Stochastic Distributed Learning with Gradient Quantization and Variance Reduction

King Abdullah University of Science and Technology
THE UNIVERSITY of EDINBURGH
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
MIPT MOSCOW INSTITUTE OF PHYSICS AND TECHNOLOGY

Samuel Horváth[1] Dmitry Kovalev[1] Konstantin Mishchenko[1] Peter Richtárik[1, 2, 3] Sebastian U. Stich[4]

[1]KAUST [2]University of Edinburgh [3]MIPT [4]EPFL

The Problem

Consider distributed optimization problem

$$\min_{x\in\mathbb{R}^d}\left[f(x) := \frac{1}{n}\sum_{i=1}^n f_i(x)\right] + R(x), \qquad (1)$$

where $f_i(x)$ for $i = 1,\dots,n$ are stored on i-th computing node and given as

$$f_i(x) = \frac{1}{m}\sum_{j=1}^m f_{ij}(x); \ m \text{ is large.} \qquad (2)$$

- n is the number of nodes,
- m is the number of functions stored on each node,
- $R(x)$ is a proper closed convex regularizer.

Quantization

Communication between computing nodes is often much more costly than local computations. We perform compression of communicated vectors via quantization.

Definition (ω-quantization)

A random operator $Q : \mathbb{R}^d \to \mathbb{R}^d$ with the properties

$$\mathbb{E}[Q(x)] = x, \qquad \mathbb{E}\left[\|Q(x)\|_2^2\right] \le (\omega+1)\|x\|_2^2$$

for all $x \in \mathbb{R}^d$ is a ω-quantization operator.

Example 1 (Random Dithering).

$$Q(x) = \|x\|_p \cdot \frac{1}{s}\cdot \operatorname{sign}(x)\circ\alpha, \quad \alpha_i = \left\lfloor s\frac{|x_i|}{\|x\|_p} + \xi_i\right\rfloor$$

for random vector $\xi \sim_{\text{u.a.r.}} [0,1]^d$, parameter $p \ge 1$, levels of rounding $s \in \{1,2,3,\dots\}$, where $\|x\|_p$ is a p-norm of x, $\circ$ is a Hadamard product. Random dithering is an ω-quantization for

$$\omega = \mathcal{O}\left(\frac{d^{1/p}+d^{1/2}}{s}\right).$$

Example 2 (Random Sparsification).

$$Q(x) = \frac{d}{r}\cdot\xi\circ x$$

for random variable $\xi \sim_{\text{u.a.r.}} \{y \in \{0,1\}^d : \|y\|_0 = r\}$ and sparsity parameter $r \in \{1,\dots,d\}$. Random sparsification is an ω-quantization for

$$\omega = \frac{d}{r} - 1.$$

Example 3 (Block Quantization). The vector $x \in \mathbb{R}^d$ is first split into t blocks: $x^\top = [v_1^\top,\dots,v_t^\top]$, $v_i \in \mathbb{R}^{d_i}$, $\sum_{i=1}^t d_i = d$. Then each block v_i is quantized using random dithering with $p = 2$, $s = 1$. Block quantization is an ω-quantization for

$$\omega = \max_{i\in\{1,\dots,t\}}\sqrt{d_i} + 1.$$

DIANA with Variance Reduction

Motivated by the idea of *compressed gradient differences* [1], we propose **the first variance reduced method for solving (1) and (2) that only computes gradients of $f_{ij}(x)$ and exchanges only quantized vector updates among workers.**

Algorithm 1 VR-DIANA based on L-SVRG (Variant 1), SAGA (Variant 2)

1. **Input:** learning rates $\alpha > 0$ and $\gamma > 0$, initial vectors $x^0, h_1^0,\dots,h_n^0$, $h^0 = \frac{1}{n}\sum_{i=1}^n h_i^0$
2. **for** $k = 0,1,\dots$ **do**
3. sample random $u^k = \begin{cases}1, & \text{with probability } \frac{1}{m}\\ 0, & \text{with probability } 1-\frac{1}{m}\end{cases}$
4. broadcast x^k, u^k to all workers
5. **for** $i = 1,\dots,n$ **do** ▷ worker side
6. pick random $j_i^k \sim_{\text{u.a.r.}} \{1,\dots,m\}$
7. $\mu_i^k = \frac{1}{m}\sum_{j=1}^m \nabla f_{ij}(w_{ij}^k)$
8. $g_i^k = \nabla f_{ij_i^k}(x^k) - \nabla f_{ij_i^k}(w_{ij_i^k}^k) + \mu_i^k$
9. $\hat\Delta_i^k = Q(g_i^k - h_i^k)$
10. $h_i^{k+1} = h_i^k + \alpha\hat\Delta_i^k$
11. **for** $j = 1,\dots,m$ **do**
12. ▷ Variant 1 (L-SVRG): update epoch
13. gradient if $u^k = 1$
14. $w_{ij}^{k+1} = \begin{cases}x^k, & \text{if } u^k = 1\\ w_{ij}^k, & \text{if } u^k = 0\end{cases}$
15. ▷ Variant 2 (SAGA): update gradient table
16. $w_{ij}^{k+1} = \begin{cases}x^k, & j = j_i^k\\ w_{ij}^k, & j \ne j_i^k\end{cases}$
17. **end for**
18. **end for**
19. $g^k = h^k + \frac{1}{n}\sum_{i=1}^n\hat\Delta_i^k$ ▷ gather quantized updates
20. $x^{k+1} = \operatorname{prox}_{\gamma R}(x^k - \gamma g^k)$
21. $h^{k+1} = h^k + \frac{\alpha}{n}\sum_{i=1}^n\hat\Delta_i^k$
22. **end for**

Experiments: Different Stepsizes α

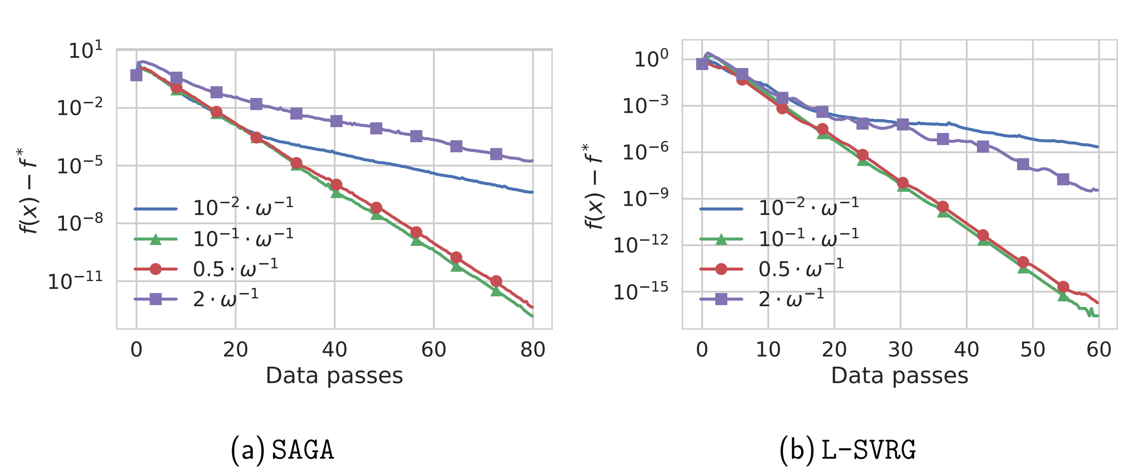

Figure 1: Comparison of VR methods with different parameter α for solving gisette with block size 2000, ℓ_2-penalty $\lambda_2 = 2\cdot 10^{-1}$, and ℓ_2 random dithering.

Convergence of VR-DIANA

We make the following technical assumptions:

Assumption 1. Functions $f_{ij} : \mathbb{R}^d \to \mathbb{R}$ are L-smooth.

Assumption 2. Functions $f_{ij} : \mathbb{R}^d \to \mathbb{R}$ are convex.

Assumption 3. Function $f : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex for $\mu > 0$.

Further by x^* we denote the optimal solution of (1).

Theorem 1 (Strongly convex case)

Let Assumptions 1, 2 and 3 hold. Let $\gamma = \frac{1}{L(1+36(\omega+1)/n)}$, $\alpha = \frac{1}{\omega+1}$. Then the number of iterations VR-DIANA needs to achieve precision $\mathbb{E}\left[\|x^k - x^*\|_2^2\right] \le \varepsilon$ is

$$\mathcal{O}\left(\left(\kappa + \kappa\frac{\omega}{n} + m + \omega\right)\log\frac{1}{\varepsilon}\right).$$

Further let x^a be a randomly chosen iterate of Algorithm 1, i.e.

$$x^a \sim_{\text{u.a.r.}} \{x^0, x^1,\dots,x^{k-1}\}.$$

Theorem 2 (Convex case)

Let Assumptions 1 and 2 hold. Let $\gamma = \frac{1}{2L\sqrt{m}\left(1+\frac{36(\omega+1)}{n}\right)}$, $\alpha = \frac{1}{\omega+1}$. Then the number of iterations VR-DIANA needs to achieve precision $\mathbb{E}[f(x^a) - f(x^*)] \le \varepsilon$ is

$$\mathcal{O}\left(\frac{\left(1+\frac{\omega}{n}\right)\sqrt{m} + \frac{\omega}{\sqrt{m}}}{\varepsilon}\right),$$

where $B_f(x,y) = f(x) - f(y) - \langle\nabla f(y), x-y\rangle$ is a Bregman divergence.

Theorem 3 (Non-convex case)

Let Assumption 1 hold and $R \equiv 0$. Let $\gamma = \frac{1}{10L\left(1+\frac{\omega}{n}\right)^{1/2}(m^{2/3}+\omega+1)}$, $\alpha = \frac{1}{\omega+1}$. Then the number of iterations VR-DIANA needs to achieve precision $\mathbb{E}\left[\|\nabla f(x^a)\|_2^2\right] \le \varepsilon$ is

$$\mathcal{O}\left(\left(1+\frac{\omega}{n}\right)^{1/2}\frac{m^{2/3}+\omega}{\varepsilon}\right).$$

References

[1] Konstantin Mishchenko, Eduard Gorbunov, Martin Takáč, and Peter Richtárik. Distributed learning with compressed gradient differences. *arXiv preprint arXiv:1901.09269*, 2019.

[2] Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. QSGD: Communication-efficient SGD via gradient quantization and encoding. In *Advances in Neural Information Processing Systems*, pages 1709–1720, 2017.

Experiments: Comparison with Existing Methods

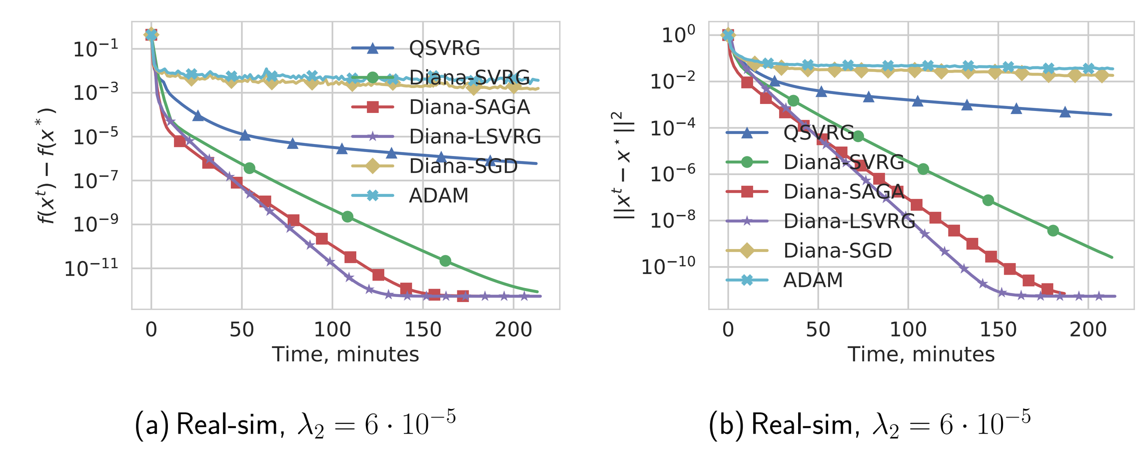

Figure 2: Comparison of VR-DIANA, Diana-SGD [1], QSVRG [2] and TernGrad-Adam with $n = 12$ workers on real-sim in suboptimality (left) and distance from the optimum (right). ℓ_∞ dithering is used for every method except for QSVRG, which uses ℓ_2 dithering.

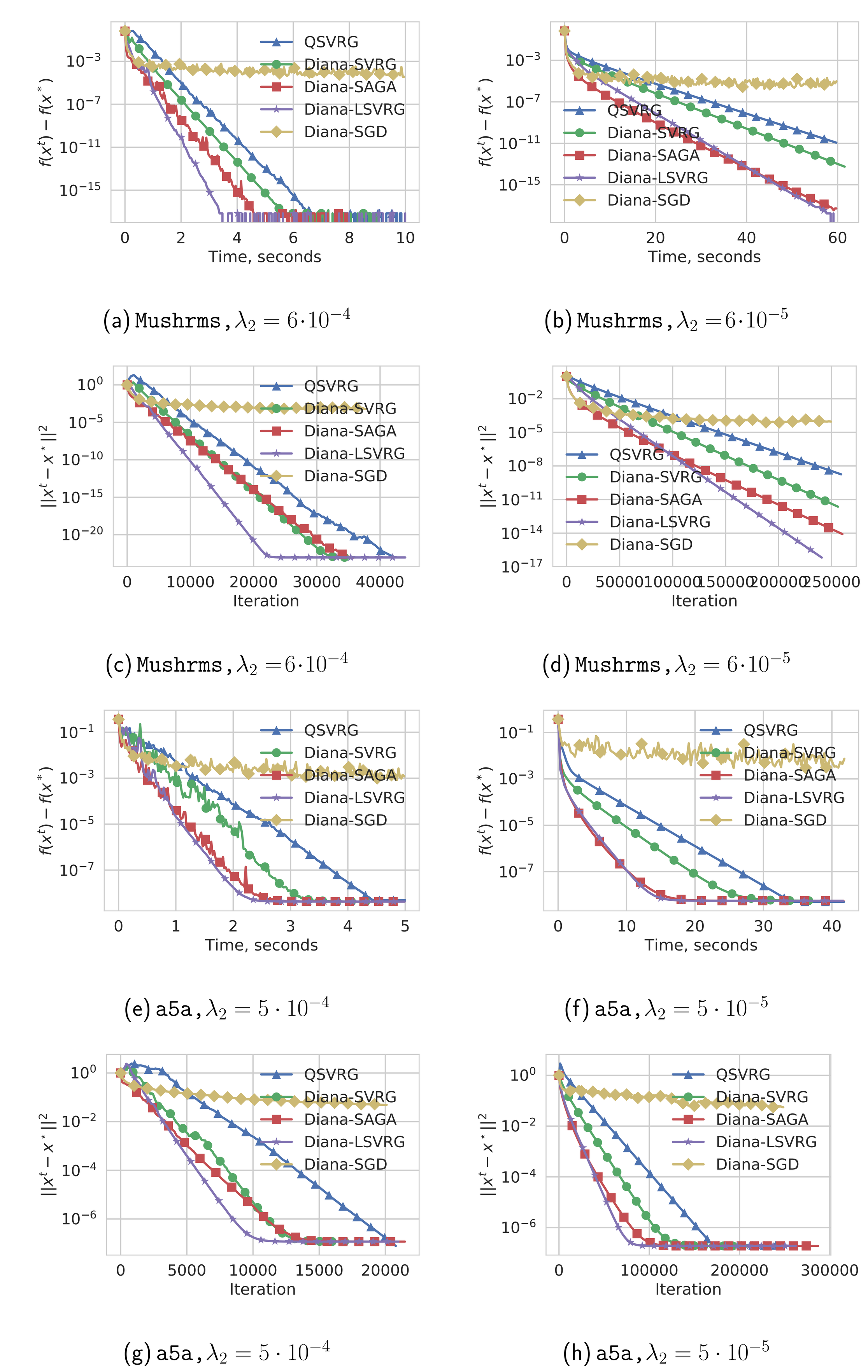

Figure 3: Comparison of VR-DIANA and Diana-SGD [1] against QSVRG [2] on mushrooms and a5a in suboptimality (top) and distance to the solution (bottom).