ETH zürich

Neighbourhood watch: Variance Reduction using nearest-neighbours

Aurélien Lucchi
Collaborators: B. McWilliams, H. Hassani, G. Krummenacher and T. 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)

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

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

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

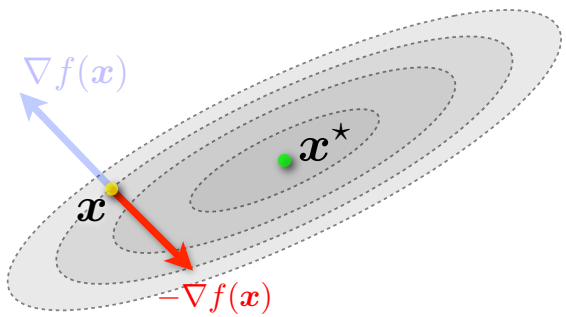


Figure : Credit: Martin Jaggi

Variance reduction

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

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

- Convergence analysis considers

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

Bound for SAGA-style updates

- Introduce a **correction term** for f_i' denoted $g_i(\phi^t)$, so that

$$v^t := f_i'(x^t) - g_i(\phi^t) + \frac{1}{n}\sum_i g_i(\phi^t).$$

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

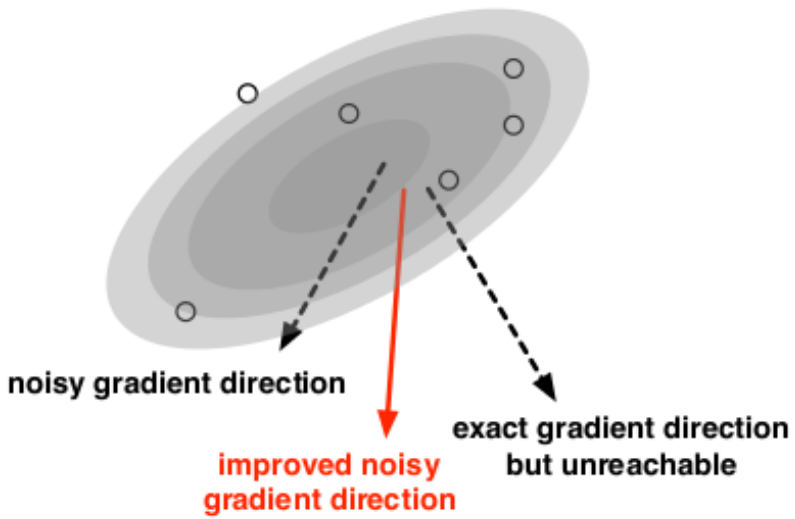


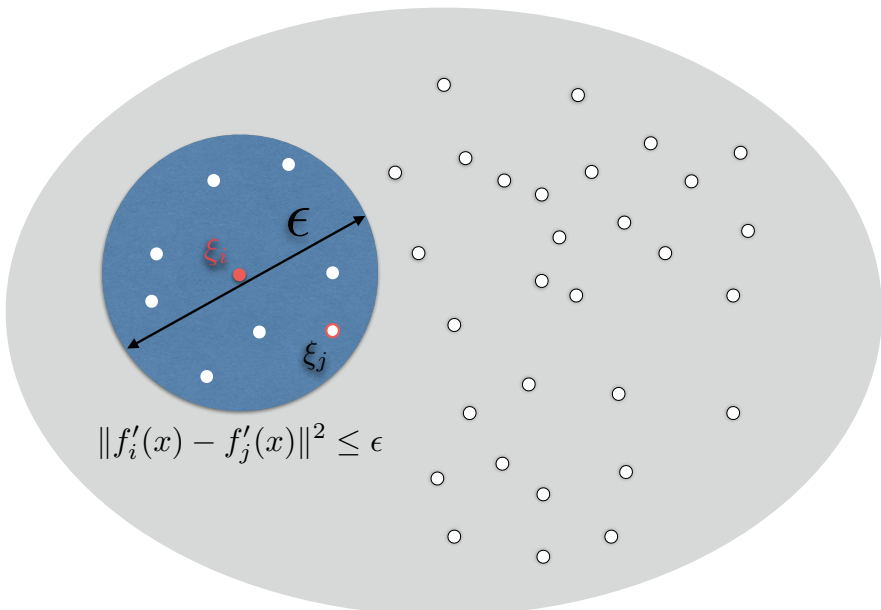
Figure : Credit: Chong Wang, 2013

Generalizing SAGA updates

- Propose to generalize the correction to a weighted (convex) sum of the following type:

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

- Set $\tau_{ij} = \delta_{ij}$ to recover SAGA
- Exploit some clustering structure in how to chose τ
- 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



Neighbourhood watch

- Convergence within an ϵ ball to the optimum
- Open question: How can we get $\epsilon \to 0$?
- Experiment on three datasets show significant gains compared to SGD and SAGA.

Thank you very much for your attention

Come to our poster!