SVRG and Katyusha are Better Without the Outer Loop

("Don't Jump Through Hoops and Remove Those Loops")

Dmitry Kovalev **Samuel Horváth** Peter Richtárik

KAUST
King Abdullah University of Science and Technology

Introduction

Empirical Risk Minimization

$$\min_{x \in \mathbb{R}^d} f(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x)$$

μ strongly-convex

n is big

convex
L-smooth

Baselines

Gradient Descent

$$x^{k+1} = x^k - \eta \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(x^k)$$

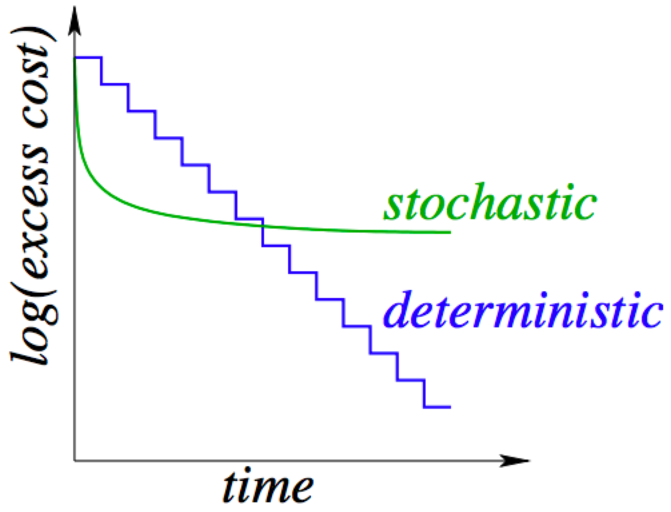
$\kappa = \frac{L}{\mu}$

Complexity – stochastic gradient computations

$\|x^k - x^\star\|^2 \leq \epsilon$

$$\mathcal{O}\left(n\kappa \log \frac{1}{\epsilon}\right)$$

Baselines



Stochastic Gradient Descent

$$x^{k+1} = x^k - \eta_k \nabla f_{i_k}(x^k)$$

Complexity

$$\mathbb{E}\|x^k - x^\star\|^2 \leq \epsilon + \sigma$$

$$\mathcal{O}\left(\underbrace{(n+\kappa)\log\frac{1}{\epsilon}}\right)$$

just to neighborhood, scales with $\mathcal{O}\left(\frac{1}{\mu n}\sum_{i=1}^{n}\|\nabla f_i(x^\star)\|^2\right)$

(Gower et al. 2019)

image credits
(https://www.cs.ubc.ca/~schmidtm/Documents/2014_Google_SAG.pdf)

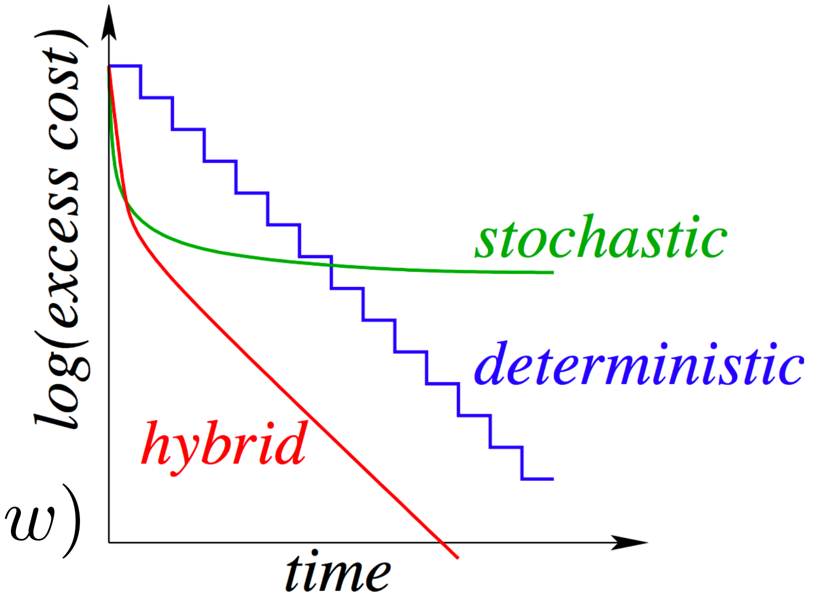
Variance Reduction

Control variates

$$Z = X + \beta(Y - \mathbb{E}(Y))$$

Optimization $\boxed{\nabla f_i(x)}$

Variance Reduction



SVRG $Z^k = \nabla f_{i^k}(x^k) - f_{i^k}(w) + \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(w)$

(Johnson & Zhang 2013)

SAGA $Z^k = \nabla f_{i^k}(x^k) - f_{i^k}(w_{i^k}) + \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(w_i)$

(Defazio et al. 2014)

Complexity

$$\mathbb{E}\|x^k - x^\star\|^2 \leq \epsilon$$

First method **SAG** (Roux et al. 2012)

$$\mathcal{O}\left((n+\kappa)\log\frac{1}{\epsilon}\right)$$

image credits
(https://www.cs.ubc.ca/~schmidtm/Documents/2014_Google_SAG.pdf)

Comparison

Algorithm 1 SVRG

Parameters: stepsize $\eta > 0$, inner-cycle size m
Initialization: $x^0 = w^0 \in \mathbb{R}^d$
for $k = 0, 1, 2, \dots$ **do**
  Sample $i \in \{1, \dots, n\}$ uniformly at random
  $g^k = \nabla f_i(x^k) - \nabla f_i(w^k) + \nabla f(w^k)$
  $x^{k+1} = x^k - \eta g^k$
  if $k \mod m = 0$ **then**
    $w^{k+1} = x^k$
  else
    $w^{k+1} = w^k$
  end if
end for

Algorithm 2 SAGA

Parameters: stepsize $\eta > 0$
Initialization: $x^0 = w_j^0 \in \mathbb{R}^d, \forall j \in [n]$
for $k = 0, 1, 2, \dots$ **do**
  Sample $i \in \{1, \dots, n\}$ uniformly at random
  $g^k = \nabla f_i(x^k) - \nabla f_i(w_i^k) + \frac{1}{n} \sum_{j=1} \nabla f_j(w_j^k)$
  $x^{k+1} = x^k - \eta g^k$
  for $j = 1, 2, \dots, n$ **do**
    if $i = j$ **then**
      $w_j^{k+1} = x^k$
    else
      $w_j^{k+1} = w_j^k$
    end if
  end for
end for

Advantages and Disadvantages

SAGA
- Bad: High storage requirements: n vectors
- Good: Adaptive to strong convexity

usually not known to practitioners

SVRG
- Bad: The inner-loop size (m) depends on the condition number
- Good: Low storage requirements: O(1) vectors

Can we construct a method combining the advantages?

New Method

New Method: L-SVRG

Algorithm 3 Loopless SVRG (L-SVRG)

Parameters: stepsize $\eta > 0$, probability $p \in (0, 1]$
Initialization: $x^0 = w^0 \in \mathbb{R}^d$
for $k = 0, 1, 2, \ldots$ **do**
  Sample $i \in \{1, \ldots, n\}$ uniformly at random
  $g^k = \nabla f_i(x^k) - \nabla f_i(w^k) + \nabla f(w^k)$
  $x^{k+1} = x^k - \eta g^k$

$$w^{k+1} = \begin{cases} x^k & \text{with probability } p \\ w^k & \text{with probability } 1 - p \end{cases}$$

end for

One vector to keep in memory

Random inner-loop size $m = \frac{1}{p}$

p can be anything between c/n and c/κ for some constant c

Analysis

Lyapunov Function

$$\Phi^k \overset{\text{def}}{=} \left\|x^k - x^*\right\|^2 + \mathcal{D}^k$$

$$\mathcal{D}^k \overset{\text{def}}{=} \frac{4\eta^2}{pn} \sum_{i=1}^{n} \left\|\nabla f_i(w^k) - \nabla f_i(x^*)\right\|^2$$

Strong Convexity + Unbiasedness

$$
\begin{aligned}
\mathrm{E}\left[\left\|x^{k+1}-x^*\right\|^2\right] &= \mathrm{E}\left[\left\|x^k-x^*-\eta g^k\right\|\right]^2 \\
&\overset{\text{Alg. 3}}{=} \left\|x^k-x^*\right\|^2+\mathrm{E}\left[2\eta\left\langle g^k, x^*-x^k\right\rangle\right]+\eta^2\mathrm{E}\left[\left\|g^k\right\|^2\right] \\
&\overset{(2)}{=} \left\|x^k-x^*\right\|^2+2\eta\left\langle\nabla f(x^k), x^*-x^k\right\rangle+\eta^2\mathrm{E}\left[\left\|g^k\right\|^2\right] \\
&\overset{(4)}{\leq} \left\|x^k-x^*\right\|^2+2\eta\left(f^*-f(x^k)-\frac{\mu}{2}\left\|x^k-x^*\right\|\right)+\eta^2\mathrm{E}\left[\left\|g^k\right\|^2\right] \\
&= \left\|x^k-x^*\right\|^2(1-\eta\mu)+2\eta\left(f^*-f(x^k)\right)+\eta^2\mathrm{E}\left[\left\|g^k\right\|^2\right].
\end{aligned}
$$

Bounding Variance

$$\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$$

$$\begin{aligned}
\mathrm{E}\left[\left\|g^k\right\|^2\right] &\overset{\text{Alg. 3}}{=} \mathrm{E}\left[\left\|\nabla f_i(x^k) - \nabla f_i(x^*) + \nabla f_i(x^*) - \nabla f_i(w^k) + \nabla f(w^k)\right\|^2\right] \\
&\overset{(25)}{\leq} 2\mathrm{E}\left[\left\|\nabla f_i(x^k) - \nabla f_i(x^*)\right\|^2\right] + 2\mathrm{E}\left[\left\|\nabla f_i(x^*) - \nabla f_i(w^k) - \mathrm{E}\left[\nabla f_i(x^*) - \nabla f_i(w^k)\right]\right\|^2\right] \\
&\overset{(3),(24)}{\leq} 4L(f(x^k) - f^*) + 2\mathrm{E}\left[\left\|\nabla f_i(w^k) - \nabla f_i(x^*)\right\|^2\right] \\
&\overset{(10)}{=} 4L(f(x^k) - f^*) + \frac{p}{2\eta^2}\mathcal{D}^k.
\end{aligned}$$

$$\mathbb{E}\|X - E[X]\|^2 \leq \mathbb{E}\|X\|^2$$

$$\begin{aligned}
\mathrm{E}\left[\mathcal{D}^{k+1}\right] &\overset{\text{Alg. 3}}{=} (1-p)\mathcal{D}^k + p\frac{4\eta^2}{pn}\sum_{i=1}^{n}\left\|\nabla f_i(x^k) - \nabla f_i(x^*)\right\|^2 \\
&\overset{(3)}{\leq} (1-p)\mathcal{D}^k + 8L\eta^2(f(x^k) - f^*).
\end{aligned}$$

Putting it all Together

$$
\begin{aligned}
\mathrm{E}\left[\left\|x^{k+1}-x^*\right\|^2+\mathcal{D}^{k+1}\right] &\overset{(12),(14)}{\leq} (1-\mu\eta)\left\|x^k-x^*\right\|^2+2\eta(f^*-f(x^k))+\eta^2\mathrm{E}\left[\left\|g^k\right\|^2\right] \\
&\quad +(1-p)\mathcal{D}^k+8L\eta^2(f(x^k)-f^*) \\
&\overset{(13)}{\leq} (1-\mu\eta)\left\|x^k-x^*\right\|^2+(1-p)\mathcal{D}^k+(2\eta-8L\eta^2)(f^*-f(x^k)) \\
&\quad +\eta^2\left(4L(f(x^k)-f^*)+\frac{p}{2\eta^2}\mathcal{D}^k\right) \\
&= (1-\mu\eta)\left\|x^k-x^*\right\|^2+\left(1-\frac{p}{2}\right)\mathcal{D}^k+(2\eta-12L\eta^2)(f^*-f(x^k))
\end{aligned}
$$

Now we use the fact that $\eta \leq \frac{1}{6L}$ and obtain the desired inequality:

$$
\mathrm{E}\left[\left\|x^{k+1}-x^*\right\|^2+\mathcal{D}^{k+1}\right] \leq (1-\mu\eta)\left\|x^k-x^*\right\|^2+\left(1-\frac{p}{2}\right)\mathcal{D}^k.
$$

Main Result

p can be anything between c/n and c/κ for some constant c

Theorem 4.5. Let $\eta = \frac{1}{6L}$, $p = \frac{1}{n}$. Then $\mathrm{E}\left[\Phi^k\right] \leq \varepsilon \Phi^0$ as long as

$$k \geq \mathcal{O}\left(\left(n + \frac{L}{\mu}\right) \log \frac{1}{\varepsilon}\right).$$

Acceleration

Katyusha (Allen-Zhu, 2017)

Algorithm 4 Loopless Katyusha (`L-Katyusha`)

Parameters: θ_1, θ_2, probability $p \in (0,1]$

Initialization: Choose $y^0 = w^0 = z^0 \in \mathbb{R}^d$, stepsize $\eta = \frac{\theta_2}{(1+\theta_2)\theta_1}$ and set $\sigma = \frac{\mu}{L}$

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

$x^k = \theta_1 z^k + \theta_2 w^k + (1 - \theta_1 - \theta_2) y^k$

Sample $i \in \{1, \ldots, n\}$ uniformly at random

$g^k = \nabla f_i(x^k) - \nabla f_i(w^k) + \nabla f(w^k)$

$z^{k+1} = \frac{1}{1+\eta\sigma}\left(\eta\sigma x^k + z^k - \frac{\eta}{L} g^k\right)$

$y^{k+1} = x^k + \theta_1 (z^{k+1} - z^k)$

$$\boxed{w^{k+1} = \begin{cases} y^k & \text{with probability } p \\ w^k & \text{with probability } 1-p \end{cases}}$$

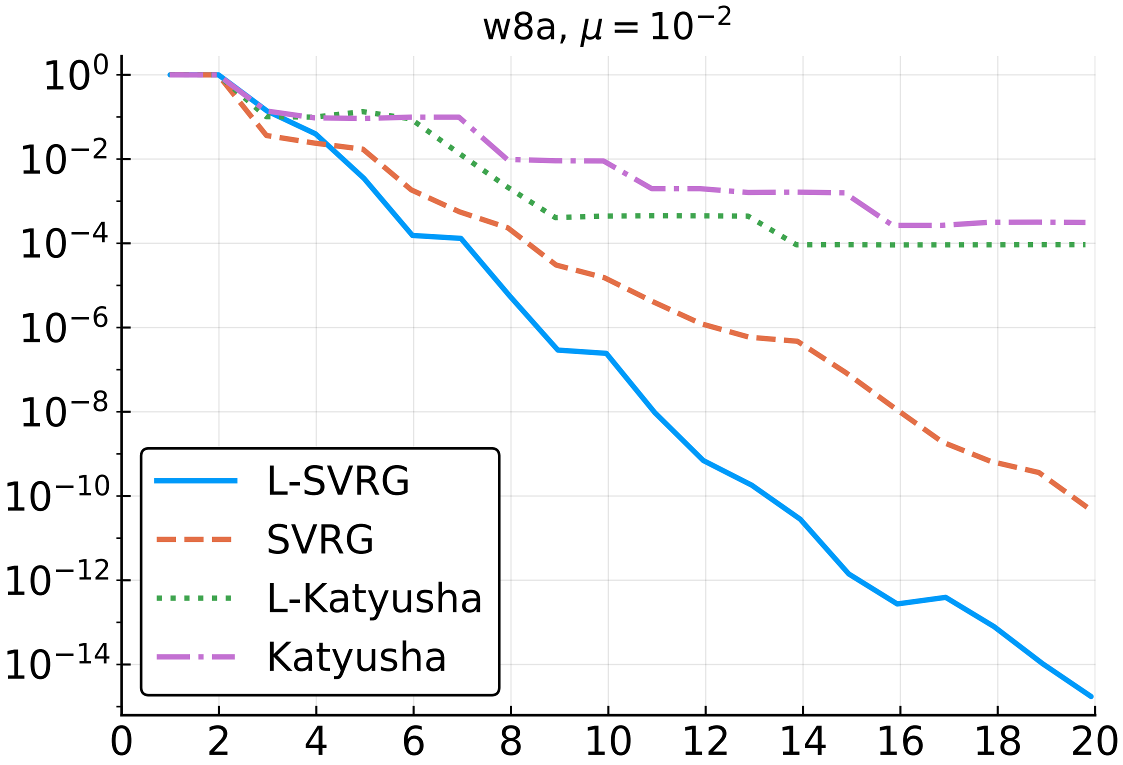
end for

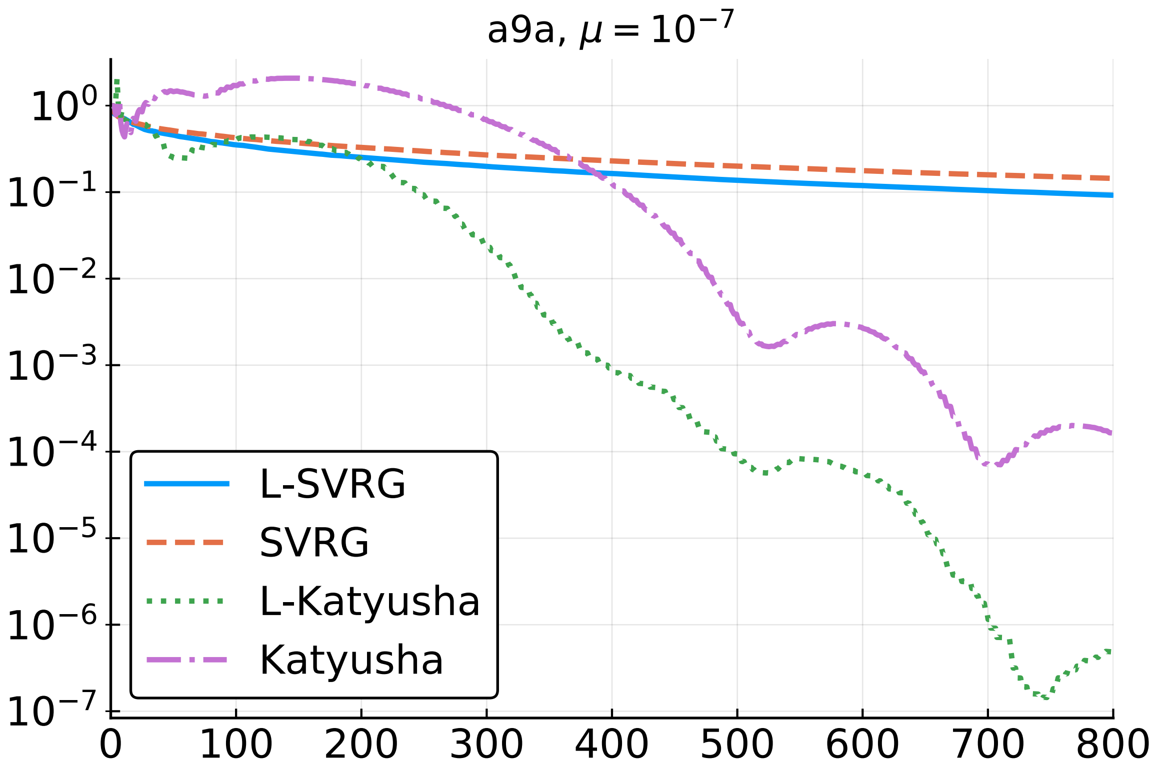
$$\boxed{p = \frac{1}{n} \implies \mathcal{O}\left(n + \sqrt{n\kappa}\log\frac{1}{\epsilon}\right)}$$

- The same rate was shown by Allen-Zhu for case $m = n$ (inner-loop size)

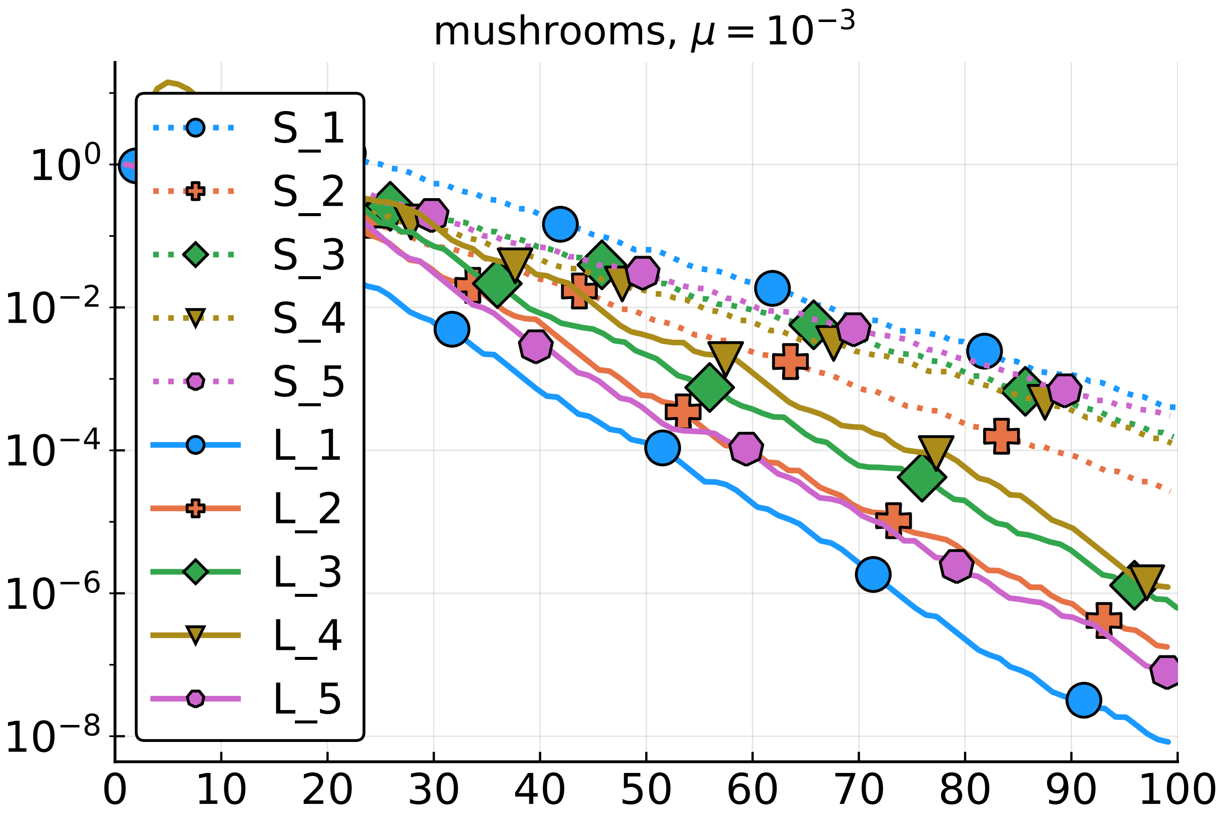
Experiments

Experiment 1





Experiment 2



Conclusions

Conclusions

- We proposed loopless variants of SVRG (Johnson and Zhang, 2013) and Katyusha (Allen-Zhu, 2017)
  - simplified analysis (shorter)
  - more insightful analysis (Lyapunov function)
  - robust to parameter settings (can choose p from a large interval)
  - better in practice

- A step towards parameter-adaptive methods (Lei & Jordan, 2019),
  - L-SVRG does not need to know the condition number

Thank you!