SAN: Stochastic Average Newton Algorithm for Minimizing Finite Sums

Jiabin Chen[1,6], Rui Yuan[2,5,6], Guillaume Garrigos[3], Robert M. Gower[4,5,6]

[1]Baidu Inc., [2]Meta AI, [3]Université de Paris, Sorbonne Université, CNRS, LPSM
[4]CCM, Flatiron Institute, [5]LTCI, Télécom Paris, [6]Institut Polytechnique de Paris.

International Conference on Artificial Intelligence and Statistics (AISTATS), 2022

Context

- Minimizing a finite sum with $n, d \gg 1$

$$w^* \in \arg\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i(w) \stackrel{\text{def}}{=} f(w) \tag{1}$$

Context

- Minimizing a finite sum with $n, d \gg 1$

$$w^* \in \arg\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i(w) \overset{\text{def}}{=} f(w) \tag{1}$$

- First-order methods: *SVRG* [Johnson and Zhang, 2013], *SAG* [Schmidt et al., 2017], etc. *Issue:* require parameter tuning, and/or the knowledge of the parameters of the problem

Context

- Minimizing a finite sum with $n, d \gg 1$

$$w^* \in \arg\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i(w) \stackrel{\text{def}}{=} f(w) \tag{1}$$

- First-order methods: *SVRG* [Johnson and Zhang, 2013], *SAG* [Schmidt et al., 2017], etc.
  Issue: require parameter tuning, and/or the knowledge of the parameters of the problem
- Second-order methods: *Stochastic Quasi-Newton* [Gower et al., 2016], *IQN* [Mokhtari et al., 2018], *SNM* [Kovalev et al., 2019]
  Issues: not incremental, or too expensive even for GLMs ($\mathcal{O}(d^2)$ per iteration)

Context

- Minimizing a finite sum with $n, d \gg 1$

$$w^* \in \arg\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i(w) \overset{\text{def}}{=} f(w) \tag{1}$$

- First-order methods: *SVRG* [Johnson and Zhang, 2013], *SAG* [Schmidt et al., 2017], etc.
  Issue: require parameter tuning, and/or the knowledge of the parameters of the problem
- Second-order methods: *Stochastic Quasi-Newton* [Gower et al., 2016], *IQN* [Mokhtari et al., 2018], *SNM* [Kovalev et al., 2019]
  Issues: not incremental, or too expensive even for GLMs ($\mathcal{O}(d^2)$ per iteration)

Develop a second order method for solving (1) that is *incremental*, *efficient*, scales well with the dimension d, and that requires no *knowledge from the problem*, neither *parameter tuning*.

SAN: Stochastic Average Newton (1/2)

1) Rewrite the optimality conditions $\nabla f(w) = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(w) = 0$ as follows

$$\frac{1}{n}\sum_{i=1}^{n} \alpha_i = 0, \tag{2}$$

$$\alpha_i = \nabla f_i(w), \quad \forall i \in \{1, \ldots, n\}. \tag{3}$$

SAN: Stochastic Average Newton (1/2)

1) Rewrite the optimality conditions $\nabla f(w) = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(w) = 0$ as follows

$$\frac{1}{n}\sum_{i=1}^{n} \alpha_i = 0, \tag{2}$$

$$\alpha_i = \nabla f_i(w), \quad \forall i \in \{1, \ldots, n\}. \tag{3}$$

Motivation:

- Each gradient lies on a separate equation.
- This motivates us to **sample** one equation per iteration, and **project** our current iterate on the **linearization** of this equation.

SAN: Stochastic Average Newton (2/2)

$(n+1)$ *equations:* (2) : $\frac{1}{n}\sum_{i=1}^{n}\alpha_i = 0,$ (3) : $\alpha_i = \nabla f_i(w), \quad \forall i \in \{1,\ldots,n\}$

2) Subsampled Newton Raphson [Yuan et al., 2021]

SAN: Stochastic Average Newton (2/2)

$(n+1)$ equations: (2) : $\frac{1}{n}\sum_{i=1}^{n} \alpha_i = 0,$ (3) : $\alpha_i = \nabla f_i(w), \quad \forall i \in \{1, \ldots, n\}$

2) Subsampled Newton Raphson [Yuan et al., 2021]

- With probability $\frac{1}{n+1}$, *sample* equation (2) and *project* onto its set of solutions:

$$\alpha_1^{k+1}, \ldots, \alpha_n^{k+1} = \underset{\alpha_1, \ldots, \alpha_n \in \mathcal{R}^d}{\arg\min} \sum_{i=1}^n \|\alpha_i - \alpha_i^k\|^2$$

$$\text{s.t. } \frac{1}{n}\sum_{i=1}^n \alpha_i = 0$$

SAN: Stochastic Average Newton (2/2)

$(n+1)$ *equations:* (2) : $\frac{1}{n}\sum_{i=1}^{n}\alpha_i = 0,$ (3) : $\alpha_i = \nabla f_i(w), \quad \forall i \in \{1,\ldots,n\}$

2) Subsampled Newton Raphson [Yuan et al., 2021]

- With probability $\frac{1}{n+1}$, *sample* equation (2) and *project* onto its set of solutions:

$$\alpha_1^{k+1},\ldots,\alpha_n^{k+1} = \underset{\alpha_1,\ldots,\alpha_n\in\mathcal{R}^d}{\arg\min} \sum_{i=1}^{n}\|\alpha_i-\alpha_i^k\|^2$$
$$\text{s.t. } \frac{1}{n}\sum_{i=1}^{n}\alpha_i = 0$$

- With probability $\frac{1}{n+1}$, *sample* the j-th equation of (3), and *project* onto the set of solutions of its *linearization* at w_k:

$$\alpha_j^{k+1}, w^{k+1} = \underset{\alpha_j, w\in\mathcal{R}^d}{\arg\min} \|\alpha_j-\alpha_j^k\|^2 + \|w-w^k\|^2_{\nabla^2 f_j(w^k)}$$
$$\text{s.t. } \nabla f_j(w^k) + \nabla^2 f_j(w^k)(w-w^k) = \alpha_j$$

What's the point by doing this ?

(see paper for technique details and additional properties)

It turns out that SAN

- is *incremental*, i.e. samples only one single data point per iteration

What's the point by doing this ?

(see paper for technique details and additional properties)

It turns out that SAN

- is *incremental*, i.e. samples only one single data point per iteration
- is *efficient* and scales well with the dimension d, i.e. costs $\mathcal{O}(d)$ per iteration for generalized linear models

What's the point by doing this ?

(see paper for technique details and additional properties)

It turns out that SAN

- is *incremental*, i.e. samples only one single data point per iteration
- is *efficient* and scales well with the dimension d, i.e. costs $\mathcal{O}(d)$ per iteration for generalized linear models
- requires no parameter tuning (*e.g. learning rate*), neither knowledge from the problem (*no smoothness constant*)

What's the point by doing this ?

(see paper for technique details and additional properties)

It turns out that SAN

- is *incremental*, i.e. samples only one single data point per iteration
- is *efficient* and scales well with the dimension d, i.e. costs $\mathcal{O}(d)$ per iteration for generalized linear models
- requires no parameter tuning (*e.g. learning rate*), neither knowledge from the problem (*no smoothness constant*)

☞ We provide a *global linear convergence theory* of SAN

What's the point by doing this ?

(see paper for technique details and additional properties)

It turns out that SAN

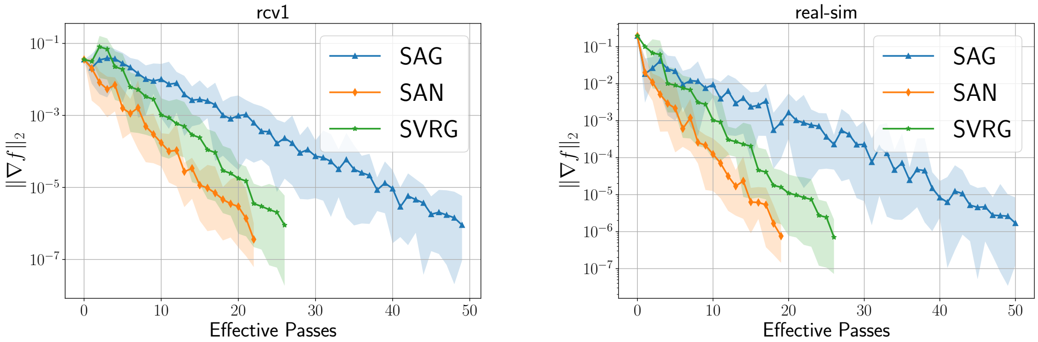
- is *incremental*, i.e. samples only one single data point per iteration
- is *efficient* and scales well with the dimension d, i.e. costs $\mathcal{O}(d)$ per iteration for generalized linear models
- requires no parameter tuning (*e.g. learning rate*), neither knowledge from the problem (*no smoothness constant*)

☞ We provide a *global linear convergence theory* of SAN

☞ Using our approach, we develop other new stochastic Newton methods, e.g., *SANA* and *SNRVM*

Experiments for SAN
(see paper for additional experiments)

Logistic regression for binary classification with the datasets from LibSVM



(a) rcv1 ($d : 47236$, $n : 20242$)   (b) real-sim ($d : 20958$, $n : 72309$)

Figure: Experiments for SAN applied for generalized linear model.

Details are in our paper:

SAN: Stochastic Average Newton Algorithm for Minimizing Finite Sums

Jiabin Chen, Rui Yuan, Guillaume Garrigos, Robert M. Gower

Thank you

Robert M. Gower, Donald Goldfarb, and Peter Richtárik. Stochastic block BFGS: Squeezing more curvature out of data. *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems 26*, pages 315–323. Curran Associates, Inc., 2013.

Dmitry Kovalev, Konstantin Mishchenko, and Peter Richtarik. Stochastic Newton and cubic Newton methods with simple local linear-quadratic rates. *arxiv:1912.01597*, 2019.

Aryan Mokhtari, Mark Eisen, and Alejandro Ribeiro. Iqn: An incremental quasi-newton method with local superlinear convergence rate. *SIAM Journal on Optimization*, 28(2):1670–1698, 2018. doi: 10.1137/17M1122943.

Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1):83–112, Mar 2017.

Rui Yuan, Alessandro Lazaric, and Robert M. Gower. Sketched newton-raphson, 2021.