Debiasing Averaged Stochastic Gradient Descent to handle missing values

Séminaire Parisien de Statistiques

Aymeric Dieuleveut[3]

with Aude Sportisse [1] Claire Boyer [1,2] 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

8th February 2021

Motivation: Large-scale incomplete data

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

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

NA: Not Available.

Outline

1. SGD with missing data
2. Convergence results
   - Without missing values: rates and proofs
   - Convergence of Algorithm 1
   - Rates for empirical risk? Beyond one pass?
   - Adaptation to estimated missing probabilities
3. Experiments

features

output

Setting

- $(X_{i:}, y_i)_{i \geq 1} \in \mathbb{R}^d \times \mathbb{R}$ 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 $\epsilon_i \in \mathbb{R}$.
- Loss function: $f_i(\beta) = \left(\langle X_{i:}, \beta \rangle - y_i\right)^2 / 2$.
- **True risk minimization**:

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

- Stochastic gradient method.
  - At the heart of Machine Learning.
  - Very well suited for large d and n.

Objective - missing data

- **Problem: $(X_{i:})$'s partially known**
  1. What should we estimate?

- **True risk minimization**:

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

  2. How to adapt algorithms to the missing data case?

Optimization without missing values

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

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

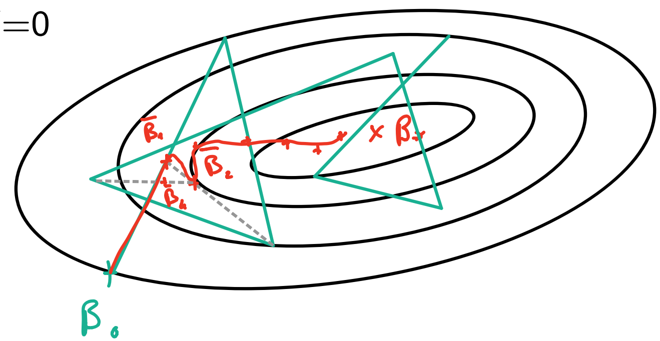
where α is the step-size and $\mathbb{E}\left[g_k(\beta_{k-1}) | \mathcal{F}_{k-1}\right] = \nabla F(\beta_{k-1})$, $\mathcal{F}_{k-1} = \sigma(X_{1:}, y_1, \ldots, X_{k-1:}, y_{k-1})$ the filtration.

- **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.



Optimization without missing values

- **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.

- **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.

2 questions

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

for observation i

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

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

$j = 1 \dots d.$

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

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

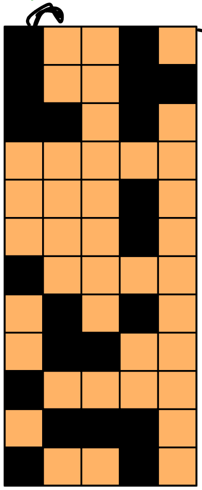
$\odot$ element-wise product, $1_d = (1 \dots 1)^T \in \mathbb{R}^d$, $\mathtt{NA} \times 0 = 0$, $\mathtt{NA} \times 1 = \mathtt{NA}$.

- ***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 work[4]

- Expectation Maximization algorithm[1] (maximization of the observed likelihood)
  ✗ parametric assumptions: Gaussian assumption for the covariates, no solution available for large dimension $\cancel{p}$. d.
- Matrix completion (predicting NA before applying usual algorithms)
  ✗ it can lead to bias and underestimation of the variance of the estimate[2].
- **Imputing naively by** 0 and modifying the usual algorithms to **account for the imputation error**: in particular, a modified SGD[3].

---

[1] 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.

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

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

[4] 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 $(X_{i:}^{\mathrm{NA}}, y_i)$

$$\begin{pmatrix} 1 & 2 & NA \\ 2 & NA & 5 \end{pmatrix}$$

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

$$\tilde{X}_{i:} = X_{i:}^{\mathrm{NA}} \odot D_{i:} = X_{i:} \odot D_{i:} \text{ imputed covariates} \longrightarrow \begin{pmatrix} 1 & 2 & 0 \\ 2 & 0 & 5 \end{pmatrix}$$

$\tilde{X}_1$ $\tilde{X}_2$

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

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

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

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

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

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

$\mathbb{E}\,\mathcal{B}(p) = p$!

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

One obtains

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

$$\nabla F(\beta) = (x^T\beta - y)\, x$$

quadratic in X

$$\mathbb{E}\left[y P^{-1} \tilde{X}\right] = y X$$

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 β_* with missing data.
- ✓ We consider a **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.

Outline

Optimization **without** missing values

convergence rates and proof techniques

If F is convex and L-smooth.[5]

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

If F is convex and L-smooth, μ-strongly convex.

✗ Convergence rate: $\mathcal{O}((\mu k)^{-1})$, with μ known.

If F is convex and quadratic, e.g., for least-squares regression[6].

✓ Convergence rate: $\mathcal{O}(k^{-1})$

- ? Why do we get a faster rate for quadratic functions?
- ? What does it require?

---

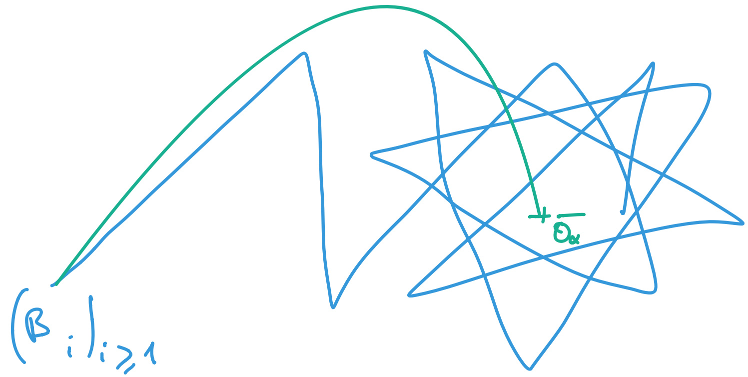
[5] Arkadi Nemirovski et al. "Robust stochastic approximation approach to stochastic programming". In: *SIAM Journal on optimization* 19.4 (2009), pp. 1574–1609.

[6] 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.

Faster rates for Least Squares regression

- Typical proof for convex:

$$\mathbb{E}\left[\|\beta_k - \beta_*\|^2 \,\middle|\, \beta_{k-1}\right] \leq \|\beta_{k-1} - \beta_*\|^2 - 2\gamma\langle \nabla F(\beta_{k-1}), \beta_{k-1} - \beta_* \rangle + \gamma^2 \|\underbrace{g_n(\beta_{k-1})}_{\nabla f \,+\, \varepsilon_k}\|^2$$



$$\bar{\beta}_n \longrightarrow \bar{\beta}_\alpha = \beta_*$$

true only for quadratic f.!

$$\mathbb{E}_{\pi_\gamma}(\nabla f'(\beta)) = 0$$

↳ I can lose a factor of N on the noise control!

Faster rates for Least Squares regression

- Typical proof for quadratic:

PJ.

$$\beta_k = \beta_{k-1} - \alpha \nabla F(\beta_{k-1}) + \alpha \varepsilon .$$

$$\alpha H(\beta_d - \beta_*) = \beta_{k-1} - \beta_k + \alpha \varepsilon$$

$$\left(\overline{\beta}_n - \beta_*\right) = H^{-1} \frac{\beta_0 - \beta_*}{\alpha n} + \frac{1}{n} \sum_{i=1}^{n} \underline{(H^{-1} \varepsilon_i)}$$

$$\mathbb{E} \| H^{-1} \varepsilon \|^2 = \operatorname{tr}\left(H^{-2} \overbrace{\varepsilon \varepsilon^T}^{\preccurlyeq H}\right) \text{ is bounded}$$

Summary

SGD

↓

Least squares ⟶ unbiased gradient oracle with NA

(*) this line

↓

Fast rate of cvgce

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.
- $\mathbb{E}[\tilde{g}_k(\beta^\star)\tilde{g}_k(\beta^\star)^T] \preccurlyeq C(\beta^\star) = c(\beta^\star)H$. $\longrightarrow$ structure

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

$\frac{\sigma^2}{n}$

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** $\alpha \leq \frac{1}{2L}$, ensures that, for any $k \geq 0$:

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

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

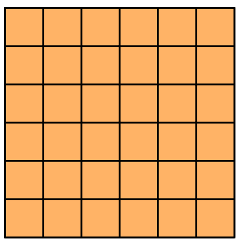
- Optimal rate for least-squares regression.
- In the complete case: same bound as Bach and Moulines.
- Bound on the iterates for the **ridge regression** ($\beta \to R(\beta) + \lambda\|\beta\|^2$ is 2λ-strongly convex).

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

$(X^T\beta - y)\,X$

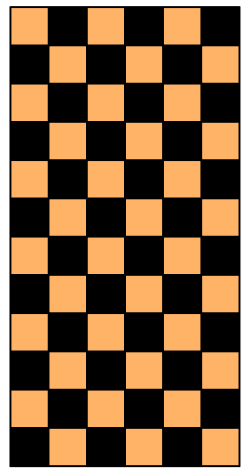
Theoretical results

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?



$k\,p$ obs. full.

$$\frac{\sigma^2}{k\,p}$$



k obs

p 50%

$$\frac{\sigma^2}{k\,p^2}$$

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?

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

Open Questions: Lower bound!

Possible Approach Gaussian assumptions on the data distribution: use the distribution of the full data knowing observed data.

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

$$
X = \begin{array}{c} \\ \\ \\ \\ \\ \\ \end{array}
\begin{array}{ccc} X_1 & X_2 & X_3 \\ \end{array}
$$

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

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 $\boxed{k_{co} \sim \mathcal{B}(k, p^d).}$
  - statistical lower bound: $\frac{\mathrm{Var}(\epsilon_k) d}{k_{co}}$.
  - in expectation, lower bound on the risk larger than $\frac{\mathrm{Var}(\epsilon_k) d}{k p^d}$. $\frac{1}{kp^d}$
- 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.

Outline

1. SGD with missing data
2. Convergence results
   - Without missing values: rates and proofs
   - Convergence of Algorithm 1
   - Rates for empirical risk? Beyond one pass?
   - Adaptation to estimated missing probabilities
3. Experiments

Open Question: rates for ERM?

- **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 observation?
  - ✗ 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.

[7] Ohad Shamir. "Without-Replacement Sampling for Stochastic Gradient Methods". In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. NIPS'16. Barcelona, Spain: Curran Associates Inc., 2016, pp. 46–54. ISBN: 9781510838819.

Open Question: rates for ERM?

- **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 observation?
  - ✗ 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.

Possible Approach: similar to wo replacement sampling for ERM.[7]

[7] Shamir, "Without-Replacement Sampling for Stochastic Gradient Methods".

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".

Outline

1. SGD with missing data
2. Convergence results
   - Without missing values: rates and proofs
   - Convergence of Algorithm 1
   - Rates for empirical risk? Beyond one pass?
   - Adaptation to estimated missing probabilities
3. Experiments

$\frac{1}{k p_m^2}$

$\boxed{p_m \lesssim \frac{1}{\sqrt{k}}}$ $\Longrightarrow$ vacuous

$p_m \gtrsim \frac{1}{\sqrt{k}}$

Theoretical results

$n p_m$

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 $k \geq 0$:

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

$\hat{p}_j \gtrsim \frac{p_j}{2}$

$e^{-n p_m}$

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

Proof Sketch

Open questions

OQ: Tighter convergence rate with estimated probabilities:

- Without strong convexity
- Better dependence w.r.t. p.

Approach: Proof related to *stability* approaches.

Open questions

OQ: Tighter convergence rate with estimated probabilities:

- Without strong convexity
- Better dependence w.r.t. p.

Approach: Proof related to *stability* approaches.

OQ: working in a distributed or federated framework

- Each participant has its own missing value probability
- Each participant has its own objective function.

Approach: Federated Learning algorithms. Estimation of probabilities based on a global prior + local estimation.

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 erasing incomplete data, but not as good as pk complete observations
- ✓ Convergence with strong-convexity and estimated probabilities (preserved k^{-1}, degraded dependence in p)

Partial answers & open questions:

- ✓ Matching lower bound?
- ✓ ERM, Beyond one pass? impossible to minimize ER to arbitrary precision, but a guarantee for the first pass seems possible.
- ✓ Better dependence in p for estimated probabilities case?
- ✓ Distributed & multi-agent frameworks are crucial.

? In practice?

Outline

1. SGD with missing data
2. Convergence results
   - Without missing values: rates and proofs
   - Convergence of Algorithm 1
   - Rates for empirical risk? Beyond one pass?
   - Adaptation to estimated missing probabilities
3. Experiments

Experiments

Synthetic data: convergence rate

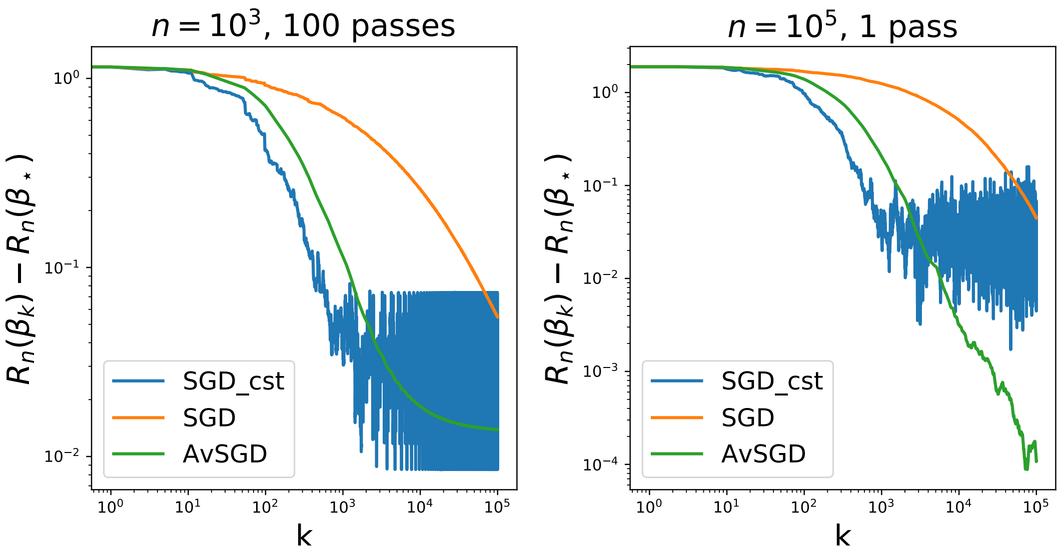


Figure: Empirical excess risk $(R_n(\beta_k) - R_n(\beta^\star))$.

- 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

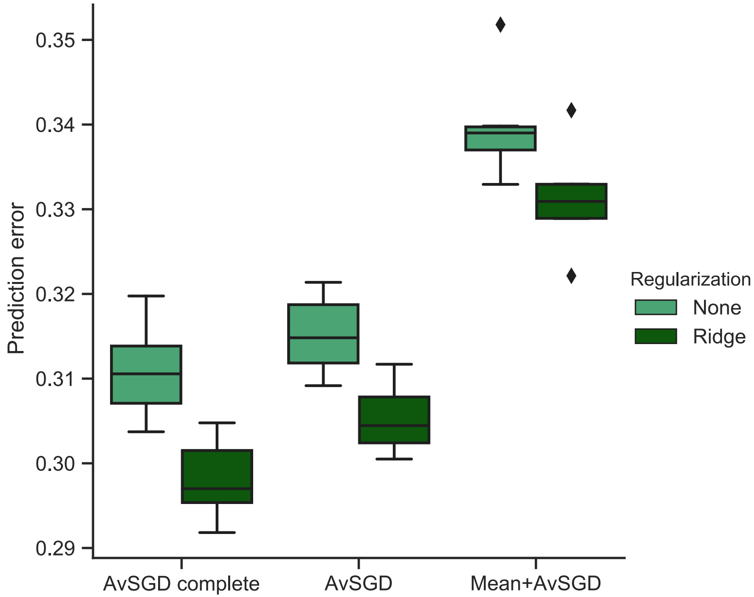


Figure: Prediction error $\|\hat{y} - y\|^2 / \|y\|^2$ boxplots.

- 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

✓ A new algorithm with a fast rate to perform SGD with missing data.

✓ Python implementation of regularized regression with missing values for large scale data.

✓ More details in the paper[9]!

Many perspectives:

- Dealing with more general loss function.
- More complex missing-data patterns such as MAR and MNAR.
- Lower bounds
- Distributed case
- Bounds on the empirical risk, tighter bound for estimated p.

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