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

8$^{th}$ February 2021
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 → keep only 5% of the rows.:(**
  - ⇒ **Simpler algorithmic solutions?**

### Traumabase: 15,000 patients/ 250 var/ 15 hospitals

<table>
<thead>
<tr>
<th>Center</th>
<th>Age</th>
<th>Sex</th>
<th>Weight</th>
<th>Height</th>
<th>Heart rate</th>
<th>Lactates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beaujon</td>
<td>54</td>
<td>m</td>
<td>85</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Lille</td>
<td>33</td>
<td>m</td>
<td>80</td>
<td>1.8</td>
<td>180</td>
<td>4.8</td>
</tr>
<tr>
<td>Pitie</td>
<td>26</td>
<td>m</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>3.9</td>
</tr>
</tbody>
</table>

**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
Setting

• \((X_i, y_i)_{i \geq 1} \in \mathbb{R}^d \times \mathbb{R}\) i.i.d. observations

• Linear regression model

\[ y_i = X_i^T \beta^* + \epsilon_i, \]

parametrized by \(\beta^* \in \mathbb{R}^d\), 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:**

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

• **Stochastic gradient method.**
  
  . At the heart of Machine Learning.
  
  . Very well suited for large \(d\) and \(n\).
Objective - missing data

- Problem: \((X_i, y_i)\)’s partially known
  1. What should we estimate?

- True risk minimization:

\[
\beta^* = \underset{\beta \in \mathbb{R}^d}{\arg \min} \left\{ R(\beta) := \mathbb{E}(X_i, y_i) [f_i(\beta)] \right\}
\]

2. How to adapt algorithms to the missing data case?
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 $\alpha$ is the step-size and $\mathbb{E}[g_k(\beta_{k-1})|\mathcal{F}_{k-1}] = \nabla F(\beta_{k-1})$, $\mathcal{F}_{k-1} = \sigma(X_1, y_1, \ldots, X_{k-1}, y_{k-1})$ 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

- **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 $\alpha$ is the step-size and $E[g_k(\beta_{k-1})|\mathcal{F}_{k-1}] = \nabla F(\beta_{k-1})$, $\mathcal{F}_{k-1} = \sigma(X_1:, y_1, \ldots, X_{k-1}:, y_{k-1})$ 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.
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 \( X_i^{NA} :\in (\mathbb{R} \cup \{NA\})^d \) instead of \( X_i :\)

\[
X_i^{NA} := X_i : \odot D_i : + NA(1_d - D_i :),
\]

\( \odot \) element-wise product, \( 1_d = (1 \ldots 1)^T \in \mathbb{R}^d \), \( NA \times 0 = 0 \), \( NA \times 1 = NA \).
**Missing values setting**

**Mechanism assumption**

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

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

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

- Expectation Maximization algorithm\(^1\) (maximization of the observed likelihood)\(^\text{X}\) parametric assumptions: Gaussian assumption for the covariates, no solution available for large dimension \(p\).\(^\text{X}\)
- Matrix completion (predicting NA before applying usual algorithms)\(^\text{X}\) 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\(^3\).

---


Dealing with missing values

Our strategy inspired by Ma et Needell

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

- **Imputing the missing values by 0:**
  \[
  \tilde{X}_i = X_{i;NA}^{i} \odot D_i = X_i \odot D_i: \text{imputed covariates}
  \]

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

Our strategy inspired by Ma et Needell

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

- **Imputing the missing values by 0.**
  \[
  \tilde{X}_{i:} = X_{i:}^{\text{NA}} \odot D_{i:} = X_{i:} \odot D_{i:} \text{ imputed covariates}
  \]

- Using a **debiased gradient** for the **averaged SGD**:
  Find \(\tilde{g}_k(\beta_{k-1})\) such that 
  \[
  \mathbb{E}[\tilde{g}_k(\beta_{k-1}) | \mathcal{F}_{k-1}] = \nabla R(\beta_{k-1})
  \]
  - \(\mathcal{F}_{k-1} = \sigma(X_{1:}, y_1, D_{1:}, \ldots, X_{k-1:}, y_{k-1}, D_{k-1:})\)
  - \(\nabla R(\beta_{k-1}) = \mathbb{E}_{(x_k:y_k)}[X_k:(X_k^T\beta_{k-1} - y_k)]\)

- No access to \(X_{k:}\), only to \(\tilde{X}_{k:}\).

- Another source of randomness: \(\mathbb{E} = \mathbb{E}(x_k:y_k), D_k: \overset{\text{indep}}{=} \mathbb{E}(x_k:y_k)\mathbb{E}_{D_k:}\)

- \(\mathbb{E}_{D_k:} |\mathcal{F}_{k-1} \sim \mathbb{E}_{D_k:}\)

  ✓ **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 \((X_{i:}^{NA}, y_i)\)

- **Imputing the missing values by 0.**
  \[
  \tilde{X}_{i:} = X_{i:}^{NA} \odot D_{i:} = X_{i:} \odot D_{i:} \text{ imputed covariates}
  \]

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

Thus

\[
\mathbb{E}_{D_k}: [\tilde{X}_k:] = \mathbb{E}_{D_k}: \begin{bmatrix}
(\delta_{k1}X_{k1}) \\
\vdots \\
(\delta_{kd}X_{kd})
\end{bmatrix} = \begin{bmatrix}
p_{1}X_{k1} \\
\vdots \\
p_{d}X_{kd}
\end{bmatrix}
\]

Thus

\[
\mathbb{E}_{D_k}: [P^{-1}\tilde{X}_k:] := \begin{bmatrix}
p_{1}^{-1} & \cdots & \\
\vdots & \ddots & \\
p_{d}^{-1} & \cdots & p_{d}
\end{bmatrix} \begin{bmatrix}
p_{1}X_{k1} \\
\vdots \\
p_{d}X_{kd}
\end{bmatrix} = X_k:
\]
Dealing with missing values

Our strategy inspired by Ma et Needell

Online-streaming: for a new observation \((X^\text{NA}_i, y_i)\)

- **Imputing the missing values by 0.**
  \[
  \tilde{X}_i = X^\text{NA}_i \odot D_i = X_i \odot D_i \quad \text{imputed covariates}
  \]

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

One obtains

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

\[
\nabla F(\beta) = \begin{pmatrix} x^T \beta - y \end{pmatrix} x
\]

\[
\mathbb{E}\left[ y^T P \tilde{X} \right] = y^T X
\]
Algorithm 1 Averaged SGD for Heterogeneous Missing Data

**Input:** data $\tilde{X}, y, \alpha$ (step size)

Initialize $\beta_0 = 0_d$.

Set $P = \text{diag} ((p_j)_{j \in \{1, \ldots, d\}}) \in \mathbb{R}^{d \times d}$.

**for** $k = 1$ **to** $n$ **do**

$$\tilde{g}_k(\beta_{k-1}) = P^{-1} \tilde{X}_k: (\tilde{X}_k^T P^{-1} \beta_{k-1} - y_k) - (I - P)P^{-2} \text{diag} (\tilde{X}_k: \tilde{X}_k^T) \beta_{k-1}$$

$$\beta_k = \beta_{k-1} - \alpha \tilde{g}_k(\beta_{k-1})$$

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

**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 $\beta_*$ 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.
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**
Optimization **without** missing values

convergence rates and proof techniques

If $F$ is convex and $L$-smooth.\(^5\)

\[ \times \quad \text{Convergence rate: } \mathcal{O}(k^{-1/2}) \]

If $F$ is convex and $L$-smooth, $\mu$-strongly convex.

\[ \times \quad \text{Convergence rate: } \mathcal{O}((\mu k)^{-1}), \text{ with } \mu \text{ known.} \]

If $F$ is convex and quadratic, e.g., for least-squares regression\(^6\).

\[ \checkmark \quad \text{Convergence rate: } \mathcal{O}(k^{-1}) \]

? Why do we get a faster rate for quadratic functions?

? What does it require?

---


Faster rates for Least Squares regression

- Typical proof for convex:

\[ \mathbb{E} \left[ \frac{\| \beta_k - \beta_* \|^2}{\| \beta_{k-1} - \beta_* \|^2} \right] \]

\[ - 2 \gamma \left\langle \nabla F(\beta_{k-1}), \beta_{k-1} - \beta_* \right\rangle \]

\[ - \frac{\gamma^2}{\beta} \| g_n(\beta_{n-1}) \|^2 \]

\[ \nabla f + \Sigma_n \]

\[ \beta_n \to \beta_* = \beta_{\star} \quad \text{linearly fast} \]

\[ \beta_{\pi_0}^* \left( \nabla f(\beta) \right) = 0 \]

\[ \Rightarrow \text{I can lose a factor of } N \text{ and raise cut-off!} \]
Faster rates for Least Squares regression

• Typical proof for quadratic:

\[ P \hat{\beta}_k = \beta_{n+1} - \alpha \nabla F(\beta_{k+1}) + \alpha \varepsilon. \]

\[ \alpha H (\hat{\beta}_d - \hat{\beta}_x) = \beta_{n+1} - \beta_n + \alpha \varepsilon \]

\[ \left( \overline{\beta_m} - \beta_x \right) = H^{-1} \frac{\overline{B_0} - \overline{B}_x}{\alpha n} + \frac{1}{n} \hat{\varepsilon} \left( H^{-1} \varepsilon \right) \]

\[ \| \tilde{\varepsilon} \| \leq \| H^\top \tilde{\varepsilon} \|^2 = \text{tr} \left( H^\top \tilde{\varepsilon} \tilde{\varepsilon}^\top \right) \] is bounded
Summary

SGD

Least squares $\rightarrow$ unbiased gradient oracle with NA

Fast rate of convergence
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, \(H := \mathbb{E}(x_k,y_k)\mathbb{E}(x_k:X_k^T)\) invertible.

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

\((\tilde{g}_k(\beta^*))_k\) with \(\beta^*\) is \(\mathcal{F}_k\)-measurable and \(\forall k \geq 0\),
- \(\mathbb{E}[\tilde{g}_k(\beta^*) | \mathcal{F}_{k-1}] = 0\) a.s.
- \(\mathbb{E}[\|\tilde{g}_k(\beta^*)\|^2 | \mathcal{F}_{k-1}]\) is a.s. finite.
- \(\mathbb{E}[\tilde{g}_k(\beta^*)\tilde{g}_k(\beta^*)^T] \leq C(\beta^*) = c(\beta^*)H\).

**Lemma: \((\tilde{g}_k(\beta^*))_k\) 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
Theoretical results
Convergence results

Theorem: convergence rate of $O(k^{-1})$, 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 $k \geq 0$:

$$E[R(\bar{\beta}_k) - R(\beta^*)] \leq \frac{1}{2k} \left( \frac{\sqrt{c(\beta^*)d}}{1 - \sqrt{\alpha}L} + \frac{||\beta_0 - \beta^*||}{\sqrt{\alpha}} \right)^2,$$

- $L := \sup_{k,D}$ Lipschitz constants of $\tilde{g}_k$
- $p_m = \min_{j=1,...,d} p_j$ minimal probability to be observed
- $c(\beta^*) = \frac{\text{Var}(\epsilon_k)}{p_m^2} + \left( \frac{2 + 5p_m(1 - p_m)}{p_m^3} \right)^2 \gamma^2 \|\beta^*\|^2$
  \[\text{increasing with the missing values rate}\]
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 ($\beta \rightarrow R(\beta) + \lambda \|\beta\|^2$ is $2\lambda$-strongly convex).

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

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.
Theoretical results

What impact of missing values?

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

\[
X = \begin{pmatrix}
X_1 & X_2 & X_3 \\
12 & 28 & 31 \\
\text{NA} & 23 & 89 \\
32 & 6 & 24 \\
\vdots & \vdots & \vdots \\
\text{NA} & 3 & 7 \\
\end{pmatrix}
\]

\[
X = \begin{pmatrix}
X_1 & X_2 & X_3 \\
12 & 28 & 31 \\
\text{NA} & 23 & 89 \\
32 & 6 & 24 \\
\vdots & \vdots & \vdots \\
\text{NA} & 3 & 7 \\
\end{pmatrix}
\]
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: $\frac{\text{Var}(\epsilon_k)d}{k_{co}}$.
  - in expectation, lower bound on the risk larger than $\frac{\text{Var}(\epsilon_k)d}{kp^d}$.

- keeping all the observations, averaged SGD: upper bound $O\left(\frac{\text{Var}(\epsilon_k)d}{kp^2} + \frac{C(X, \beta^*)}{kp^3}\right)$.

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**: \( \beta^n_\star = \arg \min_{\beta \in \mathbb{R}^d} \{ R_n(\beta) := \frac{1}{n} \sum_{i=1}^{n} f_i(\beta) \} \)

  How to choose the \( k \)-th observation?
  
  - \( \times \) \( k \) uniformly at random \( \Rightarrow \) we use a data several times.
  - \( \times \) \( k \) not chosen uniformly at random \( \Rightarrow \) sampling not uniform and bias in the gradient.

---

Open Question: rates for ERM?

- **Empirical risk:** \( \beta^n_* = \arg \min_{\beta \in \mathbb{R}^d} \{ R_n(\beta) := \frac{1}{n} \sum_{i=1}^{n} f_i(\beta) \} \)
  
  How to choose the \( k \)-th observation?
  
  - \( \times \) \( k \) uniformly at random ⇒ we use a data several times.
  - \( \times \) \( k \) not chosen uniformly at random ⇒ 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”.  

Theoretical results
Comparison with related work

Comparison with Ma et Needell\textsuperscript{8}:

- $\mu$-strongly convex problem
- no averaged iterates

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

$\times$ $\mu$ generally out of reach.

$\times$ only homogeneous MCAR data.

$\times$ main theorem mathematically invalid (empirical risk).

\textsuperscript{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
Finite-sample setting: $n$ is fixed

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

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

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 $k \geq 0$:

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

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

$$e^{-np_m}$$
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.

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 \( (R_n(\beta_k) - R_n(\beta^*)) \).

• Multiple passes (left): saturation.

• One pass (right): saturation for SGD\textsubscript{cst}, \( O(n^{-1/2}) \) for SGD, \( O(n^{-1}) \) 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.