# Variance reduction for stochastic gradient methods

Axel Böhm

September 13, 2021

A common Task in (supervised) machine learning:

$$\min_{x \in \mathbb{R}^d} f(x) := \frac{1}{n} \sum_{i=1}^{n} \underbrace{\text{loss for } i\text{-th sample}} + \underbrace{\psi(x)}_{\text{regularizer}}$$

where the $i$-th sample is $(a_i, y_i)$.

- linear regression: $f_i(x) = (a_i^T x - y_i)^2$, and $\psi = 0$
- logistic regression: $f_i(x) = \log(1 + e^{-y_i a_i^T x})$, and $\psi = 0$ "sigmoid function" and logistic loss.
- Lasso: $f_i$ as for linear regression but $\psi(x) = \|x\|_i$
- SVM: $f_i(x) = \max\{0, 1 - y_i a_i^T x\}$ and $\psi(x) = \|x\|^2$

# Stochastic gradient descent

**Algorithm** SGD

**for** $k = 1, 2, \ldots$ **do**
    pick $i_k$ uniform at random in $[n]$
    $x_{k+1} = x_k - \alpha_k \nabla f_{i_k}(x_k)$

We already noticed that:

- large stepsizes fail to suppress variability in the stoch. gradients  leads to oscilations
- decreasing stepsizes mitigate this problem but slows down convergence (too conservative)

# Recall SGD (template)

$$x_{k+1} = x_k - \alpha_k g_k$$

- $g_k$ is an unbiased estimator of the true gradient $\nabla F(x_k)$
- convergence depends on variance $\mathbb{E}[\|g_k - \nabla F(x_k)\|] \leq \sigma_g$
- vanilla SGD $g_k = \nabla f_{i_k}(x_k)$
  **issue:** $\sigma_g$ is non-negligible even close to the solution
- **Q:** can we choose $g_k$ in a different way to reduce variability?

# A simple idea

Consider

- estimator $X$ for parameter $\mu$ ($\mathbb{E}[X] = \mu$ and $\mathbb{V}[X] = \sigma^2$)
- want to keep unbiased but reduce variance
- find $Y$ such that $\mathbb{E}[Y] = 0$ but $\mathrm{Cov}(X, Y)$ is large and $\tilde{X} = X - Y$
- remains unbiased

$$\mathbb{V}[\tilde{X}] = \mathbb{V}[X] + \mathbb{V}[Y] - 2\,\mathrm{Cov}[X, Y]$$

- can be much smaller than $[X]$ if $X, Y$ are highly correlated

# Stochastic average gradient (SAG), 2013

- maintain table containing gradients $g_i$ of $f_i$
- at step $k = 1, 2, \dots$ pick random $i_k \in [n]$ and

$$g_i^k := \nabla f_{i_k}(x^k)$$

set $g_i^k = g_i^{k-1}$ for all $i \neq i_k$ (remain the same)
- Update

$$x^{k+1} = x^k - \alpha_k \frac{1}{n} \sum_{i=1}^n g_i^k$$

- assuming gradients do not change too much along trajectory
- gradient estimator no longer unbiased
- Isn't it expensive to average these gradients?

$$x^{k+1} = x^k - u\alpha_k \left( \frac{g_{i_k}^k}{n} - \frac{g_{i_k}^{k-1}}{n} + \underbrace{\frac{1}{n} \sum_{i=1}^n g_i^k}_{\text{old table average}} \right)$$

# SAG variance reduction

Gradient estimator in SAG:

$$x^{k+1} = x^k - \alpha_k \frac{1}{n}\Bigg( \underbrace{g_{i_k}^k}_{X} - \underbrace{g_{i_k}^{k-1} - \sum_{i=1}^{n} g_i^k}_{Y} \Bigg)$$

- While $\mathbb{E}[X] = \nabla f(x^k)$, but $\mathbb{E}[Y] \neq 0 \to$ biased estimator
- $X$ and $Y$ are correlated as $X - Y \to 0$ as
- $x^k$ and $x_{k-1}$ both converge to $x^*$ and
- the last term converges to $\nabla f(x^*) = 0$

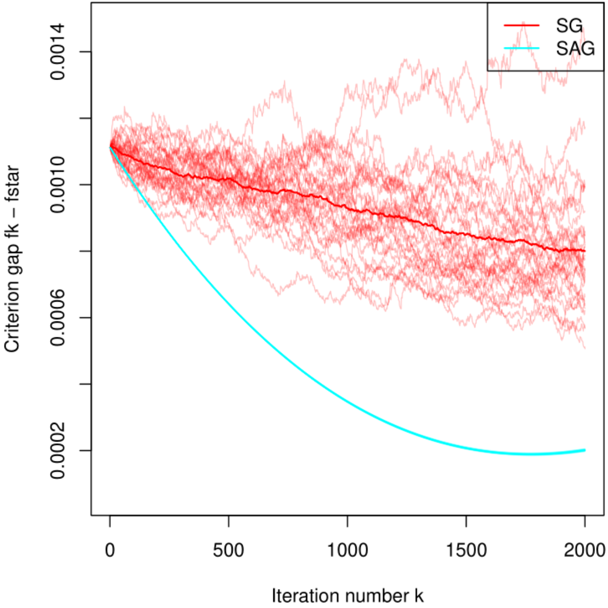
## Convergence

As always, initialization plays a role: $D^2 := \|x^0 - x^*\|^2$.

$$\begin{aligned} \text{SAG:} \quad & \frac{n}{k}(f(x^0) - f^*) + \frac{L}{k}D^2 \\ \text{GD:} \quad & \frac{L}{k}D^2 \\ \text{SAG:} \quad & \frac{L}{\sqrt{k}}D^2 \end{aligned}$$

- Achieves linear convergence in the *strongly* convex setting.
- proofs are difficult (and computer-aided)

Same gradient oracle cost as SGD, but same converge rate as GD.


SG
SAG
Criterion gap fk – fstar
0.0002
0.0006
0.0010
0.0014
0
500
1000
1500
2000
Iteration number k

# SAGA, 2014

Very similar:

- maintain table containing gradients $g_i$ of $f_i$
- at step $k = 1, 2, \ldots$ pick random $i_k \in [n]$ and

$$g_i^k := \nabla f_{i_k}(x^k)$$

set $g_i^k = g_i^{k-1}$ for all $i \neq i_k$ (remain the same)

- Update

$$x^{k+1} = x^k - u\alpha_k \left( g_{i_k}^k - g_{i_k}^{k-1} + \frac{1}{n} \sum_{i=1}^{n} g_i^k \right)$$

- estimator now **unbiased**!

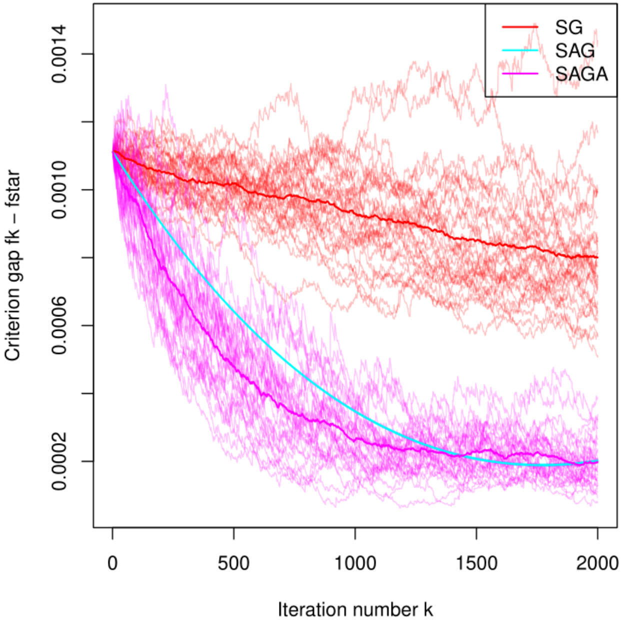
# For Comparison

SAGA gradient estimate:

$$g_{i_k}^k - g_{i_k}^{k-1} + \frac{1}{n}\sum_{i=1}^n g_i^k$$

SAGA gradient estimate:

$$\frac{1}{n}g_{i_k}^k - \frac{1}{n}g_{i_k}^{k-1} + \frac{1}{n}\sum_{i=1}^n g_i^k$$


SG
SAG
SAGA
Criterion gap fk – fstar
0.0002
0.0006
0.0010
0.0014
0
500
1000
1500
2000
Iteration number k

# Stochastic Variance Reduced Gradient (SVRG), 2013

**Algorithm** SVRG

1: **for** $k = 1, 2, \dots$ **do**
2: Set $x^0 = \tilde{x} = \tilde{x}^k$
3: Compute $\tilde{\mu} :- \nabla f(\tilde{x})$
4: **for** $l = 1, 2, \dots, m$ **do**
5: pick $i_l$ uniform at random in $[n]$
6: Set $x^{l+1} = x^l - \alpha(\nabla f_{i_l}(x^l) - \nabla f_{i_l}(\tilde{x}) + \tilde{\mu})$

- Does not need to store full table of gradients.
- requires *batch* gradient computation every *epoch*
- convergence rates similar to SAGA, but simpler analysis.
- per iteration cost of SVRG is comparable to that of SGD if $m \geq n$