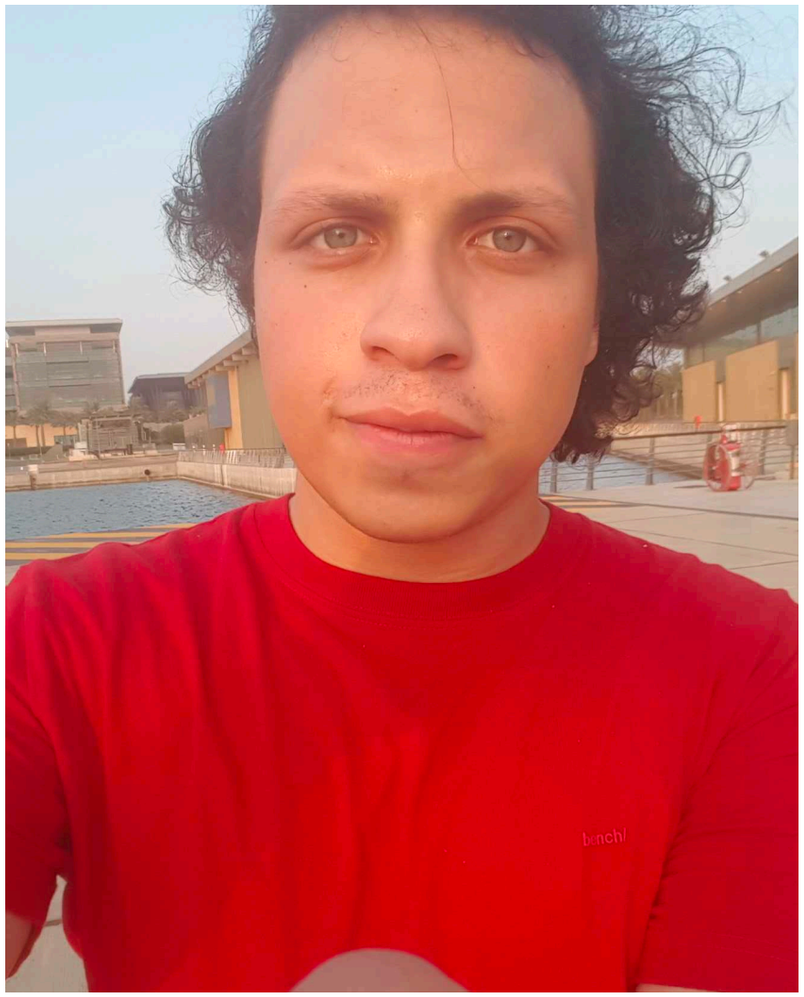
Proximal and Federated Random Reshuffling

Konstantin Mishchenko, Ahmed Khaled, Peter Richtárik

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology



Ahmed Khaled

Peter Richtárik

Talk outline

1. **Problem formulation**
2. **Sampling, shuffling and fixed order**
3. **Theoretical results**
4. **Experiments**

The problem

$$\min_{x\in\mathbb{R}^d} \frac{1}{N}\sum_{m=1}^{M} F_m(x), \qquad F_m(x) = \sum_{j=1}^{N_{m_j}} f_{mj}$$

The problem

Data in total

Data locally

$$\min_{x\in\mathbb{R}^d} \frac{1}{N}\sum_{m=1}^{M} F_m(x), \qquad F_m(x) = \sum_{j=1}^{N_{m_j}} f_{mj}$$

The problem

Data in total

Data locally

$$\min_{x\in\mathbb{R}^d} \frac{1}{N}\sum_{m=1}^{M} F_m(x), \qquad F_m(x) = \sum_{j=1}^{N_{m_j}} f_{mj}$$

Dimension

$$\mu \preccurlyeq \nabla^2 f_{mj} \preccurlyeq L$$

The problem reformulated

$$\min_{\boldsymbol{x}\in\mathbb{R}^{d\cdot M}} \frac{1}{N}\sum_{m=1}^{M} F_m(x_m) + \psi(\boldsymbol{x}),$$

$$\psi(\boldsymbol{x}) = \begin{cases} 0, & \text{if } x_1 = \cdots = x_M, \\ +\infty, & \text{otherwise} \end{cases}$$

The problem reformulated

$$\min_{\boldsymbol{x}\in\mathbb{R}^{d\cdot M}} \frac{1}{N}\sum_{m=1}^{M} F_m(x_m) + \psi(\boldsymbol{x}),$$

$$\psi(\boldsymbol{x}) = \begin{cases} 0, & \text{if } x_1 = \cdots = x_M, \\ +\infty, & \text{otherwise} \end{cases}$$

The problem reformulated

$$\min_{\boldsymbol{x}\in\mathbb{R}^{d\cdot M}} \frac{1}{N}\sum_{m=1}^{M} F_m(x_m) + \psi(\boldsymbol{x}),$$

$$\psi(\boldsymbol{x}) = \begin{cases} 0, & \text{if } x_1 = \cdots = x_M, \\ +\infty, & \text{otherwise} \end{cases}$$

$$\min_{\boldsymbol{x}\in\mathbb{R}^{d\cdot M}} \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{x}) + \psi(\boldsymbol{x})$$

$$f_i(\boldsymbol{x}) = \sum_{m=1}^{M} f_{mi}(x_m)$$

Talk outline

1. **Problem formulation**
2. **Sampling, shuffling and fixed order**
3. **Theoretical results**
4. **Experiments**

What is SGD?

Leon Bottou “Stochastic Gradient Descent Tricks”

What is SGD?

Leon Bottou "Stochastic Gradient Descent Tricks"

4.1 Preparing the data

Randomly shuffle the training examples.

Although the theory calls for picking examples randomly, it is usually faster to zip sequentially through the training set. But this does not work if the examples are grouped by class or come in a particular order. Randomly shuffling the examples eliminates this source of problems. Section 1.4.2 provides an additional discussion.

What is SGD?

SGD

Random Reshuffling

Stochastic Algorithms

$$\min_x \frac{1}{n} \sum_{i=1}^{n} f_i(x)$$

$$\psi(x) \equiv 0$$

Stochastic Algorithms

SGD

$$x_{t+1} = x_t - \gamma_t \nabla f_i(x_t), \quad i \sim U(\{1, \dots, n\})$$

Stochastic Algorithms

SGD

$$x_{t+1} = x_t - \gamma_t \nabla f_i(x_t), \quad i \sim U(\{1, \dots, n\})$$

IG

$$x_t^{i+1} = x_t^i - \gamma_{t,i} \nabla f_{i+1}(x_t^i), \quad x_{t+1}^0 = x_t^n$$

Stochastic Algorithms

SGD

$$x_{t+1} = x_t - \gamma_t \nabla f_i(x_t), \quad i \sim U(\{1, \dots, n\})$$

IG

$$x_t^{i+1} = x_t^i - \gamma_{t,i} \nabla f_{i+1}(x_t^i), \quad x_{t+1}^0 = x_t^n$$

RR/SO

$$x_t^{i+1} = x_t^i - \gamma_{t,i} \nabla f_{\pi_i}(x_t^i), \quad \{\pi_0, \dots, \pi_{n-1}\} = \{1, \dots, n\}$$

Stochastic Algorithms

SGD

$$x_{t+1} = x_t - \gamma_t \nabla f_i(x_t), \quad i \sim U(\{1, \dots, n\})$$

IG

$$x_t^{i+1} = x_t^i - \gamma_{t,i} \nabla f_{i+1}(x_t^i), \quad x_{t+1}^0 = x_t^n$$

RR/SO

$$x_t^{i+1} = x_t^i - \gamma_{t,i} \nabla f_{\pi_i}(x_t^i), \quad \{\pi_0, \dots, \pi_{n-1}\} = \{1, \dots, n\}$$

Hard to analyze: biased!

Recent results on RR

J. Haochen and S. Sra
Random Shuffling Beats SGD after Finite Epochs
International Conference on Machine Learning, 2019

S. Rajput, A. Gupta, and D. Papailiopoulos
Closing the convergence gap of SGD without replacement
International Conference on Machine Learning, 2020

Recent results on RR

J. Haochen and S. Sra
Random Shuffling Beats SGD after Finite Epochs
International Conference on Machine Learning, 2019

S. Rajput, A. Gupta, and D. Papailiopoulos
Closing the convergence gap of SGD without replacement
International Conference on Machine Learning, 2020

A. Khaled, K. Mishchenko, P. Richtárik
Random Reshuffling: Simple Analysis with Vast Improvements
NeurIPS, 2020

SGD

$$x_{t+1} = x_t - \gamma_t \nabla f_i(x_t), \quad i \sim U(\{1, \ldots, n\})$$

Strongly convex: $$\mathbb{E}[\|x_t - x^*\|^2] = \mathcal{O}\left(\frac{1}{t}\right)$$

R. Gower, N. Loizou, X. Qian, A. Sailanbayev, E. Shulgin, P. Richtárik

SGD: General analysis and improved rates

International Conference on Machine Learning, 2019

Random Reshuffling and Shuffle Once

$$x_t^{i+1} = x_t^i - \gamma_{t,i} \nabla f_{\pi_i}(x_t^i), \quad x_{t+1} = x_{t+1}^0 = x_t^n$$

Strongly convex: $$\mathbb{E}[\|x_t - x^*\|^2] = \mathcal{O}\left(\frac{1}{nt^2}\right)$$

A. Khaled, K. Mishchenko, P. Richtárik

Random Reshuffling: Simple Analysis with Vast Improvements

NeurIPS, 2020

Talk outline

1. Problem formulation
2. Sampling, shuffling and fixed order
3. Theoretical results
4. Experiments

Reminder: objective

$$\min_{\boldsymbol{x}\in\mathbb{R}^{d\cdot M}} \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{x}) + \psi(\boldsymbol{x})$$

Reminder: objective

$$\min_{\boldsymbol{x}\in\mathbb{R}^{d\cdot M}} \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{x}) + \psi(\boldsymbol{x})$$

$$\mathrm{prox}_{\gamma\psi}(x) \stackrel{\mathrm{def}}{=} \arg\min_z \left\{ \gamma\psi(z) + \frac{1}{2}\|z-x\|^2 \right\}$$

Reminder: objective

$$\min_{\boldsymbol{x}\in\mathbb{R}^{d\cdot M}} \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{x}) + \psi(\boldsymbol{x})$$

$$\operatorname{prox}_{\gamma\psi}(x) \stackrel{\text{def}}{=} \arg\min_z \left\{ \gamma\psi(z) + \frac{1}{2}\|z-x\|^2 \right\}$$

$$\operatorname{prox}_{\gamma\psi}(\boldsymbol{x}) = \begin{pmatrix} \overline{x} \\ \vdots \\ \overline{x} \end{pmatrix} \qquad \overline{x} = \frac{1}{M}\sum_{m=1}^{M} x_m$$

Prox-SGD

$$i \sim U(\{1, \ldots, n\})$$

$$x_{t+1} = \text{prox}_{\gamma\psi}(x_t - \gamma \nabla f_i(x_t))$$

Prox-SGD

$$i \sim U(\{1, \ldots, n\})$$

$$x_{t+1} = \text{prox}_{\gamma\psi}(x_t - \gamma \nabla f_i(x_t))$$

$$\begin{cases} x_{t+1}^m = x_t - \gamma \nabla f_{mi}(x_t), \\ x_{t+1} = \frac{1}{M} \sum_{m=1}^{M} x_{t+1}^m \end{cases}$$

Prox-RR

Easy!

$$x_t^{i+1} = \mathrm{prox}_{\gamma\psi}(x_t^i - \gamma\nabla f_{\pi_i}(x_t))$$

Prox-RR

Easy!

$$x_t^{i+1} = \mathrm{prox}_{\gamma f_{\pi_i}}(x_t^i - \gamma \nabla f_{\pi_i}(x_t))$$

Not so fast...

Prox-RR

Easy!

$$x_t^{i+1} = \mathrm{prox}_{\gamma\psi}(x_t^i - \gamma \nabla f_{\pi_i}(x_t))$$

Not so fast...

$f_i(x) = \langle c_i, x \rangle$

$$x_{t+1} = \mathrm{prox}_{\gamma n \psi}\left(x_t - \gamma \sum_{i=0}^{n-1} \nabla f_{\pi_i}(x_t^i)\right)$$

$$= \mathrm{prox}_{\gamma n \psi}(x_t - \gamma n \nabla f(x_t))$$

Prox-RR

$$x_t^{i+1} = x_t^i - \gamma \nabla f_{\pi_i}(x_t^i)$$

$$x_{t+1} = \operatorname{prox}_{\gamma n \psi}(x_t^n)$$

Prox-RR

$$\mathbb{E}\left[\|x_T - x_*\|^2\right] \leq (1+\gamma\mu n)^{-T}\|x_0 - x_*\|^2 + \frac{2\gamma^2\sigma_{\text{rad}}^2}{\mu}$$

$$\sigma_{\text{rad}}^2 \leq \frac{L_{\max}}{2} n\left(n\|\nabla f(x_*)\|^2 + \frac{1}{2}\sigma_*^2\right)$$

Prox-RR

$$\mathbb{E}\left[\|x_T - x_*\|^2\right] \leq (1+\gamma\mu n)^{-T}\|x_0 - x_*\|^2 + \frac{2\gamma^2\sigma_{\mathrm{rad}}^2}{\mu}$$

$$\sigma_{\mathrm{rad}}^2 \leq \frac{L_{\max}}{2} n\left(n\|\nabla f(x_*)\|^2 + \frac{1}{2}\sigma_*^2\right)$$

IID data: $\|\nabla f(x_*)\|^2 = 0$

Prox-RR

$$\mathbb{E}\left[\|x_T - x_*\|^2\right] \leq (1+\gamma\mu n)^{-T}\|x_0 - x_*\|^2 + \frac{2\gamma^2\sigma_{\text{rad}}^2}{\mu}$$

$$\sigma_{\text{rad}}^2 \leq \frac{L_{\max}}{2} n\left(n\|\nabla f(x_*)\|^2 + \frac{1}{2}\sigma_*^2\right)$$

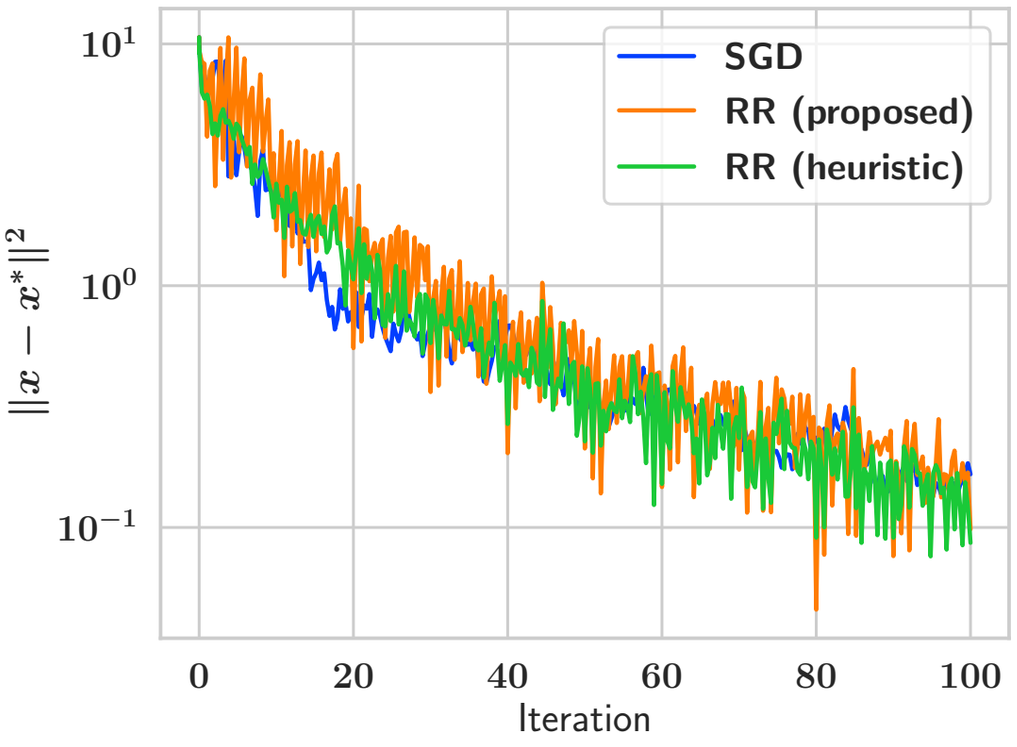
Data heterogeneity

IID data: $\|\nabla f(x_*)\|^2 = 0$

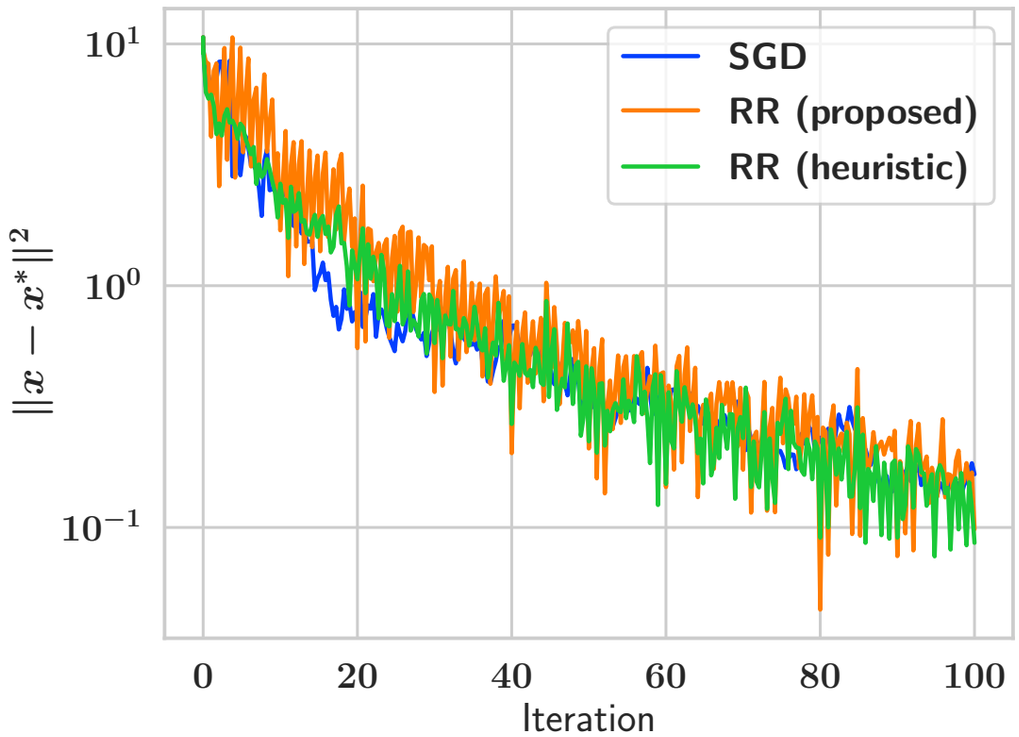
Talk outline

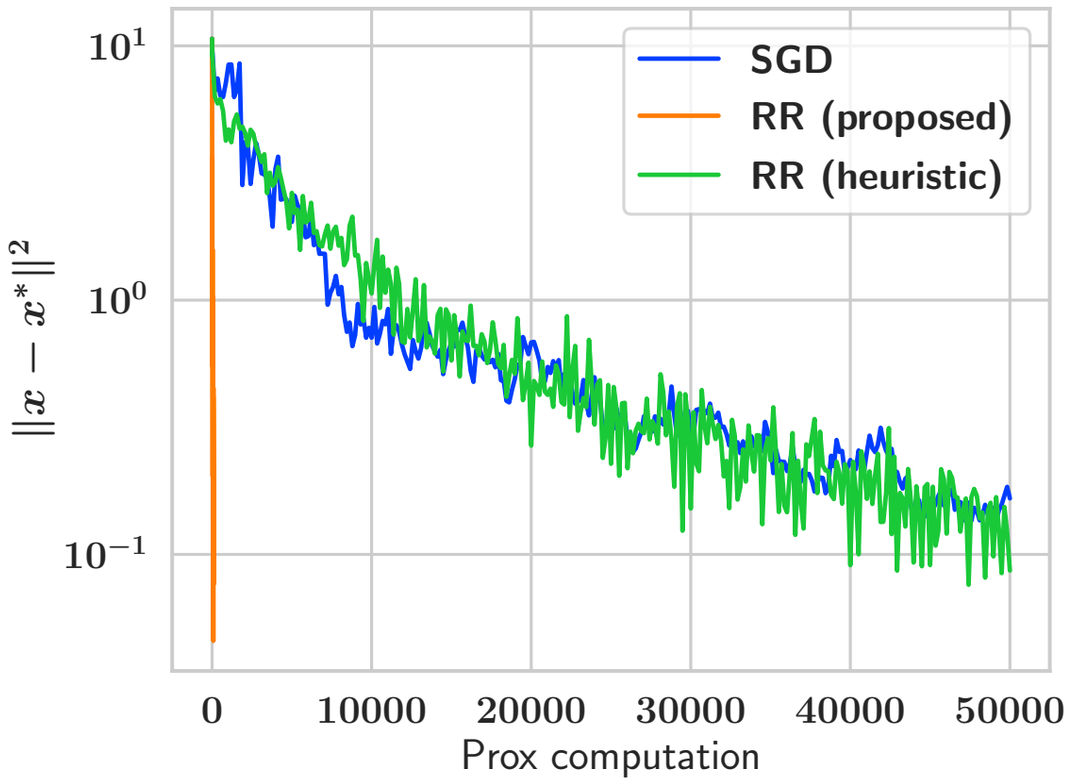
1. **Problem formulation**
2. **Sampling, shuffling and fixed order**
3. **Theoretical results**
4. **Experiments**

Experiments: logistic regression w/ l2 regularization



Experiments: logistic regression w/ l2 regularization





More things in the paper

1. **Importance sampling**
2. **Decreasing stepsize**
3. **Regularized Fed RR**
4. **Bounds for iid and non-iid data**