Neighbourhood watch: Variance Reduction using nearest-neighbours

Aurélien Lucchi Brian McWilliams Hamed Hassani Gabriel Krummenacher Thomas Hofmann

Institute of Machine Learning, ETH Zürich, Switzerland

Task

- We are interested in minimizing a loss function $f(x), x \in \mathbb{R}^d$ defined as

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x).$$

The minimizer of this function is written as $x^* = \arg\min_x f(x)$.

Stochastic Gradient Descent (SGD)

- Gradient descent used to tackle this problem uses the following update rule:

$$x^{t+1} = x^t - \eta^t f'(x^t), \tag{1}$$

where η^t is a chosen step-size for iteration t.

- SGD picks a random i and uses a stochastic update:

$$x^{t+1} = x^t - \eta^t f_i'(x^t). \tag{2}$$

Convergence (in expectation) is guaranteed if $\mathbf{E}_i[f_i'(x^t)] = f'(x^t)$ but SGD suffers from a **high variance**.

Variance reduction

- SAGA [1], SVRG [2], S2GD [3],... belong to a family of generalized SGD algorithms that exhibit lower variance and use the following update:

$$x^{t+1} = x^t - \eta v^t, \qquad \text{where} \quad \mathbf{E}[v^t] = f'(x^t) \tag{3}$$

- Convergence analysis considers

$$\begin{aligned}\mathbf{E}\|x^{t+1} - x^*\|^2 &= \mathbf{E}\left\|x^t - \eta v^t - x^*\right\|^2 \\ &= \|x^t - x^*\|^2 - 2\eta\left\langle x^t - x^*, f'(x^t)\right\rangle + \eta^2 \underbrace{\mathbf{E}\left\|v^t\right\|^2}_{\text{Variance}}\end{aligned} \tag{4}$$

The (negative) middle term is what guarantees progress for any gradient descent procedure on a (strongly) convex objective.

Bound for SAGA-style updates

- Introduce a **correction term** for f_i' denoted $g_i(\phi^t)$, so that $g(\phi^t) := \frac{1}{n}\sum_i g_i(\phi^t)$ denotes its expectation. We write

$$v^t := f_i'(x^t) - g_i(\phi^t) + g(\phi^t). \tag{5}$$

- We use the shorthand notation $\delta h(x) := h(x) - h(x^*)$. The variance term can be decomposed and bounded as follows:

$$\begin{aligned}\mathbf{E}\|v^t\|^2 &= \mathbf{E}[v^t]^2 + \mathbf{E}\|v^t - \mathbf{E}[v^t]\|^2 = \mathbf{E}\|v^t - f'(x^t)\|^2 + \|f'(x^t)\|^2 \\ &\le (1+\beta)\mathbf{E}\|\delta f_i'(x^t)\|^2 + (1+\beta^{-1})\mathbf{E}\|\delta g_i(\phi^t)\|^2 \\ &\quad - \beta\|f'(x^t)\|^2 - (1+\beta^{-1})\|g(\phi^t)\|^2,\end{aligned} \tag{6}$$

where $\beta > 0$.

- **The variance term vanishes as we convergence to the optimum.**

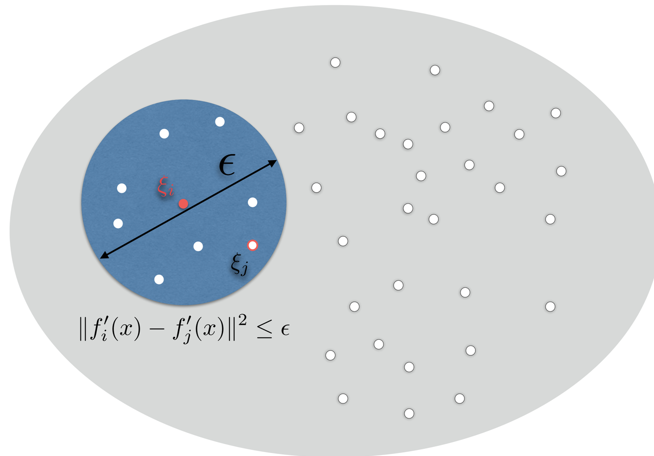
Generalizing SAGA updates

- Propose to generalize the correction to a weighted (convex) sum of the following type (exploiting some clustering structure in how to chose τ):

$$g_i(\phi^t) = \sum_j \tau_{ij} f_j'(\phi_j^t), \quad \text{with} \quad \sum_j \tau_{ij} = 1\ (\forall i), \quad \tau_{ij} \ge 0\ (\forall i,j), \tag{7}$$

where $\tau_{ij} = \delta_{ij}$ is the special case of the original SAGA algorithm

- In our algorithm, we use $\tau_{ij} \in \{0,1\}$, **i.e. selection of one "neighbor"** j whose $\ell 2$-distance from i is less than ϵ



Algorithm

1. **INPUTS :**
2. $\mathcal{D}$: Training set of n examples.
3. η : Step size
4. ϵ : Neighborhood size
5. **OUTPUT :** x^T
6. Cluster datapoints in $\mathcal{D}$ whose distance is less than ϵ
7. **for** $t = 1 \dots T$ **do**
8. Randomly pick $i \in 1 \dots n$
9. $x^t = x^{t-1} - \eta(\nabla f_i(x^{t-1}) - (g_i(\phi^t) - \mathbf{E}[g_i(\phi^t)])$
10. **end for**

Convergence properties

- Pivot the analysis around $f_j'(x^*)$ instead of $f_i'(x^*)$.
- The variance converges to $2(1+\beta)(\mathbf{E}\|f_i'(x^*) - f_j'(x^*)\|^2)$ as $x_t \to x^*$
- Proof convergence sketch:
  - Define the following Lyapunov function:

$$T^t := \|x^t - x^*\|^2 + \alpha\left[\delta f(\phi^t) - \langle f_i'(x^*), \phi^t - x^*\rangle\right] \tag{8}$$

  - Show that

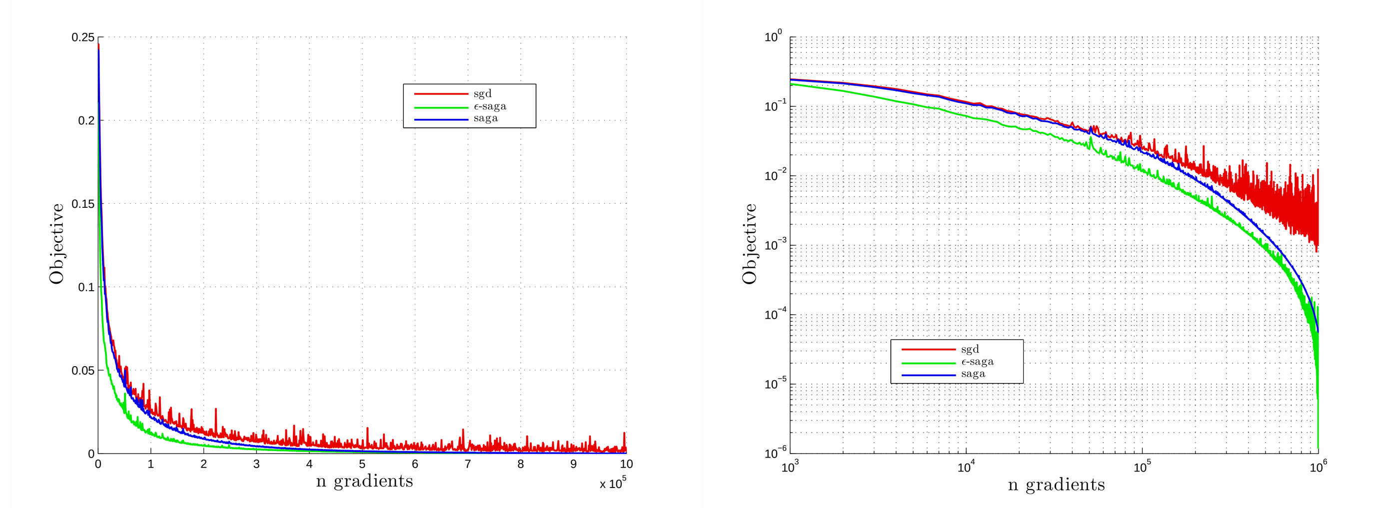
$$\mathbf{E}\|x^{t+1} - x^*\|^2 \le \rho^t T^0 + O(\eta^2\epsilon), \qquad \rho < 1$$

- In general, there is a **constant non vanishing error** of $O(\eta^2\epsilon)$
- Open question: How can we get $\epsilon \to 0$?

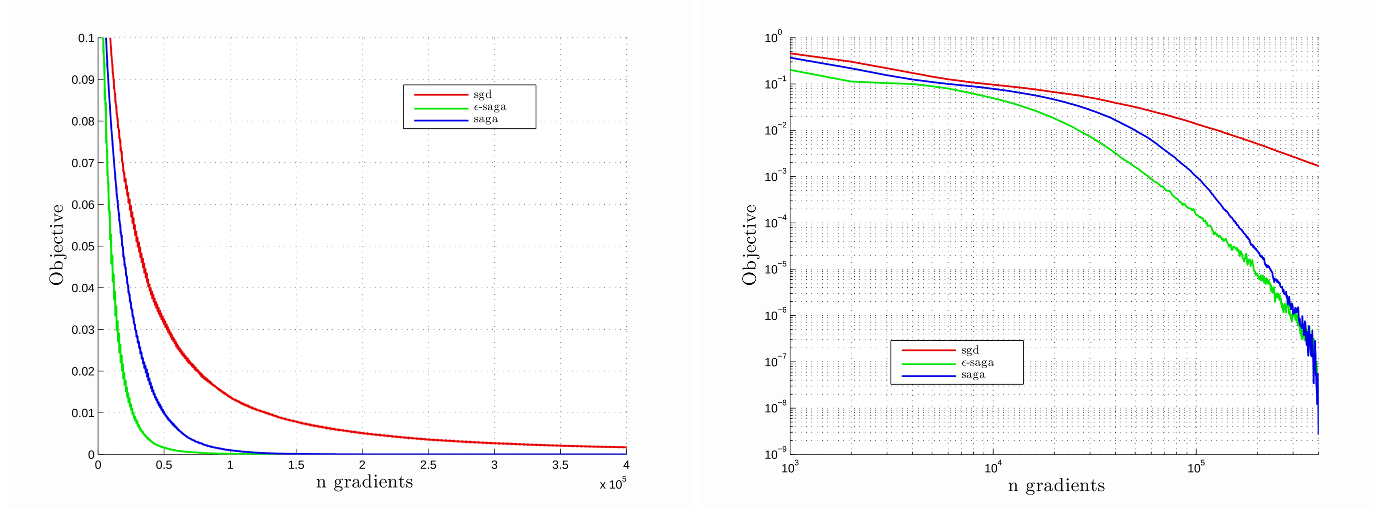
Results

- For SGD, we used $\eta_t = \frac{\eta_0 T_0}{T_0 + t}$
- Both SAGA and ϵ-SAGA (our method) use a constant step-size
- We added an $\ell 2$-regularizer with parameter $\lambda = 10^{-4}$

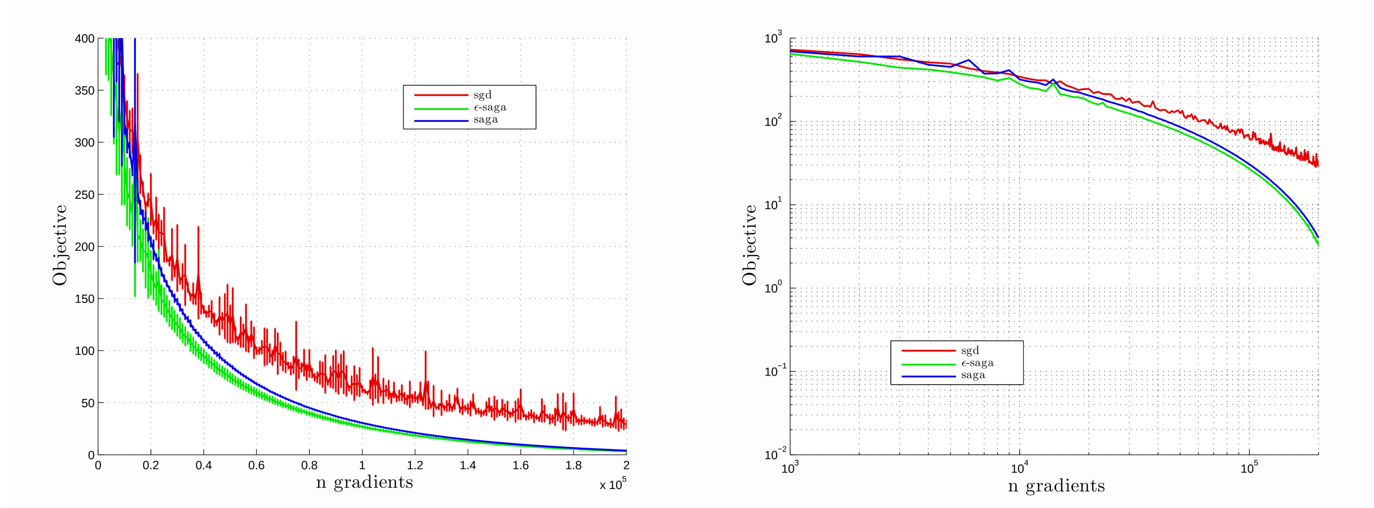
COV dataset *(logistic regression)*



IJCN dataset *(logistic regression)*



Year dataset *(least-squares regression)*



Future work

- Strengthen theoretical guarantees
- Study conditions required for $\epsilon \to 0$
- Generalization bound

References

[1] A. Defazio, F. Bach, and S. Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, pages 1646–1654, 2014.

[2] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pages 315–323, 2013.

[3] J. Konečný and P. Richtárik. Semi-stochastic gradient descent methods. *arXiv preprint arXiv:1312.1666*, 2013.