Debiasing Stochastic Gradient Descent to handle missing values

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

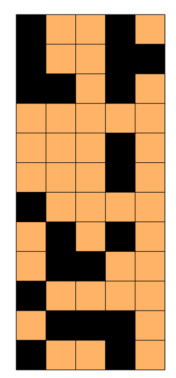
[1]Sorbonne Université [2]Ecole Polytechnique [3]INRIA

Context

- **Large-scale** data analysis: large number of observations n, large dimension of the observations p.
- **Missing data** occurs more frequently. E.g. in clinical data: failure of the measuring device, no time to measure in emergency situations, aggregating datasets from multiple hospitals, etc.
- **Stochastic gradient descent** (SGD): key role in machine learning.

Missing data

- Problem: Missing values in the covariates $X_{k:}$.
- $D_{k:} \in \{0,1\}^d$, $D_{kj} = \begin{cases} 0 \text{ if the var. } j \text{ of obs. } k \text{ missing} \\ 1 \text{ otherwise.} \end{cases}$
- **Heterogeneous MCAR** data: $\neq$ missing proba. for each covariate. $D = (\delta_{kj})$ with $\delta_{kj} \sim \mathcal{B}(p_j)$.



- Access to $X_{k:}^{\text{NA}} \in (\mathbb{R} \cup \{\texttt{NA}\})^d$ instead of $X_{k:}$

$$X_{k:}^{\text{NA}} := X_{k:} \odot D_{k:} + \texttt{NA} \odot (\mathbf{1}_d - D_{k:}),$$

Setting: linear regression with missing covariates

$(X_{k:}^{\text{NA}}, y_k)) \in \mathbb{R}^d \times \mathbb{R}$ i.i.d. observations.

$$y_k = (X_{k:}^{\text{NA}})^T \beta^\star + \epsilon_k,$$

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

How to perform linear regression with missing covariates that handle large-scale or streaming data?

Existing works in linear models

- Expectation Maximization algorithm [1].
  ✗ parametric (Gaussian) assumption for the covariates.
- Naive imputation e.g. by the mean [2].
  ✗ Bias in the estimate.
- **Imputing naively by** 0 and modifying SGD to **account for the imputation error** ([3]), also in [4], for homogeneous MCAR values.

⇒ **Our proposal**: debiased **averaged** SGD, better rate of convergence.

Methodology

- Imputing the missing values by 0: $\tilde{X}_{k:} = X_{k:}^{\text{NA}} \odot D_{k:} = X_{k:} \odot D_{k:}$.
- Using a **debiased gradient** for the **averaged SGD**:

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

where $P = \mathrm{diag}\left((p_j)_{j\in\{1,\dots,d\}}\right) \in \mathbb{R}^{d\times d}$.

- **Averaged iterates:** $\bar{\beta}_k = \frac{1}{k+1}\Sigma\beta_i$.

Theoretical results

- Goal: **establish a convergence rate.**
- Assumptions: $(X_{k:}, y_k) \in \mathbb{R}^d \times \mathbb{R}$ 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, $\mathcal{F}_k = \sigma(X_{1:}, y_1, D_{1:} \dots, X_{k:}, y_k, D_{k:})$.

Lemma 2: structured noise induced by NA

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

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

Lemma 3: $(\tilde{g}_k(\beta^\star))_{k\geq 0}$ a.s. co-coercive

- *For any $k \geq 0$, $\tilde{g}_k$ is $L_{k,D}$-Lipschitz.*
- *There exists a primitive function $\tilde{f}_k$ which is a.s. convex.*

Theorem 1: convergence rate, online streaming

*Assume that for any k, $\|X_{k:}\| \leq \gamma$ a.s. for some $\gamma > 0$. For **any constant step-size** $\alpha \leq \frac{1}{2L}$ and for any $k \geq 0$, one has:*

$$\mathbb{E}\left[R\left(\bar{\beta}_k\right) - R(\beta^\star)\right] \leq \frac{1}{2k} \cdot \left(\underbrace{\frac{\sqrt{c(\beta^\star)d}}{1-\sqrt{\alpha L}}}_{\text{variance term}} + \underbrace{\frac{\|\beta_0 - \beta^\star\|}{\sqrt{\alpha}}}_{\text{bias term}} \right)^2,$$

- $L := \sup_{k,D}$ *Lipschitz constants of* $\tilde{g}_k$.
- $p_m = \min_{j=1,\dots,d}\ p_j$ *minimal probability to be observed.*
- $$c(\beta^\star) = \underbrace{\overbrace{\frac{\mathrm{Var}(\epsilon_k)}{p_m^2}}^{\text{classical term}} + \overbrace{\left(\frac{(2+5p_m)(1-p_m)}{p_m^3}\right)\gamma^2\|\beta^\star\|^2}^{\text{multiplicative noise (induced by naive imputation)}}}_{\text{increasing with the missing values rate}}.$$

✓ **Optimal rate for least-squares regression: $O(k^{-1})$.**

✓ In the complete case ($p_m = 1$): same bound as Bach and Moulines [5].

Additional results

- **Finite-sample setting:** n is fixed.
  - **True risk**: same convergence rate holds for **only one epoch**.
    ✗ if we use the data more than once: bias in the gradient.
  - **Empirical risk (open issue)**: $\beta_\star^n = \arg\min_{\beta\in\mathbb{R}^d}\{R_n(\beta) := \frac{1}{n}\Sigma f_i(\beta)\}$.
    ✗ data used several times or non-uniform sampling.
- **Using estimated missing probabilities** $(\hat{p}_j)_j$ in our algorithm instead of $(p_j)_j$ preserves the same order of convergence rate $\mathcal{O}(k^{-1})$.
- **Ridge Regression:** $\beta \to R(\beta) + \lambda\|\beta\|^2$ is 2λ-strongly convex: no change for the debiased gradient, convergence rate of $\mathcal{O}((\lambda k)^{-1})$.

Consequences in practice

- *Before collecting data*, **fewer complete obs. is better than more incomplete ones**, e.g. variance bound for 200 incomplete obs. (50% NA) is twice as large as for 100 complete obs.
- *After collecting data with NA*, **obs. containing NA should not be removed**: the upper-bound is p^{d-3} smaller than the lower bound of any algorithm relying only on the complete observations.

Experiments on synthetic data

- $X_{i:} \overset{i.i.d.}{\sim} \mathcal{N}(0,\Sigma)$, where Σ generated randomly with decreasing eigenvalues, $y_i = X_{i:}\beta + \epsilon_i$, for β fixed and $\epsilon_i \sim \mathcal{N}(0,1)$.
- $d = 10$, 30% missing values, L oracle value.
- **AvSGD**: our method, constant step size $\alpha = \frac{1}{2L}$.
- **SGD**: [3], decreasing step size $\alpha_k = \frac{1}{\sqrt{k+1}}$.
- **SGD_cst** [3], constant step size $\alpha = \frac{1}{2L}$.

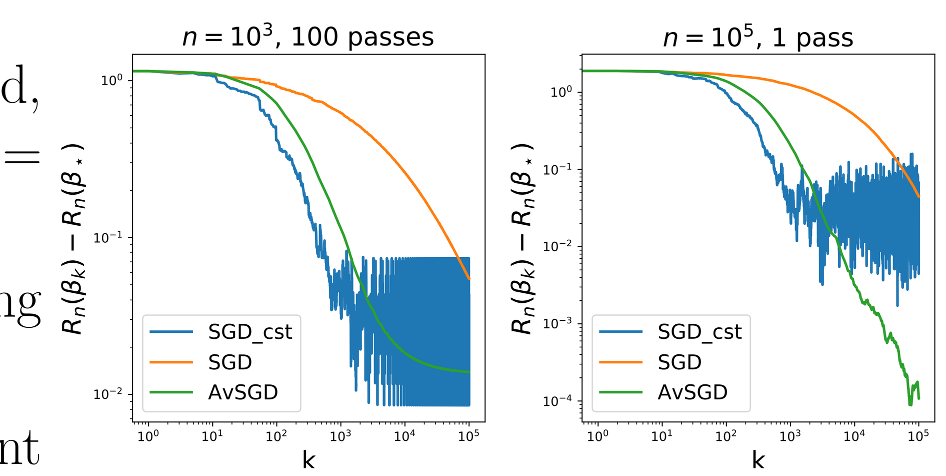


Figure 1: Empirical excess risk ($R_n(\beta_k) - R_n(\beta^\star)$).

Experiments on real data

- **Complete** dataset: 81 quant. features, 21263 superconductors.
- Introduction of 30% of heterogeneous MCAR values.
- Training/test split, with no NA in the test set.
- $\hat{y}_{n+1} = X_{n+1}^T\hat{\beta}$, with $\hat{\beta}$ computed on the training set, with **AvSGD** or with a two-steps procedure **Mean+AvSGD**.

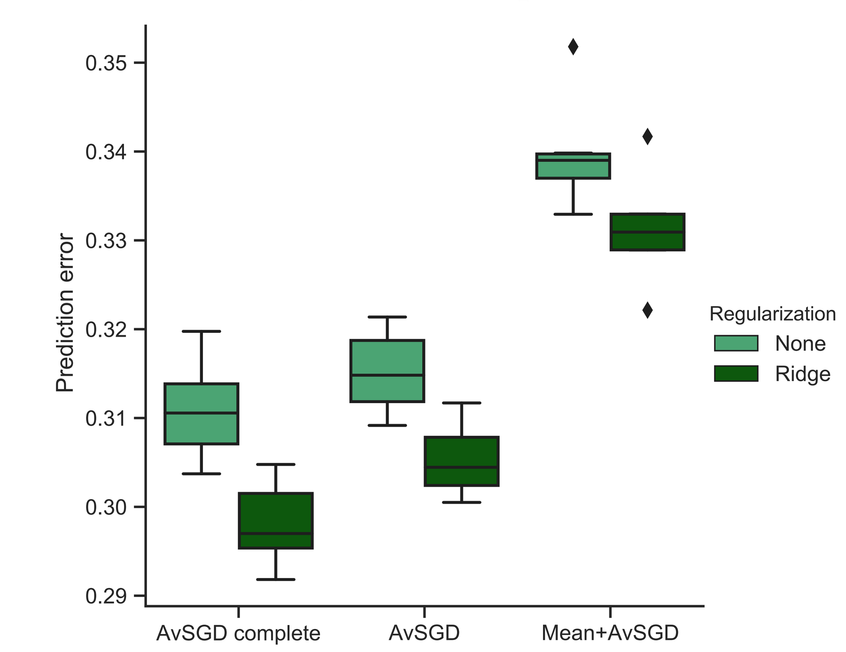


Figure 2: Prediction error $\|\hat{y} - y\|^2/\|y\|^2$.

⇒ **Further research**: MAR values, GLM models.

References

[1] A. P Dempster, N. M Laird, and D. B Rubin. Maximum likelihood from incomplete data via the em algorithm.

[2] R. JA Little and D. B Rubin. *Statistical analysis with missing data.*

[3] Anna Ma and Deanna Needell. Stochastic gradient descent for linear systems with missing data.

[4] Po-Ling Loh and Martin J Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity.

[5] F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate o (1/n).