Debiasing Averaged Stochastic Gradient Descent to handle missing values

Séminaire MIA, AgroParisTech

Aude Sportisse [1] Claire Boyer [1,2] Aymeric Dieuleveut [3]
Julie Josse [4,5]

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

[2]Département de Mathématiques et applications, Ecole Normale Supérieure

[3]Centre de Mathématiques Appliquées, Ecole Polytechnique

[4]INRIA

[5]Visiting Researcher Google Brain

5th October 2020

Large-scale and incomplete data

- **Large-scaling:** large n (number of observations), large p (dimension of the observations)
- **Incompleteness** for many reasons: "forgot to fill in the form", failure of the measuring device, no time to measure in an emergency situation, aggregating data sets from multiple hospitals,...

Traumabase: 15 000 patients/ 250 var/ 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.

Setting

- $(X_{i:}, y_i)_{i \geqslant 1} \in \mathbb{R}^d \times \mathbb{R}$ i.i.d. observations
- **Linear regression model**

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

  parametrized by $\beta^\star \in \mathbb{R}^d$, with a noise term $\epsilon_i \in \mathbb{R}$.
- **Problem:** $(X_{i:})$**'s partially known** (missing values in the covariates).
- How to estimate $\beta^\star$?

Optimization problem

- For $y_i = X_{i:}^T \beta^\star + \epsilon_i$, loss function: $f_i(\beta) = \left(\langle X_{i:}, \beta \rangle - y_i\right)^2 / 2$.
- **True risk minimization**:

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

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

Optimization without missing values

Gradient descent

- **Deterministic case**: $F : \mathbb{R}^d \to \mathbb{R}$, we consider $\min_{\beta \in \mathbb{R}^d} F(\beta)$.
- **Gradient descent (GD)**: the current iterate moves in the opposite direction of the gradient.

$$\beta_k = \beta_{k-1} - \alpha \nabla F(\beta_{k-1}),$$

with α the step size.

✓ Convergence rate: $\mathcal{O}(k^{-1})$[1] if F is convex and L-smooth, i.e. F is twice differentiable and

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

✗ costly: "full" gradient computed at each iteration.

[1] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*. Vol. 87. Springer Science & Business Media, 2013.

Optimization without missing values

Stochastic gradient descent

- **Stochastic gradient descent (SGD)**: using unbiased estimates of $\nabla F(\beta_{k-1})$.

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

where α is the step-size and $\mathbb{E}\left[g_k(\beta_{k-1})|\mathcal{F}_{k-1}\right] = \nabla F(\beta_{k-1})$, $\mathcal{F}_{k-1} = \sigma(X_{1:}, y_1, \ldots, X_{k-1:}, y_{k-1})$ the filtration.

✓ It scales with large data.

✗ Convergence rate: $\mathcal{O}(k^{-1/2})$[2] if F is convex and L-smooth.

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

Optimization without missing values

Averaged stochastic gradient descent

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

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

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

✓ It scales with large data.

✓ Convergence rate: $\mathcal{O}(k^{-1})$[3] if F is convex and L-smooth for least-squares regression.

[3] Francis Bach and Eric Moulines. "Non-strongly-convex smooth stochastic approximation with convergence rate O (1/n)". In: *Advances in neural information processing systems*. 2013, pp. 773–781.

Setting

- $(X_{i:}, y_i)_{i \geqslant 1} \in \mathbb{R}^d \times \mathbb{R}$ i.i.d. observations
- **Linear regression model:**

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

  parametrized by $\beta^\star \in \mathbb{R}^d$, with a noise term $\epsilon_i \in \mathbb{R}$.
- $(X_{i:})$'s partially known (missing values in the covariates).
- How to estimate $\beta^\star$?
- **How to derive stochastic algorithms for estimating $\beta^\star$?**

Missing values setting

Formalism

- $D_{i:} \in \{0,1\}^d$ binary mask, 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 $X_{i:}$

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

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

- **Semi-discrete nature**: mixed of continuous data (observed values) and categorical data (the missing values)
  $\Rightarrow$ **usual results can not be applied**.

Missing values setting

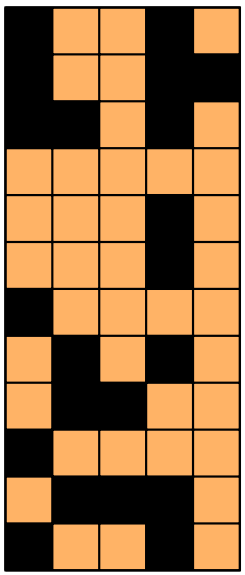
Mechanism assumption

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

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

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 work[7]

- Expectation Maximization algorithm[4] (maximization of the observed likelihood)
  ✗ parametric assumptions: Gaussian assumption for the covariates, no solution available for large dimension p.
- Matrix completion (predicting NA before applying usual algorithms)
  ✗ it can lead to bias and underestimation of the variance of the estimate[5].
- **Imputing naively by** 0 and modifying the usual algorithms to **account for the imputation error**: in particular, a modified SGD[6].

---

[4] 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.

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

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

[7] Imke Mayer et al. "R-miss-tastic: a unified platform for missing values methods and workflows". In: *arXiv preprint arXiv:1908.04822* (2019).

Dealing with missing values

Our strategy inspired by Ma et Needell

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

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

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

- 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 $(X_{i:}^{\mathrm{NA}}, y_i)$

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

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

- 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 $(X_{i:}^{\mathrm{NA}}, y_i)$

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

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

- 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 $(X_{i:}^{\mathrm{NA}}, y_i)$

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

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

- 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): $\min_{\beta\in\mathbb{R}^d} R(\beta) + \lambda\|\beta\|^2, \lambda > 0$

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 $\forall k \geqslant 0$,

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

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

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

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

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).

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

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 ?

- without missing observations: variance bound scales as $O\left(\frac{\mathrm{Var}(\epsilon_k)d}{k}\right)$.
- 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.

Theoretical results

What impact of missing values ?

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, 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.

Finite-sample setting: n is fixed

- **True risk**: same convergence rate holds for **only one epoch** (we can use only once each data).
  Otherwise: mask at step k independent from the previous constructed iterate $\Rightarrow$ bias in the gradient.

Finite-sample setting: n is fixed

- **True risk**: same convergence rate holds for **only one epoch** (we can use only once each data).
  Otherwise: mask at step k independent from the previous constructed iterate $\Rightarrow$ bias in the gradient.
- **Empirical risk**: $\beta_\star^n = \arg\min_{\beta \in \mathbb{R}^d} \left\{ R_n(\beta) := \frac{1}{n} \sum_{i=1}^n f_i(\beta) \right\}$
  How to choose the k-th obstervation ?
  - ✗ 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.

Finite-sample setting: n is fixed

- **True risk**: same convergence rate holds for **only one epoch** (we can use only once each data).
  Otherwise: mask at step k independent from the previous constructed iterate $\Rightarrow$ bias in the gradient.
- **Empirical risk**: $\beta_\star^n = \arg\min_{\beta \in \mathbb{R}^d} \left\{ R_n(\beta) := \frac{1}{n} \sum_{i=1}^n f_i(\beta) \right\}$
  How to choose the k-th obstervation ?
  - ✗ 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.

Implications:

- No unbiased gradients for the empirical risk so far.
- Keep in mind: empirical risk is in any case not observed.

Theoretical results

Comparison with related work

Comparison with Ma et Needell[8]:

- μ-strongly convex problem
- no averaged iterates

$\Rightarrow$ convergence rate of $\mathcal{O}(\frac{\log n}{\mu n})$.

✗ μ generally out of reach.

✗ only homogeneous MCAR data.

✗ main theorem mathematically invalid (empirical risk).

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

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

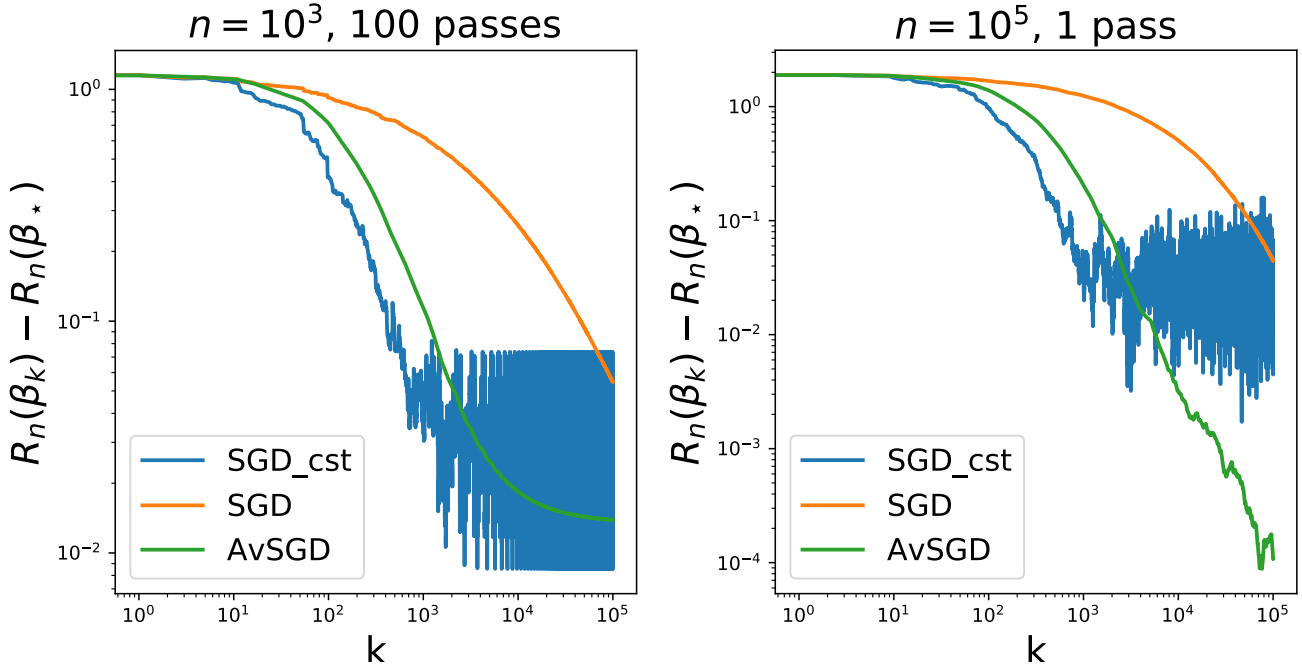


Figure: Empirical excess risk $(R_n(\beta_k) - R_n(\beta^\star))$.

- 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

Synthetic data: homogeneous vs heterogeneous

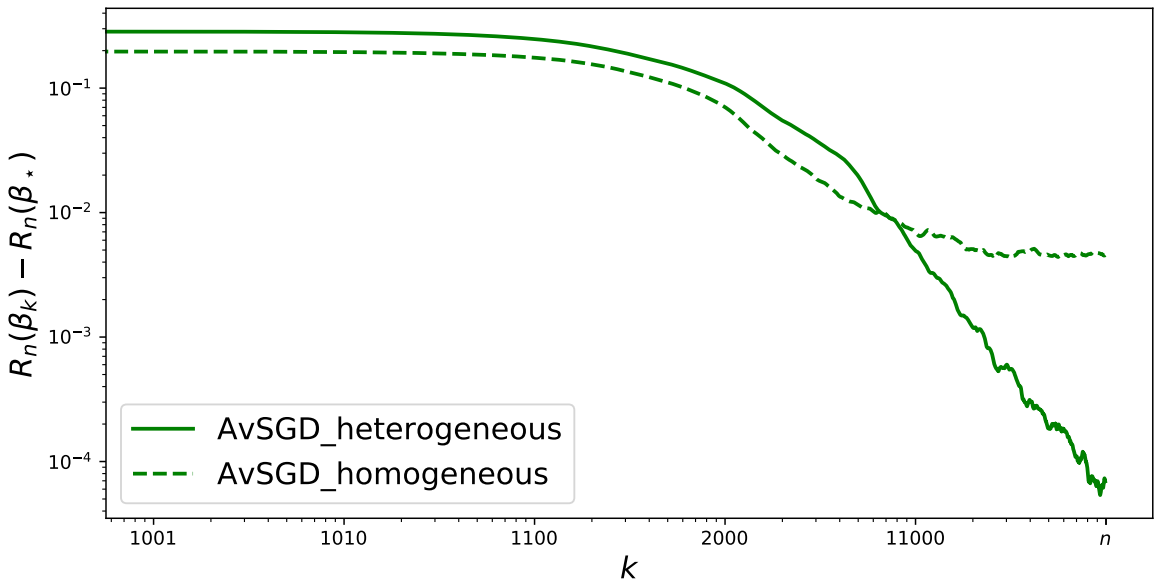


Figure: Empirical excess risk $R_n(\beta_k) - R_n(\beta^\star)$, $n = 10^5$.

- 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

- 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

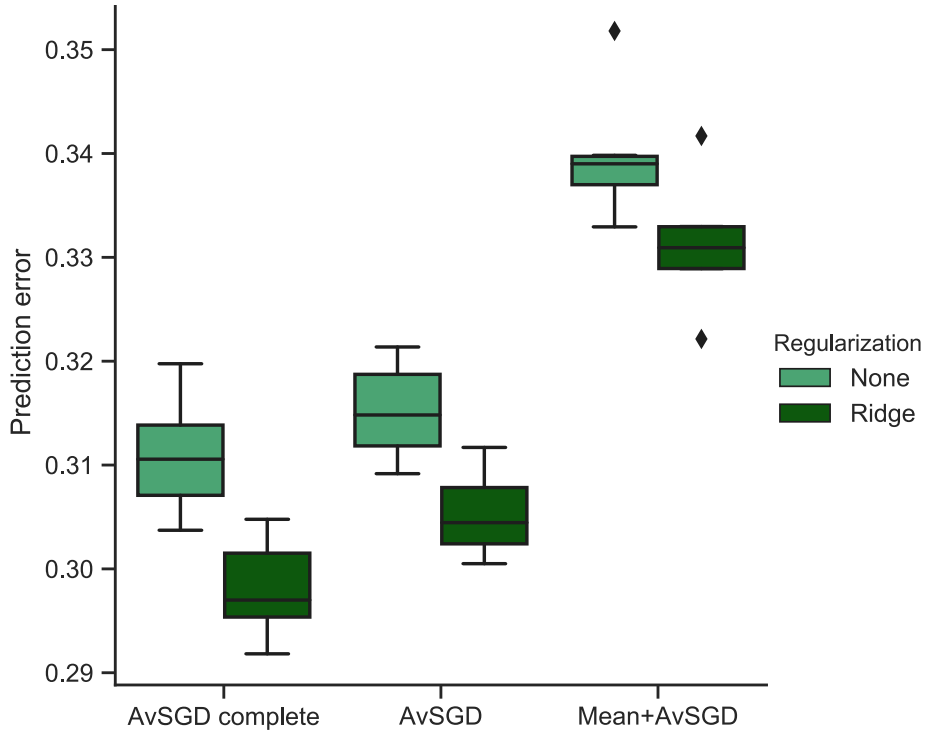


Figure: Prediction error $\|\hat{y} - y\|^2 / \|y\|^2$ boxplots.

- 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.

Conclusion

- ✓ Imputing by 0 and debiasing the gradient lead to tight and rigorous convergence guarantees for the true risk of averaged SGD.
- ✓ Python implementation of regularized regression with missing values for large scale data.
- ✓ A paper.[9]

Perspectives:

- Dealing with more general loss function.
- More complex missing-data patterns such as MAR and MNAR.

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