maximization of the observed likelihood.
  - ✗ strong assumption on the data distribution
  - ✗ computationally costly, does not scale with large data.
  - ✗ not simple to establish, no many implementations.
- **Simple imputation**: mean imputation, performing regression.



[a] Arthur P Dempster, Nan M Laird, and Donald B Rubin. "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1 (1977), pp. 1–22.

Dealing with missing values

Existing works

Aim: estimate parameters of a linear regression.

- **EM algorithm**[a]: maximization of the observed likelihood.
  - ✗ strong assumption on the data distribution
  - ✗ computationally costly, does not scale with large data.
  - ✗ not simple to establish, no many implementations.
- **Simple imputation**: mean imputation, performing regression.
  - ✗ bias in the estimates, correlation between the variables overestimated.
- **(Multiple) imputation**: `mice`[b]
  - ✗ not online, difficult to establish for Ridge regression.

---

[a] Dempster, Laird, and Rubin, "Maximum likelihood from incomplete data via the EM algorithm".

[b] Buuren and Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations in R".

Naive imputation and debiasing

- Imputing naively by 0.
- Modifying usual algorithms to account for the imputation error.

- Dantzig selector[2].
- Lasso[3].
- SGD[4].

[2] Mathieu Rosenbaum, Alexandre B Tsybakov, et al. "Sparse recovery under matrix uncertainty". In: *The Annals of Statistics* 38.5 (2010), pp. 2620–2651.

[3] Po-Ling Loh and Martin J Wainwright. "High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity". In: *Advances in Neural Information Processing Systems*. 2011, pp. 2726–2734.

[4] Anna Ma and Deanna Needell. "Stochastic Gradient Descent for Linear Systems with Missing Data". In: *arXiv preprint arXiv:1702.07098* (2017).

Dealing with missing values

Our strategy inspired by Ma et Needell

Online-streaming: for a new observation $(X_{k:}^{\mathrm{NA}}, y_k)$

- **Imputing the missing values by 0**.

$$\tilde{X}_{k:} = X_{k:}^{\mathrm{NA}} \odot D_{k:} = X_{k:} \odot D_{k:} \text{ imputed covariates}$$

- Using a **debiased gradient** for the **averaged SGD**:
  Find $\tilde{g}_k(\beta_k)$ such that $\mathbb{E}\left[\tilde{g}_k(\beta_{k-1}) \,|\, \mathcal{F}_{k-1}\right] = \nabla R(\beta_{k-1})$

Dealing with missing values

Our strategy inspired by Ma et Needell

Online-streaming: for a new observation $(X_{k:}^{\mathrm{NA}}, y_k)$

- **Imputing the missing values by 0**.

$$\tilde{X}_{k:} = X_{k:}^{\mathrm{NA}} \odot D_{k:} = X_{k:} \odot D_{k:} \text{ imputed covariates}$$

- Using a **debiased gradient** for the **averaged SGD**:
  Find $\tilde{g}_k(\beta_k)$ such that $\mathbb{E}\left[\tilde{g}_k(\beta_{k-1}) \,|\, \mathcal{F}_{k-1}\right] = \nabla R(\beta_{k-1})$

  . $\mathcal{F}_{k-1} = \sigma(X_{1:}, y_1, D_{1:} \ldots, X_{k-1:}, y_{k-1}, D_{k-1:})$

  . $\nabla R(\beta_{k-1}) = \mathbb{E}_{(X_{k:}, y_k)}[X_{k:}(X_{k:}^T \beta_{k-1} - y_k)]$

  . No access to $X_{k:}$, only to $\tilde{X}_{k:}$.

  . Another source of randomness: $\mathbb{E} = \mathbb{E}_{(X_{k:}, y_k), D_{k:}} \overset{\text{indep}}{=} \mathbb{E}_{(X_{k:}, y_k)} \mathbb{E}_{D_{k:}}$

  . $\mathbb{E}_{D_{k:}} | \mathcal{F}_{k-1} \rightsquigarrow \mathbb{E}_{D_{k:}}$

  ✓ Mask at step k independent from the previous constructed iterate.

Dealing with missing values

Our strategy inspired by Ma et Needell

Online-streaming: for a new observation $(X_{k:}^{\mathrm{NA}}, y_k)$

- **Imputing the missing values by 0**.

$$\tilde{X}_{k:} = X_{k:}^{\mathrm{NA}} \odot D_{k:} = X_{k:} \odot D_{k:} \text{ imputed covariates}$$

- Using a **debiased gradient** for the **averaged SGD**:
  Find $\tilde{g}_k(\beta_k)$ such that $\mathbb{E}\left[\tilde{g}_k(\beta_{k-1}) \,|\, \mathcal{F}_{k-1}\right] = \nabla R(\beta_{k-1})$

$$\mathbb{E}_{D_{k:}}\left[\tilde{X}_{k:}\right] = \mathbb{E}_{D_{k:}}\left[\begin{pmatrix} \delta_{k1} X_{k1} \\ \vdots \\ \delta_{kd} X_{kd} \end{pmatrix}\right] = \begin{pmatrix} p_1 X_{k1} \\ \vdots \\ p_d X_{kd} \end{pmatrix}$$

Thus

$$\mathbb{E}_{D_{k:}}\left[P^{-1}\tilde{X}_{k:}\right] := \begin{pmatrix} p_1^{-1} & & \\ & \ddots & \\ & & p_d^{-1} \end{pmatrix} \begin{pmatrix} p_1 X_{k1} \\ \vdots \\ p_d X_{kd} \end{pmatrix} = X_{k:}$$

Dealing with missing values

Our strategy inspired by Ma et Needell

Online-streaming: for a new observation $(X_{k:}^{\mathrm{NA}}, y_k)$

- **Imputing the missing values by 0**.

$$\tilde{X}_{k:} = X_{k:}^{\mathrm{NA}} \odot D_{k:} = X_{k:} \odot D_{k:} \text{ imputed covariates}$$

- Using a **debiased gradient** for the **averaged SGD**:
  Find $\tilde{g}_k(\beta_k)$ such that $\mathbb{E}\left[\tilde{g}_k(\beta_{k-1}) \,|\, \mathcal{F}_{k-1}\right] = \nabla R(\beta_{k-1})$

One obtains

$$\tilde{g}_k(\beta_{k-1}) = P^{-1}\tilde{X}_{k:}\left(\tilde{X}_{k:}^T P^{-1}\beta_{k-1} - y_k\right) - (\mathrm{I} - P)P^{-2}\mathrm{diag}\left(\tilde{X}_{k:}\tilde{X}_{k:}^T\right)\beta_{k-1}.$$

Averaged SGD for missing values

Debiasing the gradient

Algorithm 1 Averaged SGD for Heterogeneous Missing Data

Input: data $\tilde{X}, y, \alpha$ (step size)
Initialize $\beta_0 = 0_d$.
Set $P = \mathrm{diag}\left((p_j)_{j\in\{1,\dots,d\}}\right) \in \mathbb{R}^{d\times d}$.
for $k = 1$ **to** n **do**
$\tilde{g}_k(\beta_{k-1}) = P^{-1}\tilde{X}_{k:}\left(\tilde{X}_{k:}^T P^{-1}\beta_{k-1} - y_k\right) - (\mathrm{I} - P)P^{-2}\mathrm{diag}\left(\tilde{X}_{k:}\tilde{X}_{k:}^T\right)\beta_{k-1}$
$\beta_k = \beta_{k-1} - \alpha\tilde{g}_k(\beta_{k-1})$
$\bar{\beta}_k = \frac{1}{k+1}\sum_{i=0}^{k}\beta_i = \frac{k}{k+1}\bar{\beta}_{k-1} + \frac{1}{k+1}\beta_k$
end for

- $p = 1 \Rightarrow P^{-1} = I_d$ standard least squares stochastic algorithm.
- Computation cost for the gradient still weak.
- Trivially extended to ridge regularization (no change for the gradient): $\min_{\beta\in\mathbb{R}^d}\ R(\beta) + \lambda\|\beta\|^2, \lambda > 0$

SGD with NA: Take home message

- ✓ We aim to estimate $\beta^\star$ with missing data.
- ✓ We consider an heterogeneous MCAR framework.
- ✓ We provide an unbiased gradient oracle of the true risk.
- ✓ Only for Least Squares Regression.
- ✓ Requires independent points at each iteration: only for the first pass.
- ✓ Requires the knowledge of P.

Convergence?

1. SGD with missing values

2. Theoretical results
   - Without missing values: rates
   - Convergence of our algorithm
   - Adaptation to estimated missing probabilities

3. Experiments

4. Perspectives
   - Finite-sample setting
   - More general loss function
   - More general missing-data mechanisms

Optimization without missing values

- F is **convex and L-smooth**[5] i.e. if F is twice differentiable,

$$\forall \beta \in \mathbb{R}^d, 0 \leq |\text{eigenvalues}(\nabla^2 F(\beta))| \leq L.$$

  ✗ Convergence rate: $\mathcal{O}(k^{-1/2})$

- F is **μ-strongly convex and L-smooth**.
  ✗ Convergence rate: $\mathcal{O}(\mu k^{-1})$

- F is **convex and quadratic**[6].
  ✓ Convergence rate: $\mathcal{O}(k^{-1})$

---

[5] Arkadi Nemirovski et al. "Robust stochastic approximation approach to stochastic programming". In: *SIAM Journal on optimization* 19.4 (2009), pp. 1574–1609.

[6] Eric Moulines and Francis R Bach. "Non-asymptotic analysis of stochastic approximation algorithms for machine learning". In: *Advances in Neural Information Processing Systems*. 2011, pp. 451–459.

Theoretical results

Technical lemmas

- Goal: establish a convergence rate.
- Assumptions on the data: $(X_{k:}, y_k) \in \mathbb{R}^d \times \mathbb{R}$ i.i.d., $\mathbb{E}[\|X_{k:}\|^2]$ and $\mathbb{E}[y_k^2]$ finite, $H := \mathbb{E}_{(X_{k:}, y_k)}[X_{k:} X_{k:}^T]$ invertible.

Lemma: noise induced by the imputation by 0 is **structured**

$(\tilde{g}_k(\beta^\star))_k$ with $\beta^\star$ is $\mathcal{F}_k-$measurable and $\forall k \geq 0$,

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

Lemma: $(\tilde{g}_k(\beta^\star))_k$ are **a.s. co-coercive**

For any k,

- $\tilde{g}_k$ is $L_{k,D}$-Lipschitz
- there exists a random primitive function $\tilde{f}_k$ which is a.s. convex

Theorem: convergence rate of $\mathcal{O}(k^{-1})$, streaming setting

Assume that for any i, $\|X_{i:}\| \leq \gamma$ almost surely for some $\gamma > 0$. For **any constant step-size** $\alpha \leq \frac{1}{2L}$, our algorithm ensures that, for any $k \geq 0$:

$$\mathbb{E}\left[R\left(\bar{\beta}_k\right) - R(\beta^\star)\right] \leq \frac{2}{k}\left(\underbrace{\sqrt{c(\beta^\star)d}}_{\text{variance term}} + \underbrace{\frac{\|\beta_0 - \beta^\star\|}{\sqrt{\alpha}}}_{\text{bias term}}\right)^2,$$

- $L := \sup_{k,D}$ Lipschitz constants of $\tilde{g}_k$
- $p_m = \min_{j=1,\dots d}\ p_j$ minimal probability to be observed among the variables.
- $c(\beta^\star) = \underbrace{\overbrace{\frac{\mathrm{Var}(\epsilon_k)}{p_m^2}}^{\text{classical term}} + \overbrace{\left(\frac{7(1-p_m)}{p_m^3}\right)\gamma^2\|\beta^\star\|^2}^{\text{multiplicative noise (induced by naive imputation)}}}_{\text{increasing with the missing values rate}}.$

Theoretical results

Comments

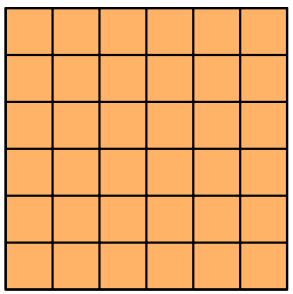
- Optimal rate for least-squares regression.
- In the complete case: same bound as Bach and Moulines.
- Bound on the iterates for the **ridge regression** ($\beta \to R(\beta) + \lambda\|\beta\|^2$ is 2λ-strongly convex).

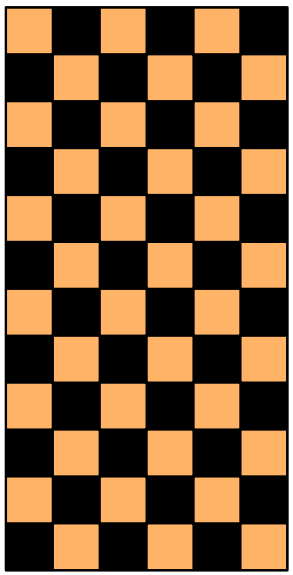
$$\mathbb{E}\left[\left\|\overline{\beta}_k - \beta^\star\right\|^2\right] \leq \frac{1}{\lambda k}\left(\sqrt{c(\beta^\star)d} + \frac{\|\beta_0 - \beta^\star\|}{\sqrt{\alpha}}\right)^2.$$

Theoretical results

What impact of missing values?

Fewer complete observations is better than more incomplete ones: is it better to access 200 incomplete observations (with a probability 50% of observing) or to have 100 complete observations?





Theoretical results

What impact of missing values?

Fewer complete observations is better than more incomplete ones: is it better to access 200 incomplete observations (with a probability 50% of observing) or to have 100 complete observations?

- without missing observations: variance bound scales as $\mathcal{O}\left(\frac{\mathrm{Var}(\epsilon_k)d}{k}\right)$.
- with missing observations: $O\left(\frac{\mathrm{Var}(\epsilon_k)d}{kp_m^2} + \frac{C(X,\beta^\star)}{kp_m^3}\right)$.
- variance bound larger by a factor p_m^{-1} for the estimator derived from k **incomplete** observations than for $k \times p_m$ **complete** observations.

The variance bound for 200 incomplete observations (with a probability 50% of observing) is twice as large as for 100 complete observations.

We do better than discarding all observations which contain missing values:

$$X = \begin{array}{c} \begin{array}{ccc} X_1 & X_2 & X_3 \end{array} \\ \begin{pmatrix} 12 & 28 & 31 \\ \mathrm{NA} & 23 & 89 \\ 32 & 6 & 24 \\ \vdots & \vdots & \vdots \\ \mathrm{NA} & 3 & 7 \end{pmatrix} \end{array} \qquad X = \begin{array}{c} \begin{array}{ccc} X_1 & X_2 & X_3 \end{array} \\ \begin{pmatrix} 12 & 28 & 31 \\ \cancel{\mathrm{NA}} & \cancel{23} & \cancel{89} \\ 32 & 6 & 24 \\ \vdots & \vdots & \vdots \\ \cancel{\mathrm{NA}} & \cancel{3} & \cancel{7} \end{pmatrix} \end{array}$$

We do better than discarding all observations which contain missing values:

Example in the homogeneous case with p the proportion of being observed.

- keeping only the complete observations, any algorithm:
  - number of complete observations $k_{co} \sim \mathcal{B}(k, p^d)$.
  - statistical lower bound: $\frac{\mathrm{Var}(\epsilon_k) d}{k_{co}}$.
  - in expectation, lower bound on the risk larger than $\frac{\mathrm{Var}(\epsilon_k) d}{k p^d}$.
- keeping all the observations, averaged SGD: upper bound $O\left(\frac{\mathrm{Var}(\epsilon_k) d}{k p^2} + \frac{C(X, \beta^\star)}{k p^3}\right)$.

Our strategy has an upper-bound p^{d-3} smaller than the lower bound of any algorithm relying only on the complete observations.

Theoretical results

Result with estimated missing probabilities

Finite-sample setting: n is fixed

Algorithm and main result: requirement of $(p_j)_{j=1,\dots,d}$.
$\rightarrow$ estimator $\bar{\beta}_k$

In practice: estimated missing probabilities $(\hat{p}_j)_{j=1,\dots,d}$
$\rightarrow$ estimator $\bar{\hat{\beta}}_k$. (finite-sample setting: first half of the data to evaluate $(\hat{p}_j)$, second half to build $\bar{\hat{\beta}}_k$).

Result with estimated missing probabilities (simplified version)

Under additional assumptions of **bounded iterates** and **strong convexity** of the risk, Algorithm 1 ensures that, for any $k \geq 0$:

$$\mathbb{E}\left[R(\bar{\hat{\beta}}_k) - R(\bar{\beta}_k)\right] = \mathcal{O}(1/kp_m^6),$$

with $p_m = \min_{j \in \{1,\dots,d\}} p_j$.

Convergence rates: Take home message

New results:

- ✓ Fast convergence rate because the noise is structured. Optimal w.r.t. k.
- ✓ Dependence with p: much better than deleting incomplete data, but not as good as pk complete observations
- ✓ Convergence with strong-convexity and estimated probabilities (preserved $1/k$, degraded dependence in p)

Open questions:

- ✓ **What about empirical risk?** [to be continued.]

1. SGD with missing values

2. Theoretical results
   - Without missing values: rates
   - Convergence of our algorithm
   - Adaptation to estimated missing probabilities

3. Experiments

4. Perspectives
   - Finite-sample setting
   - More general loss function
   - More general missing-data mechanisms

Experiments

Synthetic data: setting

- $X_{i:} \overset{i.i.d.}{\sim} \mathcal{N}(0, \Sigma)$, where Σ with uniform random eigenvectors and decreasing eigenvalues, $\epsilon_i \sim \mathcal{N}(0, 1)$
- $y_i = X_{i:}\beta + \epsilon_i$, for β fixed
- $d = 10$, 30% missing values.

- **AvSGD** averaged iterates with a constant step size $\alpha = \frac{1}{2L}$ [a].
- **SGD** [b] with iterates $\beta_{k+1} = \beta_k - \alpha_k \tilde{g}_{i_k}(\beta_k)$, and decreasing step size $\alpha_k = \frac{1}{\sqrt{k+1}}$.
- **SGD_cst** [b] with a constant step size $\alpha = \frac{1}{2L}$ [a]

[a] L is considered to be known.

[b] Ma and Needell, "Stochastic Gradient Descent for Linear Systems with Missing Data".

Experiments

Synthetic data: convergence rate

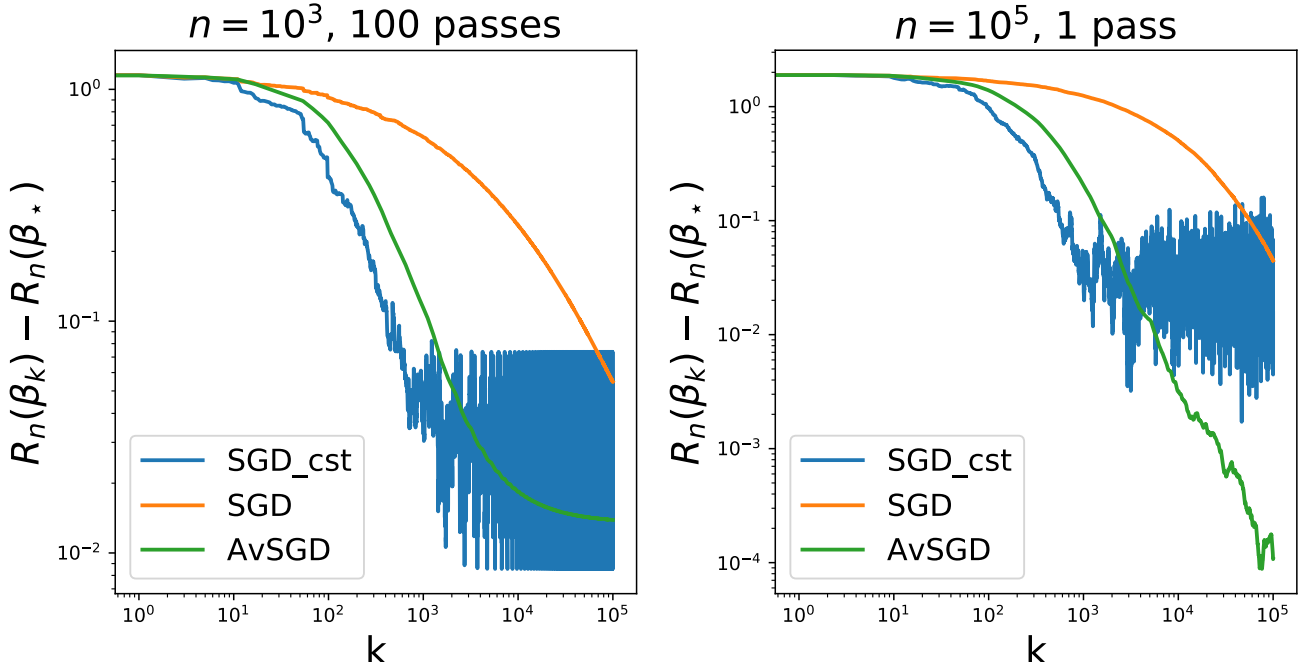


Figure: Empirical excess risk $(R_n(\beta_k) - R_n(\beta^\star))$.

- Multiple passes (left): saturation.
- One pass (right): saturation for **SGD_cst**, $\mathcal{O}(n^{-1/2})$ for **SGD**, $\mathcal{O}(n^{-1})$ for **AvSGD**.

Experiments

Real dataset: Superconductivity, prediction task

- Goal: predict the critical temperature of each superconductor. **Complete** dataset: 81 quantitatives features, 21263 superconductors.
- Introduction of 30% of heterogeneous MCAR missing values, probabilities of being observed vary between 0.7 and 1.
- Dataset divided into training and test set, with no missing values in the test set.
- Prediction of the critical temperature: $\hat{y}_{n+1} = X_{n+1}^T \hat{\beta}$ with the coefficient
  - $\hat{\beta} = \beta_n^{\mathrm{AvSGD}}$ by applying **AvSGD** on the training set.
  - $\hat{\beta} = \beta_n^{\mathrm{EM}}$ by applying the EM algorithm on the training set.
  - $\hat{\beta} = \bar{\beta}_n^{\mathrm{AvSGD}}$ by imputing the missing data naively by the mean in the training set, and applying the averaged SGD without missing data (**Mean+AvSGD**)

Experiments

Real dataset: Superconductivity, prediction task

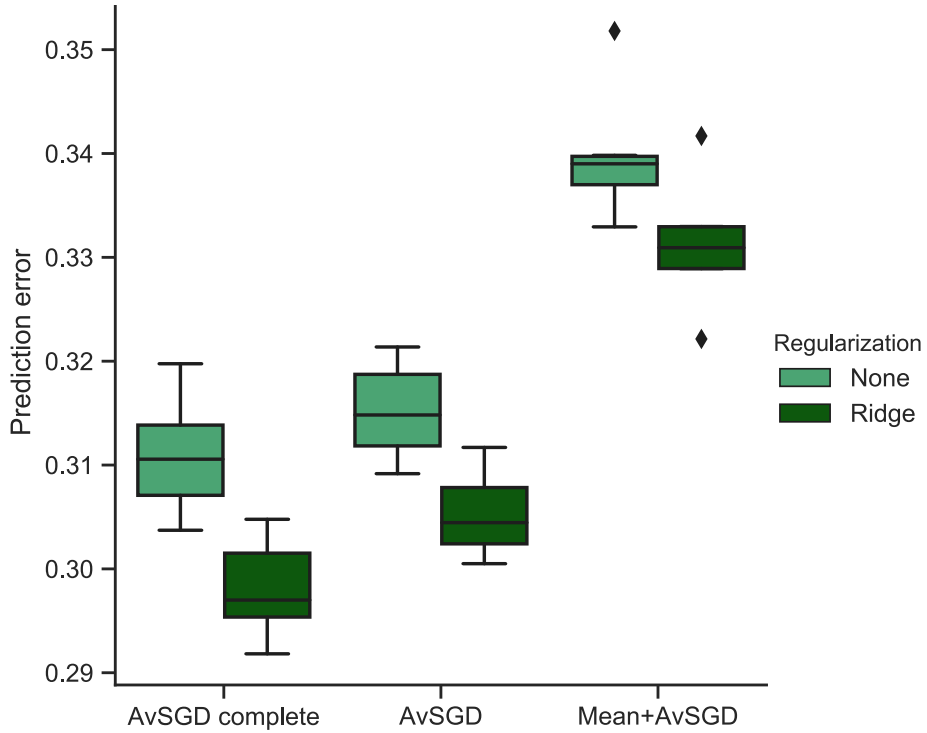


Figure: Prediction error $\|\hat{y} - y\|^2 / \|y\|^2$ boxplots.

- EM out of range (due to large number of covariates).
- **AvSGD** performs well, very close to the one obtained from the complete dataset (**AvSGD complete**) with or without regularization.

Take home message

New results:

✓ SGD for dealing with heterogeneous MCAR data with a least squares loss.

Open questions:

✓ Dealing with **more general loss function.** [to be continued.]

✓ More **complex missing-data patterns** such as MAR and MNAR. [to be continued.]

1. SGD with missing values

2. Theoretical results
   - Without missing values: rates
   - Convergence of our algorithm
   - Adaptation to estimated missing probabilities

3. Experiments

4. Perspectives
   - Finite-sample setting
   - More general loss function
   - More general missing-data mechanisms

Open question 1: Finite-sample setting

Context:

- **Finite-sample setting**: n is fixed
- Minimizing the **empirical risk**:

$$\beta_\star^n = \arg\min_{\beta \in \mathbb{R}^d} \left\{ R_n(\beta) := \frac{1}{n} \sum_{i=1}^{n} f_i(\beta) \right\}$$

Point to discuss:

- True risk in the finite sample setting?
- Unbiased gradients for the empirical risk?
- Interest of the empirical risk in presence of missing data?

Open question 1: Finite-sample setting

Remark 1: convergence rate for the true risk when n is fixed

- ✓ Same convergence rate holds.
- ✓ But only for **one epoch** (= use only once each data).
  Otherwise: $D_k \not\perp\!\!\!\perp \beta_{k-1} \rightarrow$ bias in the gradient.

Open question 1: Finite-sample setting

Remark 2: no unbiased gradients for the empirical risk?

How to choose the k-th obsservation ?

- ✗ k uniformly at random $\Rightarrow$ we use a data several times.
- ✗ k not chosen uniformly at random $\Rightarrow$ sampling not uniform and bias in the gradient.

Open question 1: Finite-sample setting

Remark 2: no unbiased gradients for the empirical risk?

How to choose the k-th obsertvation ?

- ✗ k uniformly at random $\Rightarrow$ we use a data several times.
- ✗ k not chosen uniformly at random $\Rightarrow$ sampling not uniform and bias in the gradient.

Idea? Interesting paper on without-replacement sampling[a]

- Draw a new permutation on $\{1, \ldots, n\}$ uniformly at random and process the individual in that order.
- Results on the convergence rate preserved for SGD.
- To be adapted for the averaged SGD with NA.

[a] Shamir, "Without-replacement sampling for stochastic gradient methods".

Open question 1: Finite-sample setting

Remark 3: is the empirical risk natural with missing values?

- Without NA: interest of the empirical risk is known exactly $\Rightarrow$ we can minimize it with precision.
- Without NA: empirical risk **not observed**.

Open question 2: other loss function

✓ Gradient debiased for least-squares loss.

Open question: What for the logistic loss?

$$f_i(\beta) = \frac{1}{n} \sum_i \log(1 + \exp(-y_i X_i^T \beta)), y_i \in \{1, -1\}$$

Gradient: $\nabla f_i(\beta) = \frac{-y_i X_i}{1+\exp(y_i X_i^T \beta)}$

- Debiasing the gradient?
- Deriving the theoretical results?

Open question 2: other loss function

- **Gradient:** $\nabla f_k(\beta) = \frac{-y_k X_k}{1+\exp(y_k X_k^T \beta)}$.
- **Iteration** k: $\beta_k = \beta_{k-1} - \alpha \tilde{g}_k(\beta)$.

PB for debiasing: compute $\mathbb{E}\left[\frac{u}{1+\exp(u)}\right]$, when u is Gaussian.

Ideas:

✓ Partially debiasing (only the numerator),

✓ Approximating the gradient and debiasing this approximation.

→ Only debiasing the numerator: $\tilde{g}_k(\beta) = \frac{-y_k X_k}{p(1+\exp(y_k X_k^T \beta))}$

→ Debiasing the approximation of the gradient:

$$\frac{-y_k X_k}{1+\exp(y_k X_k^T \beta)} \approx \frac{-y_k X_k}{2} + \frac{X_k^T \beta X_k}{4}$$

Open question 2: other loss function

- **Hessian:** $H_k(\beta) = \frac{\exp(y_k X_k^T \beta)}{1+\exp(y_k X_k^T \beta)} X_k^T X_k$.
- **Iteration k of Bach and Moulines:**
  $\beta_k = \beta_{k-1} - \alpha(\nabla f_k(\bar{\beta}_{k-1}) + H_k(\bar{\beta}_{k-1})(\beta_{k-1} - \bar{\beta}_{k-1})$

$\rightarrow$ Use the partially debiasing of the gradient and debiasing the quadratic part $X_k^T X_k$ of the Hessian.

$\rightarrow$ Debiasing the approximation of the gradient and debiasing the upper-bound of the Hessian $H_k(\beta) \leq \frac{1}{4} X_k^T X_k$

Open question 3: other missing-data mechanisms

✓ Heterogeneous MCAR data:

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

Open question: What for the MAR or MNAR data? We can not debiase the gradient using

$$\mathbb{E}_{D_k, X_k, y_k} \neq \mathbb{E}_{D_k} \mathbb{E}_{X_k, y_k}$$

- MCAR: $D_k \perp\!\!\!\perp X_k, y_k$
- MAR: $D_k \perp\!\!\!\perp X_k^{\text{mis}} | X_k^{\text{obs}}, y_k$
- MNAR: other cases.

Open question 3: other missing-data mechanisms

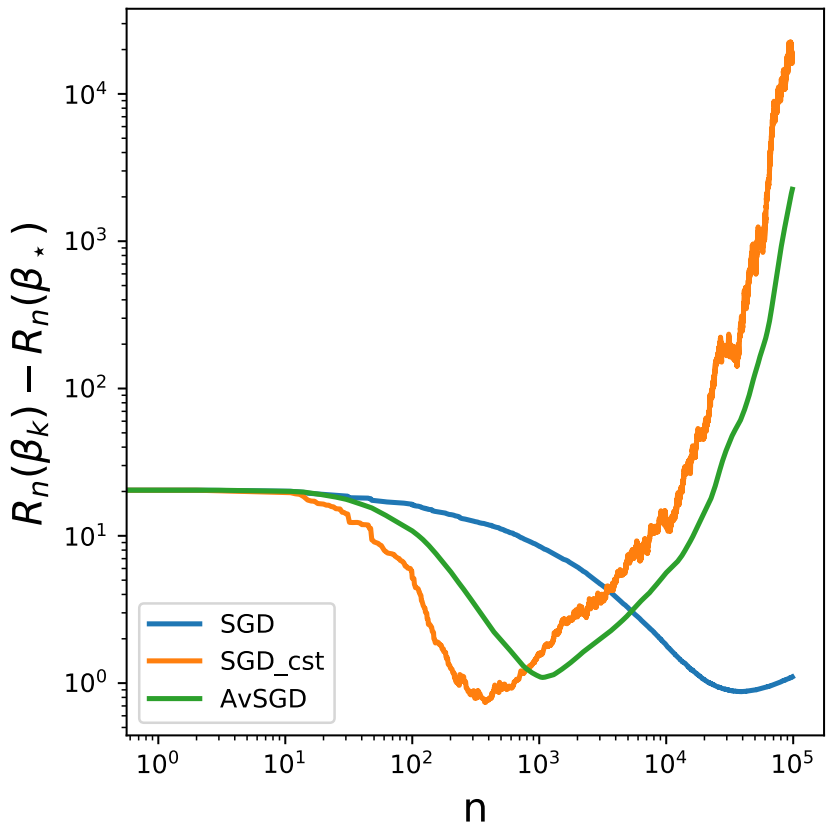


Figure: 1 pass, assuming MAR data

Thanks !

Missing values setting

Mechanism assumption

- **Why** data is missing?
- Missing-data mechanism[7]: $\mathcal{L}(D|X)$
- Example: Income & Age, with missing values on Income.
- **MCAR**: the missing-data pattern is independent of the data.
- **MAR**: the missing-data pattern depends on the observed values.
- **MNAR**: the missing-data pattern depends on the missing values (and potentially on the observed values too).

[7] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*. Vol. 793. John Wiley & Sons, 2019.

Algorithm true for $\alpha \leq \frac{1}{2L} \rightarrow$ we can overestimate L (but not underestimate)
We take $\alpha = \frac{1}{2L}$, where L is chosen:

- **oracle value**: we proved $L \leq \frac{1}{p_m^2} \max_k \|X_{k:}\|^2$ a.s.
- **estimated value**: $\hat{L}_n^{\mathrm{NA}} = \frac{1}{\hat{p}_m^2} \max_{1 \leq k \leq n} \frac{\|\tilde{X}_{k:}\|^2 d}{\sum_j D_{kj}}$, with $\hat{p}_m = \min_{1 \leq j \leq d} \hat{p}_j$, and $\hat{p}_j = \frac{\sum_k D_{kj}}{n}$. ($\|\tilde{X}_{k:}\|^2$ divided by the proportion of the NA in the row).

$d = 2$. Accounting for the effects of X_{k1}^2, X_{k2}^2, $X_{k1}X_{k2}$.

- augmented design matrix: $(X_{:1}|X_{:2}|X_{:1}X_{:2}|X_{:1}^2|X_{:2}^2)^T$.
- Debiased gradient: $U^{\odot -1} \odot \tilde{X}_{k:}\tilde{X}_{k:}^T\beta_k - \mathrm{diag}(U)^{\odot -1} \odot \tilde{X}_{k:}y_k$

$$U = \begin{pmatrix} p_1 & p_1p_2 & p_1p_2 & p_1 & p_1p_2 \\ p_1p_2 & p_2 & p_1p_2 & p_1p_2 & p_2 \\ p_1p_2 & p_1p_2 & p_1p_2 & p_1p_2 & p_1p_2 \\ p_1 & p_1p_2 & p_1p_2 & p_1 & p_1p_2 \\ p_1p_2 & p_2 & p_1p_2 & p_1p_2 & p_2 \end{pmatrix},$$

$U^{\odot -1}$: formed of the inverse coefficients of U.

$d = 2$. Accounting for the effects of X_{k1}^2, X_{k2}^2, $X_{k1}X_{k2}$.

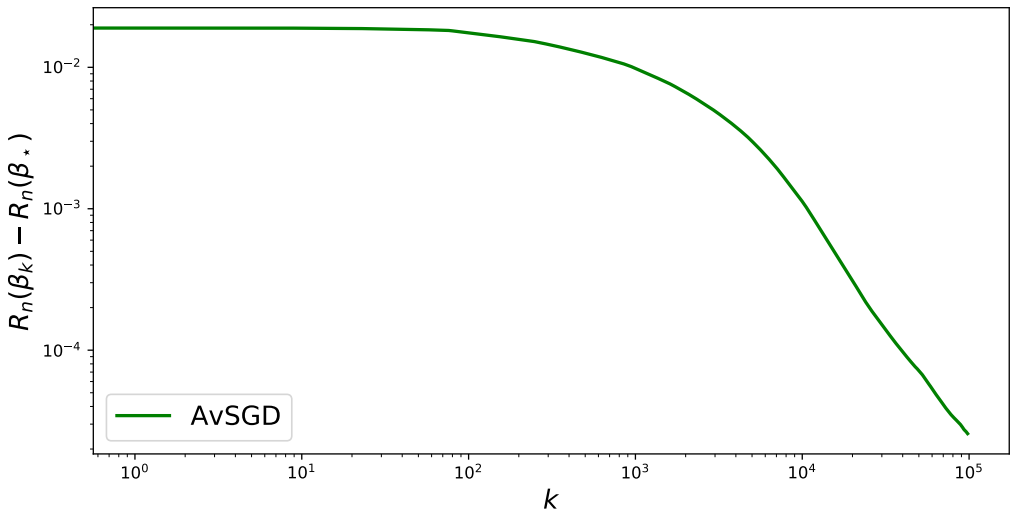


Figure: Empirical excess risk $(R_n(\beta_k) - R_n(\beta^\star))$ given n for synthetic data ($n = 10^5$, $d = 10$) when the model accounts mixed effects.

Algorithm

Polynomial features

For real data (Superconductivity dataset) 3 algorithms to compare :

- the averaged SGD on complete data (blue)
- the proposed debiased averaged SGD (orange)
- the averaged SGD run on imputed-by-0 data without any debiasing (green)

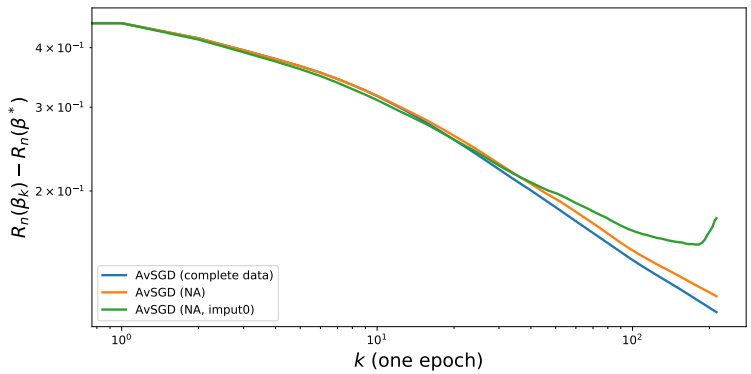


Figure: Empirical excess risk $(R_n(\beta_k) - R_n(\beta^\star))$ given n for the superconductivity dataset ($n = 21263$) (containing 81 initial features) and $d = 3403$ with polynomial features of degree 2.

Experiments

Synthetic data: homogeneous vs heterogeneous

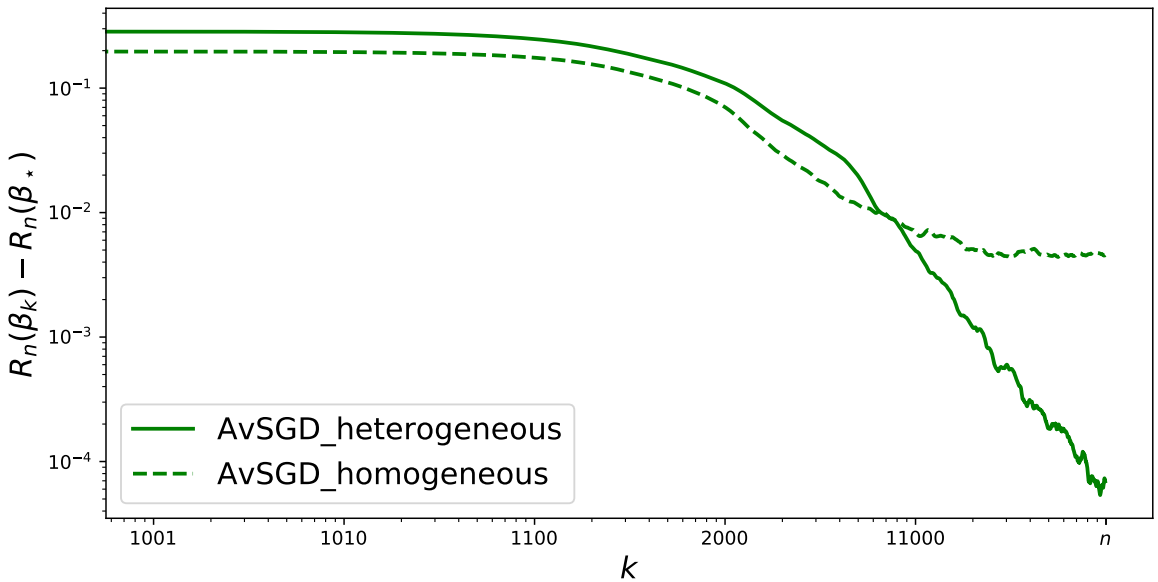


Figure: Empirical excess risk $R_n(\beta_k) - R_n(\beta^\star)$, $n = 10^5$.

- Missing values introduced with different missingness probabilities.
- Taking into account the heterogeneity in the algorithm (plain line): good rate of convergence for **AvSGD**.
- Ignoring the heterogeneity (dashed line): stagnation far from the optimum in termes of empirical risk.

Experiments

Real dataset: Traumabase, model estimation

- Goal: model the level of platelet upon arrival at the hospital from the clinical data of 15785 patients.
- Explanatory variables selected by doctors: seven quantitative (missing) variables.
- Model estimation: do the effect of the variables on the platelet make sense ?
- Similar results than EM algorithm but effects of HR and Δ.Hemo are not in agreement with the doctors opinion.

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

Experiments

Real dataset, Superconductivity, prediction task

Comparison to **two-step heuristics** (no theoretical guarantees):

- the covariates imputed
  - mean (naive)
  - IterativeImputer (estimates each feature from all the others)
- linear regression (LR) performed on the competed dataset

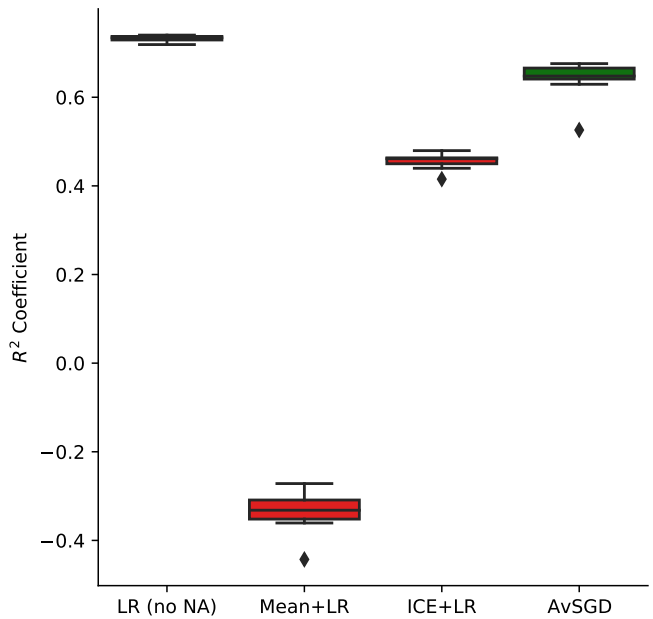


Figure: R^2 coefficients, 60% MCAR values.

Missing data setting

Missing-data patterns dependent

In our setting: independent missing-data patterns

$$D_{.j} \perp\!\!\!\perp D_{.j'}, j \neq j'$$

Dependent missing-data patterns

$$\tilde{g}_k(\beta) := (W \odot (\tilde{X}_{k:}\tilde{X}_{k:}^T))\beta - y_k P^{-1}\tilde{X}_{k:}$$

with $W \in \mathbb{R}^{d \times d}$, and $W_{ij} := 1/\mathbb{E}[\delta_{ki}\delta_{kj}]$ for $1 \leq i, j \leq d$

Key assumption for fast rate?

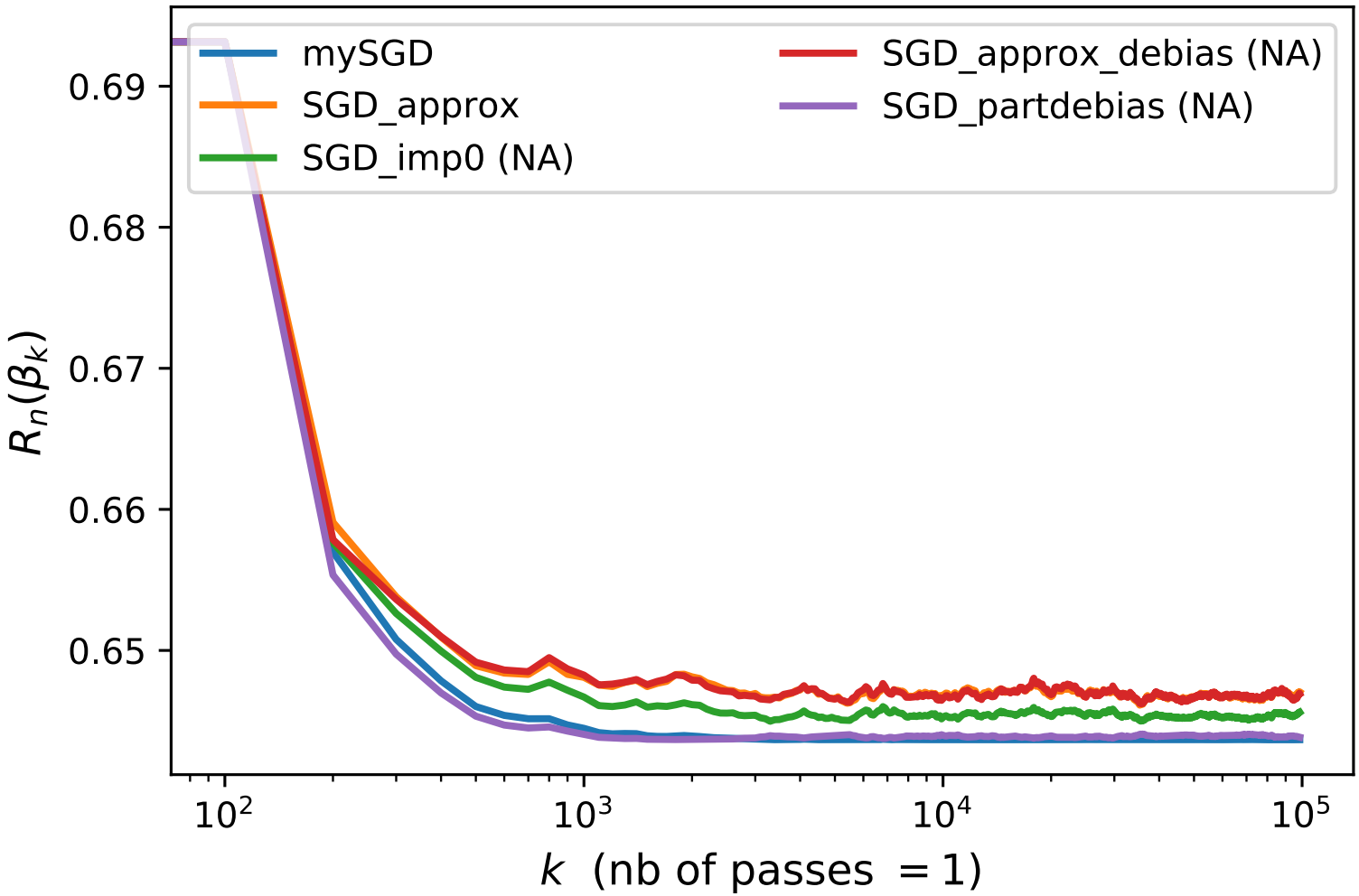
- $\nabla R(\beta_{k-1}) = \mathbb{E}_{(X_{k:},y_k)}[X_{k:}(X_{k:}^T\beta_{k-1} - y_k)] = H(\beta_{k-1} - \beta^\star)$.
- We do: $\beta_k = \beta_{k-1} - \alpha\tilde{g}_k(\beta_{k-1})$.
- **What is the noise induced by using the unbiased stochastic gradient $\tilde{g}_k$?**

$$\nabla R(\beta_{k-1}) - \tilde{g}_k(\beta_{k-1}) = \underbrace{(H - X_{k:}X_{k:}^T)(\beta_{k-1} - \beta^\star)}_{\text{multiplicative noise}} - \underbrace{X_{k:}\epsilon_k}_{\text{additive noise}} =: \zeta_k$$

- Assumption on the additive noise?

$$\beta_k = \beta_{k-1} - \alpha H(\beta_k - \beta^\star) + \alpha\zeta_k$$
$$H(\beta_k - \beta^\star) = \beta_{k-1} - \beta_k + \alpha\zeta_k$$
$$(\bar{\beta}_k - \beta^\star) = H^{-1}\frac{(\beta_0 - \beta^\star)}{\alpha n} + H^{-1}\sum_{k=1}^{n}\frac{\zeta_k}{n}$$

Open question 2: other loss function



Open question 2: other loss function