Random Reshuffling with Variance Reduction: New Analysis and Better Rates

Optimization without borders
Sochi, Sirius

Grigory Malinovsky

July 2021

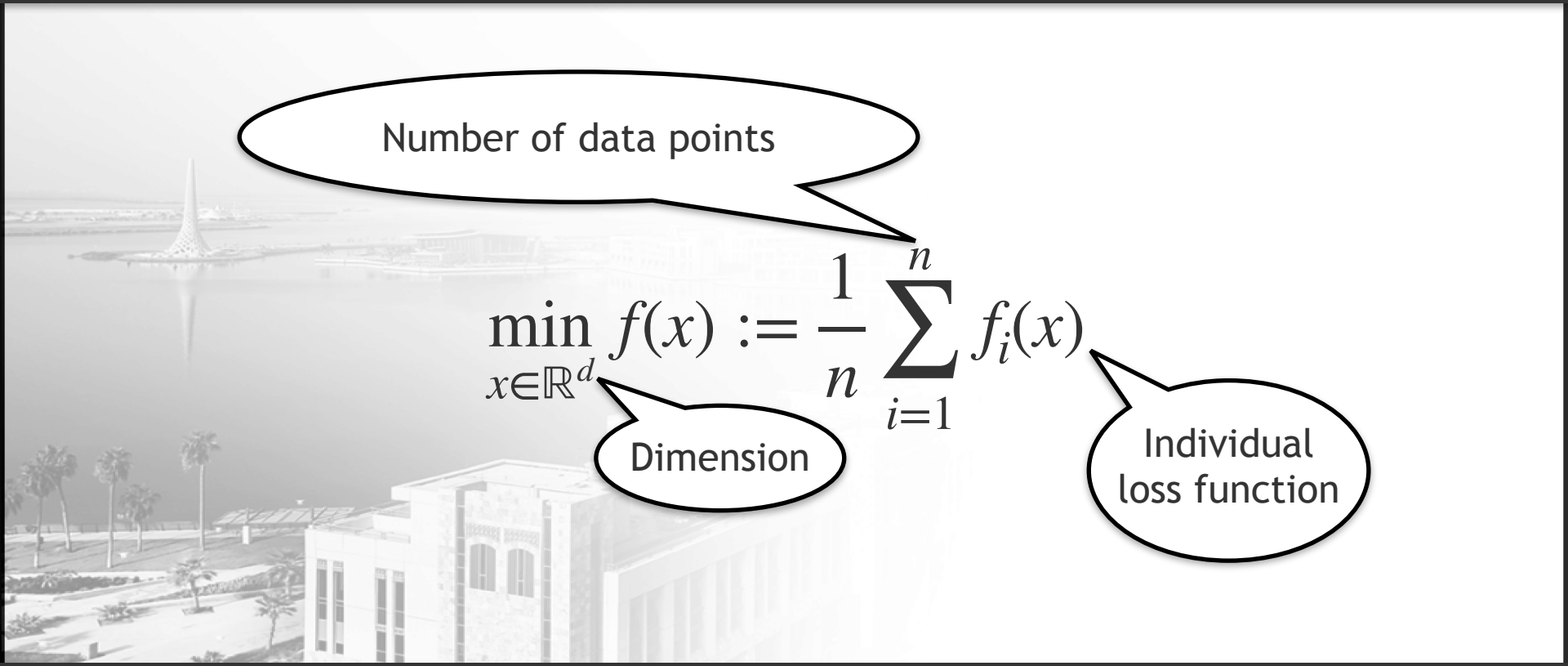
Co-authors

Grigory Malinovsky
(KAUST)

Alibek Sailanbayev
(KAUST)

Peter Richtárik
(KAUST)

The problem


Number of data points
$\min_{x \in \mathbb{R}^d} f(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x)$
Dimension
Individual loss function

Stochastic first order methods

Stochastic gradient descent

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

Permutation-based methods

Cyclic gradient descent

$$x_t^{i+1} = x_t^i - \gamma_t^i \nabla f_i(x_t^i) \quad x_{t+1}^0 = x_t^n$$

Random Reshuffling/Shuffle Once

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

π_i is a random permutation of $\{1,\dots,n\}$

Convergence of different methods

Rates of convergence

SGD

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

RR/SO

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

Cyclic GD

$$\left\| x_T - x_* \right\|^2 \leq (1-\gamma\mu)^{nT} \left\| x_0 - x_* \right\|^2 + \frac{\gamma^2 L n^2 \sigma_*^2}{\mu}.$$

Gradient variance at optimum

$$\sigma_*^2 := \frac{1}{n}\sum_{i=1}^{n} \left\| \nabla f_i(x_*) \right\|^2$$

Inner product reformulation

Reformulation

$$f(x) := \frac{1}{n}\sum_{i=1}^{n}\left(f_i(x) + \langle a_i, x\rangle\right) = \sum_{i=1}^{n}\tilde{f}_i(x)$$

$$\sum_{i=1}^{n} a_i = 0; \quad \tilde{f}_i(x) = f_i(x) + \langle a_i, x\rangle; \quad \nabla\tilde{f}_i(x) = \nabla f_i(x) + a_i$$

Application to permutation-based methods

$$g_t^i\left(x_t^i, y_t\right) = \nabla f_{\pi_i}\left(x_t^i\right) + a_i \quad a_i = -\nabla f_{\pi_i}\left(y_t\right) + \nabla f(y_t)$$

$$\sum_{i=1}^{n} a_i = -\sum_{i=1}^{n}\nabla f_{\pi_i}\left(y_t\right) + \sum_{i=1}^{n}\nabla f\left(y_t\right) = 0$$

$$g_t^i\left(x_t^i, y_t\right) = \nabla f_{\pi_i}\left(x_t^i\right) - \nabla f_{\pi_i}\left(y_t\right) + \nabla f\left(y_t\right)$$

The key lemma

Assume that each f_i is L-smooth and convex. If we apply the linear perturbation reformulation using vectors of the form $a_i = -\nabla f_{\pi_i}\left(y_t\right) + \nabla f(y_t)$, then the gradient variance of the reformulated problem at the optimum x_* can be bounded via the distance of the control vector y_t to x_* as follows:

$$\tilde{\sigma}_*^2 = \frac{1}{n}\sum_{i=1}^{n}\left\|\nabla\tilde{f}_i(x_*)\right\|^2 \leq 4L^2\left\|y_t - x_*\right\|^2$$

Description of the algorithms

Algorithm 1 Algorithms Det-Shuffle, Rand-Shuffle, Rand-Reshuffle

Input: Stepsize $\gamma > 0$, initial iterate $x_0 \in \mathbb{R}^d$, number of epochs T
Option Det-Shuffle**:** Choose a deterministic permutation $\{\pi_0, \ldots, \pi_{n-1}\}$ of $\{1, \ldots, n\}$
Option Rand-Shuffle**:** Choose a random permutation $\{\pi_0, \ldots, \pi_{n-1}\}$ of $\{1, \ldots, n\}$
for $t = 0, 1, \ldots T-1$ **do**
  Option Rand-Reshuffle**:** Choose a random permutation $\{\pi_0, \ldots, \pi_{n-1}\}$ of $\{1, \ldots, n\}$
  $x_t^0 = x_t, y_t = x_t$
  for $i = 0, \ldots, n-1$ **do**
    $g_t^i(x_t^i, y_t) = \nabla f_{\pi_i}(x_t^i) - \nabla f_{\pi_i}(y_t) + \nabla f(y_t)$
    $x_t^{i+1} = x_t^i - \gamma g_t^i(x_t^i, y_t)$
  end for
  $x_{t+1} = x_t^n$
end for

Theoretical guarantees

Algorithm	μ-strongly convex f_i	μ-strongly convex f	convex f	memory	citation
RR-SAGA	–	$\mathcal{O}\left(\kappa^2 \log \frac{1}{\epsilon}\right)$	–	$\mathcal{O}(dn)$	Ying et al. (2020)
AVRG	–	$\mathcal{O}\left(\kappa^2 \log \frac{1}{\epsilon}\right)$	–	$\mathcal{O}(d)$	Ying et al. (2020)
RR/SO-SVRG	$\mathcal{O}\left(\kappa\sqrt{\frac{\kappa}{n}} \log \frac{1}{\epsilon}\right)$ (in Big Data regime)	$\mathcal{O}\left(\kappa \log \frac{1}{\epsilon}\right)$ (in Big Data regime) $\mathcal{O}\left(\kappa\sqrt{\kappa} \log \frac{1}{\epsilon}\right)$ (in general regime)	$\mathcal{O}\left(\frac{L}{\varepsilon}\right)$	$\mathcal{O}(d)$	this paper
Cyclic SAGA	$\mathcal{O}\left(\kappa^2 \log \frac{1}{\epsilon}\right)$	–	–	$\mathcal{O}(dn)$	Park & Ryu (2020)
IAG (Cyclic SAG)	–	$\mathcal{O}\left(n\kappa^2 \log \frac{1}{\epsilon}\right)$	–	$\mathcal{O}(dn)$	Gürbüzbalaban et al. (2017)
DIAG (Cyclic Finito)	$\mathcal{O}\left(\kappa \log \frac{1}{\epsilon}\right)$	–	–	$\mathcal{O}(dn)$	Mokhtari et al. (2018)
Cyclic SVRG	–	$\mathcal{O}\left(\kappa\sqrt{\kappa} \log \frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\frac{L}{\varepsilon}\right)$	$\mathcal{O}(d)$	this paper

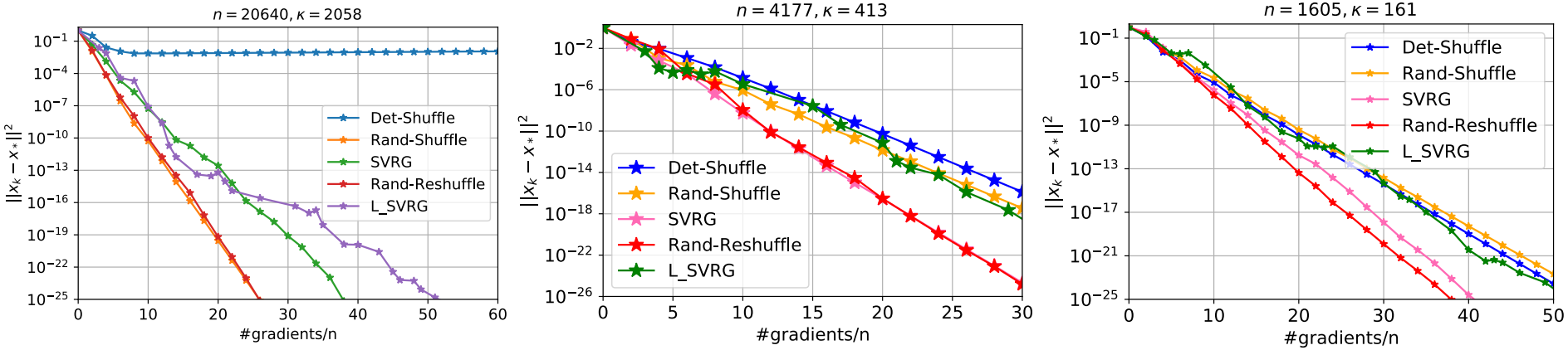
Comparison of the variance-reduced convergence results and implementations.

Experiments

In our experiments we solve the regularized ridge regression problem, which has the finite sum form with

$$f_i(x) = \frac{1}{2} \left\| A_{i,:}x - y_i \right\|^2 + \frac{\lambda}{2}\|x\|^2$$

where $A \in \mathbb{R}^{n \times d}$, $y \in \mathbb{R}^n$ and $\lambda > 0$ is a regularization parameter. Note that this problem is strongly convex and satisfies the smoothness assumption for $L = \max_i \left\| A_{i,:} \right\|^2 + \lambda$ and $\mu = \lambda_{\min}\left(A^\top A\right)/n + \lambda$ where $\lambda_{\min}$ is the smallest eigenvalue.



Comparison of methods on cadata, abalone and a1a datasets, we set the regularization constant $\lambda = 10/n$ and carefully chosen stepsizes.


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