ICML | 2019
Thirty-sixth International Conference on Machine Learning

SGD: General Analysis and Improved Rates

Peter Richtárik

Coauthors

Robert Gower

Nicolas Loizou

Xun Qian

Egor Shulgin

Alibek Sailanbayev

1. The Problem & Motivation

The Problem: Empirical Risk Minimization

f is μ–quasi strongly convex

\# training data

Smooth loss associated with data point *i*

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

Parameters describing the model

Motivation 1: Remove Strong Assumptions on Stochastic Gradients

- We get rid of unreasonable assumptions on the 2nd moment / variance of stochastic gradients:

$$\mathrm{E}\|g^k - \nabla f(x^k)\|^2 \leq \sigma^2$$

$$\mathrm{E}\|g^k\|^2 \leq \sigma^2$$

Lan, Nemirovski, Juditsky, Shapiro 2009

Such assumptions may not hold even for unconstrained minimization of strongly convex functions

Nguyen et al (ICML 2018)

Nguyen et al (arXiv:1811.12403)

- We do not need any assumptions!

Instead, we use expected smoothness assumption which follows from convexity and smoothness

Gower, Richtárik and Bach (arXiv:1706.01108)

Motivation 2: Develop SGD with Flexible Sampling Strategies

First analysis for SGD in the arbitrary sampling paradigm

(extends, simplifies and improves upon previous results)

Moulines & Bach (NIPS 2011)

Needell, Srebro and Ward (MAPR 2016)

Needell & Ward (2017)

Byproduct:

- First SGD analysis that recovers rate of GD in a special case
- First formula for optimal minibatch size for SGD
- Importance sampling for minibatch SGD

2. Stochastic Reformulation of Finite-Sum Problems

Stochastic Reformulation

Sampling vector $v = (v_1, \ldots, v_n)$

Random variable with mean 1

Linearity of expectation

$$f(x) \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} f_i(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[v_i] f_i(x) = \mathbb{E}\Bigg[\underbrace{\frac{1}{n}\sum_{i=1}^{n} v_i f_i(x)}_{f_v(x)}\Bigg]$$

Original Finite-Sum Problem

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

Stochastic Reformulation

$$\min_{x\in\mathbb{R}^d} \mathbb{E} f_v(x)$$

Minimizing the expectation over **random linear combinations** of the original functions

SGD Applied to Stochastic Reformulation

$$\min_{x\in\mathbb{R}^d} \mathbb{E}\left[f_v(x) \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n v_i f_i(x)\right]$$

stepsize

sample $v^k \sim \mathcal{D}$

$$x^{k+1} = x^k - \gamma^k \nabla f_{v^k}(x^k)$$

By varying $\mathcal{D}$, we obtain different existing and new variants of SGD

We perform a general analysis for any distribution $\mathcal{D}$

Stochastic Reformulations of Deterministic Problems: Related Work

Linear systems / convex quadratic minimization	Richtárik and Takáč (arXiv:1706.01108) **Stochastic reformulations of linear systems: algorithms and convergence theory**
Convex feasibility	Necoara, Patrascu and Richtárik (arXiv:1801.04873) **Randomized projection methods for convex feasibility problems: conditioning and convergence rates**
Variance reduction for finite-sum problems	Gower, Richtárik and Bach (arXiv:1706.01108) **Stochastic quasi-gradient methods: variance reduction via Jacobian sketching**

Sampling Without Replacement

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

$$S \subseteq \{1, 2, \dots, n\}$$

Random set $\tau \stackrel{\text{def}}{=} \mathbb{E}|S|$

$p_i \stackrel{\text{def}}{=} \mathrm{Prob}(i \in S)$

$$v_i = \begin{cases} \frac{1}{p_i} & i \in S \\ 0 & i \notin S \end{cases}$$

Sampling vector

$\mathrm{E}[v_i] = 1$

$$\nabla f_v(x) = \frac{1}{n}\sum_{i \in S} \frac{1}{p_i} \nabla f_i(x)$$

$\mathrm{E}[\nabla f_v(x)] = \nabla f(x)$

Minibatch SGD Without Replacement

$$x^{k+1} = x^k - \gamma^k \nabla f_{v^k}(x^k)$$

First time SGD is proposed and analyzed in the **arbitrary sampling** paradigm

Richtárik and Takáč (arXiv:1310.3438; Opt Letters 2016)

Example: Single Element Sampling

$$|S| = 1 \text{ with probability } 1$$

$$S = \begin{cases} \{1\} & \text{with probability} \quad p_1 \\ \{2\} & \text{with probability} \quad p_2 \\ & \vdots \\ \{n\} & \text{with probability} \quad p_n \end{cases}$$

SGD

$$x^{k+1} = x^k - \gamma^k \frac{1}{np_{i^k}} \nabla f_{i^k}(x^k)$$

Sampling With Replacement

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

$$s = \begin{cases} 1 & \text{with probability} \quad q_1 \\ 2 & \text{with probability} \quad q_2 \\ \vdots & \sum_{i=1}^n q_i = 1 \\ n & \text{with probability} \quad q_n \end{cases}$$

Sample several copies independently: $s_1, s_2 \dots, s_\tau$

$$v_i = \frac{1}{\tau q_i}\sum_{t=1}^{\tau} 1_{(s_t = i)}$$

Sampling vector
$\mathrm{E}[v_i] = 1$

$1_A = \begin{cases} 1 & \text{event } A \text{ holds} \\ 0 & \text{otherwise} \end{cases}$

$$\nabla f_v(x) = \frac{1}{n}\sum_{t=1}^{\tau} \frac{1}{\tau q_{s_t}} \nabla f_{s_t}(x)$$

$\mathrm{E}[\nabla f_v(x)] = \nabla f(x)$

Minibatch SGD With Replacement

$$x^{k+1} = x^k - \gamma^k \nabla f_{v^k}(x^k)$$

See also Algorithm 3 in Gorbunov et al (arXiv:1905.11261)

3. Expected Smoothness

Expected Smoothness

We will write: $(f, \mathcal{D}) \sim ES(\mathcal{L})$

Can hold as an identity for quadratics:

Richtárik and Takáč (1706.01108); Equation (30)

$\nabla f_v(x) = \frac{1}{n}\sum_{i=1}^n v_i \nabla f_i(x)$

Minimizer of *f*

$$\mathbb{E}\left[\|\nabla f_v(x) - \nabla f_v(x^*)\|^2\right] \leq 2\mathcal{L}\left(f(x) - f(x^*)\right)$$

Lemma f_i convex & L-smooth

$(f, \mathcal{D}) \sim ES(\mathcal{L}) \qquad \mathcal{L} = L \cdot \lambda_{\max}\left(\mathrm{E} v v^\top\right)$

Expected smoothness constant

See also: Gower, Bach & Richtárik (1805.02632); Section 3

Depends on *f* and *v*

A poor but simple bound
(we'll give much better bounds later)

Bounding the 2nd Moment

Gradient noise:

$\sigma^2 \overset{\text{def}}{=} \mathbb{E}\left[\|\nabla f_v(x^*)\|^2\right]$

Lemma $(f, \mathcal{D}) \sim ES(\mathcal{L})$

$$\mathbb{E}\left[\|\nabla f_v(x)\|^2\right] \leq 4\mathcal{L}\left(f(x) - f(x^*)\right) + 2\sigma^2$$

$\sigma^2 = 0$ ⇒ Weak growth condition

Richtárik and Takáč (1706.01108); Equation (30)

Nguyen et al (ICML 2018)

Vaswani, Bach and Schmidt (AISTATS 2019)

Generalization to proximal case (and variance reduction): $\min_{x\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n f_i(x) + \psi(x)$

Gorbunov et al (arXiv:1905.11261); Assumption 4.1

$\|\nabla f_v(x)\|^2$ ⇒ $\|\nabla f_v(x) - \nabla f(x^*)\|^2$

$f(x) - f(x^*)$ ⇒ $f(x) - f(x^*) - \langle \nabla f(x^*), x - x^* \rangle$

Computation of Expected Smoothness

Sampling (with Replacement)	Expected Smoothness	Expected Gradient Noise
General Random subset $S \subseteq \{1,2,\dots,n\}$	$c \equiv \frac{\mathbf{P}_{ij}}{p_i p_j} \; i \neq j$; f is L–smooth $L = \frac{1}{n}\sum_{i=1}^{n} L_i$; f_i is L_i–smooth $\mathcal{L} = cL + \frac{1}{n}\max_i \frac{(1-p_i c)\, L_i}{p_i}$	$\mathbf{P}_{ij} = \mathrm{Prob}(i,j \in S)$; $h_i = \nabla f_i(x^*)$ $\sigma^2 = \frac{1}{n^2}\sum_{i,j=1}^{n} \frac{\mathbf{P}_{ij}}{p_i p_j}\langle h_i, h_j \rangle$
Single Element $S = \{i\}$ with probability p_i	$\mathbf{P}_{ij} = 0 \Rightarrow c = 0$ $\mathcal{L} = \frac{1}{n}\max_i \frac{L_i}{p_i}$; $p_i = \mathrm{Prob}(i \in S)$	$\sigma^2 = \frac{1}{n^2}\sum_{i=1}^{n} \frac{1}{p_i}\|h_i\|^2$
Independent Minibatch $S = \bigcup_{i=1}^{n} S_i$; $S_i = \begin{cases} \{i\} & \text{with probability } p_i \\ \emptyset & \text{with probability } 1-p_i \end{cases}$ $S_1, \dots, S_n$ are independent	$\mathbf{P}_{ij} = p_i p_j \Rightarrow c = 1$ $\mathcal{L} = L + \frac{1}{n}\max_i \frac{(1-p_i)L_i}{p_i}$	$\sigma^2 = \frac{1}{n^2}\sum_{i=1}^{n} \frac{1-p_i}{p_i}\|h_i\|^2$
Uniform Minibatch S chosen uniformly random from all subsets of size τ	f is L–smooth; $\tau \overset{\text{def}}{=} \mathbf{E}[\lvert S \rvert] = \sum_i p_i$ $\mathcal{L} = \frac{n(\tau-1)}{\tau(n-1)}L + \frac{n-\tau}{\tau(n-1)}\max_i L_i$	$\sigma^2 = \frac{1}{n\tau} \cdot \frac{n-\tau}{n-1}\sum_{i=1}^{n}\|h_i\|^2$

4. Convergence Analysis: Linear Rate

Main Result (Linear Convergence to a Neighborhood of the Solution)

Assumption: f is μ–quasi strongly convex
$f(x^*) \geq f(x) + \langle \nabla f(x), x^* - x \rangle + \frac{\mu}{2}\|x^* - x\|^2$

Gradient noise:
$\sigma^2 \overset{\text{def}}{=} \mathbb{E}\left[\|\nabla f_v(x^*)\|^2\right]$

Theorem $(f, \mathcal{D}) \sim ES(\mathcal{L})$

$$\mathbb{E}\left\|x^k - x^*\right\|^2 \leq (1-\gamma\mu)^k \left\|x^0 - x^*\right\|^2 + \frac{2\gamma\sigma^2}{\mu}$$

Fixed stepsize: $\gamma^k \equiv \gamma \leq \frac{1}{2\mathcal{L}}$ | $\sigma = 0$ can choose $\gamma = \frac{1}{\mathcal{L}}$

Corollary $\gamma = \min\left\{\frac{1}{2\mathcal{L}}, \frac{\epsilon\mu}{4\sigma^2}\right\}$

$$k \geq \max\left\{\frac{2\mathcal{L}}{\mu}, \frac{4\sigma^2}{\epsilon\mu^2}\right\} \log\left(\frac{2\left\|x^0 - x^*\right\|^2}{\epsilon}\right)$$

$$\mathbb{E}\left\|x^k - x^*\right\|^2 \leq \epsilon$$

Optimal Minibatch Size

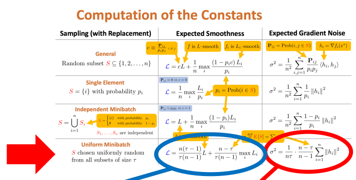
iterations

stochastic gradient evaluations in 1 iteration
$\tau = \mathbf{E}|S|$

$$\min_{1 \leq \tau \leq n} \mathcal{C}(\tau) \overset{\text{def}}{=} \max\left\{\frac{2\mathcal{L}}{\mu}, \frac{4\sigma^2}{\epsilon\mu^2}\right\} \times \tau$$

Corollary $\gamma = \min\left\{\frac{1}{2\mathcal{L}}, \frac{\epsilon\mu}{4\sigma^2}\right\}$

$$k \geq \max\left\{\frac{2\mathcal{L}}{\mu}, \frac{4\sigma^2}{\epsilon\mu^2}\right\} \log\left(\frac{2\left\|x^0 - x^*\right\|^2}{\epsilon}\right) \qquad \mathbb{E}\left\|x^k - x^*\right\|^2 \leq \epsilon$$



Optimal minibatches for different methods:

Qu et al (ICML 2016)

Bibi et al (arXiv:1806.05633)

$$\mathcal{L} = \frac{n(\tau-1)}{\tau(n-1)}L + \frac{n-\tau}{\tau(n-1)}\max_i L_i \qquad \sigma^2 = \frac{1}{n\tau} \cdot \frac{n-\tau}{n-1}\sum_{i=1}^{n}\|h_i\|^2$$

Optimal Minibatch Size

f is μ–quasi strongly convex

$f(x^*) \geq f(x) + \langle \nabla f(x), x^* - x\rangle + \frac{\mu}{2}\|x^* - x\|^2$

f is L–smooth

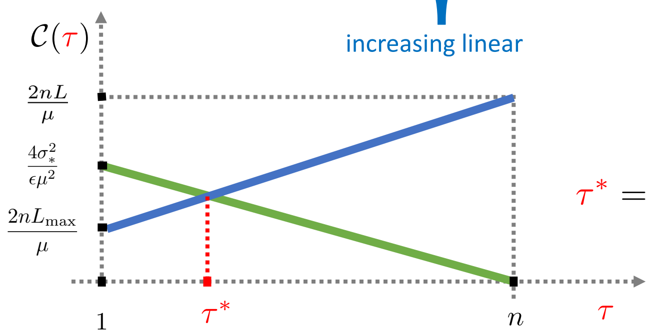
f_i is L_i–smooth

error tolerance

$\sigma_*^2 = \frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(x^*)\|^2$

$$\min_{1 \leq \tau \leq n} \mathcal{C}(\tau) \overset{\text{def}}{=} \frac{2}{\mu(n-1)} \max\left\{ \underbrace{n(\tau - 1)L + (n-\tau)\max_i L_i}_{\text{increasing linear}},\ \underbrace{(n-\tau)\frac{2\sigma_*^2}{\epsilon\mu}}_{\text{decreasing linear}} \right\}$$

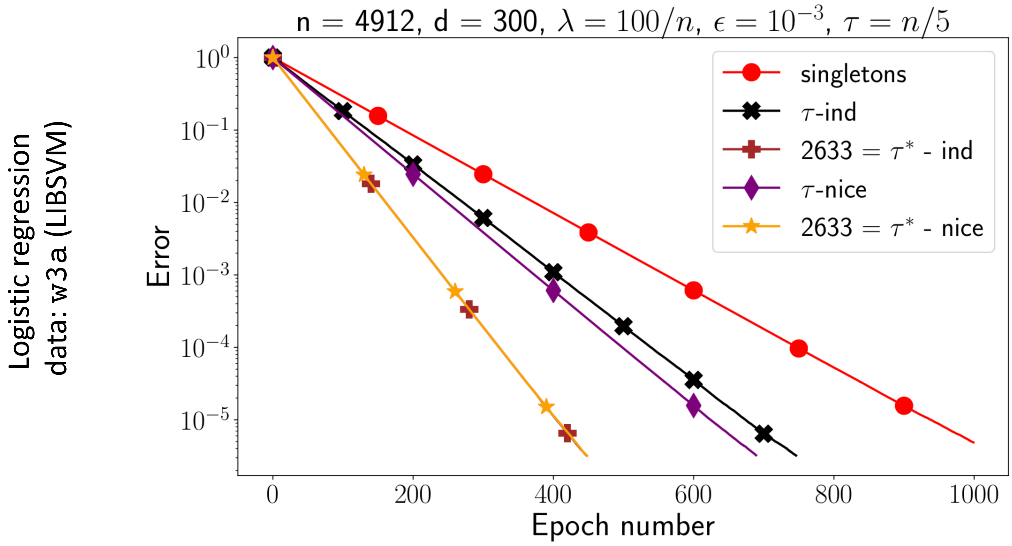
minibatch size



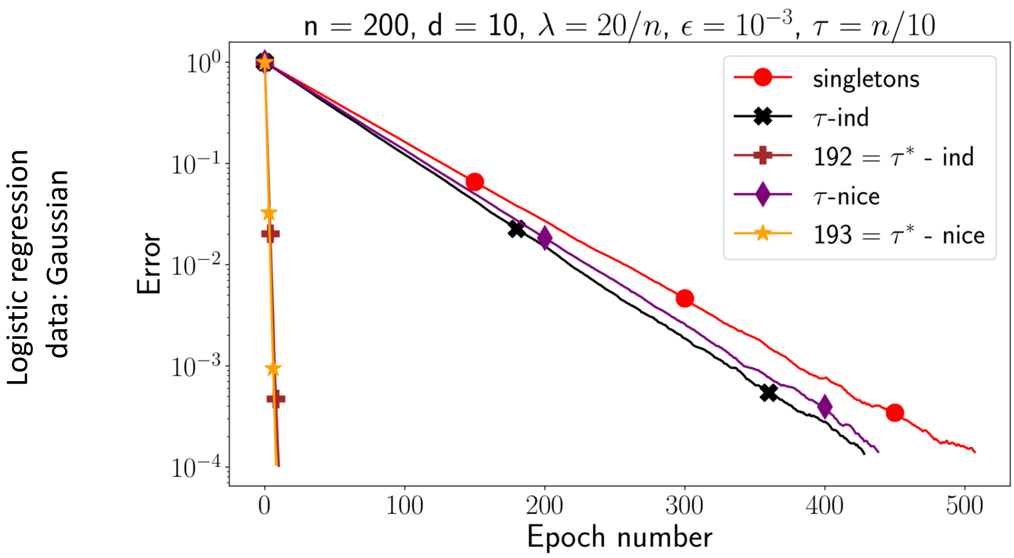
$$\tau^* = \frac{n(\theta + L - L_{\max})}{\theta + nL - L_{\max}}$$

$$\theta = \frac{2\sigma_*^2}{\epsilon\mu}$$

Optimal Minibatch Size: LIBSVM data



Optimal Minibatch Size: Synthetic Data



Importance Sampling for Minibatches

Details in: Paper

Richtárik and Takáč (Opt Let 2016) | Csiba and Richtárik (JMLR 2018) | Gower, Richtárik and Bach (arXiv:1805.02632) | Hanzely and Richtárik (AISTATS 2019)

5. Convergence Analysis: Sublinear Rate

Learning Schedule: Constant & Decreasing

Assumption: f is μ–quasi strongly convex

$$f(x^*) \geq f(x) + \langle \nabla f(x), x^* - x \rangle + \frac{\mu}{2} \|x^* - x\|^2$$

Theorem $(f, \mathcal{D}) \sim ES(\mathcal{L})$

$$\gamma^k = \begin{cases} \frac{1}{2\mathcal{L}} & \text{for } k \leq 4\lceil \mathcal{L}/\mu \rceil \\ \frac{2k+1}{(k+1)^2\mu} & \text{for } k > 4\lceil \mathcal{L}/\mu \rceil \end{cases}$$

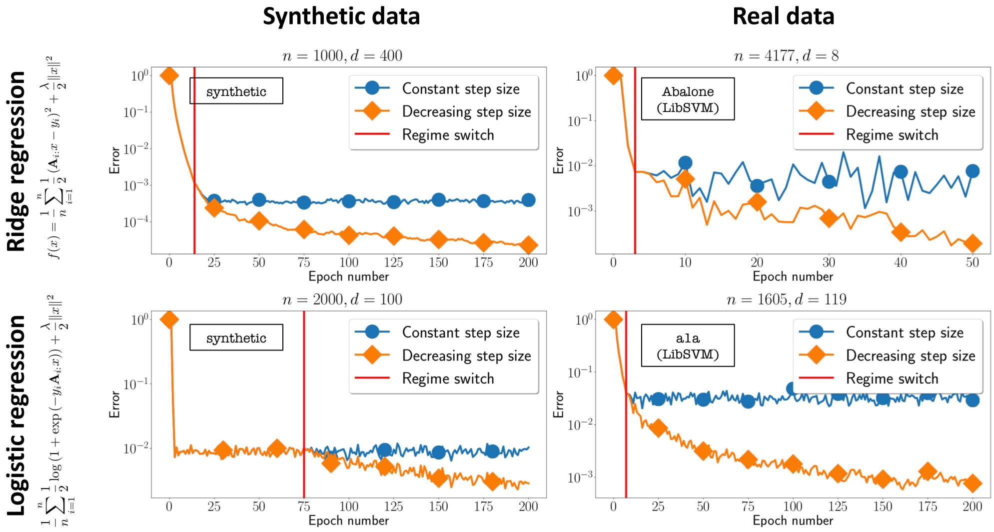
Gradient noise:

$$\sigma^2 \stackrel{\text{def}}{=} \mathbb{E}\left[\|\nabla f_v(x^*)\|^2\right]$$

$$\mathbb{E}\left\|x^k - x^*\right\|^2 \leq \frac{8\sigma^2}{\mu^2 k} + \frac{16\lceil \mathcal{L}/\mu \rceil^2}{e^2 k^2} \left\|x^0 - x^*\right\|^2$$

for $k \geq \frac{4\mathcal{L}}{\mu}$

Learning Schedule: Constant & Decreasing



Regularizer parameter: $\lambda = \frac{1}{n}$

6. Summary of Contributions

Summary of Contributions

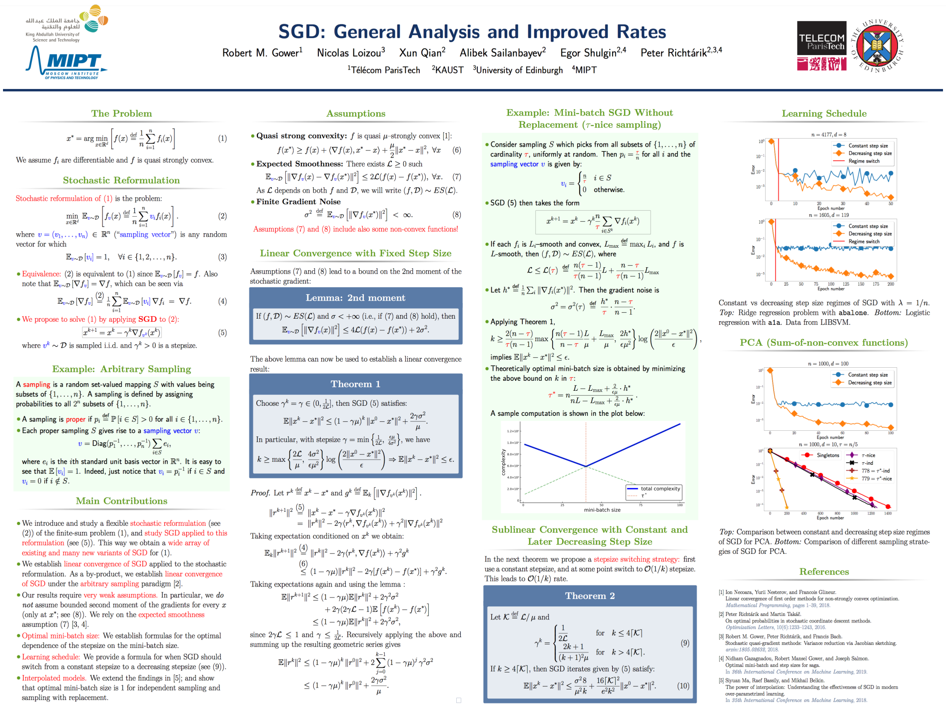
1. New conceptual tool: stochastic reformulation of finite-sum problems
2. First SGD analysis in the arbitrary sampling paradigm
3. Linear rate for smooth quasi-strongly functions to a neighborhood of the solution without the need for any noise assumptions!
4. First SGD analysis which recovers the rate for GD as a special case
5. First formulas for optimal minibatch size for SGD
6. First importance sampling for minibatches for SGD
7. A powerful learning schedule switching strategy with a sublinear rate
8. Tight extensions of previous results (Richárik-Takáč 2017, Viswani-Bach-Schmidt 2018)

Extra Material:
Brief History of Arbitrary Sampling

#	Paper	Algorithm	Comment
1	**R. & Takáč (OL 2016; arXiv 2013)** On optimal probabilities in stochastic coordinate descent methods	NSync	**Arbitrary sampling (AS) first introduced** Analysis of coordinate descent under strong convexity
2	**Qu, R. & Zhang (NeurIPS 2015)** Quartz: Randomized dual coordinate ascent with arbitrary sampling	QUARTZ	**First AS SGD method for min *P*** Primal-dual stochastic fixed point method; variance reduced
3	**Csiba & R. (arXiv 2015)** Primal method for ERM with flexible mini-batching schemes and non-convex losses	Dual-free SDCA	**First primal-only AS SGD method for min *P*** Variance-reduced
4	**Qu & R. (OMS 2016)** Coordinate descent with arbitrary sampling I: algorithms and complexity	ALPHA	**First accelerated coordinate descent method with AS** Analysis for smooth convex functions
5	**Qu & R. (OMS 2016)** Coordinate descent with arbitrary sampling II: expected separable overapproximation		**First dedicated study of ESO inequalities needed for analysis of AS methods** $\mathbb{E}_S\left[\left\|\sum_{i\in S}\mathbf{A}_i h_i\right\|^2\right] \leq \sum_{i=1}^n p_i v_i \|h_i\|^2$
6	**Chambolle, Ehrhardt, R. & Schoenlieb (SIOPT 2018)** Stochastic primal-dual hybrid gradient algorithm with arbitrary sampling and imaging applications	SPDHGM	**Chambolle-Pock method with AS**
7	**Hanzely, Mishchenko & R. (NeurIPS 2018)** SEGA: Variance reduction via gradient sketching	SEGA	**Variance-reduce coordinate descent with AS**
8	**Hanzely & R. (AISTATS 2019)** Accelerated coordinate descent with arbitrary sampling and best rates for minibatches	ACD	**First accelerated coordinate descent method with AS** Analysis for smooth strongly convex functions Importance sampling for minibatches
9	**Horváth & R. (ICML 2019)** Nonconvex variance reduced optimization with arbitrary sampling	SARAH, SVRG, SAGA	**First non-convex analysis of an AS method** First optimal mini-batch sampling
10	**Gower, Loizou, Qian, Sailanbayev, Shulgin & R. (ICML 2019)** SGD: general analysis and improved rates	SGD-AS	**First AS variant of SGD (without variance reduction)** Optimal minibatch size
11	**Qian, Qu & R. (ICML 2019)** **SAGA with arbitrary sampling**	**SAGA-AS**	**First AS variant of SAGA**

The End

POSTER #195


SGD: General Analysis and Improved Rates
Robert M. Gower[1] Nicolas Loizou[3] Xun Qian[2] Alibek Sailanbayev[2] Egor Shulgin[2,4] Peter Richtárik[2,3,4]
[1]Télécom ParisTech [2]KAUST [3]University of Edinburgh [4]MIPT
The Problem
Stochastic Reformulation
Example: Arbitrary Sampling
Main Contributions
Assumptions
Linear Convergence with Fixed Step Size
Lemma: 2nd moment
Theorem 1
Example: Mini-batch SGD Without Replacement (τ-nice sampling)
Sublinear Convergence with Constant and Later Decreasing Step Size
Theorem 2
Learning Schedule
PCA (Sum-of-non-convex functions)
References