

ICML | 2019

Thirty-sixth International Conference on Machine Learning

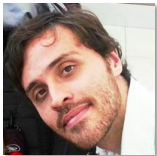


SGD: General Analysis and Improved Rates

Peter Richtárik



Coauthors



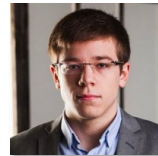
Robert Gower



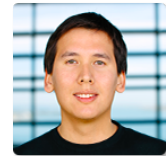
Nicolas Loizou



Xun Qian



Egor Shulgin

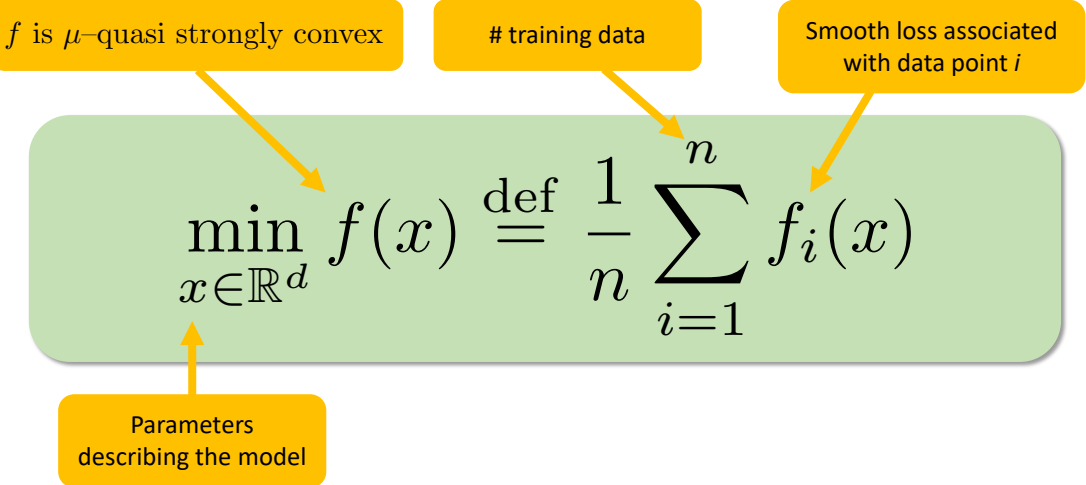


Alibek Sailanbayev



1. The Problem & Motivation

The Problem: Empirical Risk Minimization



Motivation 1: Remove Strong Assumptions on Stochastic Gradients

- We get rid of **unreasonable assumptions** on the 2nd 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 $v = (v_1, \dots, v_n)$
 $f_v(x)$

Original Finite-Sum Problem

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



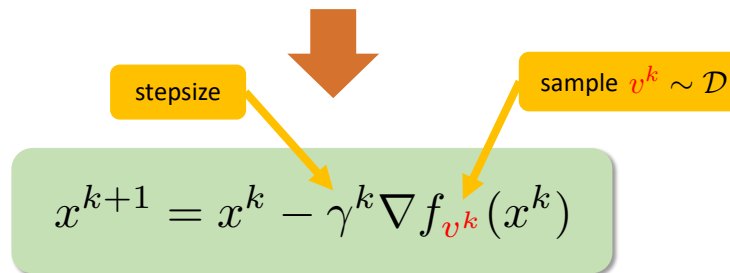
Stochastic Reformulation

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

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

SGD Applied to Stochastic Reformulation

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



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)

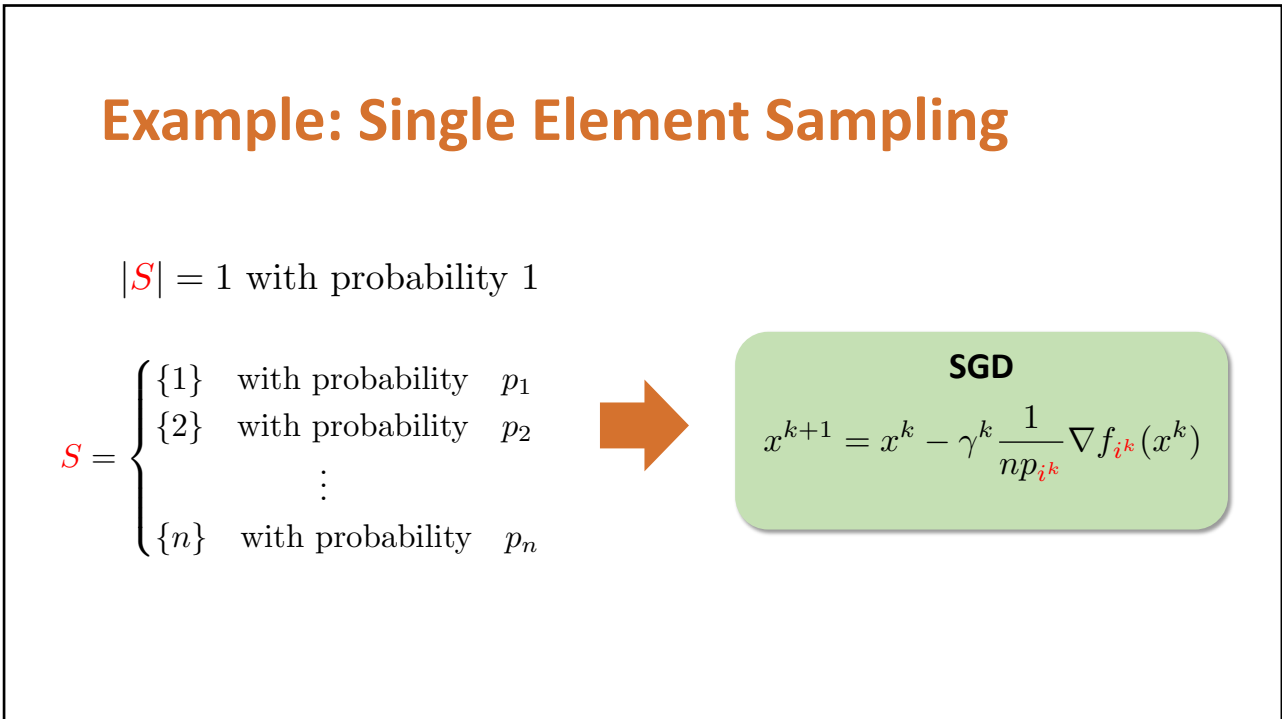
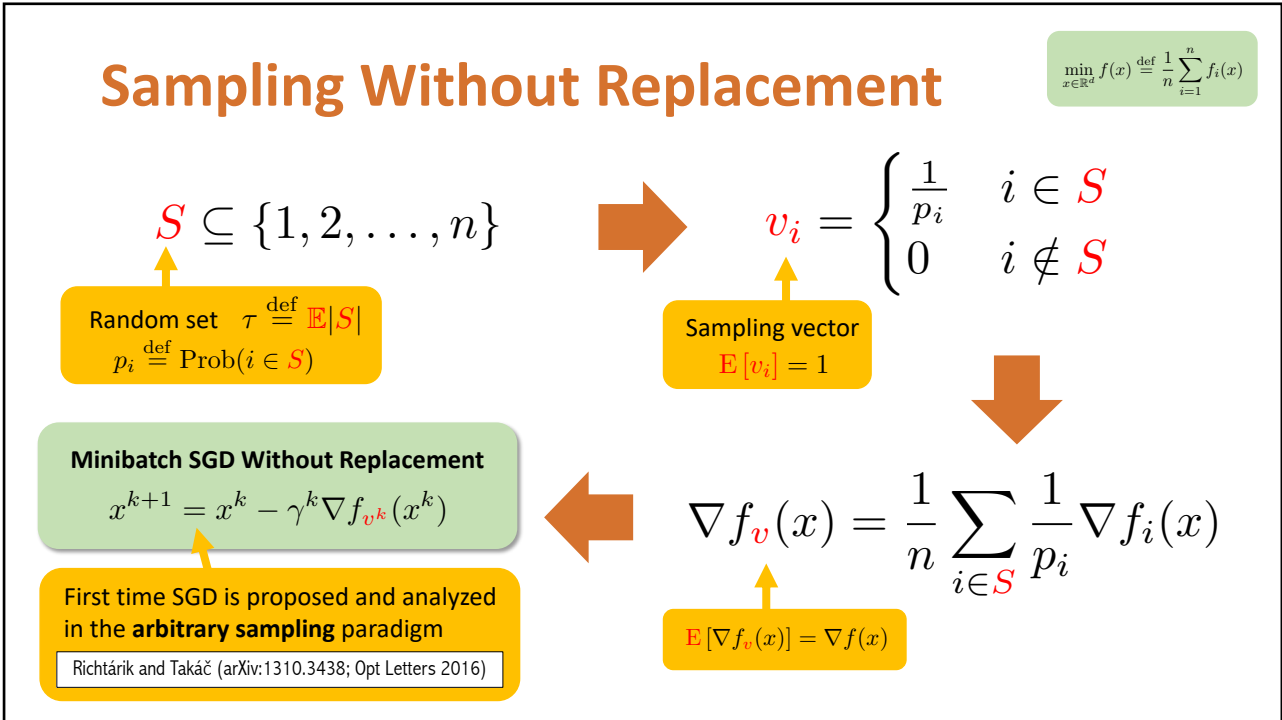
Randomized projection methods for convex feasibility problems: conditioning and convergence rates

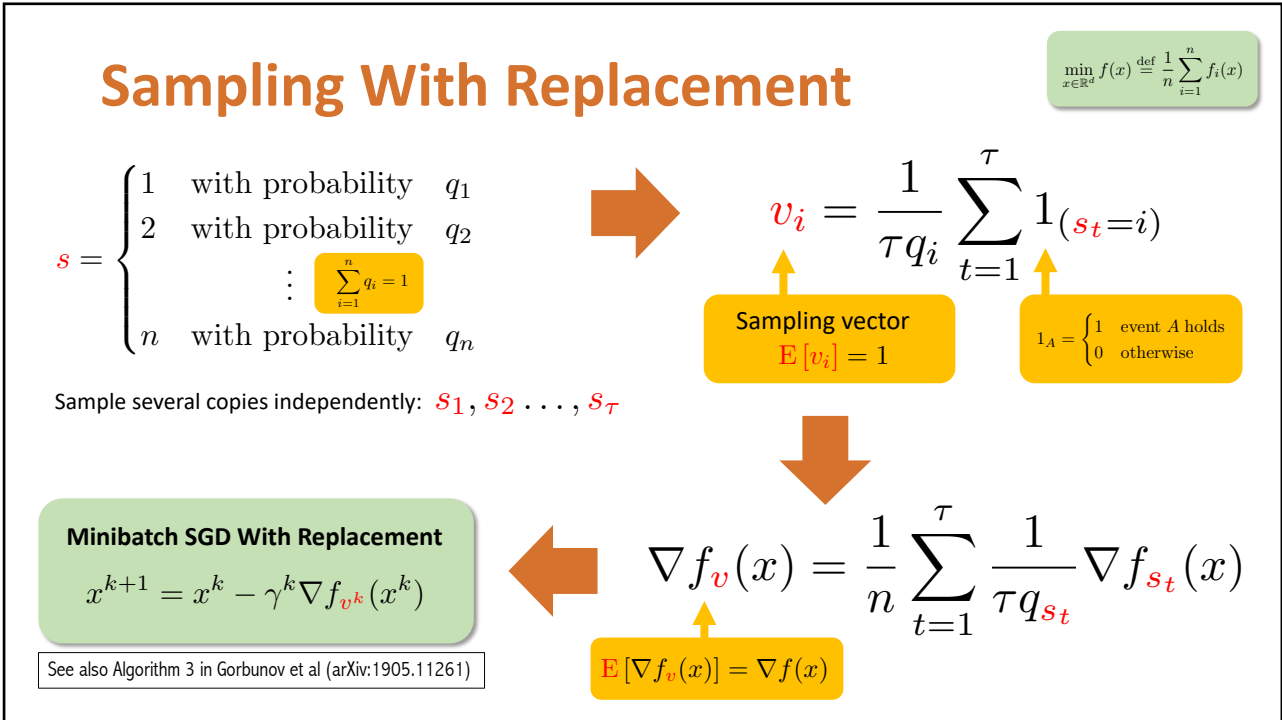
Variance reduction for finite-sum problems



Gower, Richtárik and Bach (arXiv:1706.01108)

Stochastic quasi-gradient methods: variance reduction via Jacobian sketching





3. Expected Smoothness

Expected Smoothness

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

Minimizer of f

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

Can hold as an identity for quadratics:
Richtárik and Takáč (1706.01108); Equation (30)

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

Lemma f_i convex & L -smooth

↓

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

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

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

↓

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

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

$\sigma^2 = 0 \rightarrow$ 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

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

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

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

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

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 $\gamma = \frac{1}{\mathcal{L}}$

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

iterations

stochastic gradient evaluations in 1 iteration

$$\tau = \mathbb{E}|S|$$

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

Corollary

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

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

Computation of the Constants

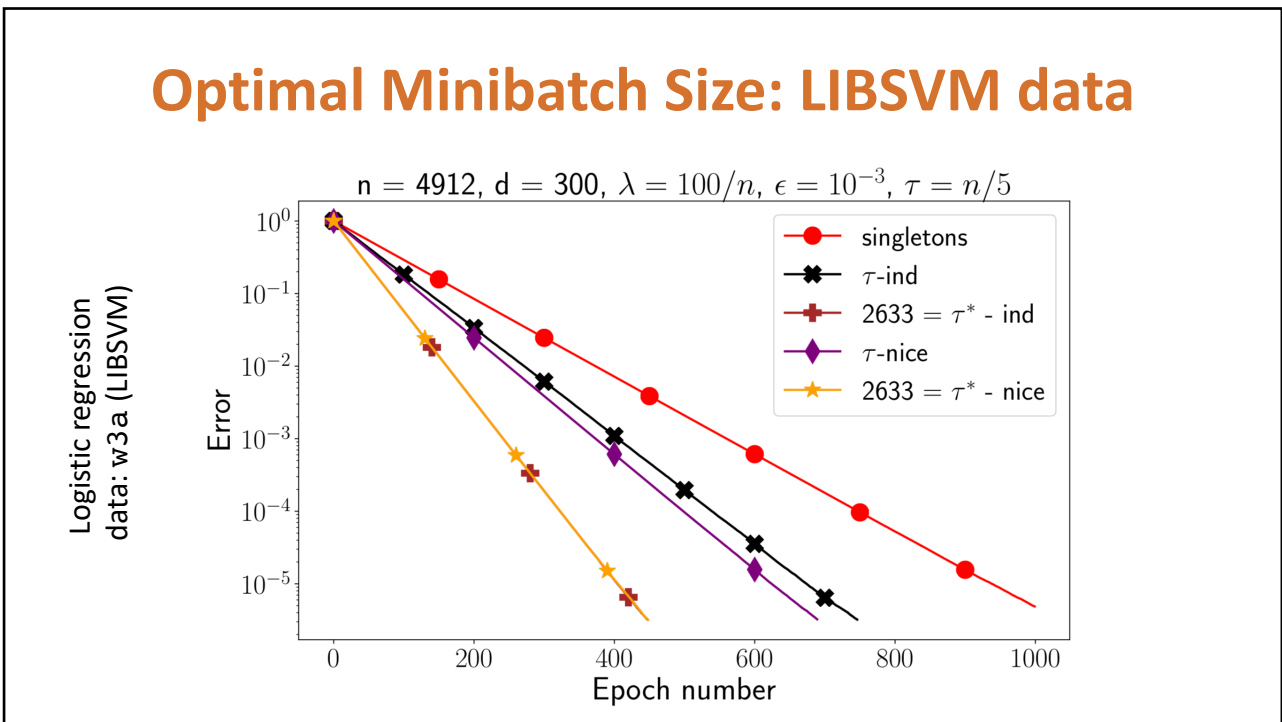
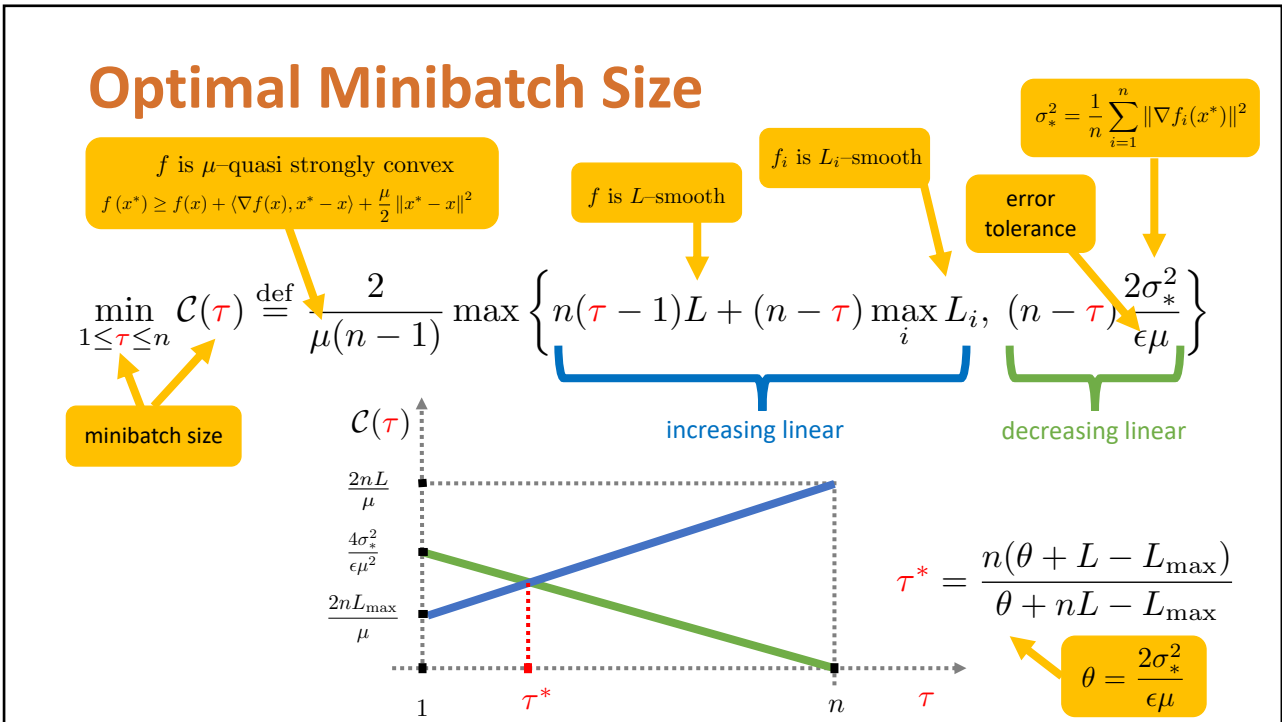
Sampling (with Replacement)	Expected Smoothness	Expected Gradient Noise
General Random subset $S \subseteq \{1, 2, \dots, n\}$	$\mathcal{L} = \frac{1}{\tau} \max_{i \in S} L_i$	$\sigma^2 = \frac{1}{\tau} \sum_{i \in S} \mathbb{E} \ h_i\ ^2$
Single Element $S = \{i\}$ with probability p_i	$\mathcal{L} = \frac{1}{\tau} \sum_{i=1}^n p_i L_i$	$\sigma^2 = \frac{1}{\tau} \sum_{i=1}^n p_i \ h_i\ ^2$
Independent Minibatch $S = \{i_1, \dots, i_\tau\}$ w/o replacement	$\mathcal{L} = \frac{1}{\tau} \sum_{i=1}^n L_i$	$\sigma^2 = \frac{1}{\tau} \sum_{i=1}^n \ h_i\ ^2$
Uniform Minibatch S drawn uniformly random from all subsets of size τ	$\mathcal{L} = \frac{n-1}{\tau(n-1)} L + \frac{n-\tau}{\tau(n-1)} \max_i L_i$	$\sigma^2 = \frac{1}{n\tau} \cdot \frac{n-\tau}{n-1} \sum_{i=1}^n \ h_i\ ^2$

Optimal minibatches for different methods:

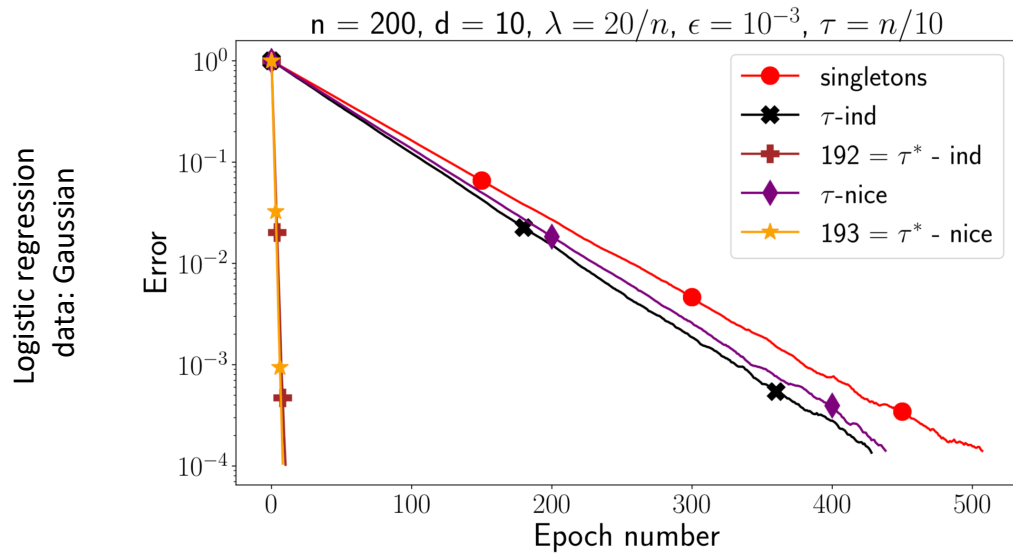
Qu et al (ICML 2016)

Bibi et al (arXiv:1806.05633)

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



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

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

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

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

Gradient noise:

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

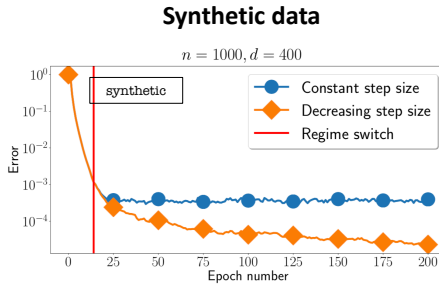
$$\mathbb{E} \|x^k - x^*\|^2 \leq \frac{8\sigma^2}{\mu^2 k} + \frac{16 \lceil \mathcal{L}/\mu \rceil^2}{e^2 k^2} \|x^0 - x^*\|^2$$

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

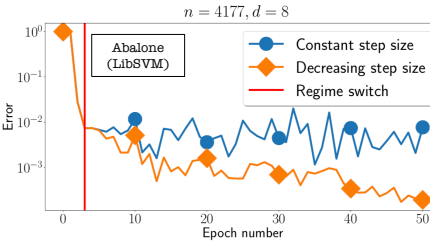
Learning Schedule: Constant & Decreasing

Ridge regression

$$f(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{2} (\lambda_i x_i - y_i)^2 + \frac{\lambda}{2} \|x\|^2$$

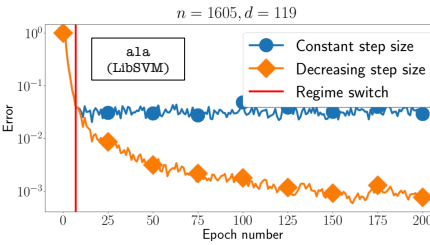
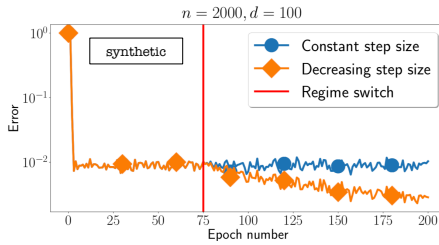


Real data



Logistic regression

$$\frac{1}{n} \sum_{i=1}^n \frac{1}{2} \log(1 + \exp(-y_i \lambda_i x_i)) + \frac{\lambda}{2} \|x\|^2$$



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 $\mathbb{E}_S \left[\left\ \sum_{i \in S} A_i h_i \right\ ^2 \right] \leq \sum_{i=1}^n p_i v_i \ h_i\ ^2$
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. Gower¹ Nicolas Loizou³ Xun Qian² Alibek Saitbayev² Egor Shulgin^{2,4} Peter Richtárik^{2,3,4}



The Problem
x* = arg min_{x in R^n} f(x) = 1/2 ||sum_{i=1}^n f_i(x)||^2
We assume f_i are differentiable and f is quasi-strongly convex.

Stochastic Reformulation
Stochastic reformulation of (1) is the problem:
min_{x in R^n} E_{S, \omega} [f(x) + 1/2 ||sum_{i in S} \omega_i f_i(x)||^2]
where \omega = (\omega_1, ..., \omega_n) in R^n ('sampling vector') is any random vector for which E_{S, \omega}[\omega_i] = 1, \forall i in {1, 2, ..., n}.

Assumptions
(1) f is \mu-strongly convex.
(2) f is L-smooth.
(3) f is \sigma-smooth.
(4) f is \gamma-smooth.

Example: Arbitrary Sampling
A sampling is a random set-valued mapping S with values being subsets of {1, ..., n}. A sampling is defined by assigning probabilities to all subsets of {1, ..., n}.

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).

Assumptions
Quasi strong convexity: f is quasi-\mu-strongly convex [1]:
f(x*) \ge f(x) + (\nabla f(x), x* - x) + \mu/2 ||x* - x||^2, \forall x.
Expected Smoothness: There exists L \ge 0 such
E_{S, \omega} [||\sum_{i in S} \omega_i \nabla f_i(x) - \nabla f(x)||^2] \le 2L f(x) - f(x*), \forall x.
Finite Gradient Noise
\sigma^2 \triangleq E_{S, \omega} [||\sum_{i in S} \omega_i \nabla f_i(x) - \nabla f(x)||^2] < \infty.

Linear Convergence with Fixed Step Size
Assumptions (7) and (8) lead to a bound on the 2nd moment of the stochastic gradient:
Lemma: 2nd moment
If (f, D) \sim ES(C) and \sigma < +\infty (i.e., if (7) and (8) hold), then
E_{S, \omega} [||\sum_{i in S} \omega_i \nabla f_i(x) - \nabla f(x)||^2] \le 4C(Lf(x) - f(x*)) + 2\sigma^2.

Theorem 1
Choose \gamma^* = \gamma \in (0, 1/2], then SGD (5) satisfies:
E ||x^k - x^*||^2 \le (1 - \gamma\mu^*)^k ||x^0 - x^*||^2 + \frac{2\sigma^2}{\mu^*}
In particular, with stepsize \gamma = \min{\frac{1}{2L}, \frac{1}{2\sigma^2}}, we have
k \ge \max{\frac{2C}{\mu^*}, \frac{4\sigma^2}{\mu^*}} \log{\frac{2||x^0 - x^*||^2}{\epsilon}} \Rightarrow E ||x^k - x^*||^2 \le \epsilon.

Proof. Let x^k \triangleq x^k - x^* and \sigma^k \triangleq E_{S, \omega} [||\sum_{i in S} \omega_i \nabla f_i(x^k) - \nabla f(x^k)||^2].
||x^{k+1}||^2 = ||x^k - \gamma \sum_{i in S} \omega_i \nabla f_i(x^k)||^2
= ||x^k||^2 - 2\gamma (\sum_{i in S} \omega_i \nabla f_i(x^k), x^k) + \gamma^2 \sum_{i in S} \omega_i^2 ||\nabla f_i(x^k)||^2
Taking expectation conditioned on x^k we obtain:
E ||x^{k+1}||^2 | x^k = ||x^k||^2 - 2\gamma (f(x^k) - f(x^*)) + \gamma^2 \sigma^k.
Taking expectations again and using the lemma:
E ||x^{k+1}||^2 \le (1 - \gamma\mu) E ||x^k||^2 + 2\gamma^2 \sigma^2 + 2\gamma^2 C L (E ||x^k||^2 - ||x^k||^2) (E ||x^k||^2 - f(x^*))
\le (1 - \gamma\mu) E ||x^k||^2 + 2\gamma^2 \sigma^2.
since 2\gamma C \le 1 and \gamma \le 1/2. Recursively applying the above and summing up the resulting geometric series gives
E ||x^k||^2 \le (1 - \gamma\mu)^k ||x^0||^2 + 2 \sum_{i=0}^{k-1} (1 - \gamma\mu)^i \gamma^2 \sigma^2
\le (1 - \gamma\mu)^k ||x^0||^2 + \frac{2\sigma^2}{\mu}.

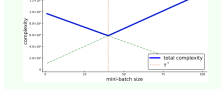
Example: Mini-batch SGD Without Replacement (r-nice sampling)
Consider sampling S which picks from all subsets of {1, ..., n} of cardinality r, uniformly at random. Then p_i = 1/r for all i and the sampling vector v_i is given by:
v_i = { 1/r if i in S, 0 otherwise.

SGD (5) then takes the form:
x^{k+1} = x^k - \gamma \sum_{i in S} v_i \nabla f_i(x^k)

If each f_i is L-smooth and convex, L_max \triangleq \max_i L_i, and f is L-smooth, then (f, D) \sim ES(C), where
C \le L(f) \triangleq \frac{L(L-1)}{r} k + \frac{n-r}{r(n-1)} L_max

Let k^* \triangleq \frac{1}{\mu} \sum_{i=1}^n ||\nabla f_i(x^*)||^2. Then the gradient noise is
\sigma^2 = \sigma^2(x^*) \triangleq \frac{L^2}{r} \frac{n-r}{n-1}
Applying Theorem 1,
k \ge \frac{2C}{\mu^*} \log{\frac{2||x^0 - x^*||^2}{\epsilon}} \ge \frac{2C}{\mu^*} \log{\frac{2||x^0 - x^*||^2}{\epsilon}}

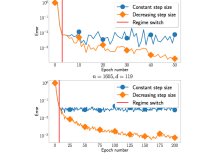
Theoretically optimal mini-batch size is obtained by minimizing the above bound on k in r:
r^* = \frac{L - L_max + \frac{2}{\mu}}{L - L_max + \frac{1}{\mu}}



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 O(1/k) stepsize. This leads to O(1/k) rate.

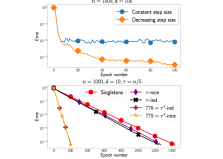
Theorem 2
Let K \triangleq L/\mu and
\gamma^* = { 1/(2K+1) for k \le 4|K|, 1/(2k+1) for k > 4|K|.
If k \ge 4|K|, then SGD iterates given by (5) satisfy:
E ||x^k - x^*||^2 \le \frac{\sigma^2}{\mu k} + \frac{16|K|}{k^2} ||x^0 - x^*||^2.

Learning Schedule



Constant vs decreasing step size regimes of SCD with \lambda = 1/n. Top: Ridge regression problem with abalone. Bottom: Logistic regression with a.a. Data from LIBSVM.

PCA (Sum-of-non-convex functions)



Top: Comparison between constant and decreasing step size regimes of SCD for PCA. Bottom: Comparison of different sampling strategies of SGD for PCA.

References

[1] Ion Necoara, Yuesi Yin, and Francis Clarke. Linear convergence of fast cubic methods for nonstrongly convex optimization. Mathematical Programming, pages 1–16, 2018.
[2] Peter Richtárik and Martin Teufel. On central gradient descent in stochastic nonconvex descent methods. Optimization Letters, 10(1):223–236, 2016.
[3] Robert M. Gower, Peter Richtárik, and Frank Bach. Gradient descent algorithms: Beyond convexity. In Advances in Machine Learning, pages 1–16, 2018.
[4] Robert M. Gower, Peter Richtárik, and Frank Bach. Gradient descent algorithms: Beyond convexity. In Advances in Machine Learning, pages 1–16, 2018.
[5] Robert M. Gower, Peter Richtárik, and Frank Bach. Gradient descent algorithms: Beyond convexity. In Advances in Machine Learning, pages 1–16, 2018.