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

TELECOM ParisTech

THE UNIVERSITY OF EDINBURGH

KAUST King Abdullah University of Science and Technology

MIPT MOSCOW INSTITUTE OF PHYSICS AND TECHNOLOGY

KAUST King Abdullah University of Science and Technology

1. The Problem & Motivation

The Problem: Empirical Risk Minimization

f is μ–quasi strongly convex

training data

Smooth loss associated with data point *i*

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

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
$v = (v_1, \dots, v_n)$

Random variable with mean 1

Linearity of expectation

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

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

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

stepsize

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

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

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

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

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

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

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

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

Sampling vector

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

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

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

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

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

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

$$\sum_{i=1}^n q_i = 1$$

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

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

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

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

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

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

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

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

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)

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

Expected smoothness constant

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

Depends on *f* and *v*

Lemma f_i convex & L-smooth

⬇

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

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

Bounding the 2nd Moment

Gradient noise:

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

Lemma

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

$$\Downarrow$$

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

$\sigma^2 = 0$ ➡ 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 $S \subseteq \{1,2,\dots,n\}$	$c \equiv \frac{\mathbf{P}_{ij}}{p_i p_j} \quad i \neq j$ 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 $S = \bigcup_{i=1}^{n} S_i$ $S_i = \begin{cases} \{i\} & \text{with probability} \quad p_i \\ \emptyset & \text{with probability} \quad 1-p_i \end{cases}$ $S_1, \dots, S_n$ are independent	$\mathbf{P}_{ij} = p_i p_j \Rightarrow c = 1$ $\mathcal{L} = L + \frac{1}{n}\max_i \frac{(1-p_i)L_i}{p_i}$ f is L–smooth	$\sigma^2 = \frac{1}{n^2}\sum_{i=1}^{n} \frac{1-p_i}{p_i}\|h_i\|^2$
Uniform Minibatch S chosen uniformly random from all subsets of size τ	$\tau \overset{\text{def}}{=} \mathbf{E}[S] = \sum_i p_i$ $\mathcal{L} = \frac{n(\tau-1)}{\tau(n-1)}L + \frac{n-\tau}{\tau(n-1)}\max_i L_i$	$\sigma^2 = \frac{1}{n\tau}\cdot\frac{n-\tau}{n-1}\sum_{i=1}^{n}\|h_i\|^2$

4. Convergence Analysis: Linear Rate

Main Result

(Linear Convergence to a Neighborhood of the Solution)

Assumption: f is μ–quasi strongly convex

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

Gradient noise:

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

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

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

Fixed stepsize: $\gamma^k \equiv \gamma \leq \frac{1}{2\mathcal{L}}$

$\sigma = 0 \Rightarrow$ can choose $\gamma = \frac{1}{\mathcal{L}}$

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

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

Optimal Minibatch Size

iterations

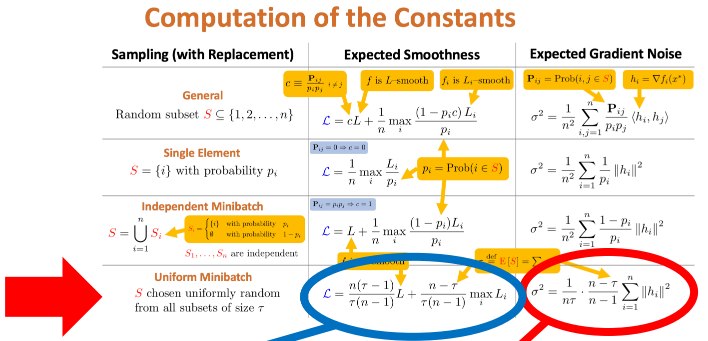
stochastic gradient evaluations in 1 iteration

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

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

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

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



Computation of the Constants

Sampling (with Replacement)	Expected Smoothness	Expected Gradient Noise
General Random subset $S \subseteq \{1,2,\ldots,n\}$	$c \equiv \frac{\mathbf{P}_{ij}}{p_i p_j}\ i \neq j$ f is L–smooth 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 $S = \bigcup_{i=1}^n S_i$ $S_i = \begin{cases}\{i\} & \text{with probability } p_i \\ \emptyset & \text{with probability } 1-p_i\end{cases}$ $S_1,\ldots,S_n$ 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	$\tau \overset{\text{def}}{=} \mathrm{E}[S] = \sum$ $\mathcal{L} = \frac{n(\tau-1)}{\tau(n-1)}L + \frac{n-\tau}{\tau(n-1)}\max_i L_i$	$\sigma^2 = \frac{1}{n\tau}\cdot\frac{n-\tau}{n-1}\sum_{i=1}^n \|h_i\|^2$

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

Optimal minibatches for different methods:

Qu et al (ICML 2016)

Bibi et al (arXiv:1806.05633)

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

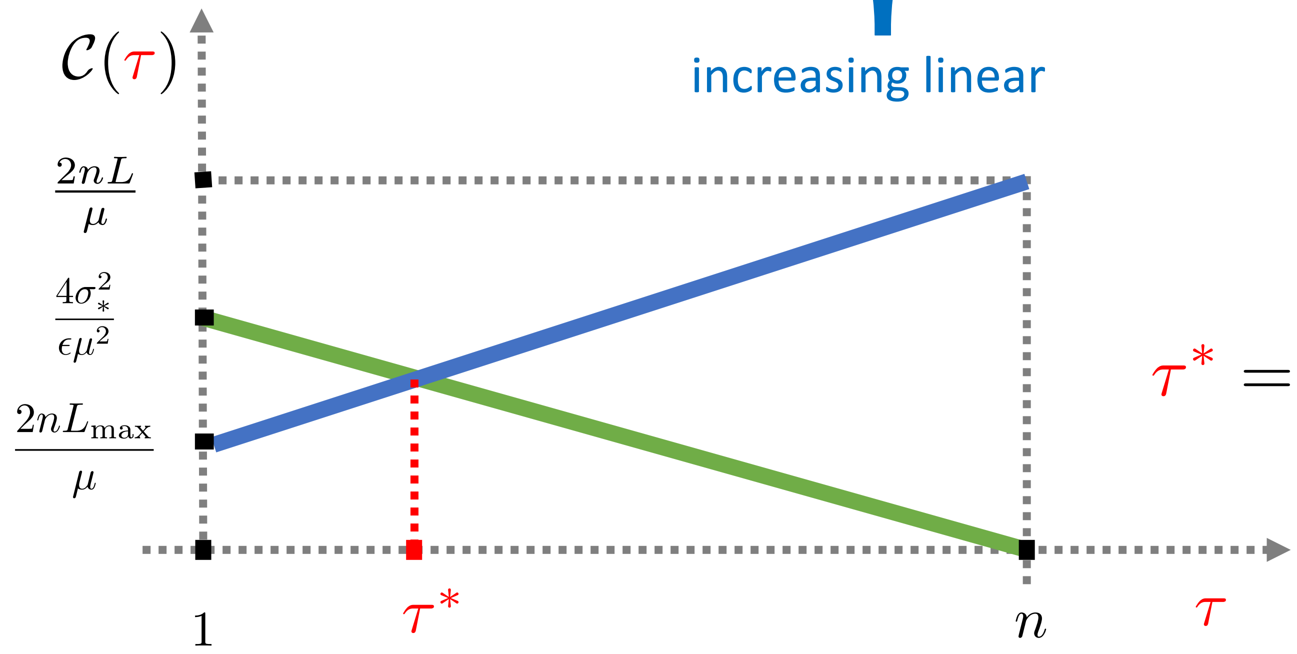
f_i is L_i–smooth

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

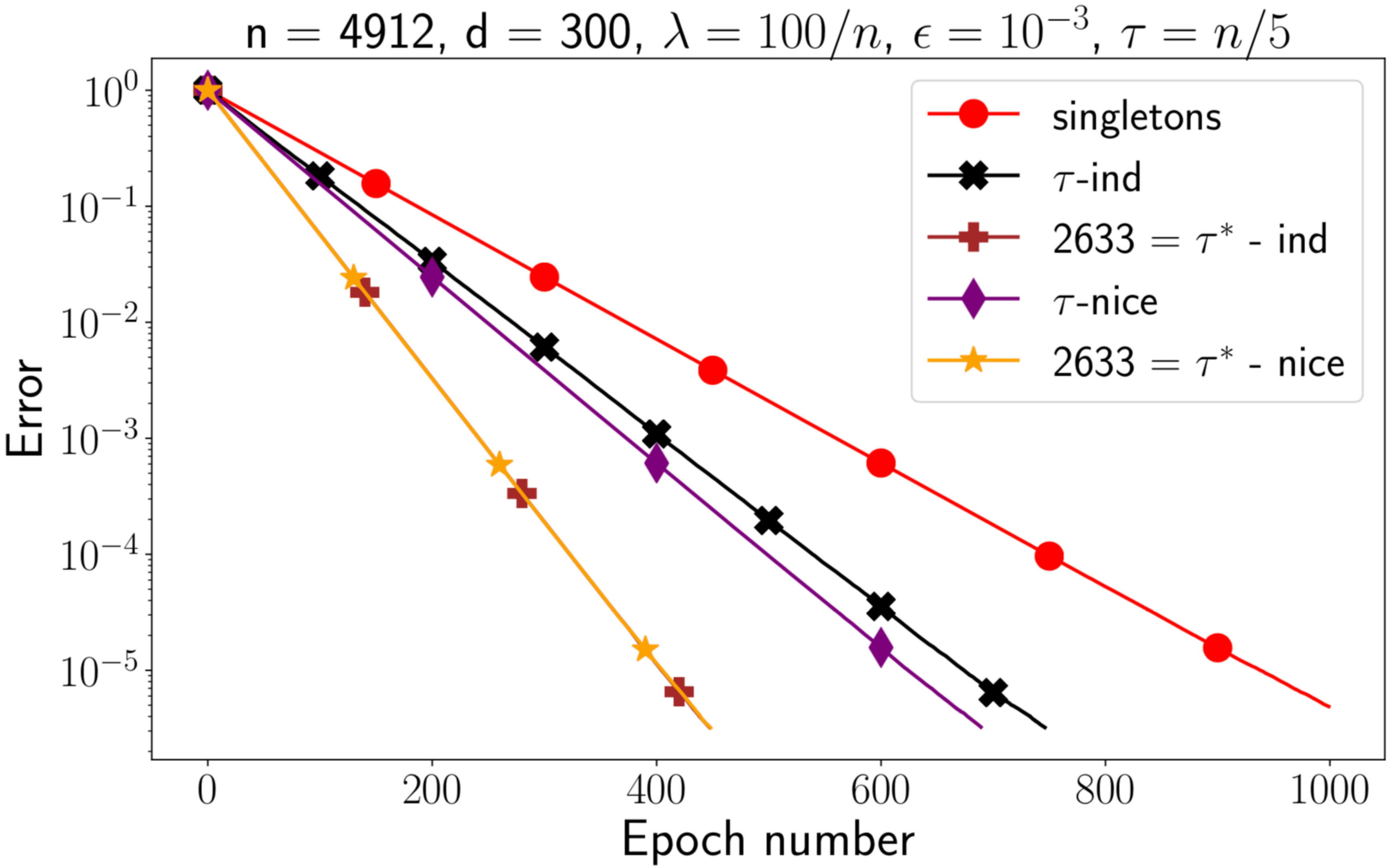


$$\tau^* = \frac{n(\theta + L - L_{\max})}{\theta + nL - L_{\max}}$$

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

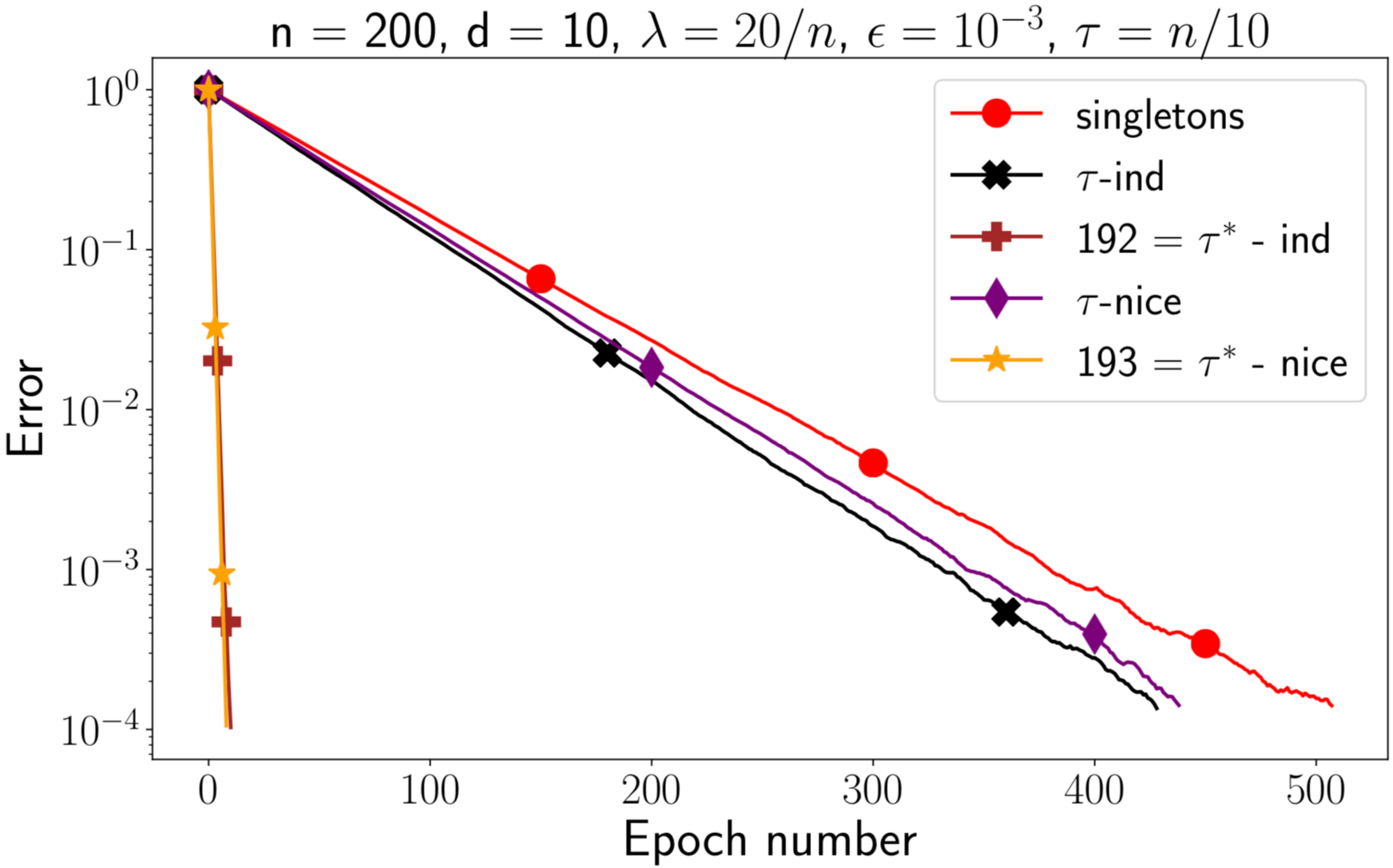
Optimal Minibatch Size: LIBSVM data

Logistic regression
data: w3a (LIBSVM)



Optimal Minibatch Size: Synthetic Data

Logistic regression
data: Gaussian



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:

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

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

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

Learning Schedule: Constant & Decreasing

Synthetic data

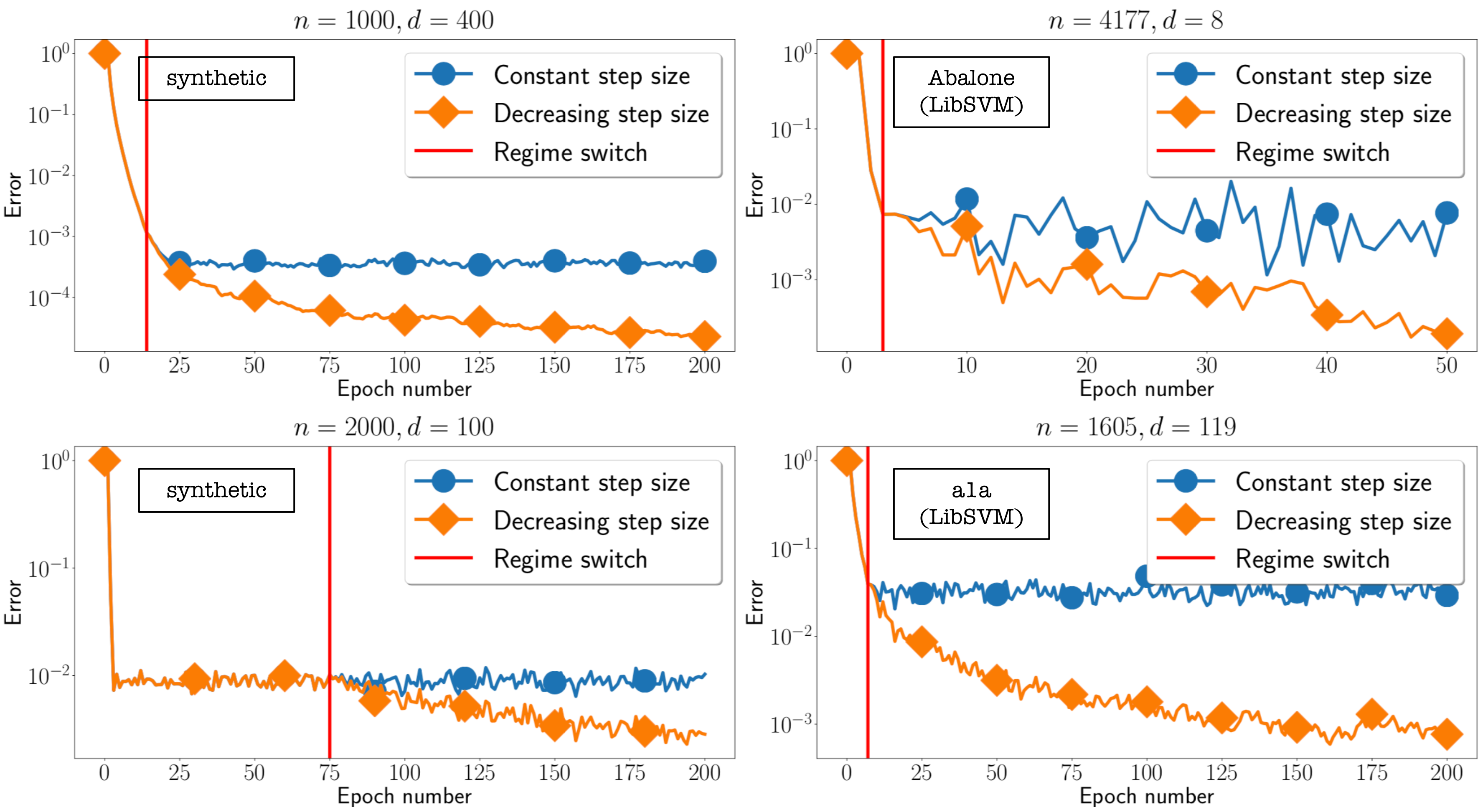
Real data

Ridge regression

$f(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}(\mathbf{A}_{i:}x - y_i)^2 + \frac{\lambda}{2}\|x\|^2$

Logistic regression

$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\log(1+\exp(-y_i\mathbf{A}_{i:}x)) + \frac{\lambda}{2}\|x\|^2$



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)

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

$$x^* = \arg\min_{x\in\mathbb{R}^d}\left[f(x) \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n f_i(x)\right] \tag{1}$$

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

Stochastic Reformulation

Stochastic reformulation of (1) is the problem:

$$\min_{x\in\mathbb{R}^d} \mathbb{E}_{v\sim\mathcal{D}}\left[f_v(x) \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n v_i f_i(x)\right]. \tag{2}$$

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

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

- Equivalence: (2) is equivalent to (1) since $\mathbb{E}_{v\sim\mathcal{D}}[f_v] = f$. Also note that $\mathbb{E}_{v\sim\mathcal{D}}[\nabla f_v] = \nabla f$, which can be seen via

$$\mathbb{E}_{v\sim\mathcal{D}}[\nabla f_v] \overset{(2)}{=} \tfrac{1}{n}\sum_{i=1}^n \mathbb{E}_{v\sim\mathcal{D}}[v_i]\nabla f_i \;=\; \nabla f. \tag{4}$$

- We propose to solve (1) by applying **SGD** to (2):

$$\boxed{x^{k+1} = x^k - \gamma^k \nabla f_{v^k}(x^k)} \tag{5}$$

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

Example: Arbitrary Sampling

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

- A sampling is proper if $p_i \overset{\text{def}}{=} \mathbb{P}[i\in S] > 0$ for all $i \in \{1,\dots,n\}$.
- Each proper sampling S gives rise to a sampling vector v:

$$v = \text{Diag}(p_1^{-1},\dots,p_n^{-1})\sum_{i\in S} e_i,$$

where e_i is the ith standard unit basis vector in $\mathbb{R}^n$. It is easy to see that $\mathbb{E}[v_i] = 1$. 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 establish linear convergence 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* assume bounded second moment of the gradients for every x (only at x^*; see (8)). We rely on the expected smoothness assumption (7) [3, 4].
- Optimal mini-batch size: 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)).
- Interpolated models. We extend the findings 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 μ–strongly convex [1]:

$$f(x^*) \geq f(x) + \langle\nabla f(x), x^* - x\rangle + \frac{\mu}{2}\|x^* - x\|^2, \;\forall x \tag{6}$$

- **Expected Smoothness:** There exists $\mathcal{L} \geq 0$ such

$$\mathbb{E}_{v\sim\mathcal{D}}\left[\|\nabla f_v(x) - \nabla f_v(x^*)\|^2\right] \leq 2\mathcal{L}(f(x) - f(x^*)), \;\forall x. \tag{7}$$

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

- **Finite Gradient Noise**

$$\sigma^2 \overset{\text{def}}{=} \mathbb{E}_{v\sim\mathcal{D}}\left[\|\nabla f_v(x^*)\|^2\right] \;<\; \infty. \tag{8}$$

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

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 $(f,\mathcal{D}) \sim ES(\mathcal{L})$ and $\sigma < +\infty$ (i.e., if (7) and (8) hold), then

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

The above lemma can now be used to establish a linear convergence result:

Theorem 1

Choose $\gamma^k = \gamma \in (0, \frac{1}{2\mathcal{L}}]$, then SGD (5) satisfies:

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

In particular, with stepsize $\gamma = \min\left\{\frac{1}{2\mathcal{L}}, \frac{\epsilon\mu}{4\sigma^2}\right\}$, we have

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

Proof. Let $r^k \overset{\text{def}}{=} x^k - x^*$ and $g^k \overset{\text{def}}{=} \mathbb{E}_k\left[\|\nabla f_{v^k}(x^k)\|^2\right]$.

$$\begin{aligned}\|r^{k+1}\|^2 &\overset{(5)}{=} \|x^k - x^* - \gamma\nabla f_{v^k}(x^k)\|^2\\ &= \|r^k\|^2 - 2\gamma\langle r^k, \nabla f_{v^k}(x^k)\rangle + \gamma^2\|\nabla f_{v^k}(x^k)\|^2\end{aligned}$$

Taking expectation conditioned on x^k we obtain:

$$\begin{aligned}\mathbb{E}_k\|r^{k+1}\|^2 &\overset{(4)}{=} \|r^k\|^2 - 2\gamma\langle r^k, \nabla f(x^k)\rangle + \gamma^2 g^k\\ &\overset{(6)}{\leq} (1-\gamma\mu)\|r^k\|^2 - 2\gamma[f(x^k) - f(x^*)] + \gamma^2 g^k.\end{aligned}$$

Taking expectations again and using the lemma :

$$\begin{aligned}\mathbb{E}\|r^{k+1}\|^2 &\leq (1-\gamma\mu)\mathbb{E}\|r^k\|^2 + 2\gamma^2\sigma^2\\ &\quad + 2\gamma(2\gamma\mathcal{L} - 1)\mathbb{E}\left[f(x^k) - f(x^*)\right]\\ &\leq (1-\gamma\mu)\mathbb{E}\|r^k\|^2 + 2\gamma^2\sigma^2,\end{aligned}$$

since $2\gamma\mathcal{L} \leq 1$ and $\gamma \leq \frac{1}{2\mathcal{L}}$. Recursively applying the above and summing up the resulting geometric series gives

$$\begin{aligned}\mathbb{E}\|r^k\|^2 &\leq (1-\gamma\mu)^k\|r^0\|^2 + 2\sum_{j=0}^{k-1}(1-\gamma\mu)^j\gamma^2\sigma^2\\ &\leq (1-\gamma\mu)^k\|r^0\|^2 + \frac{2\gamma\sigma^2}{\mu}.\end{aligned}$$

□

Example: Mini-batch SGD Without Replacement (τ-nice sampling)

- Consider sampling S which picks from all subsets of $\{1,\dots,n\}$ of cardinality τ, uniformly at random. Then $p_i = \frac{\tau}{n}$ for all i and the sampling vector v is given by:

$$v_i = \begin{cases}\frac{n}{\tau} & i \in S\\ 0 & \text{otherwise.}\end{cases}$$

- SGD (5) then takes the form

$$\boxed{x^{k+1} = x^k - \gamma^k\frac{n}{\tau}\sum_{i\in S^k}\nabla f_i(x^k)}$$

- If each f_i is L_i–smooth and convex, $L_{\max} \overset{\text{def}}{=} \max_i L_i$, and f is L-smooth, then $(f,\mathcal{D}) \sim ES(\mathcal{L})$, where

$$\mathcal{L} \leq \mathcal{L}(\tau) \overset{\text{def}}{=} \frac{n(\tau-1)}{\tau(n-1)}L + \frac{n-\tau}{\tau(n-1)}L_{\max}$$

- Let $h^* \overset{\text{def}}{=} \frac{1}{n}\sum_i\|\nabla f_i(x^*)\|^2$. Then the gradient noise is

$$\sigma^2 = \sigma^2(\tau) \overset{\text{def}}{=} \frac{h^*}{\tau}\cdot\frac{n-\tau}{n-1}.$$

- Applying Theorem 1,

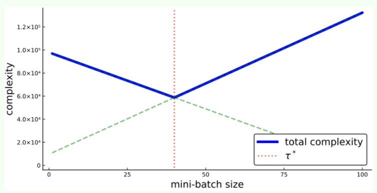
$$k \geq \frac{2(n-\tau)}{\tau(n-1)}\max\left\{\frac{n(\tau-1)}{n-\tau}\frac{L}{\mu} + \frac{L_{\max}}{\mu}, \frac{2h^*}{\epsilon\mu^2}\right\}\log\left(\frac{2\|x^0 - x^*\|^2}{\epsilon}\right),$$

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

- Theoretically optimal mini-batch size is obtained by minimizing the above bound on k in τ:

$$\tau^* = n\frac{L - L_{\max} + \frac{2}{\epsilon\mu}\cdot h^*}{nL - L_{\max} + \frac{2}{\epsilon\mu}\cdot h^*}.$$

A sample computation is shown in the plot below:



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

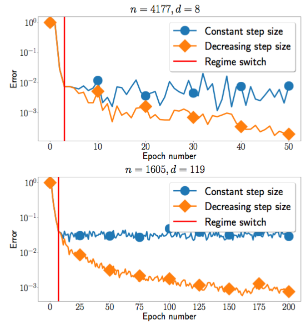
Let $\mathcal{K} \overset{\text{def}}{=} \mathcal{L}/\mu$ and

$$\gamma^k = \begin{cases}\dfrac{1}{2\mathcal{L}} & \text{for}\quad k \leq 4\lceil\mathcal{K}\rceil\\ \dfrac{2k+1}{(k+1)^2\mu} & \text{for}\quad k > 4\lceil\mathcal{K}\rceil.\end{cases} \tag{9}$$

If $k \geq 4\lceil\mathcal{K}\rceil$, then SGD iterates given by (5) satisfy:

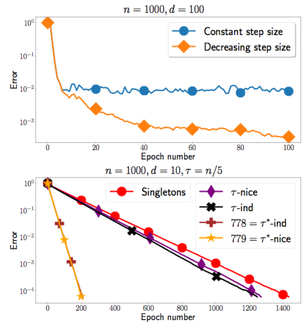
$$\mathbb{E}\|x^k - x^*\|^2 \leq \frac{\sigma^2}{\mu^2}\frac{8}{k} + \frac{16\lceil\mathcal{K}\rceil^2}{e^2k^2}\|x^0 - x^*\|^2. \tag{10}$$

Learning Schedule



Constant vs decreasing step size regimes of SGD with $\lambda = 1/n$. *Top:* Ridge regression problem with `abalone`. *Bottom:* Logistic regression with `a1a`. Data from LIBSVM.

PCA (Sum-of-non-convex functions)



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

References

[1] Ion Necoara, Yurii Nesterov, and Francois Glineur.
Linear convergence of first order methods for non-strongly convex optimization.
Mathematical Programming, pages 1–39, 2018.

[2] Peter Richtárik and Martin Takáč.
On optimal probabilities in stochastic coordinate descent methods.
Optimization Letters, 10(6):1233–1243, 2016.

[3] Robert M. Gower, Peter Richtárik, and Francis Bach.
Stochastic quasi-gradient methods: Variance reduction via Jacobian sketching.
arxiv:1805.02632, 2018.

[4] Nidham Gazagnadou, Robert Mansel Gower, and Joseph Salmon.
Optimal mini-batch and step sizes for saga.
In *36th International Conference on Machine Learning*, 2019.

[5] Siyuan Ma, Raef Bassily, and Mikhail Belkin.
The power of interpolation: Understanding the effectiveness of SGD in modern over-parametrized learning.
In *35th International Conference on Machine Learning*, 2018.

The End

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**