SGD with Coordinate Sampling: Theory and Practice

Rémi LELUC
Ecole Polytechnique, Institut Polytechnique de Paris, France

Joint Work with François Portier, paper
Published in *Journal of Machine Learning Research, 2022.*

Introduction

Underlying optimization problem

Let $f : \mathbb{R}^p \to \mathbb{R}$ be a general objective function.

• **Goal:** Solve

$$\min_{\theta \in \mathbb{R}^p} \{f(\theta) = \mathbb{E}_\xi[f(\theta, \xi)]\}$$

• **Constraints:** ∇f is hard to compute (large-scale problems) or even intractable (black-box) !

• **Central question:** Fast and Efficient procedures

Empirical Risk Minimization. data $z_1, \ldots, z_n \subset \mathcal{Z}$ and loss function $\ell : \mathbb{R}^p \times \mathcal{Z} \to \mathbb{R}$,

$$\forall \theta \in \mathbb{R}^p, \quad f(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(\theta, z_i)$$

The true gradient, $n^{-1} \sum_{i=1}^{n} \nabla \ell(\theta, z_i)$ requires n evaluations

Introduction

Noisy gradients

- Zeroth-Order (biased):

$$g(\theta) = \sum_{k=1}^{p} h^{-1}(f(\theta + he_k) - f(\theta))e_k \underset{h\to 0}{\approx} \nabla f(\theta)$$

- First-Order (unbiased):

$$g_{t+1} := \nabla_\theta \ell(\theta_t, z_{\xi_{t+1}})$$

where $\xi_{t+1} \sim \mathcal{U}([\![1, n]\!])$ is uniformly distributed.

Stochastic Gradient Descent (Robbins and Monro, 1951)

- Start $(t = 0)$ from random point $\theta_0 \in \mathbb{R}^d$.
- Evaluate noisy gradient g_{t+1}
- Update iterate $\theta_{t+1} = \theta_t - \gamma_{t+1} g_{t+1}$.

- **(SCGD): Stochastic Coordinate Gradient Descent**

$$\begin{cases} \theta_{t+1}^{(k)} = \theta_t^{(k)} & \text{if } k \neq \zeta_{t+1} \\ \theta_{t+1}^{(k)} = \theta_t^{(k)} - \gamma_{t+1} g_{t+1}^{(k)} & \text{if } k = \zeta_{t+1} \end{cases}$$

ζ_{t+1} is a random variable valued in $[\![1, p]\!]$.

- Reduction of the computing cost
- Covers all approaches that uses a gradient estimate g_{t+1}
- **2 sources of randomness:**
  (i) noisy gradient g_{t+1}
  (ii) noisy coordinate ζ_{t+1}

Contributions

- **(SCGD): Stochastic Coordinate Gradient Descent**

$$\begin{cases} \theta_{t+1}^{(k)} = \theta_t^{(k)} & \text{if } k \neq \zeta_{t+1} \\ \theta_{t+1}^{(k)} = \theta_t^{(k)} - \gamma_{t+1} g_{t+1}^{(k)} & \text{if } k = \zeta_{t+1} \end{cases}$$

- How to update the selecting policy ζ_{t+1} ?
  $\rightarrow$ We develop an algorithm MUSKETEER to leverage the data structure and move along relevant directions.
- What condition on ζ_{t+1} for convergence ?
  $\rightarrow$ We analyze the properties of SCGD algorithms (convergence of the iterates, convergence of the policy, non-asymptotic bound)

Related work

- CD using f or true gradient ∇f (Loshchilov et al., 2011; Richtárik and Takáč, 2013; Glasmachers and Dogan, 2013; Qu and Richtárik, 2016; Allen-Zhu et al., 2016; Namkoong et al., 2017)
- Most related idea: **Gauss-Southwell rule** to select the largest gradient coordinate to move the iterate (Nutini et al., 2015)
  $\rightarrow$ Here: we have stochastic g_{t+1} and ζ_{t+1}.
- **Sparsification methods** (Alistarh et al., 2017; Wangni et al., 2018), unbiased importance sampling estimate of the gradient
  $\rightarrow$ Here: no reweighting (biased) (conditioned gradient)

General framework and notation

- Only one coordinate ζ_{t+1} is selected:

$$(SCGD) \quad \theta_{t+1} = \theta_t - \gamma_{t+1} D(\zeta_{t+1}) g_{t+1}$$

  with $D(k) = e_k e_k^T = Diag(0, \dots, 0, 1, 0, \dots, 0)$.

- The distribution of ζ_{t+1}, is the **coordinate sampling policy** and is given by the probability weights vector $d_t = (d_t^{(1)}, \dots, d_t^{(p)})$

$$d_t^{(k)} = \mathbb{P}(\zeta_{t+1} = k | \mathcal{F}_t), \quad k \in [\![1, p]\!].$$

- Not the same mean field as in usual SGD. Under conditional independence between g_{t+1} and ζ_{t+1}:

$$\mathbb{E}[D(\zeta_{t+1}) g_{t+1} | \mathcal{F}_t] = \mathrm{diag}(d_t) g(\theta_t)$$

General view: Unbiased and Adaptive policies

General update rule

$$\theta_{t+1} = \theta_t - \gamma_{t+1} h(\theta_t, \omega_{t+1})$$

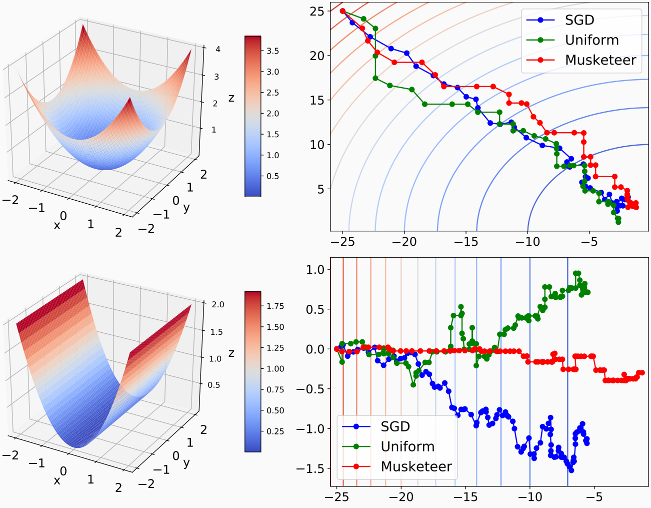
where h is a gradient generator and $(\omega_t)_{t \geq 1}$ is a sequence of random variables

- (SGD) $h(\theta, \omega_{t+1}) = g_{t+1}$
- (SCGD) $h(\theta, \omega_{t+1}) = D(\zeta_{t+1}) g_{t+1}$
- (Unbiased with importance weights as in (Wangni et al., 2018)) $h(\theta, \omega_{t+1}) = D_t^{-1} D(\zeta_{t+1}) g_{t+1}$

MUSKETEER

MUltivariate
Stochastic
Knowledge
Extraction
Through
Exploration
Exploitation
Reinforcement

Illustration/Motivation



MUSKETEER

MUSKETEER may be seen as an **adaptive bandit** problem with

$$'arms = coordinates'$$

Alternate between 2 phases

- **Exploration phase (one for all)**

$\rightarrow$ fixed d_t, draw random coordinate and move along selected direction

$\rightarrow$ cumulative gains for the visited coordinates

- **Exploitation phase (all for one)**

$\rightarrow$ share knowledge of the cumulative gains

$\rightarrow$ update the coordinate sampling probability vector d_t

MUSKETEER: Exploration phase

1) Pick a coordinate

Generate $\zeta_{t+1} \sim d_t$ and the coordinate gradient g_{t+1}

2) Update the iterate

$$\theta_{t+1}^{(\zeta_{t+1})} = \theta_t^{(\zeta_{t+1})} - \gamma_{t+1} g_{t+1}^{(\zeta_{t+1})}$$

3) Update cumulative gains

$$G_{t+1}^{(\zeta_{t+1})} = G_t^{(\zeta_{t+1})} + g_{t+1}^{(\zeta_{t+1})} / d_t^{(\zeta_{t+1})}$$

$\rightarrow$ (Variants with square) $|g_{t+1}^{(\zeta)}|$ or $g_{t+1}^{(\zeta)2}$

$\rightarrow$ Might be done T times with d_t fixed (before moving to the exploitation)

MUSKETEER: Exploitation phase

- This phase is to update the policy value of d_t
- EXP3 algorithm (Auer et al., 2002) to update the probability weights through a mixture. Given $\eta > 0$ and $\lambda \in [0, 1]$, we have for all $k \in [\![1, p]\!]$,

$$d_{t+1}^{(k)} = (1-\lambda)\frac{\exp(\eta|G_{t+1}^{(k)}|/t)}{\sum_{j=1}^{d}\exp(\eta|G_{t+1}^{(j)}|/t)} + \lambda\frac{1}{p}$$

- The mixture with $\lambda > 0$ ensure to always give a chance to everyone

MUSKETEER complete algorithm

Algorithm input: (d_0, θ_0), sequence $(\gamma_t)_{t \geq 1}$ and parameter (η, λ)

for $t = 0, 1, 2, \ldots$ **do**

Set $d = d_t$ and sample coordinate $\zeta \sim d$ and gradient g

Update iterate: $\theta_{t+1}^{(\zeta)} = \theta_t^{(\zeta)} - \gamma_{t+1} g^{(\zeta)}$

Update gain: $G_{t+1}^{(\zeta)} = G_t^{(\zeta)} + g^{(\zeta)} / d^{(\zeta)}$

Whenever $t = 0 \pmod{T}$: update weights d_{t+1} with

$$d_{t+1}^{(k)} = (1 - \lambda) \frac{\exp(\eta |G_t^{(k)}| / t)}{\sum_{j=1}^{d} \exp(\eta |G_t^{(j)}| / t)} + \lambda \frac{1}{p}$$

end for

Numerical Experiments

- We apply ERM to regularized **regression** and **classification** problems.
- Given a data matrix $X = (x_{i,j}) \in \mathbb{R}^{n \times p}$ with labels $y \in \mathbb{R}^n$ and a regularization parameter $\mu > 0$, the *Ridge regression* is

$$\min_{\theta \in \mathbb{R}^p} f(\theta) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} x_{i,j}\theta_j)^2 + \frac{\mu}{2} \|\theta\|_2^2$$

and the ℓ_2*-regularized logistic regression* is defined by

$$\min_{\theta \in \mathbb{R}^p} f(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(-y_i \sum_{j=1}^{p} x_{i,j}\theta_j)) + \mu \|\theta\|_2^2$$

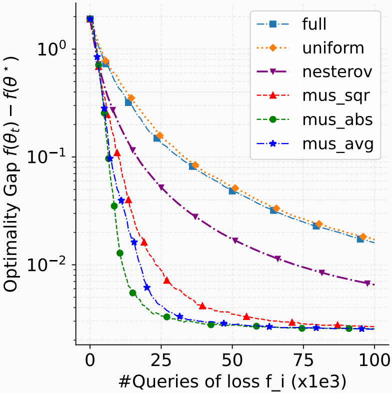
where μ is set to the classical value $\mu = 1/n$

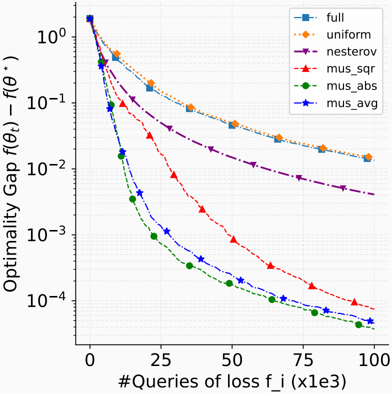
Special covariance structure

$X[:, k] \sim \mathcal{N}(0, \sigma_k^2 I_n)$ with $\sigma_k^2 = k^{-\alpha}$ for $k \in [\![1, p]\!]$

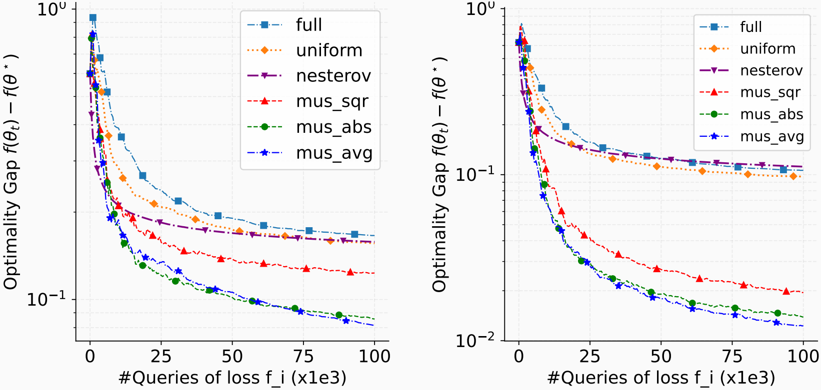
Setting $\gamma_t = 1/t$, $n = 10,000$, $p = 250$, $T = \lfloor \sqrt{p} \rfloor = 15$

ZO Ridge Regression ($\alpha = 5$ and $\alpha = 10$)



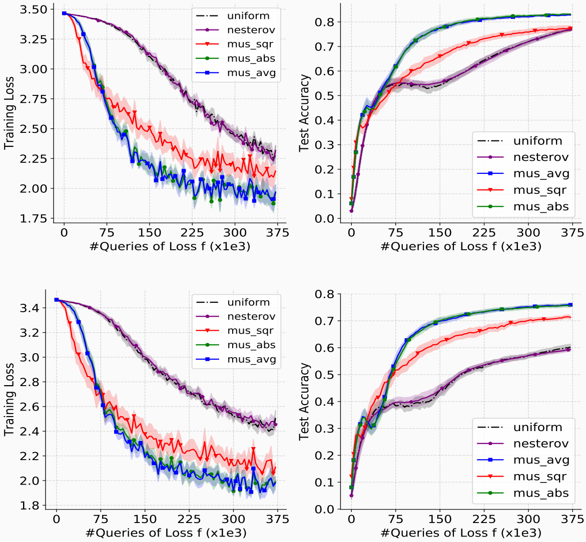


ZO Logistic Regression ($\alpha = 2$ and $\alpha = 5$)



Numerical Experiments

- **MNIST and Fashion-MNIST (ZO)** ($p = 55,050$ and $T = 234$)



Back ground

Stochastic Optimization

$$\min_{\theta \in \mathbb{R}^p} \{f(\theta) = \mathbb{E}_\xi[f(\theta, \xi)]\}$$

Gradients might be biased

There exists constant $c \geq 0$ such that

$$\forall h > 0, \theta \in \mathbb{R}^p, \quad \| \mathbb{E}_\xi[g_h(\theta, \xi)] - \nabla f(\theta)\| \leq ch.$$

- $h \geq 0$ is a parameter controlling the bias
- $c = 0$ recovers 1st-order gradient estimates
- Allows to cover general zeroth-order estimates

ZO gradient estimates

Example 1 (smoothing).

(Nesterov and Spokoiny, 2017). The smoothed gradient estimate is

$$\forall \theta \in \mathbb{R}^p, g_h(\theta, \xi) = h^{-1}[f(\theta + hU, \xi) - f(\theta, \xi)]U$$

where $U \sim \mathcal{N}(0, I)$ (Alternative version with $U \sim Unif(\mathbb{S})$)

Example 2 (finite differences).

The finite differences gradient estimate is given by

$$\forall \theta \in \mathbb{R}^p, g_h(\theta, \xi) = \sum_{k=1}^{p} g_h(\theta, \xi)^{(k)} e_k$$

where for all $k = 1, \ldots, p$ the coordinates are

$$g_h(\theta, \xi)^{(k)} = h^{-1}[f(\theta + he_k, \xi) - f(\theta, \xi)]$$

General form

There exists probability measure ν satisfying $\int_{\mathbb{R}^p} xx^\top \nu(\mathrm{d}x) = I_p$,

$$\forall h > 0, \theta \in \mathbb{R}^p, \quad \mathbb{E}_\xi[g_h(\theta, \xi)] = \int_{\mathbb{R}^p} x \left\{ \frac{f(\theta + hx) - f(\theta)}{h} \right\} \nu(\mathrm{d}x).$$

Lemma

Under the previous assumption (if f is L-smooth) the biased gradient assumption is satisfies with

$$c = (L/2)\sqrt{\int_{\mathbb{R}^p} \|x\|_2^6 \nu(\mathrm{d}x)}$$

- smoothing gradient is recovered when ν is the Gaussian measure
- Take $\nu = \sum_{k=1}^p \delta_{e_k}/p$ covers the finite differences estimate
- (MUSKETEER) Use a measure ν that evolves through time and put different weights on the different directions !

Assumption

Growth condition

There exist $0 \leq \mathcal{L}, \sigma^2 < \infty$

$$\forall h > 0, \theta \in \mathbb{R}^p \quad \mathbb{E}\left[\|g_h(\theta, \xi)\|_\infty^2\right] \leq 2\mathcal{L}\left(f(\theta) - f^\star\right) + \sigma^2.$$

Smoothness and lower bound

f is L-smooth and lower bounded by $f^\star$

Two algorithms

Gradient generator $g_t = g_{h_{t+1}}(\theta_t, \xi_{t+1})$

$$(SGD) \quad \theta_{t+1} = \theta_t - \gamma_{t+1} g_t$$
$$(SCGD) \quad \theta_{t+1} = \theta_t - \gamma_{t+1} D(\zeta_{t+1}) g_t$$

Auxiliary result: SGD case

Robbins-Monro

$$\sum_{t \geq 1} \gamma_t = +\infty \quad \text{and} \quad \sum_{t \geq 1} \gamma_t^2 < +\infty$$

small bias

$$h_t^2 = O(\gamma_t)$$

Main results

Theorem (Almost sure convergence of (biased) SGD)

Under previous assumptions, $\nabla f(\theta_t) \to 0$ *a.s. when* $t \to \infty$*.*

Theorem (Almost sure convergence of particular SCGD)

Under previous assumptions

(i) *(max gradient) if* $\zeta_{t+1} = \arg\max_{k=1,\dots,p} |\partial_k f(\theta_t)|$ *then* $\nabla f(\theta_t) \to 0$ *almost surely as* $t \to +\infty$*.*

(ii) *(gradient weights) if* $D_t \propto (|\nabla_k f(\theta_t)|^q)_{1\leq k\leq p}$ *with* $q > 0$ *then* $\nabla f(\theta_t) \to 0$ *almost surely as* $t \to +\infty$*.*

- When f coercive and unique solution $\{\theta : \nabla f(\theta) = 0\} = \{\theta^\star\}$ then almost sure convergence towards minimizer $\theta_t \to \theta^\star$.

Main results: general SCGD

Theorem (Almost sure convergence general SCGD)

Under previous assumptions, if $\beta_{t+1} = \min_{1 \leq k \leq p} d_t^{(k)}$ *is away from* 0 *then* $\nabla f(\theta_t) \to 0$ *almost surely as* $t \to +\infty$.

Main results: MUSKETEER

Theorem (Almost sure convergence)

The sequence of iterates $(\theta_t)_{t\geq 0}$ obtained by the MUSKETEER satisfies $\nabla f(\theta_t) \to 0$ almost surely as $t \to +\infty$.

Theorem (Weak convergence)

The MUSKETEER's coordinate policy $(d_t)_{t\in\mathbb{N}}$ converges weakly to the uniform distribution

Theorem (Non-asymptotic bounds, (Moulines and Bach, 2011))

Let $(\theta_t)_{t\in\mathbb{N}}$ obtained by MUSKETEER with $\gamma_t = \gamma t^{-\alpha}$ then

$$\mathbb{E}\left[f(\theta_t) - f^\star\right] = O(1/t), \quad (\alpha = 1)$$

Conclusion Future work

Contributions

• (**Theory**) Almost-sure convergence SCGD towards stationary points, non-asymptotic bounds on the optimality gap $\mathbb{E}[f(\theta_t) - f^\star]$.
• Conditions are relatively weak as f is only L-smooth (classical in non-convex problems) and the stochastic gradients are possibly biased with unbounded variance.

• (**Practice**) New algorithm, called MUSKETEER: in the image of the motto 'all for one and one for all', this procedure belongs to the SCGD framework with a particular design for the *coordinate sampling policy*.
• MUSKETEER compares the value of all past gradient estimates g_t to select a descent direction (*all for one*) and then moves the current iterate according to the chosen direction (*one for all*).

Future work

Study the asymptotic behavior of other adaptive sampling strategies

References

Alistarh, D., D. Grubic, J. Li, R. Tomioka, and M. Vojnovic (2017). Qsgd: Communication-efficient sgd via gradient quantization and encoding. In *Advances in Neural Information Processing Systems*, pp. 1709–1720.

Allen-Zhu, Z., Z. Qu, P. Richtárik, and Y. Yuan (2016). Even faster accelerated coordinate descent using non-uniform sampling. In *International Conference on Machine Learning*, pp. 1110–1119.

Auer, P., N. Cesa-Bianchi, and P. Fischer (2002). Finite-time analysis of the multiarmed bandit problem. *Machine learning 47*(2-3), 235–256.

Glasmachers, T. and U. Dogan (2013). Accelerated coordinate descent with adaptive coordinate frequencies. In *Asian Conference on Machine Learning*, pp. 72–86.

Loshchilov, I., M. Schoenauer, and M. Sebag (2011). Adaptive coordinate descent. In *Proceedings of the 13th annual conference on Genetic and evolutionary computation*, pp. 885–892.

Moulines, E. and F. R. Bach (2011). Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, pp. 451–459.

Namkoong, H., A. Sinha, S. Yadlowsky, and J. C. Duchi (2017). Adaptive sampling probabilities for non-smooth optimization. In *International Conference on Machine Learning*, pp. 2574–2583.

Nesterov, Y. and V. Spokoiny (2017). Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics 17*(2), 527–566.

Nutini, J., M. Schmidt, I. Laradji, M. Friedlander, and H. Koepke (2015). Coordinate descent converges faster with the gauss-southwell rule than random selection. In *International Conference on Machine Learning*, pp. 1632–1641.

Qu, Z. and P. Richtárik (2016). Coordinate descent with arbitrary sampling i: Algorithms and complexity. *Optimization Methods and Software 31*(5), 829–857.

Richtárik, P. and M. Takáč (2013). On optimal probabilities in stochastic coordinate descent methods. *arXiv preprint arXiv:1310.3438*.

Robbins, H. and S. Monro (1951). A stochastic approximation method. *The annals of mathematical statistics*, 400–407.

Wangni, J., J. Wang, J. Liu, and T. Zhang (2018). Gradient sparsification for communication-efficient distributed optimization. In *Advances in Neural Information Processing Systems*, pp. 1299–1309.