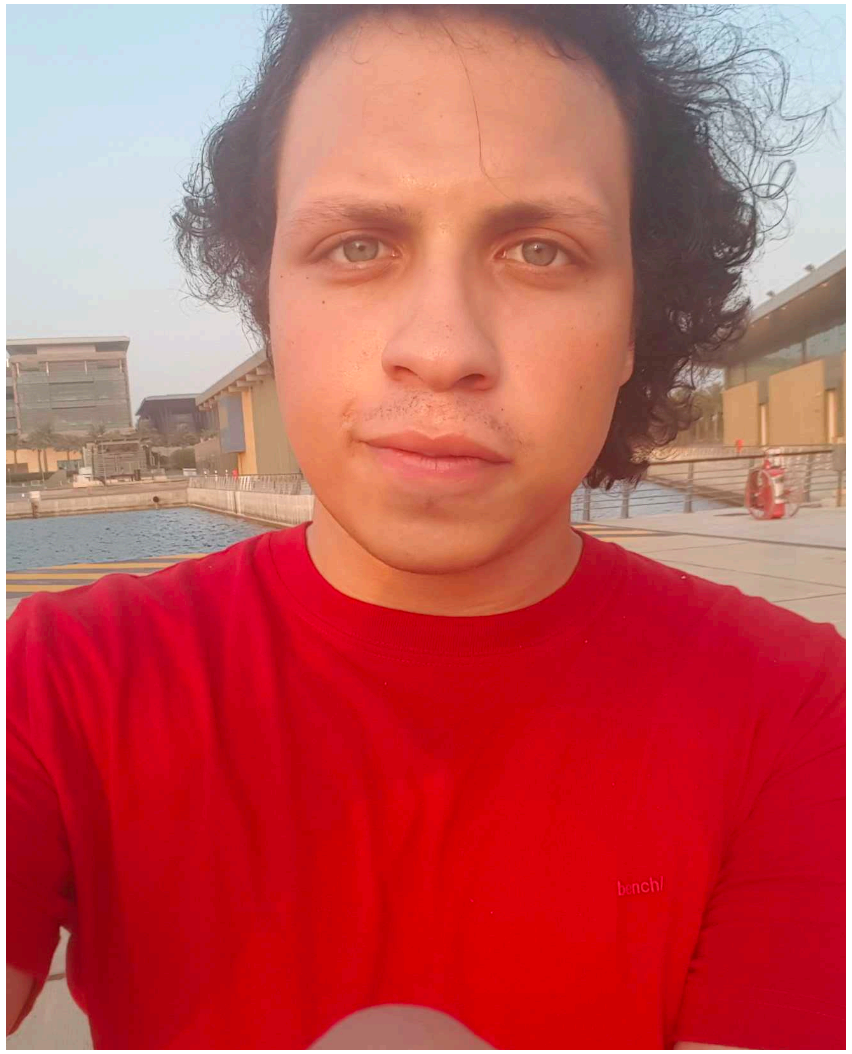
Random Reshuffling: Simple Analysis with Vast Improvements

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

$$\min_{x \in \mathbb{R}^d} f(x) \overset{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} f_i(x)$$

The problem

of data observations

L-smooth

$$\min_{x\in\mathbb{R}^d} f(x) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

Dimension

The problem

of data observations

L-smooth

$$\min_{x \in \mathbb{R}^d} f(x) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} f_i(x)$$

Dimension

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq \boxed{L} \|x - y\|$$

Talk outline

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

Stochastic Algorithms

SGD

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

Stochastic Algorithms

SGD

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

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

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

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

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

Talk outline

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

SGD

H. Robbins and S. Monro
Stochastic Approximation Method
The Annals of Mathematical Statistics, 1951

R. Gower, N. Loizou, X. Qian, A. Sailanbayev, E. Shulgin, P. Richtárik
SGD: General analysis and improved rates
International Conference on Machine Learning, 2019

SGD

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

SGD

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

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

SGD

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

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

Convex: $$\min_{k \le t} \mathbb{E}[f(x_k) - f(x^*)] = \mathcal{O}\left(\frac{1}{\sqrt{t}}\right)$$

SGD

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

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

Convex: $$\min_{k \le t} \mathbb{E}[f(x_k) - f(x^*)] = \mathcal{O}\left(\frac{1}{\sqrt{t}}\right)$$

Nonconvex: $$\min_{k \le t} \mathbb{E}[\|\nabla f(x_k)\|^2] = \mathcal{O}\left(\frac{1}{\sqrt{t}}\right)$$

Incremental Gradient

D. Bertsekas
Incremental Gradient, Subgradient, and Proximal Methods for Convex Optimization: A Survey
Optimization for Machine Learning, chapter 4, 2011

M. Gürbüzbalaban, A. Ozdaglar, and P. A. Parrilo
Convergence Rate of Incremental Gradient and Incremental Newton Methods
Mathematical Programming, 2019

Incremental Gradient

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

Incremental Gradient

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

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

Convex: $$\min_{k \leq t} \mathbb{E}[f(x_k) - f(x^*)] = \mathcal{O}\left(\frac{1}{t^{2/3}}\right)$$

Nonconvex: $$\min_{k \leq t} \mathbb{E}[\|\nabla f(x_k)\|^2] = \mathcal{O}\left(\frac{1}{t^{2/3}}\right)$$

Incremental Gradient

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

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

Convex: $$\min_{k \leq t} \mathbb{E}[f(x_k) - f(x^*)] = \mathcal{O}\left(\frac{1}{t^{2/3}}\right)$$

Nonconvex: $$\min_{k \leq t} \mathbb{E}[\|\nabla f(x_k)\|^2] = \mathcal{O}\left(\frac{1}{t^{2/3}}\right)$$

Incremental Gradient

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

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

Convex: $$\min_{k \le t} \mathbb{E}[f(x_k) - f(x^*)] = \mathcal{O}\left(\frac{1}{t^{2/3}}\right)$$

Nonconvex: $$\min_{k \le t} \mathbb{E}[\|\nabla f(x_k)\|^2] = \mathcal{O}\left(\frac{1}{t^{2/3}}\right)$$

Always slower than Gradient Descent

Random Reshuffling

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

Random Reshuffling and Shuffle Once (new!)

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

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

Convex: $\min_{k \le t} \mathbb{E}[f(x_k) - f(x^*)] = \mathcal{O}\left(\frac{1}{n^{1/3}t^{2/3}}\right)$

Nonconvex: $\min_{k \le t} \mathbb{E}[\|\nabla f(x_k)\|^2] = \mathcal{O}\left(\frac{1}{n^{1/3}t^{2/3}}\right)$

Main improvements

1. **Better rates**

Main improvements

1. **Better rates**
2. **Large stepsizes allowed** $\gamma_{t,i} \leq \frac{1}{L}$

Main improvements

1. **Better rates**
2. **Large stepsizes allowed** $\gamma_{t,i} \le \frac{1}{L}$
3. **Convergence within epoch**

$$x_*^i = x_* - \gamma \sum_{j=0}^{i-1} \nabla f_{\pi_j}(x_*), \quad x_*^n = x_*$$

Main improvements

1. **Better rates**
2. **Large stepsizes allowed** $\gamma_{t,i} \le \frac{1}{L}$
3. **Convergence within epoch**

$$x_*^i = x_* - \gamma \sum_{j=0}^{i-1} \nabla f_{\pi_j}(x_*), \quad x_*^n = x_*$$

$$\mathbb{E}[\|x_t^{i+1} - x_*^{i+1}\|^2] \le (1 - \gamma\mu)\mathbb{E}[\|x_t^i - x_*^i\|^2] + 2\gamma^2\sigma_{\mathrm{Shuffle}}^2$$

Main improvements

1. **Better rates**
2. **Large stepsizes allowed** $\gamma_{t,i} \le \frac{1}{L}$
3. **Convergence within epoch**

$$x_*^i = x_* - \gamma \sum_{j=0}^{i-1} \nabla f_{\pi_j}(x_*), \qquad x_*^n = x_*$$

$$\mathbb{E}[\|x_t^{i+1} - x_*^{i+1}\|^2] \le (1-\gamma\mu)\mathbb{E}[\|x_t^i - x_*^i\|^2] + 2\gamma^2\sigma^2_{\text{Shuffle}}$$

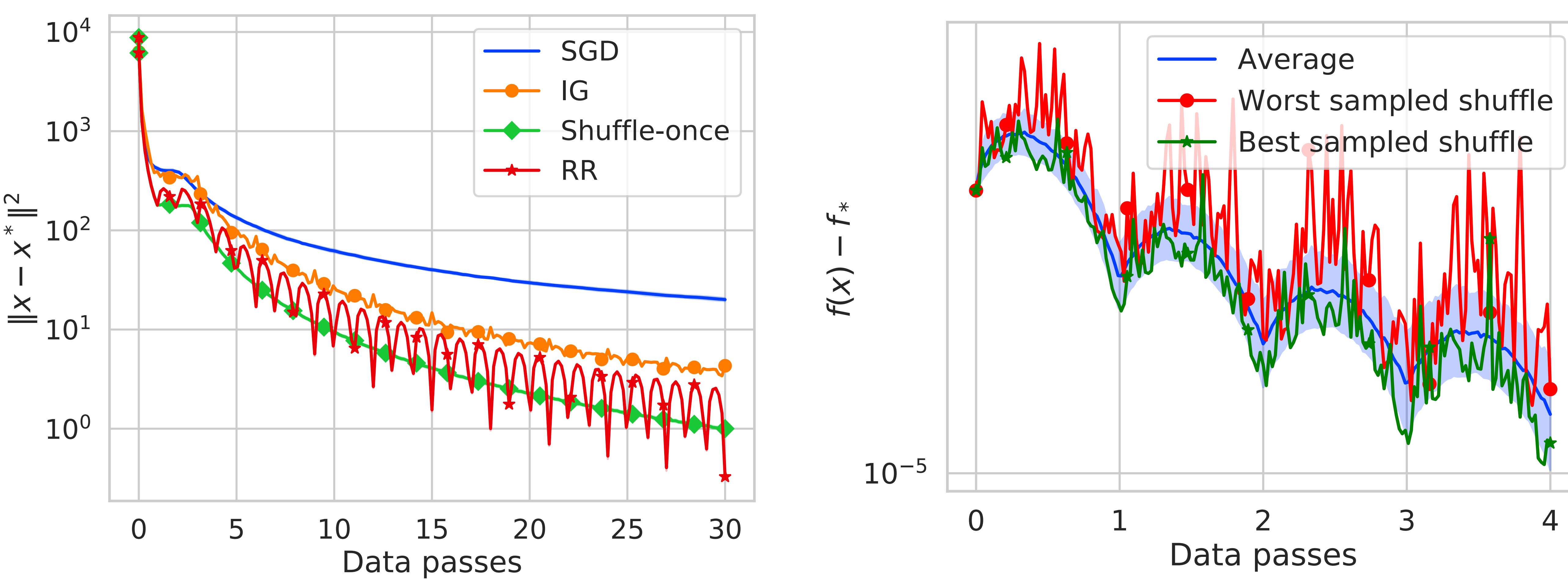
4. **“Variance” of shuffling**

$$\frac{\gamma\mu n}{8}\sigma^2_{\text{SGD}} \le \sigma^2_{\text{Shuffle}} \le \frac{\gamma L n}{4}\sigma^2_{\text{SGD}}$$

Talk outline

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

Experiments: logistic regression w/ l2 regularization



Experiments: “variance”