

**ICML | 2019**

Thirty-sixth International Conference on  
Machine Learning

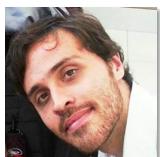


# SGD: General Analysis and Improved Rates

Peter Richtárik



## Coauthors



Robert Gower



Nicolas Loizou



Xun Qian



Egor Shulgin



Alibek Sailanbayev



# 1. The Problem & Motivation

## The Problem: Empirical Risk Minimization

$f$  is  $\mu$ -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  $2^{\text{nd}}$  moment / variance of stochastic gradients:

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

$$\mathbb{E}\|g^k\|^2 \leq \sigma^2 \quad \text{Lan, Nemirovski, Juditsky, Shapiro 2009}$$

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

$$f(x) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n f_i(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[v_i] f_i(x) = \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n v_i f_i(x) \right]$$

Random variable with mean 1      Linearity of expectation      Sampling vector  $v = (v_1, \dots, v_n)$

#### Original Finite-Sum Problem

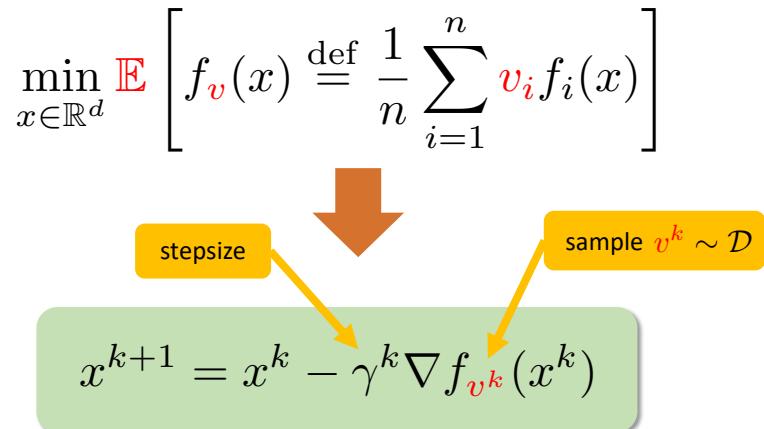
$$\min_{x \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n f_i(x)$$

#### 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



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

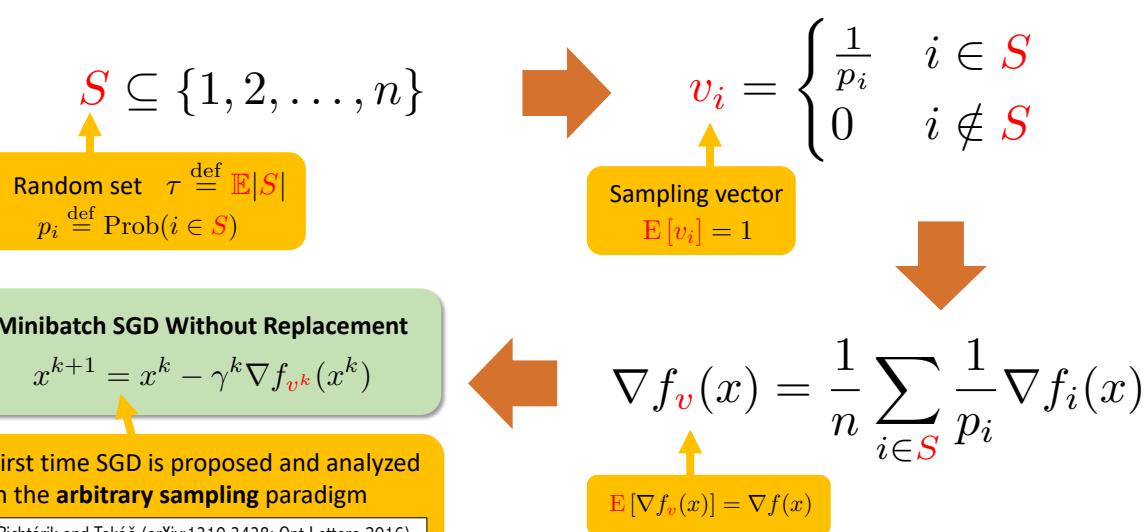
Variance reduction for finite-sum problems



Gower, Richtárik and Bach (arXiv:1706.01108)  
Stochastic quasi-gradient methods: variance reduction via Jacobian sketching

## Sampling Without Replacement

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



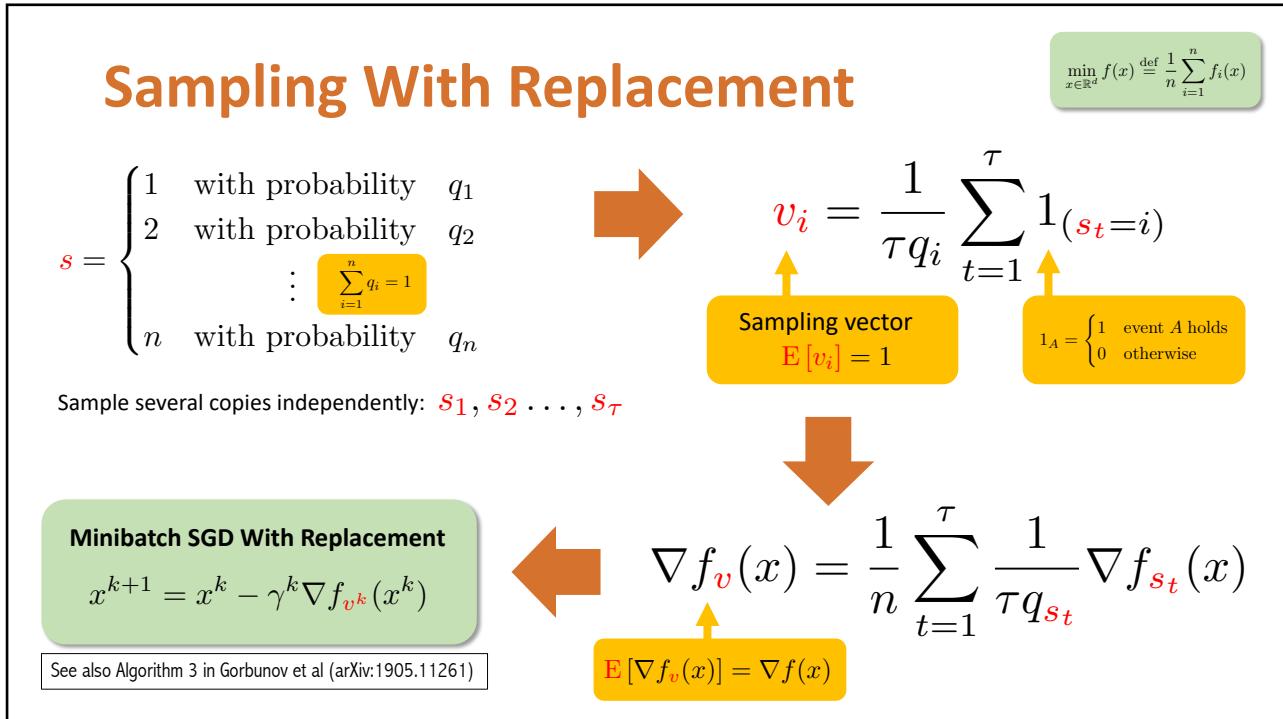
## Example: Single Element Sampling

$|S| = 1$  with probability 1

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

SGD

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



### 3. Expected Smoothness

## Expected Smoothness

$$\nabla f_{\mathbf{v}}(x) = \frac{1}{n} \sum_{i=1}^n \mathbf{v}_i \nabla f_i(x)$$

Minimizer of  $f$

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

**Lemma**  $f_i$  convex &  $L$ -smooth



$$(f, \mathcal{D}) \sim ES(\mathcal{L}) \quad \mathcal{L} = L \cdot \lambda_{\max} (\mathbf{E} \mathbf{v} \mathbf{v}^\top)$$

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

Can hold as an identity for quadratics:

Richtárik and Takáč (1706.01108); Equation (30)

Expected smoothness constant

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

Depends on  $f$  and  $\mathbf{v}$

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

## Bounding the 2<sup>nd</sup> Moment

**Lemma**

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



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

Gradient noise:

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

$$\sigma^2 = 0$$

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

$$\|\nabla f_{\mathbf{v}}(x)\|^2 \rightarrow \|\nabla f_{\mathbf{v}}(x) - \nabla f(x^*)\|^2$$

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

## Computation of Expected Smoothness

Sampling (with Replacement)	Expected Smoothness	Expected Gradient Noise
<b>General</b> $c \equiv \frac{\mathbf{P}_{ij}}{p_i p_j} \quad i \neq j$ Random subset $\mathcal{S} \subseteq \{1, 2, \dots, n\}$	$f$ is $L$ -smooth $L = \frac{1}{n} \sum_{i=1}^n L_i$ $f_i$ is $L_i$ -smooth $\mathcal{L} = cL + \frac{1}{n} \max_i \frac{(1 - p_i c) L_i}{p_i}$	$\mathbf{P}_{ij} = \text{Prob}(i, j \in \mathcal{S})$ $h_i = \nabla f_i(x^*)$ $\sigma^2 = \frac{1}{n^2} \sum_{i,j=1}^n \mathbf{P}_{ij} \langle h_i, h_j \rangle$
<b>Single Element</b> $\mathcal{S} = \{i\}$ with probability $p_i$	$\mathbf{P}_{ij} = 0 \Rightarrow c = 0$ $\mathcal{L} = \frac{1}{n} \max_i \frac{L_i}{p_i}$ $p_i = \text{Prob}(i \in \mathcal{S})$	$\sigma^2 = \frac{1}{n^2} \sum_{i=1}^n \frac{1}{p_i} \ h_i\ ^2$
<b>Independent Minibatch</b> $\mathcal{S} = \bigcup_{i=1}^n \mathcal{S}_i$ $\mathcal{S}_i = \begin{cases} \{i\} & \text{with probability } p_i \\ \emptyset & \text{with probability } 1 - p_i \end{cases}$ $\mathcal{S}_1, \dots, \mathcal{S}_n$ are independent	$\mathbf{P}_{ij} = p_i p_j \Rightarrow c = 1$ $\mathcal{L} = L + \frac{1}{n} \max_i \frac{(1 - p_i) L_i}{p_i}$ $f$ is $L$ -smooth	$\sigma^2 = \frac{1}{n^2} \sum_{i=1}^n \frac{1 - p_i}{p_i} \ h_i\ ^2$
<b>Uniform Minibatch</b> $\mathcal{S}$ chosen uniformly random from all subsets of size $\tau$	$\mathcal{L} = \frac{n(\tau - 1)}{\tau(n - 1)} L + \frac{n - \tau}{\tau(n - 1)} \max_i L_i$	$\sigma^2 = \frac{1}{n\tau} \cdot \frac{n - \tau}{n - 1} \sum_{i=1}^n \ h_i\ ^2$

## 4. Convergence Analysis: Linear Rate

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

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

Gradient noise:  
 $\sigma^2 \stackrel{\text{def}}{=} \mathbb{E} [\|\nabla f_{\mathbf{v}}(x^*)\|^2]$

**Theorem**

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

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

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

**Corollary**

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

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

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

## Optimal Minibatch Size

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

# iterations

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

**Corollary**

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

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

Optimal minibatches for different methods:

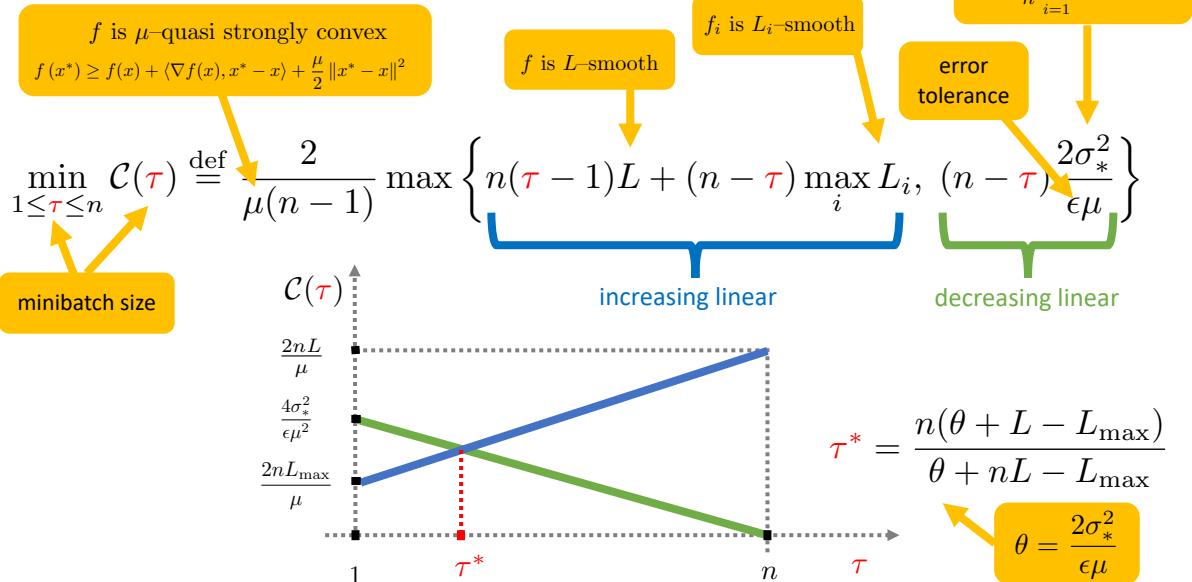
Qu et al (ICML 2016)

Bibi et al (arXiv:1806.05633)

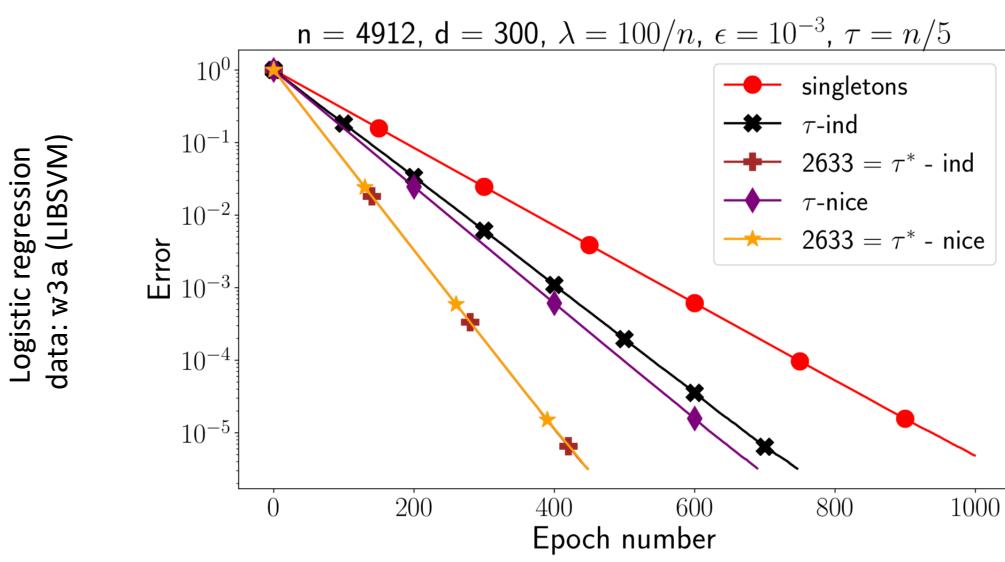
$$\mathcal{L} = \frac{n(\tau-1)}{\tau(n-1)} L + \frac{n-\tau}{\tau(n-1)} \max_i L_i \quad \sigma^2 = \frac{1}{n\tau} \cdot \frac{n-\tau}{n-1} \sum_{i=1}^n \|h_i\|^2$$

Computation of the Constants		
Sampling (with Replacement)	Expected Smoothness	Expected Gradient Noise
General	$L = \max_i L_i$	$\sigma^2 = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[h_i^T h_i]$
Random subset $S \subseteq \{1, 2, \dots, n\}$	$L = \max_i \frac{L_i}{P(S_i)}$	$\sigma^2 = \frac{1}{n} \sum_{i=1}^n P(S_i) h_i^T h_i$
Single Element $S = \{i\}$ with probability $p_i$	$L = \frac{1}{p_i} L_i$	$\sigma^2 = \frac{1}{\sum p_i} \sum p_i h_i^T h_i$
Independent Minibatch $S = \bigcup_{i=1}^m S_i$ where $S_i$ are independent	$L = L_i$	$\sigma^2 = \frac{1}{m} \sum_{i=1}^m \mathbb{E}[h_i^T h_i]$
Uniform Minibatch $S$ chooses uniformly random from all subsets of size $\tau$	$L = \frac{1}{\binom{n}{\tau}} \sum_{S \in \binom{n}{\tau}} \max_i L_i$	$\sigma^2 = \frac{1}{n\tau} \cdot \frac{n-\tau}{n-1} \sum_{i=1}^n \ h_i\ ^2$

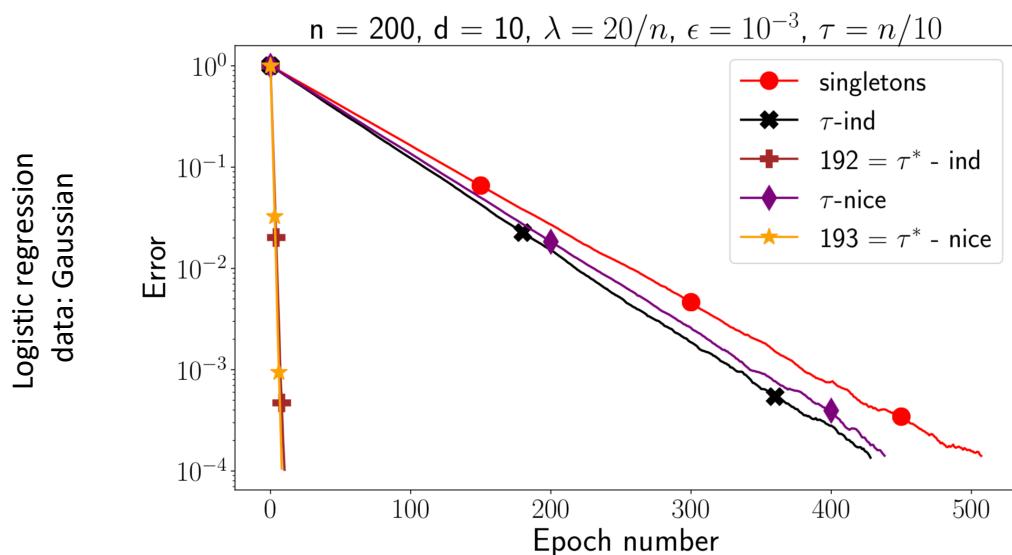
## Optimal Minibatch Size



## Optimal Minibatch Size: LIBSVM data



## Optimal Minibatch Size: Synthetic Data



## Importance Sampling for Minibatches

Details in: Paper

## 5. Convergence Analysis: Sublinear Rate

### Learning Schedule: Constant & Decreasing

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

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

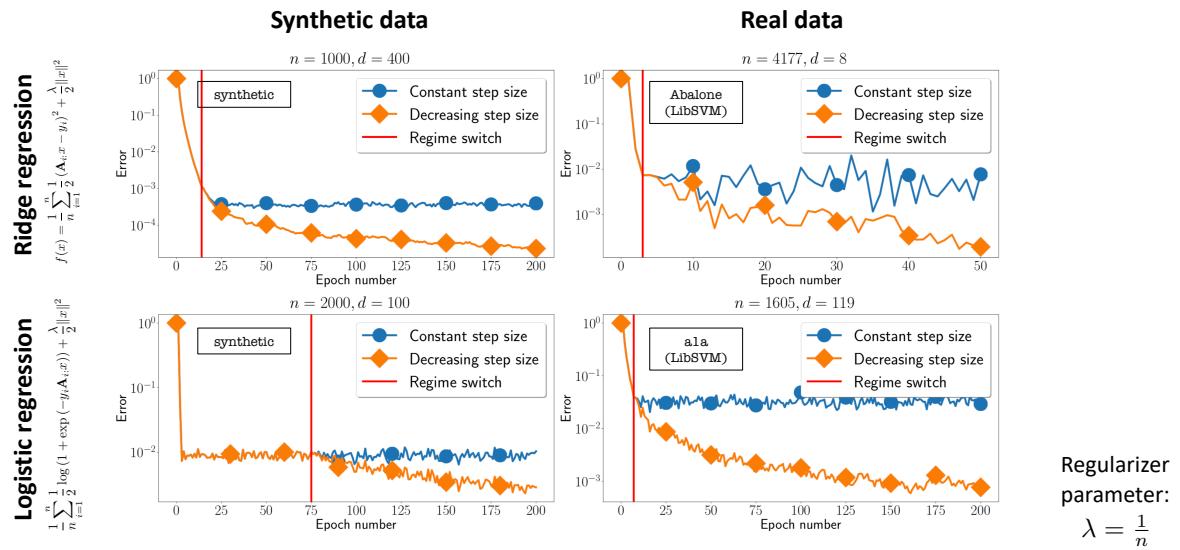
$$\gamma^k = \begin{cases} \frac{1}{2\mathcal{L}} & \text{for } k \leq 4 \lceil \mathcal{L}/\mu \rceil \\ \frac{2k+1}{(k+1)^2\mu} & \text{for } k > 4 \lceil \mathcal{L}/\mu \rceil \end{cases}$$

Gradient noise:  
 $\sigma^2 \stackrel{\text{def}}{=} \mathbb{E} [\|\nabla f_{\nu}(x^*)\|^2]$

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

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

## Learning Schedule: Constant & Decreasing



## 6. Summary of Contributions

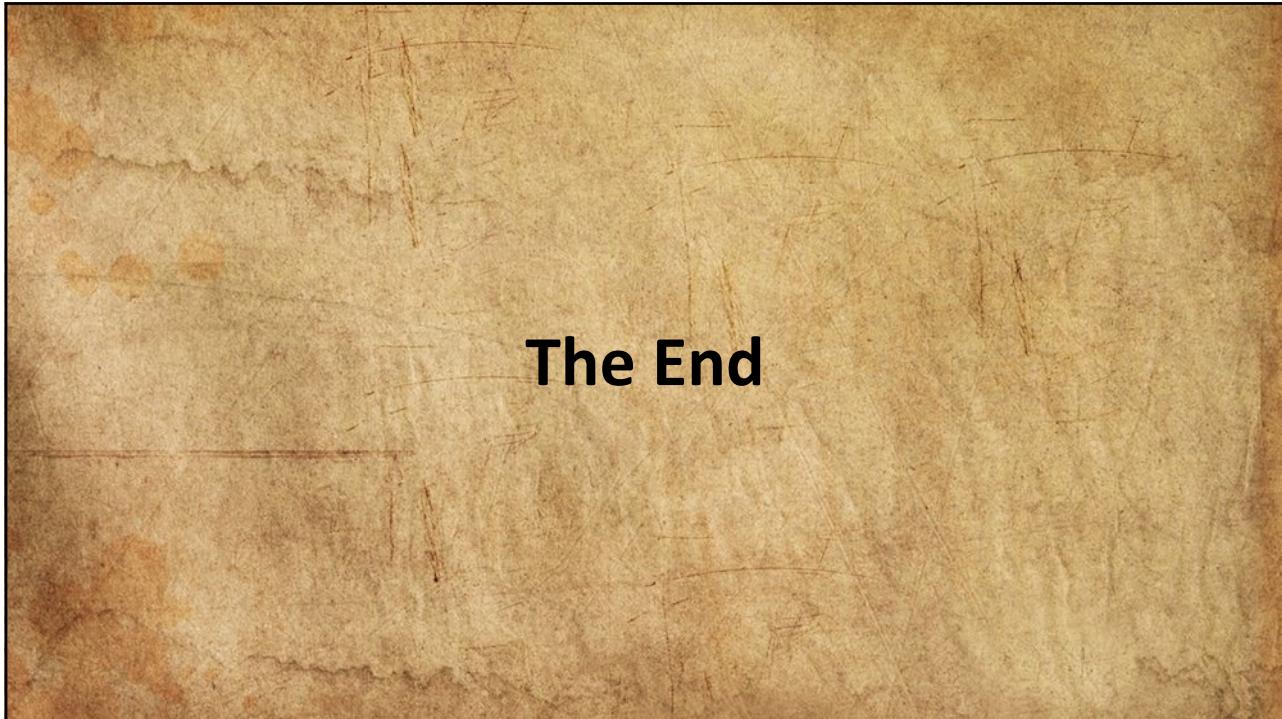
## Summary of Contributions

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 $\mathbb{E}_{\mathcal{S}} \left[ \left\  \sum_{i \in \mathcal{S}} \mathbf{A}_i h_i \right\ ^2 \right] \leq \sum_{i=1}^n p_i v_i \ h_i\ ^2$ needed for analysis of AS methods
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, Sainanbayev, 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<sup>1</sup> Nicolas Loizou<sup>3</sup> Xun Qian<sup>2</sup> Alibek Salamatbayev<sup>2</sup> Egor Shulgin<sup>2,4</sup> Peter Richtárik<sup>2,3,4</sup>

<sup>1</sup>Télécom ParisTech <sup>2</sup>KAUST <sup>3</sup>University of Edinburgh <sup>4</sup>MIPT

**The Problem**

$$x^* = \arg \min_{x \in \mathbb{R}^n} f(x) \triangleq \frac{1}{n} \sum_{i=1}^n f_i(x) \quad (1)$$

We assume  $f_i$  are differentiable and  $f$  is quasi strongly convex.

**Stochastic Reformulation**

Stochastic reformulation of (1) is the problem:

$$\min_{x \in \mathbb{R}^n} \mathbb{E}_{\mathcal{S} \sim \mathcal{D}} \left[ f_i(x) + \frac{1}{n} \sum_{i \in S} v_i f_i(x) \right], \quad (2)$$

where  $v = (v_1, \dots, v_n) \in \mathbb{R}^n$  ("sampling vector") is any random vector for which

$$\mathbb{E}_{v \sim \mathcal{D}} [v] = 1, \quad \forall i \in \{1, 2, \dots, n\}. \quad (3)$$

- **Equivalence:** (2) is equivalent to (1) since  $\mathbb{E}_{v \sim \mathcal{D}} [f_i] = f$ . Also note that  $\mathbb{E}_{v \sim \mathcal{D}} [\nabla f_i] = \nabla f$ , where  $\nabla f$  is defined by averaging probabilities to all  $2^n$  subsets of  $\{1, \dots, n\}$ .
- A sampling is *proper* if  $\mathbb{P}[i \in S] > 0$  for all  $i \in \{1, \dots, n\}$ .
- Each proper sampling  $S$  gives rise to a sampling vector  $v$ :
$$v = \text{Diag}(p_1, \dots, p_n) \sum_{i \in S} e_i,$$

where  $e_i$  is the  $i$ th standard basis vector in  $\mathbb{R}^n$ . It is easy to see that  $\mathbb{E}[v] = 1$ . Indeed, just notice that  $v_i = p_i^{-1}$  if  $i \in S$  and  $v_i = 0$  if  $i \notin S$ .

**Main Contributions**

- We introduce and study a flexible stochastic reformulation (see (2)) of the finite-sum problem (1), and study SGD applied to this reformulation (see (5)). This way we obtain a wide array of existing and many new variants of SGD for (1).
- We propose a general analysis of SGD applied to the stochastic reformulation. As a by-product, we establish linear convergence of SGD under the arbitrary sampling paradigm [2].
- Our results require very weak assumptions. In particular, we do not require  $\mathcal{D}$  to be a uniform distribution of the elements for every  $x$  (only at  $x^*$ ; see (8)). We rely on the **expected smoothness** assumption (7) [4, 5].
- **Sampling analysis.** We establish formulas for the optimal dependence of the stepsize on the mini-batch size.
- **Learning schedule.** We provide a formula for when SGD should switch from a constant stepsize to a decreasing stepsize (see (9)).
- **Optimal stepsize.** We establish that in (5) and show that optimal mini-batch size is 1 for independent sampling and sampling with replacement.

**Assumptions**

- **Quasi strong convexity:**  $f$  is quasi  $\mu$ -strongly convex [1]:
$$f(x^*) \geq f(x) + \langle \nabla f(x^*), x^* - x \rangle + \frac{\mu}{2} \|x^* - x\|^2, \quad \forall x. \quad (6)$$

- **Expected Smoothness:** There exists  $L \geq 0$  such that
$$\mathbb{E}_{v \sim \mathcal{D}} [\|\nabla f(x) - \nabla f_i(x^*)\|^2] \leq L(\|x - x^*\|), \quad \forall x.$$

As  $\mathcal{L}$  depends on both  $f$  and  $\mathcal{D}$ , we will write  $(f, \mathcal{D}) \sim \mathcal{ES}(\mathcal{L})$ .

**Finite Gradient Noise**

$$\sigma^2 \triangleq \mathbb{E}_{v \sim \mathcal{D}} [\|\nabla f_i(x^*)\|^2] < \infty. \quad (8)$$

Assumptions (7) and (8) include some non-convex function!

**Linear Convergence with Fixed Step Size**

Assumption (7) and (8) lead to a bound on the 2nd moment of the stochastic gradient:

$$\mathbb{E}_{v \sim \mathcal{D}} [\nabla f_i] \stackrel{(7)}{=} \mathbb{E}_{v \sim \mathcal{D}} [\nabla f] = \nabla f. \quad (4)$$

• We propose (1) by applying SGD to (2)

$$x^{k+1} = x^k - \gamma \nabla f_i(x^k) \quad (5)$$

where  $\gamma^k \sim \mathcal{D}$  is sampled i.i.d. and  $\gamma^k > 0$  is a stepsize.

**Example: Arbitrary Sampling**

A sampling  $S$  is a random set-valued mapping  $S$  with values being subsets of  $\{1, \dots, n\}$ . A sampling  $v$  is defined by assigning probabilities to all  $2^n$  subsets of  $\{1, \dots, n\}$ .

• A sampling is *proper* if  $\mathbb{P}[i \in S] > 0$  for all  $i \in \{1, \dots, n\}$ .

• Each proper sampling  $S$  gives rise to a sampling vector  $v$ :

$$v = \text{Diag}(p_1, \dots, p_n) \sum_{i \in S} e_i,$$

where  $e_i$  is the  $i$ th standard basis vector in  $\mathbb{R}^n$ . It is easy to see that  $\mathbb{E}[v] = 1$ . Indeed, just notice that  $v_i = p_i^{-1}$  if  $i \in S$  and  $v_i = 0$  if  $i \notin S$ .

**Assumptions**

- Consider sampling  $S$  which picks from all subsets of  $\{1, \dots, n\}$  of cardinality  $r$ , uniformly at random. Then  $p_i = \frac{1}{\binom{n}{r}}$  for all  $i$  and the sampling vector  $v$  is given by:
$$v_i = \begin{cases} \frac{1}{\binom{n}{r}} & \text{if } i \in S \\ 0 & \text{otherwise.} \end{cases}$$

**SGD (5) then takes the form**

$$x^{k+1} = x^k - \gamma^k \sum_{i \in S} \nabla f_i(x^k) \quad (8)$$

- If each  $f_i$  is  $L$ -smooth and  $\mathcal{D}$  is uniform,  $L_{\max} \triangleq \max_i L_i$ , and  $f$  is  $L$ -smooth, then  $(f, \mathcal{D}) \sim \mathcal{ES}(L)$ , where
$$L \leq \mathbb{E}[L] \triangleq \frac{n(r-1)L}{\binom{n}{r}(r-1)} L < \frac{n-r}{\binom{n}{r}(r-1)} L_{\max}$$

**Learning Schedule**

Top: Ridge regression step size regime of SGD with  $\lambda = 1/n$ . Bottom: Logistic regression with afa. Data from LIBSVM.

**Constant vs Decreasing Step Size Regime of SGD with  $\lambda = 1/n$**

Constant: Ridge regression step size regime of SGD with  $\lambda = 1/n$ . Decreasing: Logistic regression with afa. Data from LIBSVM.

**PCA (Sum-of-non-convex functions)**

**Sublinear Convergence with Constant and Later Decreasing Step Size**

In the next theorem we propose a *stepsize switching strategy*: first use a constant stepsize, and at some point switch to  $\mathcal{O}(1/k)$  stepsize. This leads to  $\mathcal{O}(1/k)$  rate.

**Theorem 2**

Let  $K \triangleq \mathcal{L}/\mu$  and

$$\gamma^k = \begin{cases} \frac{1}{2K} & \text{for } k \leq 4(K) \\ \frac{k+1}{(k+1)\mu} & \text{for } k > 4(K). \end{cases} \quad (9)$$

If  $k \geq 4(K)$ , then SGD iterates given by (5) satisfy:

$$\mathbb{E}[x^{k+1}] \leq (1 - \gamma \mu) \mathbb{E}[x^k]^2 + \frac{2\sigma^2}{\mu} + \frac{16[K]^2}{\mu^2} \|x^* - x^k\|^2. \quad (10)$$

**References**

- [1] Iain Horner, Yann Nesterov, and Francis Bach. Linear convergence of first order methods for non-strongly convex optimization. *Mathematics of Computation*, 84(291):1425–1448, 2015.
- [2] Peter Richtárik and Martin Takáč. On optimal algorithms for distributed stochastic optimization. *Mathematics of Computation*, 83(287):1473–1517, 2014.
- [3] Robert M. Gower, Peter Richtárik, and Francis Bach. Stochastic gradient descent, steepest descent, and accelerated gradient descent. *Mathematics of Computation*, 84(291):1615–1637, 2015.
- [4] Volodymyr Gaspinov, Robert M. Gower, and Joseph Schone. Optimal step size and step-size rule for SGD. *Machine Learning*, 103(10):2033–2055, 2014.
- [5] Sylvain Ma, Béatrice Bouchard, and Mikael Belil. On the stepsize rule for stochastic gradient descent. In *Advances in Neural Information Processing Systems 26*, pages 2656–2664. Neural Information Processing Systems Foundation, 2013.

**POSTER #195**

17