Optimal Mini-Batch and Step Sizes for SAGA

Nidham Gazagnadou[1,a]

joint work with Robert M. Gower[1] & Joseph Salmon[2]

[1]LTCI, Télécom Paris, Institut Polytechnique de Paris, France
[2]IMAG, Univ Montpellier, CNRS, Montpellier, France

[a]This work was supported by grants from Région Ile-de-France

The Optimization Problem

- **Goal**

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

The Optimization Problem

- **Goal**

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

where

  - n **i.i.d.** observations: $(a_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ or $\mathbb{R}^d \times \{-1, 1\}$
  - $f_i : \mathbb{R}^d \to \mathbb{R}$ is $\boldsymbol{L_i}$**–smooth** $\forall i \in [n]$
  - f is $\boldsymbol{L}$**–smooth** and $\boldsymbol{\mu}$**–strongly convex**

The Optimization Problem

- **Goal**

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

  where

  - n **i.i.d.** observations: $(a_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ or $\mathbb{R}^d \times \{-1, 1\}$
  - $f_i : \mathbb{R}^d \to \mathbb{R}$ is $\boldsymbol{L_i}$**–smooth** $\forall i \in [n]$
  - f is $\boldsymbol{L}$**–smooth** and $\boldsymbol{\mu}$**–strongly convex**

- **Covered problems**
  - Ridge regression
  - Regularized logistic regression

Reformulation of the ERM

- **Sampling vector**

  Let $v \in \mathbb{R}^n$, with distribution $\mathcal{D}$ s.t. for all i in $[n] := \{1, \ldots, n\}$

$$\mathbb{E}_{\mathcal{D}}[v_i] = 1$$

Reformulation of the ERM

- **Sampling vector**

  Let $v \in \mathbb{R}^n$, with distribution $\mathcal{D}$ s.t. for all i in $[n] := \{1, \ldots, n\}$

$$\mathbb{E}_{\mathcal{D}}[v_i] = 1$$

- **ERM stochastic reformulation**

$$\text{find } w^* \in \arg\min_{w \in \mathbb{R}^d} = \mathbb{E}_{\mathcal{D}}\left[f_v(w) := \frac{1}{n}\sum_{i=1}^{n} v_i f_i(w)\right]$$

  leading to an unbiased gradient estimate

$$\mathbb{E}_{\mathcal{D}}[\nabla f_v(w)] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\mathcal{D}}[v_i]\, f_i(w) = \nabla f(w)$$

Reformulation of the ERM

- **Sampling vector**

  Let $v \in \mathbb{R}^n$, with distribution $\mathcal{D}$ s.t. for all i in $[n] := \{1, \ldots, n\}$

$$\mathbb{E}_{\mathcal{D}}[v_i] = 1$$

- **ERM stochastic reformulation**

$$\text{find } w^* \in \arg\min_{w \in \mathbb{R}^d} = \mathbb{E}_{\mathcal{D}}\left[f_v(w) := \frac{1}{n}\sum_{i=1}^{n} v_i f_i(w)\right]$$

  leading to an unbiased gradient estimate

$$\mathbb{E}_{\mathcal{D}}[\nabla f_v(w)] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\mathcal{D}}[v_i] f_i(w) = \nabla f(w)$$

- **Arbitrary sampling includes all mini-batching strategies**

  such as sampling $b \in [n]$ elements without replacement

$$\mathbb{P}\left[v = \frac{n}{b}\sum_{i \in B} e_i\right] = \frac{1}{\binom{n}{b}}, \quad \text{for all } B \subseteq [n], \quad |B| = b.$$

Focus on b Mini-Batch SAGA

The algorithm

- Sample a mini-batch $B \subset [n] := \{1, \dots, n\}$ s.t. $|B| = b$
- Build the **gradient estimate**

$$\boldsymbol{g(w^k)} = \frac{1}{b}\sum_{i \in B} \nabla f_i(w^k) - \frac{1}{b}\sum_{i \in B} J^k_{:i} + \frac{1}{n} J^k e$$

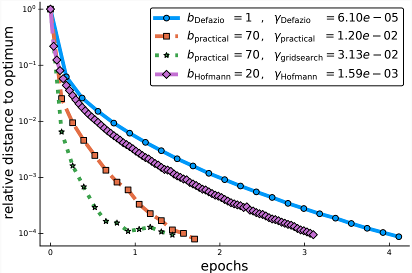
  where e is the all-ones vector and $J^k_{:i}$ the i–th column of $J^k \in R^{d \times n}$
- Take a step: $$w^{k+1} = w^k - \gamma \boldsymbol{g(w^k)}$$
- Update the **Jacobian estimate $\mathbf{J^k}$**

$$J^k_i = \nabla f_i(w^k), \quad \forall i \in B$$

Our contribution:

optimal mini-batch and step size

Example of SAGA run on real data (*slice* data set)



Key Constant: Expected Smoothness

Definition (Expected Smoothness constant)

If f is $\mathcal{L}$–smooth in expectation, then for every $w \in \mathbb{R}^d$

$$\mathbb{E}_{\mathcal{D}}\left[\|\nabla f_v(w) - \nabla f_v(w^*)\|_2^2\right] \leq 2\mathcal{L}(f(w) - f(w^*))$$

Key Constant: Expected Smoothness

Definition (Expected Smoothness constant)

If f is $\mathcal{L}$–smooth in expectation, then for every $w \in \mathbb{R}^d$

$$\mathbb{E}_{\mathcal{D}}\left[\|\nabla f_v(w) - \nabla f_v(w^*)\|_2^2\right] \leq 2\mathcal{L}(f(w) - f(w^*))$$

- **Total Complexity** of b mini-batch SAGA, for a given $\epsilon > 0$, is

  $$K_{\text{total}}(b) = \max\left\{\frac{4b(\mathcal{L}+\lambda)}{\mu}, n + \frac{n-b}{n-1}\frac{4(L_{\max}+\lambda)}{\mu}\right\}\log\left(\frac{1}{\epsilon}\right),$$

  where λ is the regularizer and $L_{\max} := \max_{i=1\dots n} L_i$
- For a step size

  $$\gamma = \frac{1}{4\max\left\{\mathcal{L}+\lambda, \frac{1}{b}\frac{n-b}{n-1}L_{\max} + \frac{\mu}{4}\frac{n}{b}\right\}}.$$

Key Constant: Expected Smoothness

Definition (Expected Smoothness constant)

If f is $\mathcal{L}$–smooth in expectation, then for every $w \in \mathbb{R}^d$

$$\mathbb{E}_{\mathcal{D}}\left[\|\nabla f_v(w) - \nabla f_v(w^*)\|_2^2\right] \leq 2\mathcal{L}(f(w) - f(w^*))$$

- **Total Complexity** of b mini-batch SAGA, for a given $\epsilon > 0$, is

  $$K_{\text{total}}(b) = \max\left\{\frac{4b(\mathcal{L}+\lambda)}{\mu}, n + \frac{n-b}{n-1}\frac{4(L_{\max}+\lambda)}{\mu}\right\}\log\left(\frac{1}{\epsilon}\right),$$

  where λ is the regularizer and $L_{\max} := \max_{i=1\ldots n} L_i$
- For a step size

  $$\gamma = \frac{1}{4\max\left\{\mathcal{L}+\lambda, \frac{1}{b}\frac{n-b}{n-1}L_{\max} + \frac{\mu}{4}\frac{n}{b}\right\}}.$$

Problem: Calculating $\mathcal{L}$ is most of the time intractable

Our Estimates of the Expected Smoothness

Theorem (Upper-bounds of $\mathcal{L}$)

When sampling b points without replacement we have

- *Simple bound*

$$\mathcal{L} \leq \mathcal{L}_{simple}(b) := \frac{n}{b}\frac{b-1}{n-1}\bar{L} + \frac{1}{b}\frac{n-b}{n-1}L_{\max}$$

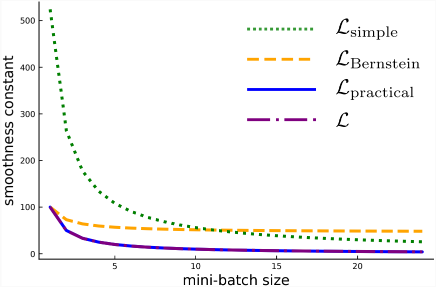
- *Bernstein bound*

$$\mathcal{L} \leq \mathcal{L}_{Bernstein}(b) := 2\frac{b-1}{b}\frac{n}{n-1}L + \frac{1}{b}\left(\frac{n-b}{n-1} + \frac{4}{3}\log d\right)L_{\max}$$

where $\bar{L} := \frac{1}{n}\sum_{i=1}^{n} L_i$ and $L_{\max} := \max_{i\in[n]} L_i$

Practical estimate

$$\mathcal{L}_{\text{practical}} := \frac{n}{b}\frac{b-1}{n-1}\boxed{L} + \frac{1}{b}\frac{n-b}{n-1}L_{\max}$$



Estimates of $\mathcal{L}$ artificial data ($n = d = 24$)

Optimal Mini-Batch from the Practical Estimate

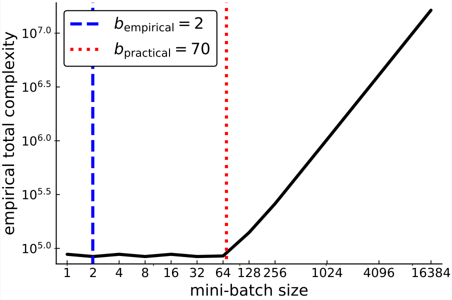
For a precision $\epsilon > 0$, the **total complexity** is

$$K_{\text{total}}(b) = \max\left\{\frac{4b(\mathcal{L}_{\text{practical}} + \lambda)}{\mu}, n + \frac{n-b}{n-1}\frac{4(L_{\max} + \lambda)}{\mu}\right\}\log\left(\frac{1}{\epsilon}\right)$$

Leading to the **optimal mini-batch** size

$$b^*_{\text{practical}} \in \arg\min_{b\in[n]} K_{\text{total}}(b)$$

$$\implies \boxed{b^*_{\text{practical}} = \left\lfloor 1 + \frac{\mu(n-1)}{4L} \right\rfloor}$$



Total complexity vs mini-batch size
(*slice* dataset, $\lambda = 10^{-1}$)

Summary

Take Home Message

- Use optimal mini-batch and step sizes available for SAGA!

What was done

- Build estimates of $\mathcal{L}$
- Give optimal settings (b, γ) for mini-batch SAGA

$$\implies \text{Faster convergence of } w^k \xrightarrow[k\to\infty]{} w^*$$

- Provide convincing numerical improvements on real datasets
- All the Julia code available at

https://github.com/gowerrobert/StochOpt.jl

References (1/2)

- F. Bach. "Sharp analysis of low-rank kernel matrix approximations". In:ArXiv e-prints (Aug. 2012). arXiv: 1208.2015 [cs.LG].
- C. C. Chang and C. J. Lin. "LIBSVM : A library for support vector machines". In: ACM Transactions on Intelligent Systems and Technology 2.3 (Apr. 2011), pp. 127.
- A. Defazio, F. Bach, and S. Lacoste-julien. "SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives". In: Advances in Neural Information Processing Systems 27. 2014, pp. 16461654.
- R. M. Gower, P. Richtrik, and F. Bach. "Stochastic Quasi-Gradient Methods: Variance Reduction via Jacobian Sketching". In: arXiv preprint arXiv:1805.02632 (2018).
- D. Gross and V. Nesme. "Note on sampling without replacing from a finite collection of matrices". In: arXiv preprint arXiv:1001.2738 (2010)
- W. Hoeffding. "Probability inequalities for sums of bounded random variables". In: Journal of the American statistical association 58.301 (1963), pp. 1330.

References (2/2)

- R. Johnson and T. Zhang. "Accelerating Stochastic Gradient Descent using Predictive Variance Reduction". In: Advances in Neural Information Processing Systems 26. Curran Associates, Inc., 2013, pp. 315323.
- H. Robbins and S. Monro. "A stochastic approximation method". In: Annals of Mathematical Statistics 22 (1951), pp. 400407.
- M. Schmidt, N. Le Roux, and F. Bach. "Minimizing finite sums with the stochastic average gradient". In: Mathematical Programming 162.1 (2017), pp. 83112.
- J. A. Tropp. "An Introduction to Matrix Concentration Inequalities". In: ArXiv e-prints (Jan. 2015). arXiv:1501.01571 [math.PR]
- J. A. Tropp. "Improved analysis of the subsampled randomized Hadamard transform". In: Advances in Adaptive Data Analysis 3.01n02 (2011), pp. 115126.
- J. A. Tropp. "User-Friendly Tail Bounds for Sums of Random Matrices". In: Foundations of Computational Mathematics 12.4 (2012), pp. 389434.