COR@L
COMPUTATIONAL OPTIMIZATION RESEARCH AT LEHIGH

LEHIGH UNIVERSITY

ISE Industrial and Systems Engineering

SARAH:
A Novel Method for Machine Learning Problems Using StochAstic Recursive GrAdient AlgoritHm

Martin Takáč

LEHIGH UNIVERSITY

August 22, 2017

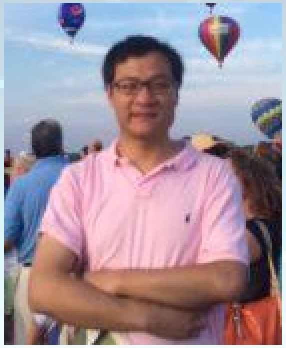
DIMACS Workshop on Distributed Optimization, Information Processing, and Learning

COR@L
COMPUTATIONAL OPTIMIZATION
RESEARCH AT LEHIGH

ISE

Industrial and
Systems Engineering

Lam Nguyen
(Lehigh)



Jie Liu
(Lehigh)

Katya Scheinberg
(Lehigh)

SARAH:

A Novel Method for Machine Learning Problems Using StochAstic Recursive GrAdient AlgoritHm

Martin Takáč

LEHIGH
UNIVERSITY

August 22, 2017

DIMACS Workshop on Distributed Optimization, Information Processing, and Learning

Empirical Loss Minimization

Nature of Data

DATA

$A_i \in \mathbb{R}^{d \times m}$

LABEL

Traffic sign - STOP

$y_i \in \mathbb{R}^m$

Impossible to know

$$(A_i, y_i) \sim Distribution$$

Goal: Predict Labels

Choose a family of prediction functions $\phi(x; w)$ parametrized by w

Example: linear predictor $\phi(x; w) = x^T w$

Task: Find a good w !

$$(A_i, y_i) \sim Distribution$$

$$\phi(A_i; w) \approx y_i$$

Choose a **loss function** to measure the success/failure

$$\ell(a, b) \qquad \ell(a, b) = ||a - b||^2$$

prediction

true label

Training Phase

$$\min_{w} \quad \mathbf{E}[\ell(\phi(A_i; w), y_i)]$$

$$(A_i, y_i) \sim Distribution$$

Sample n i.i.d. points

$$\{(A_i, y_i)\}_{i=1}^{n} \sim Distribution$$

$$\min_{w} \quad \frac{1}{n} \sum_{i=1}^{n} [\ell(\phi(A_i; w), y_i)] + Reg(w)$$

Stochastic Gradient Descent

H. Robbins and S. Monro
A Stochastic Approximation Method
Annals of Mathematical Statistics 22, pp. 400–407, 1951

Stochastic Gradient Descent

$$\min_{w} \left\{ F(w) := \frac{1}{n}\sum_{i=1}^{n} f_i(w) \right\}$$

ML applications => $n \gg 1$

Stochastic Gradient Descent

$i \in \{1, 2, \ldots, n\}$

$\mathbf{E}[\nabla f_i(w)] = \nabla F(w)$

1. choose w_0
2. for $t = 0, 1, 2, \ldots$
3. $\quad w_{t+1} = w_t - \eta_t \nabla f_i(w_t)$

Convergence

If F(w) has Lipschitz continuous gradient (for simplicity L=1):

$$F(w_{t+1}) \leq F(w_t) + \langle \nabla F(w_t), w_{t+1} - w_t \rangle + \frac{1}{2}\|w_{t+1} - w_t\|^2$$

$$w_{t+1} = w_t - \eta_t \nabla f_i(w_t)$$

$$F(w_{t+1}) \leq F(w_t) + \langle \nabla F(w_t), -\eta_t \nabla f_i(w_t) \rangle + \frac{1}{2}\|\eta_t \nabla f_i(w_t)\|^2$$

Take conditional expectation with respect to "i"

$$\mathbf{E}[F(w_{t+1})|w_t] \leq F(w_t) \boxed{- \eta_t \|\nabla F(w_t)\|^2} + \boxed{\frac{\eta_t^2}{2}}\boxed{\mathbf{E}[\|\nabla f_i(w_t)\|^2]}$$

Does NOT converge to zero, when $w_t \rightarrow w^*$!

To guarantee convergence: $\sum_t \eta_t = \infty, \sum_t \eta_t^2 < \infty$

Stochastic Gradient Descent

Pros:

- Each iteration is **independent** on n
- Achieves sublinear convergence rate (again **independent** on n)

some constants

$$\mathbf{E}[F(w_t) - F^*] \leq \frac{c}{\gamma+t} \qquad \text{if } \eta_t = \frac{d}{\gamma + t}$$

Assumptions:

- F(w) is smooth
- F(w) is strongly convex
- Second moment of stochastic gradient if bounded

Not optimal!

Can we get linear rate?

Modify stochastic gradient

Léon Bottou, Frank E Curtis, Jorge Nocedal
Optimization methods for large-scale machine learning

SVRG: Stochastic Variance Reduced Gradient

Rie Johnson, Tong Zhang

Accelerating Stochastic Gradient Descent using Predictive Variance Reduction, 2013

Lin Xiao, Tong Zhang

A Proximal Stochastic Gradient Method with Progressive Variance Reduction, 2014

Idea

- Modify stochastic gradient (decrease variance overtime)

```
1. Choose $w_0$
2. Set $\tilde{w} = w_0$
3. For $t = 0, 1, 2, \ldots, m$
4.     Choose $i_t \in \{1, 2, \ldots, n\}$
5.     $w_{t+1} = w_t - \eta(\underbrace{\nabla f_{i_t}(w_t) - \nabla f_{i_t}(\tilde{w}) + \nabla F(\tilde{w})}_{v_t})$
```

Unbiased stochastic gradient: $\mathbf{E}[v_t | w_t] = \nabla F(w_t)$

Second moment can bounded (by suboptimality):

$$\mathbf{E}[\|v_t\|^2] \leq 4L(F(w_t) - F(w^*) + F(\tilde{w}) - F(w^*))$$

Convergence of SVRG

- Choose $\eta < \frac{1}{2L}, m$

  such that $\alpha := \dfrac{1}{\mu\eta(1-2L\eta)m} + \dfrac{2L\eta}{1-2L\eta} < 1$

- Let $\tilde{w}^+ \in \{w_0, w_1, \ldots, w_{m-1}\}$

Then $\mathbf{E}[F(\tilde{w}^+) - F(w^*)] \leq \alpha\mathbf{E}[F(\tilde{w}) - F(w^*)]$

Note: For fixed η we will **not converge** to optimal solution.

Restarting

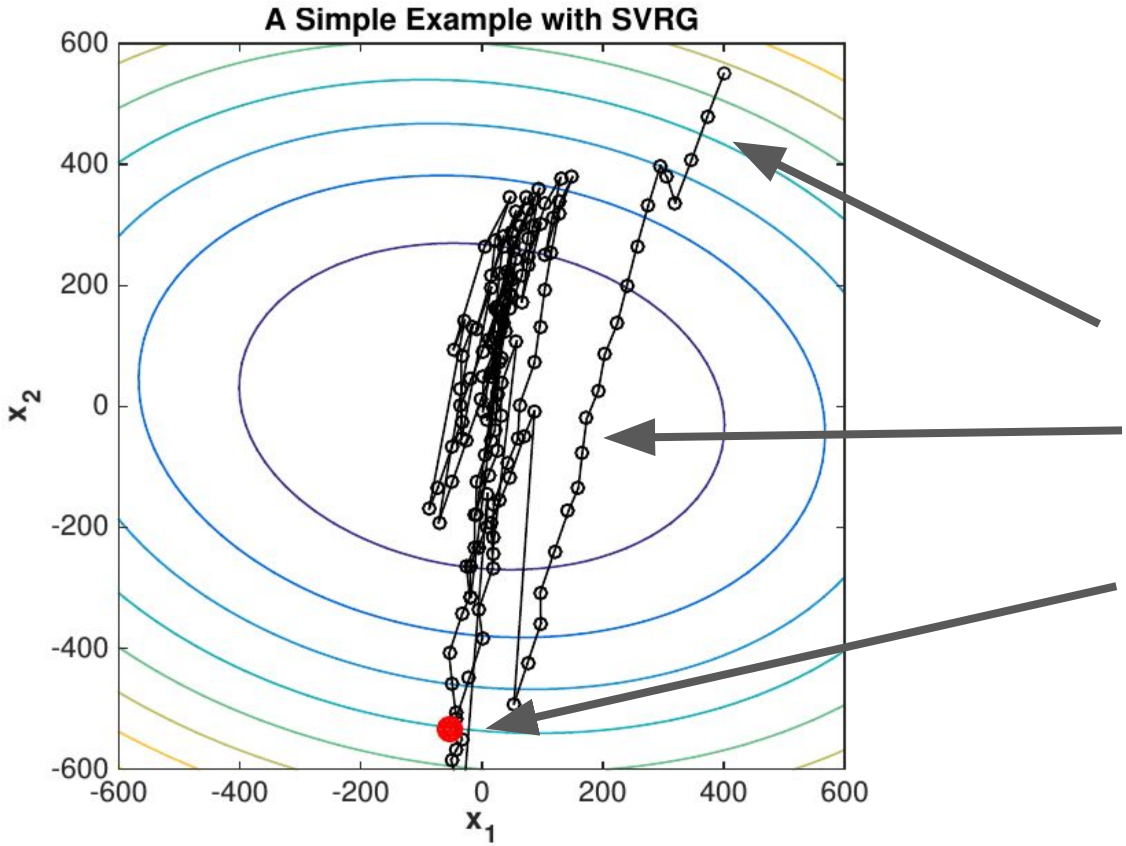
$$\tilde{w}^{(0)} \rightarrow \tilde{w}^{(1)} \rightarrow \tilde{w}^{(2)} \rightarrow \cdots \rightarrow \tilde{w}^{(s)}$$

$$\mathbf{E}[F(\tilde{w}^{(s)}) - F(w^*)] \leq \alpha^s(F(w^{(0)}) - F(w^*))$$

SVRG Algorithm

An issue:

- How to choose “m” in algorithm?
  (PS: theory too pessimistic)



“m” too small!

good “m”

“m” too large!

SAG/SAGA

- Mark Schmidt, Nicolas Le Roux, Francis Bach
  Minimizing Finite Sums with the Stochastic Average Gradient, 2013
- Aaron Defazio, Francis Bach, Simon Lacoste-Julien
  SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives, 2014

Idea

- Keep a table of “past” gradients
- In each iteration update one “gradient” in the table

$$y_{i,t} = \begin{cases} \nabla f_i(w_t), & \text{if } i_t = i \\ y_{i,t-1}, & \text{otherwise} \end{cases}$$

(SAG)

$$w_{t+1} = w_t - \eta_t \cdot \frac{1}{n} \cdot \sum_i y_{i,t}$$

(SAGA)

$$w_{t+1} = w_t - \eta \left(\nabla f_{i_t}(w_t) - y_{i_t,t-1} + \frac{1}{n} \sum_i y_{i,t-1} \right)$$

Pros:

- No need to restart
- **Linear** convergence rate!

Contra:

- Extra storage!

  Need to store “n” gradients.

Research Challenges

- SAG/SAGA - large extra storage!

  Can we eliminate it and keep linear convergence?

- SVRG - Performance sensitive on “m”

  Can we restart based on runtime criteria?

SARAH

Lam Nguyen, Jie Liu, Katya Scheinberg, Martin Takáč

SARAH: A Novel Method for Machine Learning Problems Using Stochastic Recursive Gradient

The New Stochastic Gradient

- We want

$$v_t \approx \nabla F(w_t)$$

- We also want to use only one function $f_i(w)$ to define the stochastic gradient

$$v_t = \nabla f_i(w_t) - \nabla f_i(w_{t-1}) + v_{t-1}$$

- A little bit similar to momentum

$$v_t = \nabla f_i(w_t) + 0.9 v_{t-1}$$

The Big Picture

- It do restarting (like SVRG)
- Is “similar” to SAG/SAGA, but **DOESN’T need extra storage**

```
1. Choose $w_0$, compute $v_0 = \nabla F(w_0)$
2. Set $w_1 = w_0 - \eta_0 v_0$
3. For $t = 1, 2, \ldots, m$
4.     Choose $i_t \in \{1, 2, \ldots, n\}$
5.     $v_t = \nabla f_{i_t}(w_t) - \nabla f_{i_t}(w_{t-1}) + v_{t-1}$
6.     $w_{t+1} = w_t - \eta_t v_t$
```

No extra storage is needed!

SARAH is Conditionally Biased

- Recall: $v_t = \nabla f_{i_t}(w_t) - \nabla f_{i_t}(w_{t-1}) + v_{t-1}$
- We have

$$\mathbf{E}[v_t|\mathcal{F}_t] = \nabla F(w_t) - \nabla F(w_{t-1}) + v_{t-1} \neq \nabla F(w_t)$$

Conditioned on $\{w_0, i_1, i_2, \ldots, i_{t-1}\}$

- However, we have $\mathbf{E}[v_t] = \mathbf{E}[\nabla F(w_t)]$

Theorem:

$F(w)$ is strongly convex

$$\mathbf{E}[\|v_t\|^2] \leq \begin{cases} (1 - (\frac{2}{\eta L} - 1)\mu^2\eta^2)^t \\ (1 - \frac{2\mu L \eta}{\mu + L})^t \end{cases} \cdot \mathbf{E}[\|\nabla F(w_0)\|^2]$$

$\forall i : f_i(w)$ is strongly convex

SARAH is converging (somewhere)!

SARAH Convergence

- Choose $\eta \leq \frac{2}{L}, m$

  such that $\alpha := \frac{1}{\mu\eta(m+1)} + \frac{\eta L}{2 - \eta L} < 1$

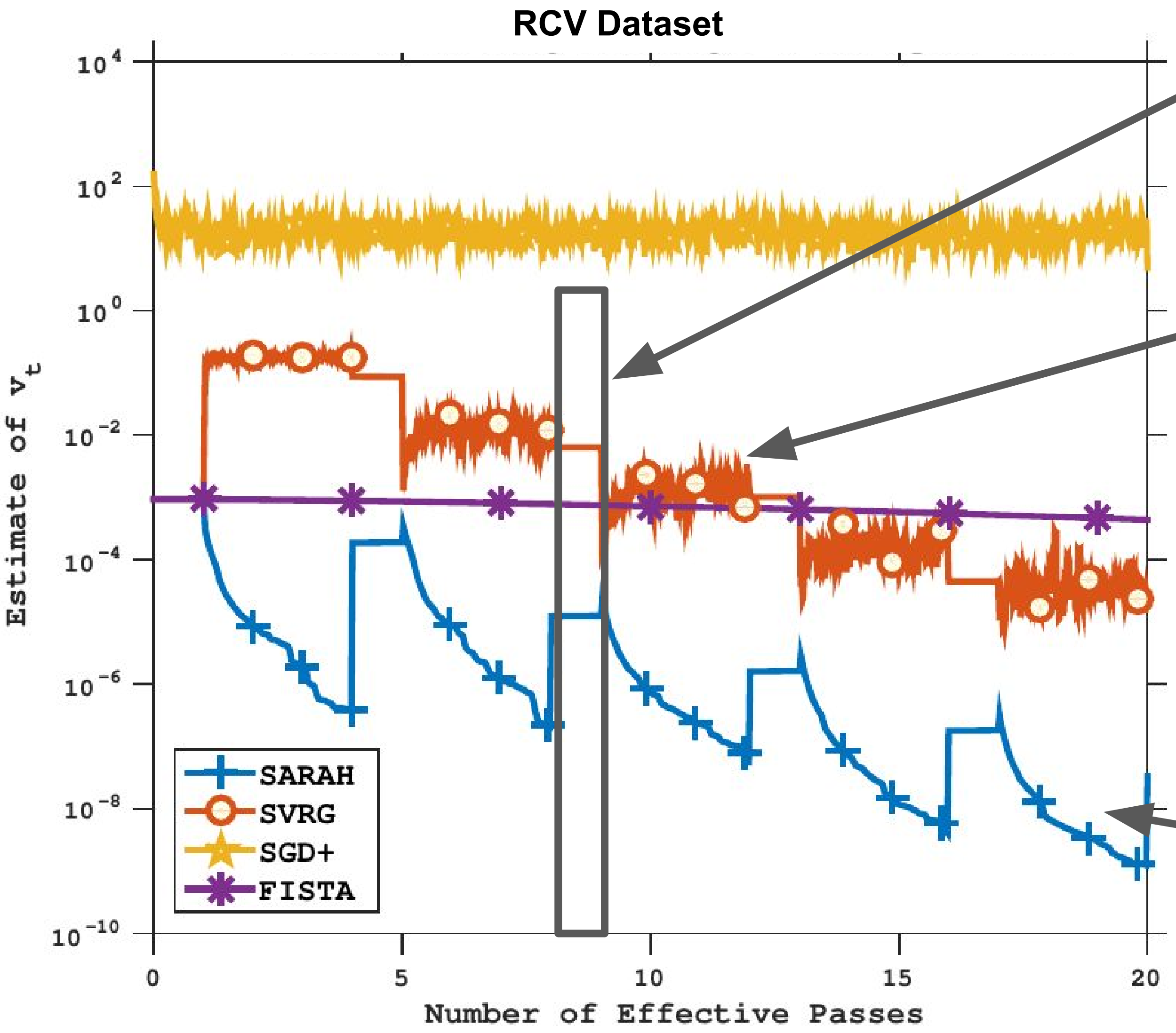
- Let $\tilde{w}^+ \in \{w_0, w_1, \ldots, w_{m-1}\}$

Then $\mathbf{E}[\|\nabla F(\tilde{w}^+)\|^2] \leq \alpha \|\nabla F(w_0)\|^2$

Ok, this is similar to SVRG (a little bit better),

but still .... **why it is cool**?

SARAH Demonstration

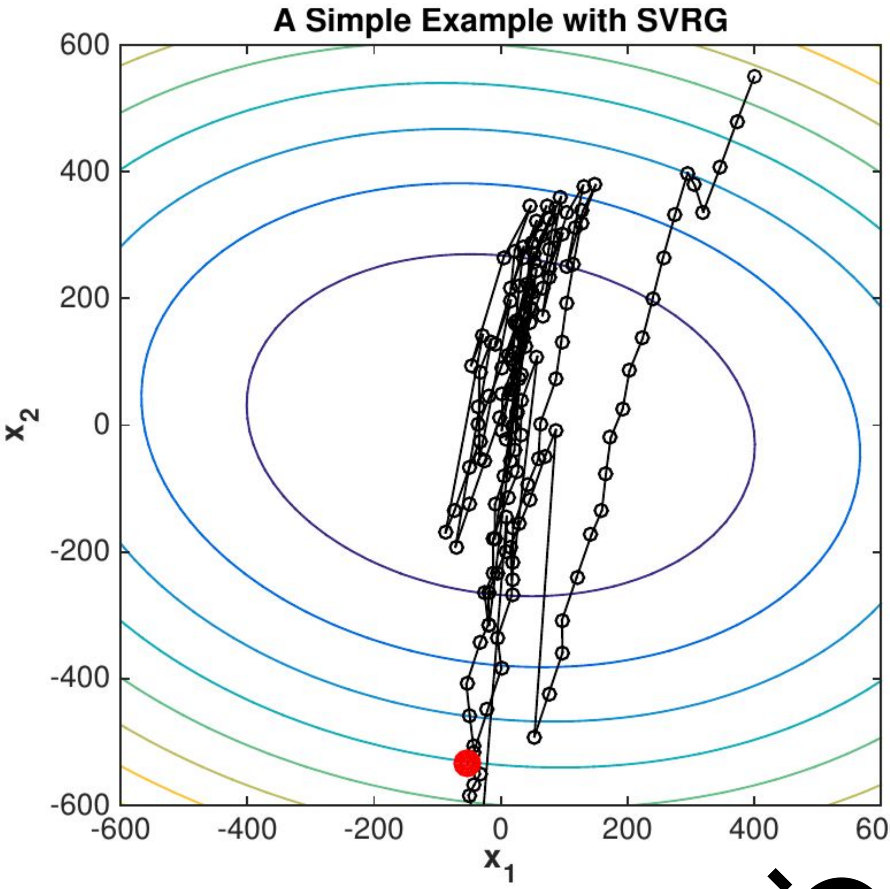


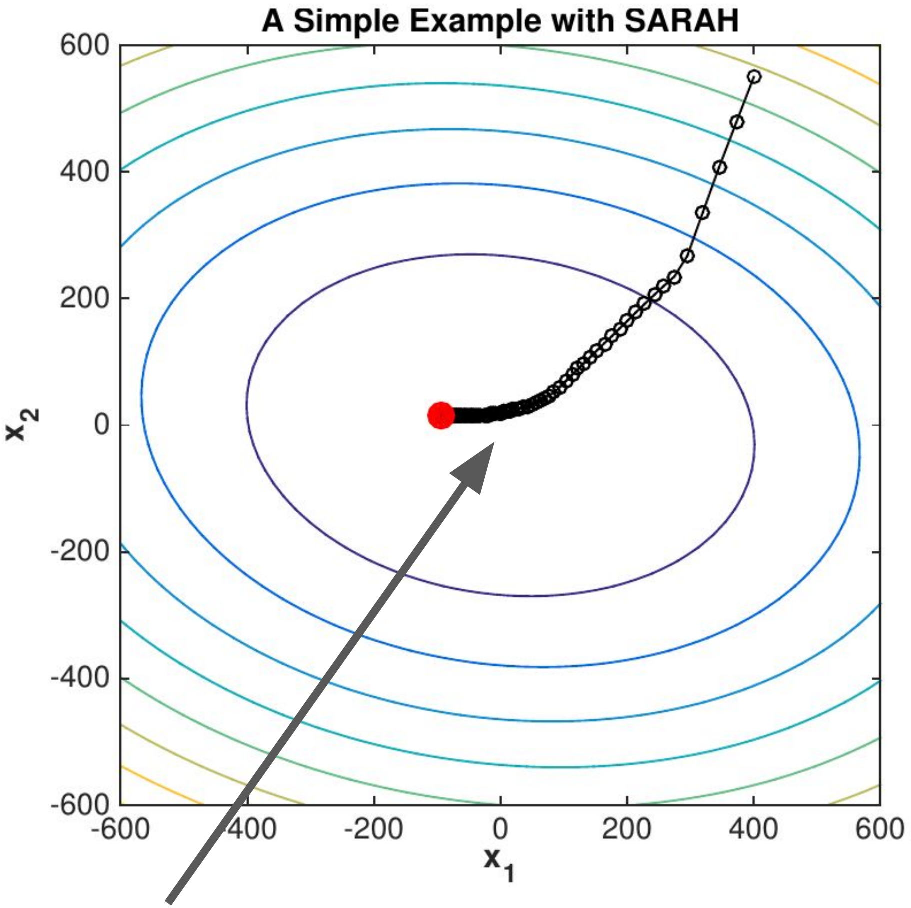
SVRG and SARAH need full gradient after restart

Variance of SVRG is decreased after each restart

Variance of SARAH goes to zero

SARAH Demonstration





Early termination

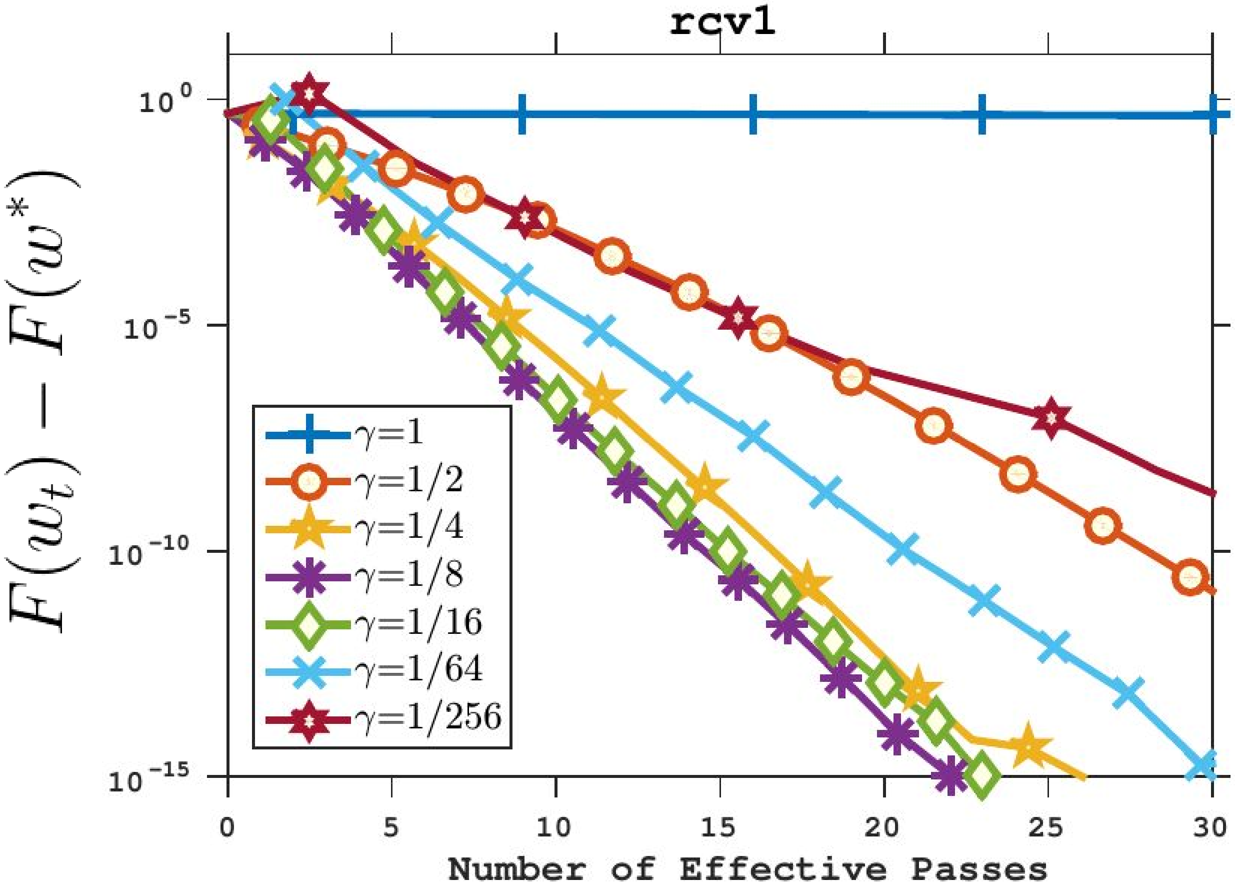
SARAH+
Practical Variant

SARAH+

Fact #1: Size of update is shrinking

It doesn't make sense to do many tiny steps!

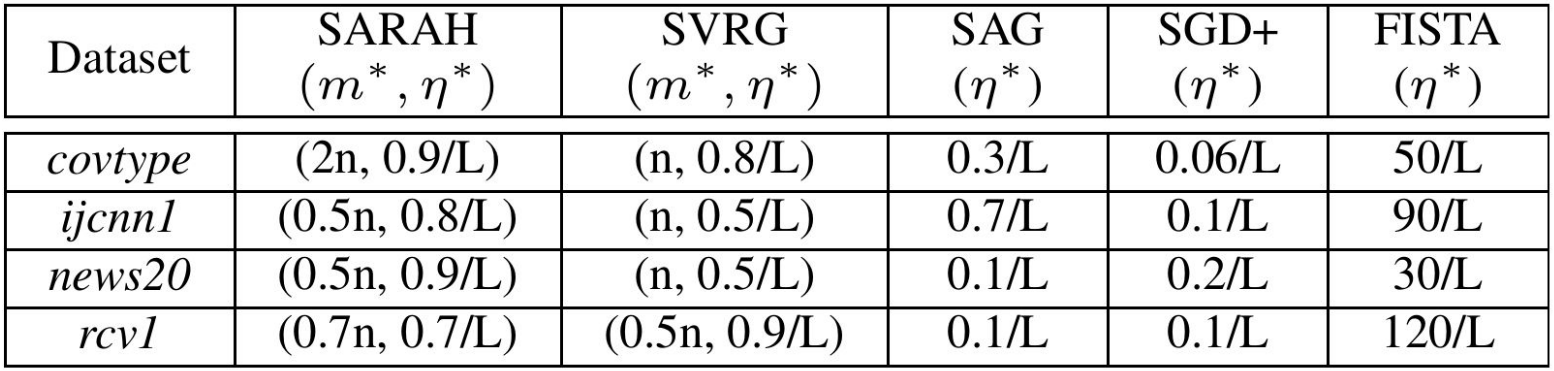
Heuristic: Restart algorithm when $\|v_t\|^2 \leq \gamma \|v_0\|^2$



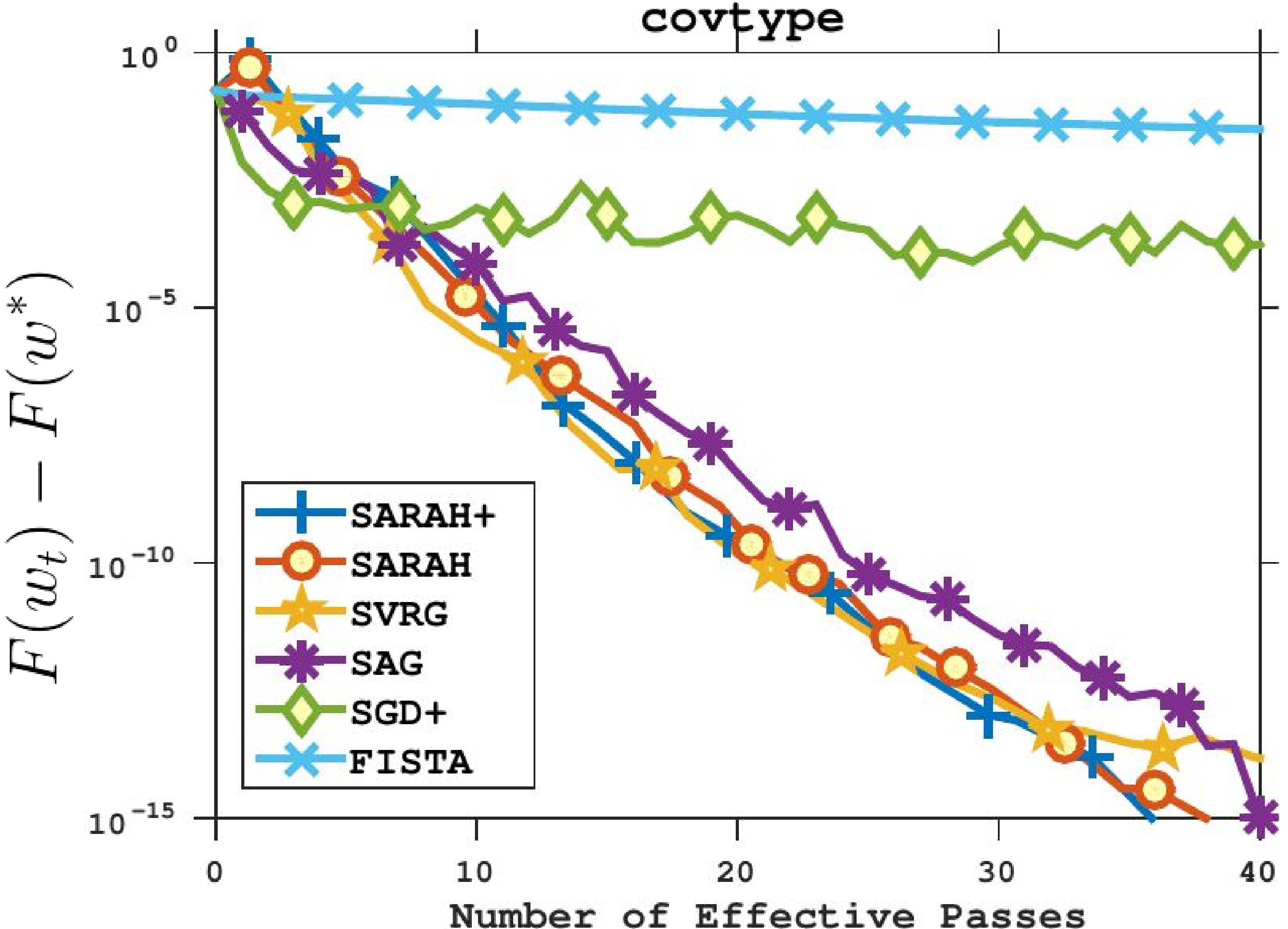
$$\gamma \approx 1/10$$

good performance across many datasets

Numerical Experiments



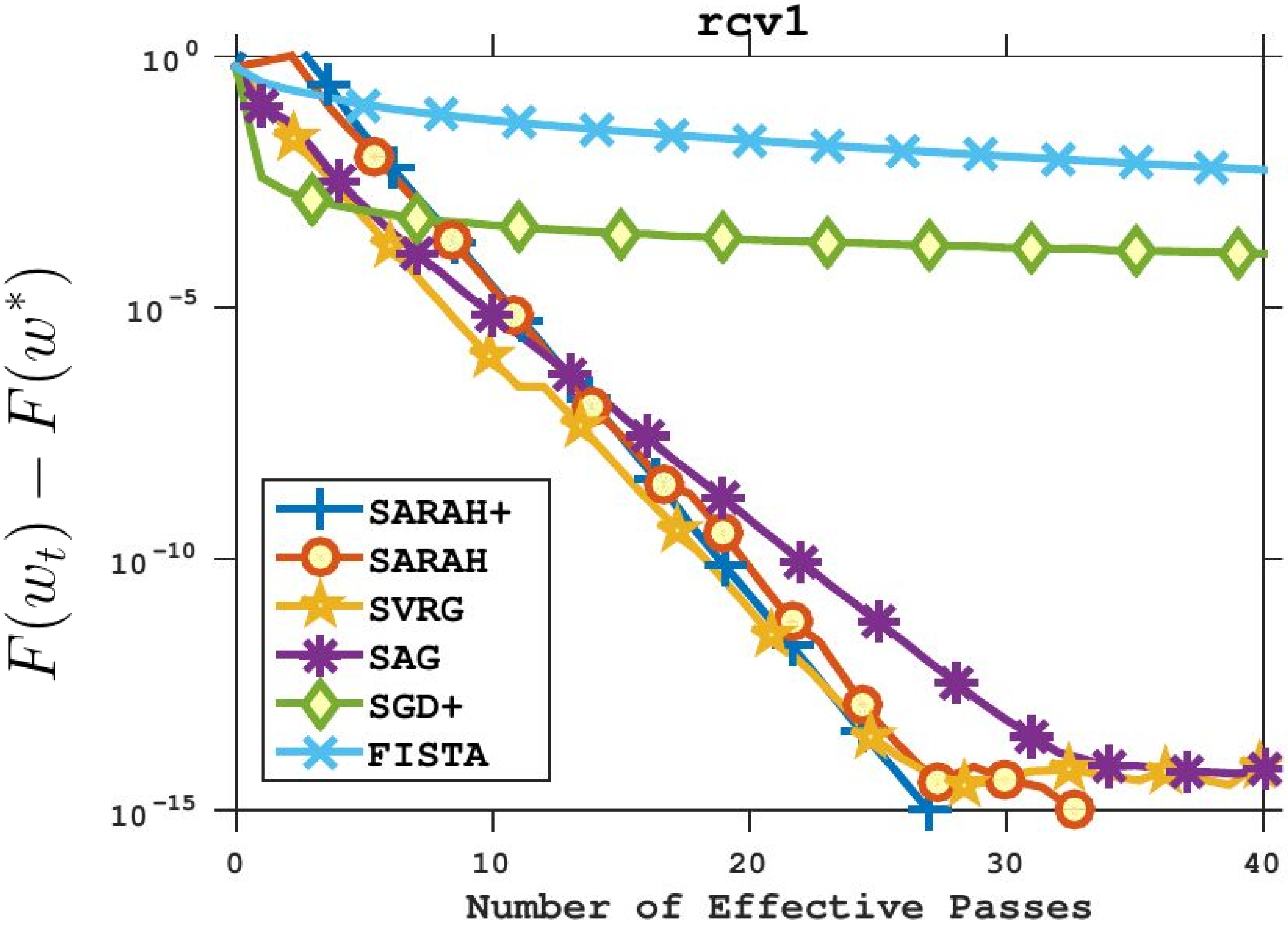
Dataset	SARAH (m^*, η^*)	SVRG (m^*, η^*)	SAG (η^*)	SGD+ (η^*)	FISTA (η^*)
covtype	(2n, 0.9/L)	(n, 0.8/L)	0.3/L	0.06/L	50/L
ijcnn1	(0.5n, 0.8/L)	(n, 0.5/L)	0.7/L	0.1/L	90/L
news20	(0.5n, 0.9/L)	(n, 0.5/L)	0.1/L	0.2/L	30/L
rcv1	(0.7n, 0.7/L)	(0.5n, 0.9/L)	0.1/L	0.1/L	120/L

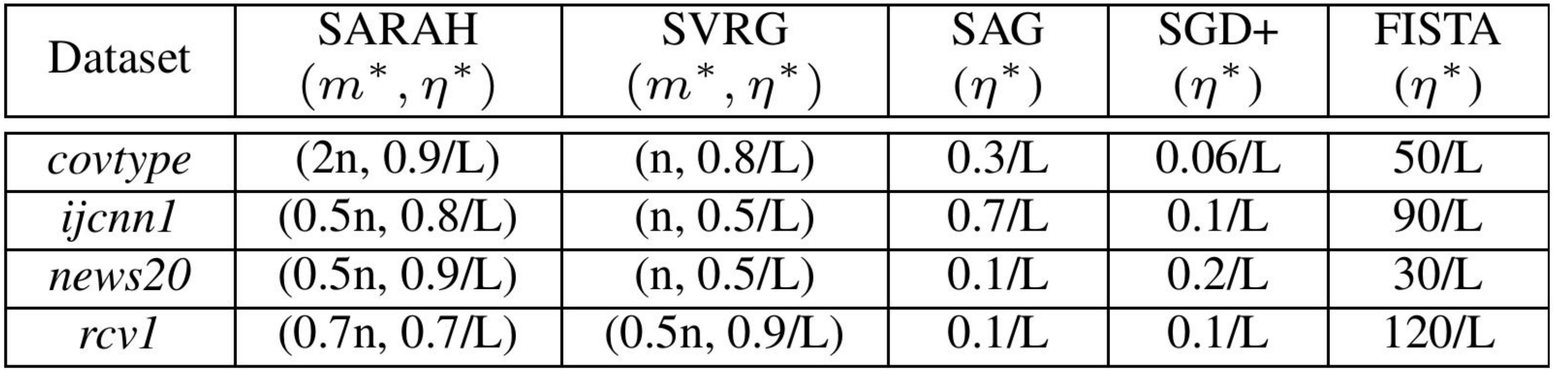


One has to tune parameters to get a good performance!

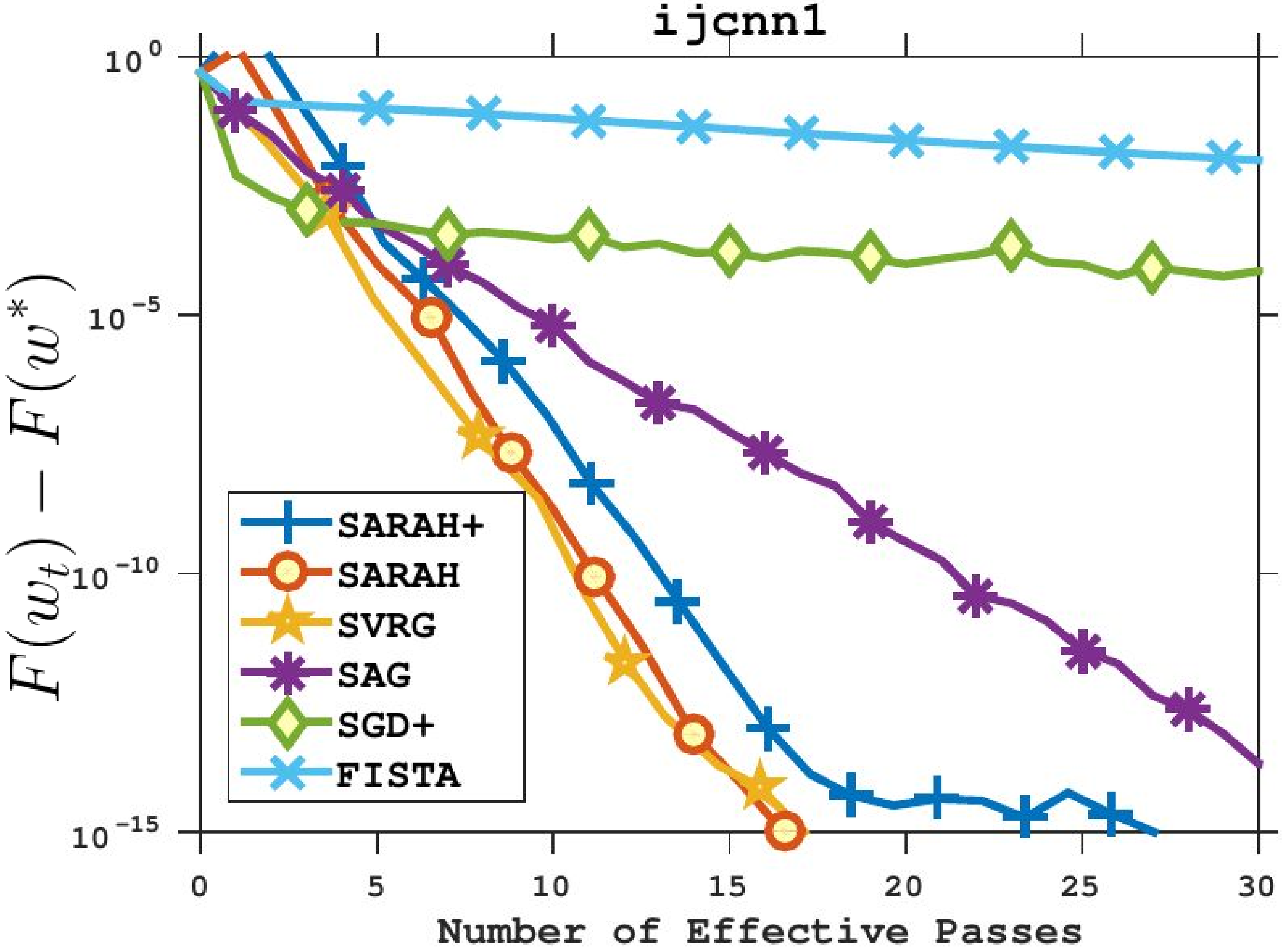
Not for SARAH+!

Dataset	SARAH (m^*, η^*)	SVRG (m^*, η^*)	SAG (η^*)	SGD+ (η^*)	FISTA (η^*)
covtype	(2n, 0.9/L)	(n, 0.8/L)	0.3/L	0.06/L	50/L
ijcnn1	(0.5n, 0.8/L)	(n, 0.5/L)	0.7/L	0.1/L	90/L
news20	(0.5n, 0.9/L)	(n, 0.5/L)	0.1/L	0.2/L	30/L
rcv1	(0.7n, 0.7/L)	(0.5n, 0.9/L)	0.1/L	0.1/L	120/L

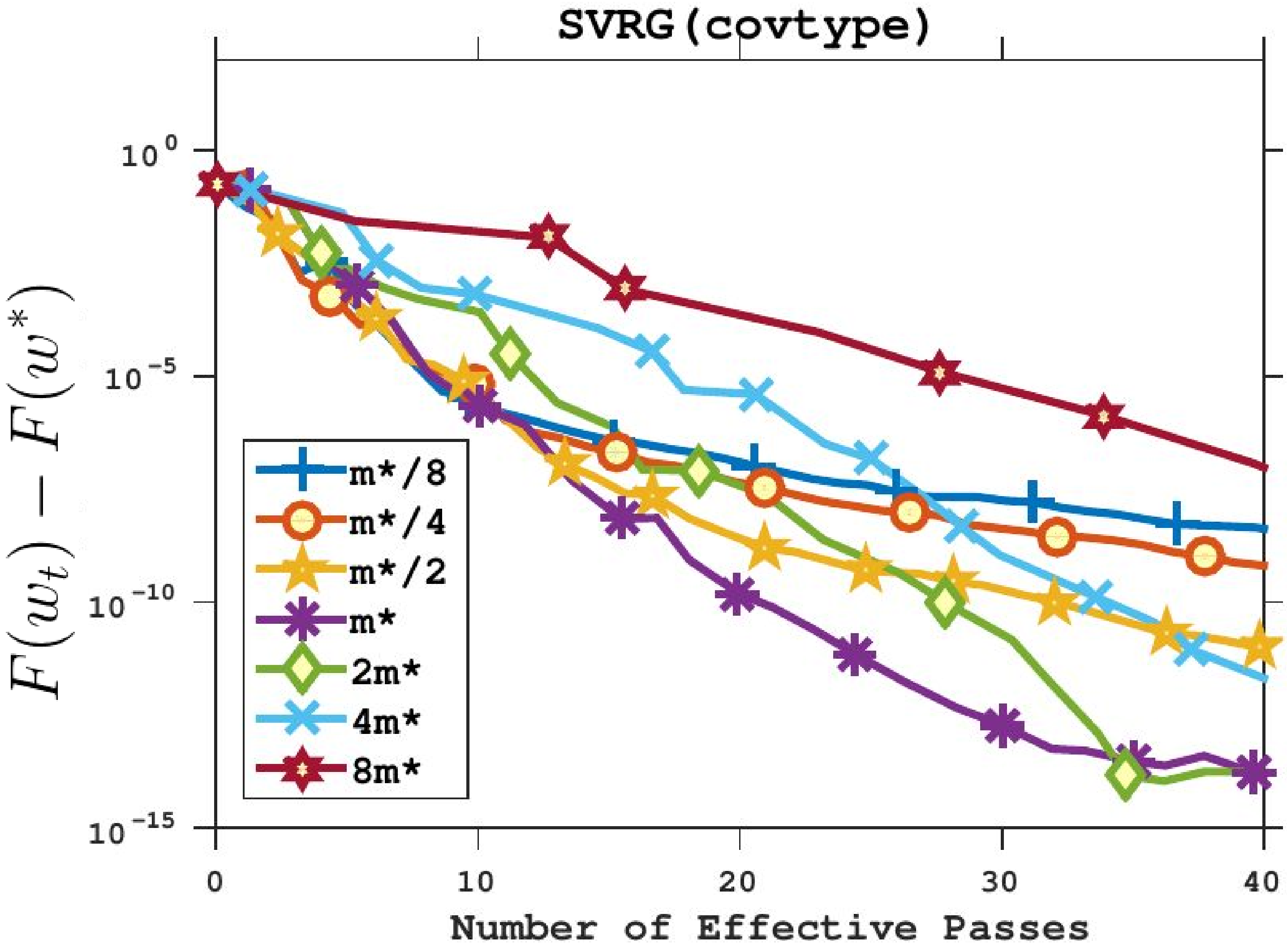




Dataset	SARAH (m^*, η^*)	SVRG (m^*, η^*)	SAG (η^*)	SGD+ (η^*)	FISTA (η^*)
covtype	(2n, 0.9/L)	(n, 0.8/L)	0.3/L	0.06/L	50/L
ijcnn1	(0.5n, 0.8/L)	(n, 0.5/L)	0.7/L	0.1/L	90/L
news20	(0.5n, 0.9/L)	(n, 0.5/L)	0.1/L	0.2/L	30/L
rcv1	(0.7n, 0.7/L)	(0.5n, 0.9/L)	0.1/L	0.1/L	120/L

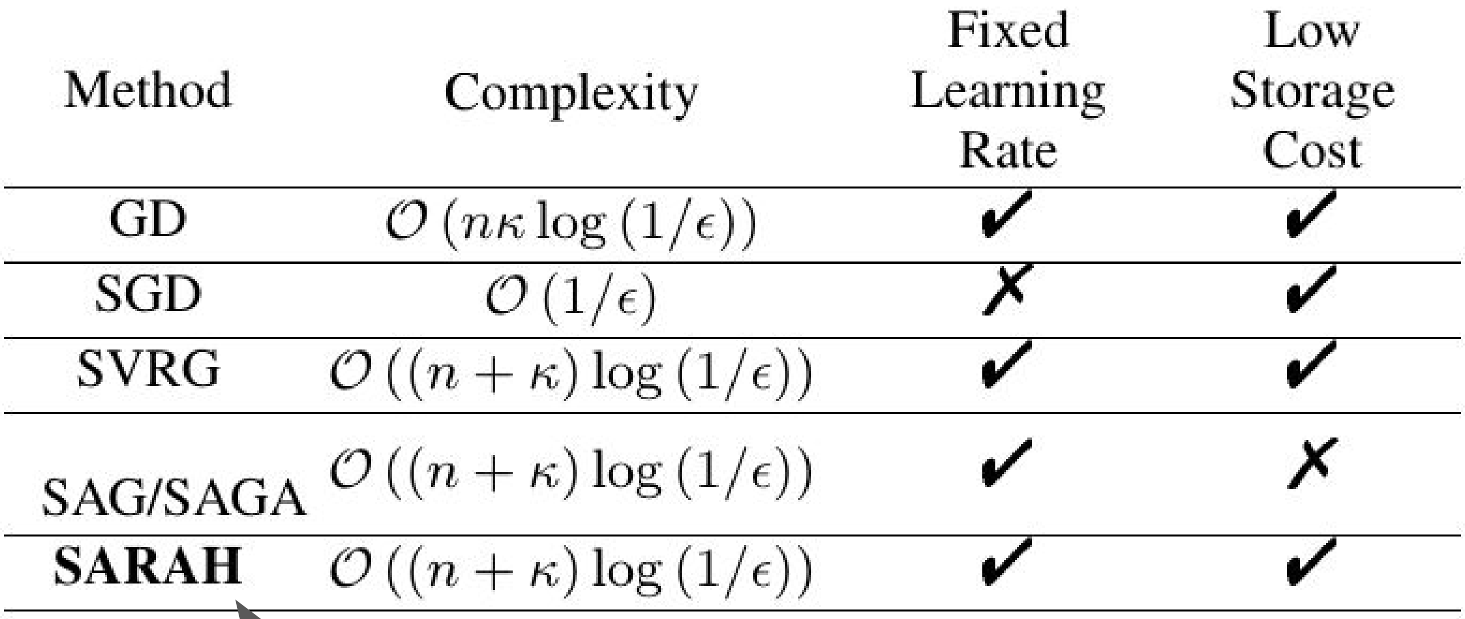


Sensitivity of SVRG on “m”



SARAH has similar behaviour!

Summary



Method	Complexity	Fixed Learning Rate	Low Storage Cost
GD	$\mathcal{O}\left(n\kappa \log\left(1/\epsilon\right)\right)$	✓	✓
SGD	$\mathcal{O}\left(1/\epsilon\right)$	✗	✓
SVRG	$\mathcal{O}\left(\left(n+\kappa\right)\log\left(1/\epsilon\right)\right)$	✓	✓
SAG/SAGA	$\mathcal{O}\left(\left(n+\kappa\right)\log\left(1/\epsilon\right)\right)$	✓	✗
SARAH	$\mathcal{O}\left(\left(n+\kappa\right)\log\left(1/\epsilon\right)\right)$	✓	✓

Practical variant available

More cases already covered:

- Smooth, convex objective function (sublinear rate)
- Smooth, non-convex objective function (sublinear rate)
- Smooth, gradient dominated function (linear rate)

Convex Case

SARAH for Convex Case

Without assuming strong convexity:

$$\mathbf{E}[\|\nabla F(w_t) - v_t\|^2] \leq \frac{\eta L}{2 - \eta L}\|v_0\|^2$$

Improvement after m iterations

$$\mathbf{E}[\|\nabla F(\tilde{w}^+)\|^2] \leq \frac{2}{\eta(m+1)}(F(w_0) - F^*) + \frac{\eta L}{2 - \eta L}\|\nabla F(w_0)\|^2$$

Non-Convex Case

SARAH for Non-Convex

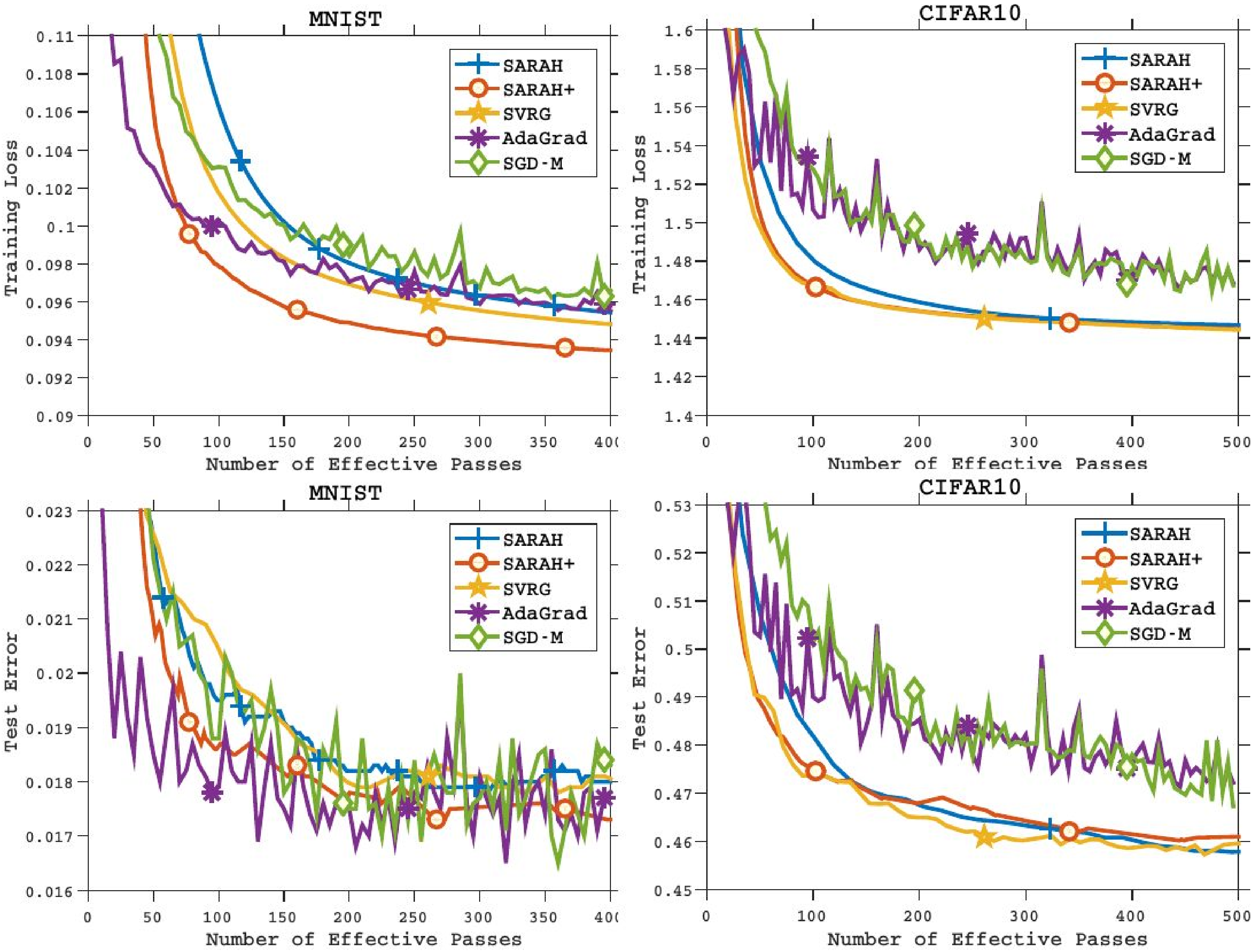
If

$$\eta \leq \frac{2}{L(\sqrt{1+4m}+1)}$$

then

$$\mathbf{E}[\|\nabla F(\tilde{w}^{+})\|^2] \leq \frac{2}{\eta(m+1)}(F(w_0) - F^*)$$

global minimum


MNIST
Training Loss
0.11
0.108
0.106
0.104
0.102
0.1
0.098
0.096
0.094
0.092
0.09
0
50
100
150
200
250
300
350
40(
Number of Effective Passes
SARAH
SARAH+
SVRG
AdaGrad
SGD-M
CIFAR10
Training Loss
1.6
1.58
1.56
1.54
1.52
1.5
1.48
1.46
1.44
1.42
1.4
0
100
200
300
400
500
Number of Effective Passes
SARAH
SARAH+
SVRG
AdaGrad
SGD-M
MNIST
Test Error
0.023
0.022
0.021
0.02
0.019
0.018
0.017
0.016
0
50
100
150
200
250
300
350
40(
Number of Effective Passes
SARAH
SARAH+
SVRG
AdaGrad
SGD-M
CIFAR10
Test Error
0.53
0.52
0.51
0.5
0.49
0.48
0.47
0.46
0.45
0
100
200
300
400
500
Number of Effective Passes
SARAH
SARAH+
SVRG
AdaGrad
SGD-M

Any Questions?