OPT2023: 15th Annual Workshop on Optimization for Machine Learning

Variance Reduced Model Based Methods: New rates and adaptive step sizes

Robert M. Gower RGOWER@FLATIRONINSTITUTE.ORG
Center for Computational Mathematics, Flatiron Institute, New York, USA

Frederik Kunstner KUNSTNER@CS.UBC.CA
University of British Columbia, Vancouver, BC, Canada

Mark Schmidt SCHMIDTM@CS.UBC.CA
University of British Columbia, Canada CIFAR AI Chair (Amii), Vancouver, BC, Canada

Abstract

Variance reduced gradients methods were introduced to control the variance of SGD (Stochastic Gradient Descent). Model-based methods are able to make use of a known lower bound on the loss, for instance, most loss functions are positive. We show how these two classes of methods can be seamlessly combined. As an example we present a Model-based Stochastic Average Gradient method MSAG, which results from using a truncated model together with the SAG method. At each iteration MSAG computes an adaptive learning rate based on a given known lower bound. When given access to the optimal objective as the lower bound, MSAG has several favorable convergence properties, including monotonic iterates, and convergence in the non-smooth, smooth and strongly convex setting. Our convergence theorems show that we can trade-off knowing the smoothness constant $L_{\max}$ for knowing the optimal objective to achieve the fast convergence of variance reduced gradient methods. Moreover our convergence proofs for MSAG are very simple, which is in contrast to the original convergence proofs of SAG.

Keywords: Variance reduced, finite sum minimization, Polyak step size, model-based method, adaptive learning rates.

1. Introduction

Consider the finite sum problem

$$x_* \in \operatorname*{argmin}_{x\in\mathbb{R}^d} \; f(x), \quad f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x), \tag{1}$$

where we assume that $f_i(x)$ is a convex differentiable function. We denote the optimal value of (1) by $f^* \in \mathbb{R}$. Let x_0 be a given initial point. We also assume throughout we have access to a lower bound $f(x) \geq \underline{f}$. Most loss functions are positive thus $\underline{f} = 0$.

Here we propose an adaptive step size for variance reduced gradient methods that leverages this lower bound $\underline{f}$. Our focus will be on the SAG (Stochastic Average Gradient) method [25], though we also show how our approach can be extended to other variance reduced methods such as SVRG in the appendix in Section B. This adaptive step size results from a model-based viewpoint of SAG, and thus we call it MSAG (Model-based SAG step size).

We present a comprehensive convergence theory for MSAG that assumes no access to the smoothness constant $L_{\max}$, and instead assumes access to the best possible lower bound $\underline{f} = f(x_*)$. Our

convergence theory includes both non-smooth, convex smooth and strongly convex functions. In the non-smooth setting we prove possibly the first $\mathcal{O}(1/\sqrt{t})$ convergence rate for an incremental method that has access to $f(x_*)$ for convex and continuously differentiable functions. In the smooth setting, we prove that MSAG has the same complexity results of SAG except without assuming access to $L_{\max}$ and with notably worse constant factors as compared to SAG. These convergence results show that, up to worse constant factors, we can exchange knowing $L_{\max}$ for knowing $f(x_*)$.

Another interesting aspect of our convergence theory is that the proof technique completely sidesteps the issue of SAG having a biased gradient. The original proof of convergence of SAG [25] is notoriously difficult because SAG uses a biased estimate of the gradient[1]. In contrast, our proofs are straight-forward, and nowhere do we need to bound an expectation of the quantity $\langle \bar{g}_k, x_k - x_* \rangle$ where $\bar{g}_k$ is this biased gradient estimate.

Since we will often not have access to $f(x_*)$ in practice, we also show how a lower bound on $f(x_*)$ can be easily estimated for convex functions by observing the iterates of SAG and accumulating some scalar quantities.

1.1. Background

This work touches on two fields of work; model-based methods and VR (variance reduced) methods.

VR methods are stochastic gradient type methods that use an estimate of the gradient whose variance reduces goes to zero [12]. Because the variance reduces to zero, in the smooth setting VR methods converge towards the solution of (1) at a faster rate of $\mathcal{O}(1/t)$ and $\mathcal{O}(\rho^t)$ for convex and strongly convex functions, respectively, where $\rho \in (0, 1)$. This is a faster rate than SGD (Stochastic Gradient Descent), which converges at a rate of $\mathcal{O}(1/\sqrt{t})$ and $\mathcal{O}(1/t)$ for convex and strongly convex functions, respectively [10]. The first VR methods with this faster rate of convergence was SVRG [16] and SAG [25]. SVRG relies on saved snapshots of the past full batch gradient to reduce the variance, and SAG relies on averages of stochastic gradients. These two different strategies encompass most VR methods, with Loopless SVRG [14] using snapshots, and SAGA [6], MISO [19] using averages. At the intersection of these two approaches are the SARAH [21] and SPIDER [9] method, which both make use of a snapshot (or resetting) and averaging.

An exception to these two strategies are the dual coordinate ascent based methods [22, 26] which instead use that the coordinate gradients have variance that reduces to zero.

As for work on VR methods with adaptive step sizes, the AI-SARAH method [27] uses local estimates of the smoothness constant instead of the global smoothness constant. Both [7] and [3] combine SVRG with the diagonal preconditioner of Adagrad [8] to achieve a $\mathcal{O}(1/t)$ rate for smooth and convex functions. In [15] the authors also achieve a $\mathcal{O}(1/t)$ rate by combining the stochastic Polyak step size [18] with a novel variance reduced scheme.

Model-based methods build a simple local model of the objective function, which is then used within a proximal point framework [2, 5]. Our approach is based on the recent MoMo (Momentum Model-based) method [23], which combines momentum with model-based truncation. In [23] the authors show how SGDM (SGD with momentum) can be interpreted as a model-based method. They then use this interpretation to combine SGDM with truncation. The result is the MoMo method, a practical new adaptive learning rate for SGD. Analogously, we show how SAG can be interpreted as a model-based method and use this interpretation to introduce truncation and an adaptive step

1. There is now some more recent work on the analysis of SAG that is more straightforward [20] using coercive operators.

size for SAG. Viewing SGDM as a model-based interpretation relies on approximating the expected true risk by a discrete measure over past sampled points. Since we consider here the finite-sum problem (1) we do not need this additional approximation.

1.2. Contributions.

Here we make the following contributions.

Variance Reduced as Model Based. VR methods build estimates of the full batch gradient. We show how these estimates can be interpreted as building a model of the full batch loss function. Using this model within the proximal point framework gives the underlying VR method. Our focus here is on the SAG method, but this interpretation extends to other methods. For instance we show in Section B how SVRG can be interpreted as a model-based method.

Truncation. Now that we see that VR methods are implicitly building a model of the loss function, we can also incorporate truncation into this model. Truncation is used when we have a known lower bound on the loss function, such as $f(x) \geq 0$. Since we know the loss function is lower bounded, our model of the loss function should satisfy this same lower bound. Imposing this lower bound on our model is what we call truncation. In Lemma 1 we give MSAG, which is an adaptive step size for SAG based on truncation. We also show in Section D that the MSAG can be derived as a step size that minimizes an upper bound on the distance to the solution.

Online lower bound estimate. We might not know a lower bound for some functions. Or sometimes, even though we know $f(x) \geq 0$, zero may be a very loose lower bound. The best lower bound is given by $f(x) \geq f(x_*)$, but excluding some particular problem instances, it is unlikely we would know $f(x_*)$. We show in Lemma 4 (appendix) that for convex loss functions, by observing the iterates of SAG with any step size, we can build an online estimate $\underline{f_t}$ such that $f(x_*) \geq \underline{f_t}$.

Convergence without knowing $L_{\max}$. In the ideal case where we know $f(x_*)$ and set the lower bound $\underline{f_t} = f(x_*)$ we give three convergence results, whose results are summarized in Table 1. We believe these are the first step sizes and convergence results of SAG where the smoothness constant $L_{\max}$ is not known. Our first convergence result in Theorem 1 shows that MSAG converges at a rate of $\mathcal{O}(1/\sqrt{t})$ for convex and continuously differentiable functions, and without assuming the gradients are $L_{\max}$–Lipschitz. We believe this is the first $\mathcal{O}(1/\sqrt{t})$ convergence result for an incremental method under only convexity and knowledge of $f(x_*)$. The only comparable result we are aware of is the $\mathcal{O}(1/\sqrt{t})$ convergence of the stochastic Polyak step size [10, 11] which holds for convex functions, but requires the stronger assumption of having access to $f_i(x_*)$ for $i = 1, \dots, n$.

In Theorem 2 we show that in the smooth setting, having access to $f(x_*)$ but not $L_{\max}$, the MSAG method converges at a rate of $\mathcal{O}(1/t)$ and $\mathcal{O}(\rho^t)$ for convex and μ–Polyak-Łojasiewicz functions, respectively. See Table (1) for details. As a consequence, MSAG can be used on problems where $L_{\max}$ is not known, such as conditional random fields [24]. One apparent weakness in our results in the smooth setting is that the constant factor in the rates is n and n^2 worse than the standard rates of SAG, in the convex and strongly convex case, respectively. For now, we do not know if these worse constants are a consequence of our analysis, or a consequence of not having access to $L_{\max}$.

2. VR and Model Based

First we introduce model-based methods with a focus on truncated models.

Method / Assumption	cvx	cvx & $L_{\max}$–smooth	μ–str. cvx & $L_{\max}$–smooth
MSAG	$nG/\sqrt{t}$	$4n^2/t$	$\left(1-\mu/8n^2L_{\max}\right)^t$
SAG	N/A	$32n/t$	$\left(1-\min\{\mu/16L_{\max}, 1/8n\}\right)^t$

Table 1: The convergence rates of our adaptive step size with SAG (MSAG in Algorithm 1) which has access to $f(x_*)$ compared to SAG which has access to $L_{\max}$. Here $G := \sup_{x:\|x-x_*\|\leq\|x_0-x_*\|} \max_{i=1,\ldots,n} \|\nabla f_i(x)\|$ is a local bound on the gradient norm, which is finite for continuously differentiable functions.

2.1. Model Based Methods

The Model-based methods [1, 5] are based on the proximal point method

$$x_{k+1} = \operatorname*{argmin}_{x\in\mathbb{R}^d} f(x) + \tfrac{1}{2\alpha_k}\|x-x_k\|^2. \tag{2}$$

Because each iteration (2) can be computationally challenging to solve, the model-based method replace the objective function with an approximate model $m_k(x)$ such that $m_k(x) \approx f(x)$ when x is close to x_k. The iterates are then updated according to

$$x_{k+1} = \operatorname*{argmin}_{x\in\mathbb{R}^d} m_k(x) + \tfrac{1}{2\alpha_k}\|x-x_k\|^2. \tag{3}$$

Gradient descent results from using the linear model

$$m_k(x) = f(x_k) + \langle \nabla f(x_k), x-x_k\rangle.$$

When $f(x)$ has a known lower bound $\underline{f} \in \mathbb{R}$, such as $f(x) \geq 0$, then the model is truncated so that it also respects this lower bound

$$m_k(x) = \left(f(x_k) + \langle \nabla f(x_k), x-x_k\rangle - \underline{f}\right)_+, \tag{4}$$

where $(a)_+ := \max\{a, 0\}$.

In the stochastic, or finite sum setting such as ours, the full function $f(x)$ is replaced by $f_i(x)$ where $i \in \{1,\ldots,n\}$ is sampled at each iteration [5]. When truncation (4) is used in the stochastic setting, the resulting method is also known as the SGDwith a stochastic Polyak stepsize [4, 13, 18].

2.2. SAG as a Model Based Method

At the iteration kth, the SAG method uses a stored table of past gradients $[g_1^k,\ldots,g_n^k] \in \mathbb{R}^{d\times n}$ to build an estimate $\bar{g}_k$ of the full batch gradient $\nabla f(x_k)$. The current iterate x_k is then updated according to

$$x_{k+1} = x_k - \tfrac{\eta_k}{n}\bar{g}_k, \quad \text{where} \quad \bar{g}_k := \frac{1}{n}\sum_{i=1}^n g_i, \tag{5}$$

where $\eta_k > 0$ is the step size. Before each step (5), an index $i_k \in \{1,\ldots,n\}$ of a data point is sampled, and the table of gradients is updated as follows

$$g_i^k = \begin{cases} \nabla f_i(x^k) & \text{If } i = i_k. \\ g_i^{k-1} & \text{if } i \neq i_k. \end{cases} \tag{6}$$

Here we will also need additional notation to keep track of which iterate was each gradient evaluated. That is, let $q_i^k \in \mathbb{R}^d$ be such that $g_i^k = \nabla f_i(q_i^k)$. These past iterates are updated analogously to the gradients as follows

$$q_i^k = \begin{cases} x^k & \text{If } i = i_k. \\ q_i^{k-1} & \text{if } i \neq i_k. \end{cases} \qquad \text{and} \qquad \mathbf{Q}_k := [q_1^k, \ldots, q_n^k] \in \mathbb{R}^{d \times n}. \tag{7}$$

We need these iterates to better describe our method, but they are not stored or updated in our forthcoming method. We refer to the collection $\mathbf{Q}_k$ of past iterates as the *memory*. The stored past gradients g_i are used to build an estimate of the full gradient. By keeping a table of past function values, we can also build an estimate of the loss function itself. Indeed, by approximating each $f_i(x)$ by its linearization around the last point q_i^k the ith data point was sampled we have that

$$f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x) \approx \frac{1}{n}\sum_{i=1}^n \left(f_i(q_i^k) + \langle \nabla f_i(q_i^k), x - q_i^k \rangle \right). \tag{8}$$

For convex functions the right-hand side is a lower bound. If we use (8) as a model for the function in (3), the resulting method is SAG (5). But we can do better if we have access to a lower bound.

Suppose then at iteration k we have access to a lower bound $\underline{f_k}$ of $f(x)$. For positive loss functions the default will be $\underline{f_k} = 0$. With a known lower bound we can improve the above estimate by truncating, resulting in the following estimate of $f(x)$ given by

$$m_k(x) = \left(\frac{1}{n}\sum_{i=1}^n \left(f_i(q_i^k) + \langle \nabla f_i(q_i^k), x - q_i^k \rangle - \underline{f_k} \right) \right)_+. \tag{9}$$

We can now use this estimate as our model in (3) and derive a closed form update.

Lemma 1 *Using the model (9), the closed form solution to (3) is given by the SAG method (5) with a stepsize of*

$$\eta_k = \left\{ \alpha_k, \frac{\left(\frac{1}{n}\sum_{i=1}^n (f_i(q_i^k) + \langle \nabla f_i(q_i^k), x_k - q_i^k \rangle - \underline{f_k}) \right)_+}{\|\bar{g}_k\|^2} \right\}, \tag{10}$$

where $\bar{g}_k$ is the SAG gradient estimate.

We call the SAG with this adaptive step size (10) the MSAG (Model-based step sizes for SAG) method. The full pseudo-code of MSAG is in Algorithm 1.

The proof follows by some re-arranging and applying Lemma 2. The cost per iteration of MSAG is essentially the same as that of SAG. We also show how SVRG can also be interpreted through this

model-based viewpoint in Section B in the appendix.

For general finite sum problems the iteration complexity of both SAG and MSAG is $\mathcal{O}(d)$, while the memory footprint is $\mathcal{O}(n \times d)$. The only additional $\mathcal{O}(d)$ computations in MSAG are the three inner products between gradients and iterates. The additional memory of MSAG is a single n dimensional vector $[f_1(q_1^k), \ldots, f_n(q_n^k)]$ and the two additional scalars (see $\bar{f}_k$ and γ_k in Algorithm 1).

Algorithm 1: MSAG: Model-based Stochastic Average Gradient.
Default settings: $\alpha_k = 1$, $\underline{f_k} = 0$.
Input: $x_0 \in \mathbb{R}^d$, $\alpha_k > 0$, $\underline{f_k}$
Init: $\bar{f}_0 = f(x_0), \bar{g}_0 = \nabla f(x_0), \gamma_0 = \langle \bar{g}_0, x_0 \rangle$, $q_i^0 = x_0$ for $i \in [n]$.
for $k = 0$ ***to*** $K-1$ **do**
  Sample $i_k \in \{1, \ldots, n\}$
  $\bar{f}_k = \bar{f}_k + \frac{1}{n} f_{i_k}(x_k) - \frac{1}{n} f_i(q_{i_k}^{k-1})$
  $\gamma_k = \gamma_{k-1} + \frac{1}{n} \langle \nabla f_{i_k}(x_k), x_k \rangle - \frac{1}{n} \left\langle \nabla f_{i_k}(q_{i_k}^{k-1}), q_{i_k}^{k-1} \right\rangle$
  $\bar{g}_k = \bar{g}_{k-1} + \frac{1}{n} \nabla f_{i_k}(x_k) - \frac{1}{n} \nabla f_i(q_{i_k}^{k-1})$
  $x_{k+1} = x_k - \min \left\{ \alpha_k, \frac{(\bar{f}_k + \langle \bar{g}_k, x_k \rangle - \gamma_k - \underline{f_k})_+}{\|\bar{g}_k\|^2} \right\} \bar{g}_k$
end
Output: x^K

3. Convergence Theorems

All of our convergence theorems start with the inequality that results from taking expectation over monotonicity inequality (28), and then using Jensen's over a special 2D function.

Proposition 1 *Let f_i be convex for $i \in \{1, \ldots, n\}$. Consider the iterates of SAG (5) with the stepsize (34) It follows that*

$$\mathbb{E}\left[\|x_{k+1} - x_*\|^2\right] \leq \mathbb{E}\left[\|x_k - x_*\|^2\right] - \frac{1}{n^2} \frac{\mathbb{E}[f(x_k) - f^*]^2}{\mathbb{E}\left[\|\bar{g}_k\|^2\right]} \tag{11}$$

Our first convergence proof only assumes that the loss functions are convex and have locally bounded gradients.

Theorem 1 *[Locally bounded gradients] Assume the gradients are bounded over the sub-level set of the starting point, that is*

$$G = \sup_{x : \|x - x_*\| \leq \|x_0 - x_*\|} \max_{i=1,\ldots,n} \|\nabla f_i(x)\| \leq \infty.$$

Note that this is the case for continuously differentiable functions. Let $\bar{x}_t := \sum_{k=0}^{t-1} x_k$. If $f_i(x)$ is convex for $i \in [n]$ then SAG with step size (34) converges according to

$$\mathbb{E}\left[f(\bar{x}_t) - f(x_*)\right] \leq \frac{nG}{\sqrt{t}} \|x_0 - x_*\|.$$

Next we consider the case where each f_i is convex and $L_{\max}$–smooth.

Theorem 2 *[Smooth functions] Let $\bar{x}_t := \sum_{k=0}^{t-1} x_k$. If f_i is convex and $L_{\max}$–smooth for $i = 1, \ldots, n$, then SAG with step size (34) converges according to*

$$\mathbb{E}\left[f(\bar{x}_t) - f^*\right] \leq \frac{2L_{\max} n(2n-1)}{t} \mathbb{E}\left[\|x_k - x_*\|^2\right] \tag{12}$$

If in addition $f(x)$ is μ–Polyak-Łojasiewicz then

$$\mathbb{E}\left[\|x_{k+1} - x_*\|^2\right] \leq \left(1 - \frac{\mu}{4L_{\max}} \frac{1}{n} \frac{1}{2n-1}\right) \mathbb{E}\left[\|x_k - x_*\|^2\right]. \tag{13}$$

References

[1] Hilal Asi and John C. Duchi. Stochastic (approximate) proximal point methods: convergence, optimality, and adaptivity. *SIAM J. Optim.*, 29(3):2257–2290, 2019. ISSN 1052-6234. doi: 10.1137/18M1230323.

[2] Hilal Asi and John C. Duchi. The importance of better models in stochastic optimization. *Proc. Natl. Acad. Sci. USA*, 116(46):22924–22930, 2019. ISSN 0027-8424. doi: 10.1073/pnas.1908018116.

[3] Bastien Batardière, Julien Chiquet, and Joon Kwon. Finite-sum optimization: Adaptivity to smoothness and loopless variance reduction, 2023.

[4] Leonard Berrada, Andrew Zisserman, and M. Pawan Kumar. Training neural networks for and by interpolation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 799–809. PMLR, 13–18 Jul 2020.

[5] Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM J. Optim.*, 29(1):207–239, 2019. ISSN 1052-6234. doi: 10.1137/18M1178244.

[6] Aaron Defazio, Francis Bach, and Simon Lacoste-julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems 27*, pages 1646–1654. 2014.

[7] Benjamin Dubois-Taine, Sharan Vaswani, Reza Babanezhad, Mark Schmidt, and Simon Lacoste-Julien. Svrg meets adagrad: Painless variance reduction. *CoRR*, abs/2102.09645, 2021.

[8] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61):2121–2159, 2011. URL http://jmlr.org/papers/v12/duchi11a.html.

[9] Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[10] Guillaume Garrigos and Robert M. Gower. Handbook of convergence theorems for (stochastic) gradient methods, 2023.

[11] Guillaume Garrigos, Robert M. Gower, and Fabian Schaipp. Function value learning: Adaptive learning rates based on the polyak stepsize and function splitting in erm, 2023.

[12] Robert M. Gower, Mark Schmidt, Francis R. Bach, and Peter Richtárik. Variance-reduced methods for machine learning. *Proc. IEEE*, 108(11):1968–1983, 2020.

[13] Robert M. Gower, Mathieu Blondel, Nidham Gazagnadou, and Fabian Pedregosa. Cutting some slack for sgd with adaptive polyak stepsizes, 2022.

[14] Thomas Hofmann, Aurélien Lucchi, and Brian McWilliams. Neighborhood watch: Stochastic gradient descent with neighbors. *CoRR*, abs/1506.03662, 2015.

[15] Xiaowen Jiang and Sebastian U. Stich. Adaptive sgd with polyak stepsize and line-search: Robust convergence and variance reduction, 2023.

[16] 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.

[17] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition, 2020.

[18] Nicolas Loizou, Sharan Vaswani, Issam Hadj Laradji, and Simon Lacoste-Julien. Stochastic Polyak step-size for SGD: An adaptive learning rate for fast convergence. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1306–1314. PMLR, 13–15 Apr 2021.

[19] Julien Mairal. Incremental majorization-minimization optimization with application to large-scale machine learning. *SIAM Journal on Optimization*, 25(2):829–855, 2015.

[20] Martin Morin and Pontus Giselsson. Cocoercivity, smoothness and bias in variance-reduced stochastic gradient methods. *Numerical Algorithms*, 91(2):749–772, apr 2022.

[21] Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *International Conference on Machine Learning*, pages 2613–2621. PMLR, 2017.

[22] Zheng Qu, Peter Richtárik, and Tong Zhang. Quartz: Randomized dual coordinate ascent with arbitrary sampling. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 865–873, Cambridge, MA, USA, 2015. MIT Press.

[23] Fabian Schaipp, Ruben Ohana, Michael Eickenberg, Aaron Defazio, and Robert M. Gower. Momo: Momentum models for adaptive learning rates, 2023.

[24] Mark Schmidt, Reza Babanezhad, Mohamed Osama Ahmed, Aaron Defazio, Ann Clifton, and Anoop Sarkar. Non-uniform stochastic average gradient method for training conditional random fields. In *AISTATS*, volume 38 of *JMLR Workshop and Conference Proceedings*. JMLR.org, 2015.

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

[26] Shai Shalev-Shwartz. Sdca without duality, regularization, and individual convexity. In *International Conference on Machine Learning*, pages 747–754. PMLR, 2016.

[27] Zheng Shi, Abdurakhmon Sadiev, Nicolas Loizou, Peter Richtárik, and Martin Takáč. Ai-sarah: Adaptive and implicit stochastic recursive gradient methods, 2022.

[28] Xiaoyu Wang, Mikael Johansson, and Tong Zhang. Generalized polyak step size for first order optimization with momentum, 2023.

Appendix A. Auxiliary Lemmas

The following Lemma was taken from [23].

Lemma 2 ([23] Lemma B.1) *Let $y_0, a \in \mathbb{R}^p$ with $a \neq 0$ and $c \in \mathbb{R}$. Let $\beta > 0$. The solution to*

$$y^+ = \arg\min_y \Big(c + \langle a, y - y_0\rangle\Big)_+ + \frac{1}{2\beta}\|y - y_0\|^2 \tag{14}$$

is given by

$$y^+ = y_0 - \min\big\{\beta, \frac{(c)_+}{\|a\|^2}\big\}a.$$

Appendix B. Model Based SVRG

We can develop an analogous model-based step size for every variance reduced method. So long as the resulting gradient estimate is such that

$$\bar{g}_k = \sum_{i=1}^n \beta_i \nabla f_i(q_i^k)$$

where $\sum_{i=1}^n \beta_i = 1$ and $\beta_i \in \mathbb{N}$. To give an example, consider the SVRG gradient estimate given by

$$\bar{g}_k = \nabla f(\tilde{x}) + \nabla f_i(x_k) - \nabla f_i(\tilde{x}) = \frac{1}{n}\sum_{j=1} \nabla f_j(\tilde{x}) + \nabla f_i(x_k) - \nabla f_i(\tilde{x}). \tag{15}$$

Note here the corresponding β coefficients sum to one since

$$\frac{1}{n}\sum_{j=1} 1 + 1 - 1 = 1.$$

Mimicking the estimate of the gradient in (15), we can build a model of the loss as follows

$$\begin{aligned}
\frac{1}{n}\sum_{j=1}^n f_j(x) &= \frac{1}{n}\sum_{j=1}^n f_j(x) + f_i(x) - f_i(x) \\
&\approx \frac{1}{n}\sum_{j=1}^n f_j(\tilde{x}) + f_i(x_k) - f_i(\tilde{x}) \\
&\quad + \langle \nabla f(\tilde{x}), x - \tilde{x}\rangle - \langle \nabla f_i(\tilde{x}), x - \tilde{x}\rangle + \langle \nabla f(x_k), x - x_k\rangle,
\end{aligned}$$

where in the approximation step we linearized the functions around either $\tilde{x}$ or x_k in such a way to mimick the structure of SVRG. When $f(x) \geq 0$ this suggests the following model

$$m_k(x) \;=\; (f(\tilde{x}) + f_i(x_k) - f_i(\tilde{x}) + \langle \nabla f(\tilde{x}) - \nabla f_i(\tilde{x}), x - \tilde{x}\rangle + \langle \nabla f_i(x_k), x - x_k\rangle)_+ \tag{16}$$

Using this model in a proximal point method gives the following update.

Lemma 3 *Let $\bar{g}_k$ be the SVRG gradient estimate given in (15). Using the model (16), the closed form solution to*

$$\operatorname*{argmin}_{x\in\mathbb{R}^d} m_k(x) + \frac{1}{2\alpha_k}\|x - x_k\|^2 \tag{17}$$

is given by

$$x_{k+1} = x_k - \min\left\{\alpha_k, \tau_k\right\}\bar{g}_k, \tag{18}$$

where

$$\tau_k := \frac{\left(f(\tilde{x}) + f_i(x_k) - f_i(\tilde{x}) + \langle \nabla f(\tilde{x}) - \nabla f_i(\tilde{x}), x_k - \tilde{x}\rangle\right)_+}{\|\bar{g}_k\|^2}. \tag{19}$$

Proof Let $v = \nabla f(\tilde{x}) - \nabla f_i(\tilde{x})$. Re-arranging (16) gives

$$\begin{aligned} m_k(x) &= \left(f(\tilde{x}) + f_i(x_k) - f_i(\tilde{x}) + \langle v, x - \tilde{x}\rangle + \langle \nabla f_i(x_k), x - x_k\rangle\right)_+ \\ &= \left(f(\tilde{x}) + f_i(x_k) - f_i(\tilde{x}) + \langle v, x_k - \tilde{x}\rangle + \langle v + \nabla f_i(x_k), x - x_k\rangle\right)_+ \\ &= \left(c + \langle v + \nabla f_i(x_k), x - x_k\rangle\right)_+, \end{aligned}$$

where $c = f(\tilde{x}) + f_i(x_k) - f_i(\tilde{x}) + \langle v, x_k - \tilde{x}\rangle$. We can now apply Lemma 2 with $a = v + \nabla f(x_k)$ and $y_0 = x_k$ which gives

$$x_{k+1} = x_k - \min\left\{\alpha_k, \frac{c_+}{\|v + \nabla f(x_k)\|^2}\right\}(v + \nabla f(x_k)).$$

Substituting out c and $v = \nabla f(\tilde{x}) - \nabla f_i(\tilde{x})$ gives the result. ■

Appendix C. Estimating f^* on the fly

As we will soon show, if $\underline{f_k}$ is the tightest possible lower bound, that is $\underline{f_k} = f(x_*) =: f^*$, then MSAG has several favourable convergence properties. But before showing these results, first we how to build an online lower bound estimate $\underline{f_k} \leq f^*$ by simply observing the iterates of SAG.

Lemma 4 *Let $f_i(x)$ be convex in x for every sample i. Furthermore let $x_* \in \operatorname*{argmin}_{x\in\mathbb{R}^d} f(x)$. Consider the x_k are the iterates of SAG (5). Let*

$$h_k := \frac{1}{n}\sum_{i=1}^{n}(f_i(q_i^k) + \left\langle \nabla f_i(q_i^k), x_k - q_i^k\right\rangle.$$

It follows that

$$f^* \geq \underline{f_{k+1}} := \frac{2\sum_{j=0}^{k}\eta_j h_j - \|x_0 - x_*\|^2 - \sum_{j=0}^{k}\eta_j^2\|\bar{g}_j\|^2}{2\sum_{j=0}^{k}\eta_j}. \tag{20}$$

Furthermore we have the recurrence

$$\underline{f_{k+1}} = \underline{f_k} + \frac{\eta_k\left(h_k - \frac{1}{2}\eta_k\|\bar{g}_k\|^2\right)}{\sum_{j=0}^{k}\eta_j}. \tag{21}$$

Note that the estimate in (20) depends on the initial distance to the solution $\|x_0 - x_*\|^2$ which we will not know. Fortunately, because $\|x_0 - x_*\|^2$ is divided by the sum of all step sizes $\sum_{j=0}^{k} \eta_j$, this term has a decreasing effect on our estimate $\underline{f_{k+1}}$ as k grows. Furthermore $\|x_0 - x_*\|^2$ can be absorbed into initial estimate $\underline{f_0}$. Indeed, in the recurrence (21) the $\|x_0 - x_*\|^2$ term no longer appears. Since we will always initialize $\underline{f_0} = 0$, we do not need to know $\|x_0 - x_*\|^2$ to use this lower bound estimate.

Proof Consider the update (5) , and switching the index $k \to j$, which is

$$x_{j+1} = x_j - \eta_j \bar{g}_j,$$

where η_j is the step size. Subtracting x_* from both sides, taking norms and expanding the squares we have that

$$\|x_{j+1} - x_*\|^2 = \|x_j - x_*\|^2 - 2\eta_j \langle \bar{g}_j, x_j - x_* \rangle + \eta_j^2 \|\bar{g}_j\|^2 . \tag{22}$$

For shorthand let $\nabla f_i := \nabla f_i(q_i^k)$ and $f_i = f_i(q_i^k)$. Now using that

$$\begin{aligned}
\langle \bar{g}_j, x_j - x_* \rangle &= \sum_{i=1}^{n} \frac{1}{n} \left\langle \nabla f_i(q_i^k), x_j - x_* \right\rangle \\
&= \sum_{i=1}^{n} \frac{1}{n} \left(\left\langle \nabla f_i(q_i^j), x_j - q_i^j \right\rangle + \left\langle \nabla f_i(q_i^j), q_i^j - x_* \right\rangle \right) \\
&\geq \sum_{i=1}^{n} \frac{1}{n} \left(\left\langle \nabla f_i(q_i^j), x_k - q_i^j \right\rangle + f_i(q_i^j) - f_i(x_*) \right) \qquad \text{(by convexity of } f_i\text{)} \\
&= h_j - f^*, \qquad (23)
\end{aligned}$$

Using (23) in (22) gives

$$\|x_{j+1} - x_*\|^2 \leq \|x_j - x_*\|^2 - 2\eta_j (h_j - f^*) + \eta_j^2 \|\bar{g}_j\|^2 . \tag{24}$$

Summing up from $j = 0, \ldots, k$ and telescoping we have that

$$\begin{aligned}
0 &\leq \|x_{k+1} - x_*\|^2 \\
&\leq \|x_0 - x_*\|^2 - 2 \sum_{j=0}^{k} \eta_j (h_j - f^*) + \sum_{j=0}^{k} \eta_j^2 \|\bar{g}_j\|^2 . \qquad (25)
\end{aligned}$$

From step (25) and re-arranging we have that

$$2f^* \Big(\sum_{j=0}^{k} \eta_j \Big) \geq 2 \sum_{j=0}^{k} \eta_j h_j - \|x_0 - x_*\|^2 - \sum_{j=0}^{k} \eta_j^2 \|\bar{g}_j\|^2 .$$

Dividing through by $\big(\sum_{j=0}^{k} \eta_j \big)$ gives the estimate

$$f^* \geq \underline{f_{k+1}} := \frac{2 \sum_{j=0}^{k} \eta_j h_j - \|x_0 - x_*\|^2 - \sum_{j=0}^{k} \eta_j^2 \|\bar{g}_j\|^2}{2 \sum_{j=0}^{k} \eta_j}.$$

Finally the recurrence follows since

$$\begin{aligned}\underline{f_{k+1}} &= \frac{2\sum_{j=0}^{k}\eta_j h_j - \|x_0 - x_*\|^2 - \sum_{j=0}^{k}\eta_j^2\|\bar{g}_j\|^2}{2\sum_{j=0}^{k}\eta_j} \\ &= \frac{\sum_{j=0}^{k}\eta_j}{\sum_{j=0}^{k}\eta_j}\frac{2\sum_{j=0}^{k}\eta_j h_j - \|x_0 - x_*\|^2 - \sum_{j=0}^{k}\eta_j^2\|\bar{g}_j\|^2}{2\sum_{j=0}^{k}\eta_j} \\ &\quad + \frac{2\eta_k h_k - \eta_k^2\|\bar{g}_k\|^2}{2\sum_{j=0}^{k}\eta_j} \\ &= \underline{f_k} + \frac{\eta_k\left(h_k - \frac{1}{2}\eta_k\|\bar{g}_k\|^2\right)}{\sum_{j=0}^{k}\eta_j}.\end{aligned}$$

■

Appendix D. Steps with Maximal Progress

Here we show that the step size of MSAG in (10) can be seen as the step size of SAG that maximize progress towards the solution. For this viewpoint to hold, we need that $\underline{f_k}$ be the tightest possible lower bound, that is $\underline{f_k} = f(x_*) =: f^*$. Consequently for this section we use MSAG step size (10) with $\alpha_k = \infty$ and $\underline{f_k} = f^*$, that is

$$\eta_k := \frac{\left(\frac{1}{n}\sum_{i=1}^{n}(f_i(q_i^k) + \langle\nabla f_i(q_i^k), x_k - q_i^k\rangle - f^*)\right)_+}{\|\bar{g}_k\|^2} = \frac{(h_k - f^*)_+}{\|\bar{g}_k\|^2} \tag{26}$$

Thus through this section an iterate x_k refers to an iterate of Algorithm 1 where $\underline{f_k} = f^*$ and $\alpha_k = \infty$. Next we show how (26) is also the step size that minimizes an upper bound on the distance to x_*. This interpretation is based on the recent Polyak momentum methods [23, 28].

D.1. Monotonicity

Now let the step size $\eta_k > 0$ be a free parameter. We can now view the next iterate x_{k+1} of SAG in (5) as a function of η_k, that is $x_{k+1}(\eta_k)$. We would like to choose η_k so that x_{k+1} is as close as possible to the optimum solution x_*, that is to minimize $\|x_{k+1}(\eta_k) - x_*\|^2$ in η_k. This is not possible because we do not know x_*. But we can minimize an upper bound of $\|x_{k+1}(\eta_k) - x_*\|^2$ if we assume that $f_i(x)$ is a convex function.

Lemma 5 *Let f_i be convex for $i \in \{1, \dots, n\}$. Consider the iterates of SAG (5). It follows that minimizing in the η_k the upper bound*

$$\|x_{k+1} - x_*\|^2 \leq \|x_k - x_*\|^2 - 2\eta_k(h_k - f^*) + \eta_k^2\|\bar{g}_k\|^2, \tag{27}$$

gives the step size (26). Thus these optimal stepsizes are equivalent to MSAG with $\underline{f_k} = f^*$ and $\alpha_k = \infty$. Plugging the stepsize (26) or MSAG (10) with $\underline{f_k} = f^*$ into (27) gives

$$\|x_{k+1} - x_*\|^2 \leq \|x_k - x_*\|^2 - \eta_k(h_k - f^*)_+. \tag{28}$$

D.2. Proof of Lemma 5

Proof . Subtracting x_* from both sides, taking norms and expanding the squares we have that

$$\|x_{k+1} - x_*\|^2 = \|x_k - x_*\|^2 - 2\eta_k \langle \bar{g}_k, x_k - x_* \rangle + \eta_k^2 \|\bar{g}_k\|^2 . \tag{29}$$

For shorthand let $\nabla f_i := \nabla f_i(q_i^k)$ and $f_i = f_i(q_i^k)$. Now using that (23) in (29) gives

$$\begin{aligned} \|x_{k+1} - x_*\|^2 &= \|x_k - x_*\|^2 - 2\eta_k \langle \bar{g}_k, x_k - x_* \rangle + \eta_k^2 \|\bar{g}_k\|^2 \\ &\leq \|x_k - x_*\|^2 - 2\eta_k (h_k - f^*) + \eta_k^2 \|\bar{g}_k\|^2 . \end{aligned} \tag{30}$$

If we now minimize the right-had side of the above in η_k, but restricted to $\eta_k \geq 0$ we arrive at (26). Inserting (26) back in we have that

$$\|x_{k+1} - x_*\|^2 \leq \|x_k - x_*\|^2 - \frac{(h_k - f^*)_+^2}{\|\bar{g}_k\|^2} = \|x_k - x_*\|^2 - \eta_k (h_k - f^*)_+ . \tag{31}$$

Alternatively using the stepsize given in (10) and noting that

$$\eta_k \leq \frac{(h_k - f^*)_+}{\|\bar{g}_k\|^2}$$

we have again that

$$\begin{aligned} \|x_{k+1} - x_*\|^2 &\leq \|x_k - x_*\|^2 - 2\eta_k (h_k - f^*) + \eta_k^2 \|\bar{g}_k\|^2 \\ &\leq \|x_k - x_*\|^2 - 2\eta_k (h_k - f^*) + \eta_k (h_k - f^*)_+ \\ &= \|x_k - x_*\|^2 - \eta_k (h_k - f^*)_+ . \end{aligned}$$

■

Appendix E. Missing Proofs from Convergence Theorems

For our forthcoming theory, it will prove convenient to first re-write the numerator of our adaptive stepsize in (26) as follows.

Lemma 6 (Memory and stepsize) *We have that*

$$h_k := \frac{1}{n} \sum_{i=1}^n (f_i(q_i^k) + \langle \nabla f_i(q_i^k), x_k - q_i^k \rangle) \;=\; f(x_k) - \bar{D}(x_k, \mathbf{Q}_k) \tag{32}$$

where

$$\bar{D}(x, \mathbf{Q}_k) := \frac{1}{n} \sum_{i=1}^n D_{f_i}(x, q_i) := \frac{1}{n} \sum_{i=1}^n f_i(x_k) - f_i(q_i^k) - \left\langle \nabla f_i(q_i^k), x_k - q_i^k \right\rangle . \tag{33}$$

Consequently our adaptive stepsize (26) is equivalent to

$$\eta_k := \frac{\left(f(x_k) - f^* - \bar{D}(x_k, \mathbf{Q}_k)\right)_+}{\|\bar{g}_k\|^2} \tag{34}$$

Proof The proof follows by just re-arranging

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\left(f_i(q_i^k)+\left\langle\nabla f_i(q_i^k),x_k-q_i^k\right\rangle\right) &= \frac{1}{n}\sum_{i=1}^{n}\left(f_i(q_i^k)+\left\langle\nabla f_i(q_i^k),x_k-q_i^k\right\rangle-f_i(x_k)+f_i(x_k)\right),\\ &= f(x_k)-\frac{1}{n}\sum_{i=1}^{n}\left(f_i(x_k)-f_i(q_i^k)-\left\langle\nabla f_i(q_i^k),x_k-q_i^k\right\rangle\right)\\ &= f(x_k)-\bar{D}(x_k,\mathbf{Q}_k). \qquad \blacksquare\end{aligned}$$

Note that (34) is not a practical way of computing the MSAG step size η_k since it apparently requires computing $f(x^k)$ at each iteration. The practical way of computing η_k is given in (10) and in Algorithm 1. We use this equivalent form in (34) only for our forthcoming convergence theory.

E.1. Change in the memory

As a reminder of the order in which the iterates are produced, we start from x_0, form the memory $\mathbf{Q}_0$, move to x_1, and then update the memory to $\mathbf{Q}_1$. That is, we go from $(x_k,\mathbf{Q}_k)$ to $(x_{k+1},\mathbf{Q}_k)$ to $(x_{k+1},\mathbf{Q}_{k+1})$. When updating the memory from $(x_{k+1},\mathbf{Q}_k)$ to $(x_{k+1},\mathbf{Q}_{k+1})$ we have the following contraction.

Proposition 2

$$\bar{D}(x_{k+1},\mathbf{Q}_{k+1})\leq\bar{D}(x_{k+1},\mathbf{Q}_k) \tag{35}$$

$$\mathbb{E}\left[\bar{D}(x_{k+1},\mathbf{Q}_{k+1})\;\middle|\;\mathbf{Q}_k\right]=\left(1-\frac{1}{n}\right)\bar{D}(x_{k+1},\mathbf{Q}_k) \tag{36}$$

Proof The proof of both items follows because $\bar{D}(x_{k+1},\mathbf{Q}_{k+1})$ is equal to $\bar{D}(x_{k+1},\mathbf{Q}_k)$ except one of the terms in the average is now zero. Specifically the i_{k+1} term. Because i_{k+1} is chosen uniformly from $\{1,\dots,n\}$ taking expectation with respect to i_{k+1} gives a contraction of $(1-1/n)$. That is the essence of the proof, which we now explicate. Note that

$$\begin{aligned}\bar{D}(x_{k+1},\mathbf{Q}_{k+1}) &= \frac{1}{n}\sum_{i=1}^{n}D_{f_i}(x_{k+1},q_i^{k+1})\\ &= \frac{1}{n}D_{f_{i_{k+1}}}(x_{k+1},q_{i_{k+1}}^{k+1})+\frac{1}{n}\sum_{i\neq i_{k+1}}^{n}D_{f_i}(x_{k+1},q_i^k).\end{aligned}$$

Since $q_{i_{k+1}}^{k+1}=x^{k+1}$ we have that $D_{f_{i_{k+1}}}(x_{k+1},q_{i_{k+1}}^{k+1})=0$. Taking expectation conditioned on $\mathbf{Q}_k$

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[D_{f_i}(x_{k+1},q_i^{k+1})\;\middle|\;\mathbf{Q}_k\right] &= \frac{1}{n}\mathbb{E}\left[\sum_{i\neq i_{k+1}}^{n}D_{f_i}(x_{k+1},q_i^k)\;\middle|\;\mathbf{Q}_k\right]\\ &= \frac{1}{n}\frac{1}{n}\sum_{j=1}^{n}\sum_{i\neq j}^{n}D_{f_i}(x_{k+1},q_i^k)\\ &= \frac{1}{n}\frac{n-1}{n}\sum_{i=1}^{n}D_{f_i}(x_{k+1},q_i^k) \;=\; \left(1-\frac{1}{n}\right)\bar{D}(x_{k+1},\mathbf{Q}_k),\end{aligned}$$

where in the last but one equality we use a double counting argument to exchange the order of the summation in i and the summation in j. ■

Proposition 3 *[Non-negative stepsize] The numerator of the step size (34) is positive with*

$$f(x_k) - f^* - \bar{D}(x_k, \mathbf{Q}_k) \geq 0. \tag{37}$$

Thus we can drop the positive part function $()_+$ in (34). Furthermore we have that

$$\bar{D}(x_{k+1}, \mathbf{Q}_k) = f(x_{k+1}) - f^*. \tag{38}$$

Proof Since $\bar{D}$ is an average of divergences,

$$\bar{D}(x_k, \mathbf{Q}_k) \coloneqq \frac{1}{n} \sum_{i=1}^{n} D_{f_i}(x_k, q_i^k),$$

where each term satisfies three-point lemma

$$D_{f_i}(x_{k+1}, q_i^k) = D_{f_i}(x_k, q_i^k) + \left\langle \nabla f_i(x_k) - \nabla f_i(q_i^k), x_{k+1} - x_k \right\rangle + D_{f_i}(x_{k+1}, x_k),$$

we have that the average satisfies

$$\bar{D}(x_{k+1}, \mathbf{Q}_k) = \bar{D}(x^k, \mathbf{Q}_k) + \langle \nabla f(x_k) - \bar{g}_k, x_{k+1} - x_k \rangle + D_f(x_{k+1}, x_k)$$

Expanding the divergence $D_f(x_{k+1}, x_k) = f(x_{k+1}) - f(x_k) - \langle \nabla f(x_k), x_{k+1} - x_k \rangle$, parts of the inner product cancel, giving the simplification

$$\begin{aligned} \bar{D}(x_{k+1}, \mathbf{Q}_k) &= \bar{D}(x^k, \mathbf{Q}_k) + f(x_{k+1}) - f(x_k) - \langle \bar{g}_k, x_{k+1} - x_k \rangle \\ &= \bar{D}(x^k, \mathbf{Q}_k) + f(x_{k+1}) - f(x_k) + \eta_k \|\bar{g}_k\|^2. \end{aligned} \tag{39}$$

The remainder of the proof now follows by induction. Our induction hypothesis is that

$$f(x_k) - f(x_*) - \bar{D}(x^k, \mathbf{Q}_k) \geq 0. \tag{40}$$

The base case follows since in Algorithm 1 we initialize $q_i^0 = x_0$ for $i = 1, \ldots, n$ we have that $\bar{D}(x^0, \mathbf{Q}_0) = 0$. Consequently

$$f(x_0) - f(x_*) - \bar{D}(x^0, \mathbf{Q}_0) = f(x_0) - f(x_*) \geq 0.$$

Now suppose that (40) holds at iteration k. Plugging in the stepsize η_k (34) into (39) gives

$$\begin{aligned} \bar{D}(x_{k+1}, \mathbf{Q}_k) &= \bar{D}(x^k, \mathbf{Q}_k) + f(x_{k+1}) - f(x_k) + f(x_k) - f^* - \bar{D}(x_k, \mathbf{Q}_k) \\ &= f(x_{k+1}) - f(x_*). \end{aligned}$$

Consequently from (35) we have that

$$f(x_{k+1}) - f(x_*) - \bar{D}(x_{k+1}, \mathbf{Q}_{k+1}) \geq f(x_{k+1}) - f(x_*) - \bar{D}(x_{k+1}, \mathbf{Q}_k) \ = 0,$$

which concludes the induction hypothesis. ■

Finally, we can show that in expectation, the memory is directly related to the suboptimality.

Lemma 7 *If follows that*

$$\mathbb{E}\left[\bar{D}(x^k, \mathbf{Q}_k)\right] = \frac{n-1}{n}\mathbb{E}\left[f(x_k) - f_{\min}\right] \tag{41}$$

Proof The proof follows from (36) in Proposition 2 and (38) in Proposition 3 since for all k we have that

$$\mathbb{E}\left[\bar{D}(x_{k+1}, \mathbf{Q}_{k+1})\right] = \frac{n-1}{n}\mathbb{E}\left[\bar{D}(x_{k+1}, \mathbf{Q}_k)\right] = \frac{n-1}{n}\mathbb{E}\left[f(x_{k+1}) - f^*\right].$$

■

E.2. Proof of Proposition 1

Proposition 1 *Let f_i be convex for $i \in \{1, \ldots, n\}$. Consider the iterates of* SAG (5) *with the stepsize* (34) *It follows that*

$$\mathbb{E}\left[\|x_{k+1} - x_*\|^2\right] \leq \mathbb{E}\left[\|x_k - x_*\|^2\right] - \tfrac{1}{n^2}\tfrac{\mathbb{E}[f(x_k) - f^*]^2}{\mathbb{E}\left[\|\bar{g}_k\|^2\right]} \tag{11}$$

Proof From (28) in Lemma 5 we have that

$$\|x_{k+1} - x_*\|^2 \leq \|x_k - x_*\|^2 - \frac{(f(x_k) - f^* - \bar{D}(x_k, \mathbf{Q}_k))^2}{\|\bar{g}_k\|^2}, \tag{42}$$

where we have dropped the positive part due to Proposition 3. Using that the mapping $m(x, y) \mapsto x^2/y$ is convex over $(x, y) \in \mathbb{R} \times \mathbb{R}_+$ and Jensen's we have that

$$\mathbb{E}\left[m(X, Y)\right] \geq \frac{\mathbb{E}\left[X\right]^2}{\mathbb{E}\left[Y\right]}, \quad \text{for all random variables } X \text{ and } Y.$$

Using this observation we have that

$$\begin{aligned}
\mathbb{E}\left[\frac{(f(x_k) - f^* - \bar{D}(x_k, \mathbf{Q}_k))^2}{\|\bar{g}_k\|^2}\right] &\geq \frac{\mathbb{E}\left[f(x_k) - f^* - \bar{D}(x_k, \mathbf{Q}_k)\right]^2}{\mathbb{E}\left[\|\bar{g}_k\|^2\right]} \\
&= \frac{\mathbb{E}\left[f(x_k) - f^* - \frac{n-1}{n}(f(x_k) - f^*)\right]^2}{\mathbb{E}\left[\|\bar{g}_k\|^2\right]} \qquad \text{Using (41)} \\
&= \frac{1}{n^2}\frac{\mathbb{E}\left[f(x_k) - f^*\right]^2}{\mathbb{E}\left[\|\bar{g}_k\|^2\right]}.
\end{aligned} \tag{43}$$

Taking expectation in (42) together with the above gives the result. ■

E.3. Proof of Theorem 1

Theorem 1 *[Locally bounded gradients] Assume the gradients are bounded over the sub-level set of the starting point, that is*

$$G = \sup_{x:\|x-x_*\|\leq\|x_0-x_*\|} \max_{i=1,\ldots,n} \|\nabla f_i(x)\| \;\leq\; \infty.$$

Note that this is the case for continuously differentiable functions. Let $\bar{x}_t := \sum_{k=0}^{t-1} x_k$. *If* $f_i(x)$ *is convex for* $i \in [n]$ *then* SAG *with step size* (34) *converges according to*

$$\mathbb{E}\left[f(\bar{x}_t) - f(x_*)\right] \;\leq\; \tfrac{nG}{\sqrt{t}} \left\|x_0 - x_*\right\|.$$

Proof Since Lemma (5) shows the iterates are bounded, that is

$$\|x_k - x_*\| \leq \|x_0 - x_*\| =: D_0,$$

we have, by Jensen's over $x \mapsto \|x\|^2$ that

$$\|\bar{g}_k\|^2 \leq \frac{1}{n}\sum_{i=1}^{n} \left\|\nabla f_i(q_i^k)\right\|^2 \leq \frac{1}{n}\sum_{i=1}^{n} \max_{\|x-x_*\|\leq D_0} \|\nabla f_i(x)\|^2 \leq G^2.$$

Using the above bound in (11) from Proposition 1 gives

$$\begin{aligned}\mathbb{E}\left[\|x_{k+1} - x_*\|^2\right] \;&\leq\; \mathbb{E}\left[\|x_k - x_*\|^2\right] - \frac{1}{n^2}\frac{\mathbb{E}\left[f(x_k) - f^*\right]^2}{\mathbb{E}\left[\|\bar{g}_k\|^2\right]} \\ &\leq\; \mathbb{E}\left[\|x_k - x_*\|^2\right] - \frac{1}{n^2}\frac{\mathbb{E}\left[f(x_k) - f^*\right]^2}{G^2}.\end{aligned}$$

Re-arranging, summing both sides over $k = 0, \ldots, t-1$ and using telescopic cancellation gives

$$\frac{1}{n^2}\sum_{k=0}^{t-1} \frac{\mathbb{E}\left[f(x_k) - f^*\right]^2}{G^2} \;\leq\; \|x_0 - x_*\|^2 - \mathbb{E}\left[\|x_{t+1} - x_*\|^2\right] \;\leq\; \|x_0 - x_*\|^2.$$

Dividing though by t, and using Jensen's twice, once with respect to $f(x)$, then once with respect to $x \mapsto x^2$ which is convex and monotone, gives

$$\begin{aligned}\mathbb{E}\left[f(\bar{x}_t) - f(x_*)\right]^2 &\leq \left(\frac{1}{t}\sum_{k=0}^{t-1} \mathbb{E}\left[f(x_k) - f(x_t)\right]\right)^2 \\ &\leq \frac{1}{t}\sum_{k=0}^{t-1} \mathbb{E}\left[f(x_k) - f^*\right]^2 \;\leq\; \frac{n^2G^2}{t}\left\|x_0 - x_*\right\|^2.\end{aligned}$$

Taking square root on both now gives the result. ■

E.4. Proof of Theorem 2

Here we consider the case where each f_i is convex and $L_{\max}$–smooth. But first we need a bound on the expected norm gradients.

Proposition 4

$$\mathbb{E}\left[\|\bar{g}_k\|^2\right] \leq 4L_{\max}\frac{2n-1}{n}\mathbb{E}\left[f(x_k)-f^*\right]. \tag{44}$$

Proof Since f_i is convex and $L_{\max}$ smooth we have that the *co-coercive* bound holds, namely

$$\|\nabla f_i(x)-\nabla f_i(y)\|^2 \;\leq\; 2L_{\max}D_{f_i}(x,y), \quad \forall x,y. \tag{45}$$

Using this co-coercive bound we have that

$$\begin{aligned}
\|\bar{g}_k\|^2 &= \|\bar{g}_k-\nabla f(x_k)+\nabla f(x_k)\|^2 \\
&\leq 2\|\bar{g}_k-\nabla f(x_k)\|^2+2\|\nabla f(x_k)\|^2 \\
&\leq 2\frac{1}{n}\sum_i\left\|\nabla f_i(q_i^k)-\nabla f_i(x_k)\right\|^2+2\|\nabla f(x_k)\|^2 \\
&\leq 4L_{\max}\left(\frac{1}{n}\sum_{i=1}^n D_{f_i}(x_k,q_i^k)+f(x_k)-f^*\right) \\
&= 4L_{\max}\left(\bar{D}(x_k,\mathbf{Q}_k)+f(x_k)-f^*\right)
\end{aligned}$$

Taking expectation and using Lemma 7 gives

$$\mathbb{E}\left[\|\bar{g}_k\|^2\right] \;\leq\; 4L_{\max}\mathbb{E}\left[\bar{D}(x_k,\mathbf{Q}_k)+f(x_k)-f^*\right] = 4L_{\max}\frac{2n-1}{n}\mathbb{E}\left[f(x_k)-f^*\right].$$

■

Theorem 2 *[Smooth functions] Let $\bar{x}_t := \sum_{k=0}^{t-1} x_k$. If f_i is convex and $L_{\max}$–smooth for $i = 1,\dots,n$, then* SAG *with step size (34) converges according to*

$$\mathbb{E}\left[f(\bar{x}_t)-f^*\right] \leq \tfrac{2L_{\max}n(2n-1)}{t}\mathbb{E}\left[\|x_k-x_*\|^2\right] \tag{12}$$

If in addition $f(x)$ is μ–Polyak-Łojasiewicz then

$$\mathbb{E}\left[\|x_{k+1}-x_*\|^2\right] \leq \left(1-\tfrac{\mu}{4L_{\max}}\tfrac{1}{n}\tfrac{1}{2n-1}\right)\mathbb{E}\left[\|x_k-x_*\|^2\right]. \tag{13}$$

Proof Using (11) from Proposition 1 together with (44) gives

$$\begin{aligned}
\mathbb{E}\left[\|x_{k+1}-x_*\|^2\right] \;&\leq\; \mathbb{E}\left[\|x_k-x_*\|^2\right]-\frac{1}{n^2}\frac{\mathbb{E}\left[f(x_k)-f^*\right]^2}{\mathbb{E}\left[\|\bar{g}_k\|^2\right]} \\
&\leq\; \mathbb{E}\left[\|x_k-x_*\|^2\right]-\frac{1}{2L_{\max}}\frac{\left(\frac{1}{n}\mathbb{E}\left[f(x_k)-f^*\right]\right)^2}{\frac{2n-1}{n}\mathbb{E}\left[f(x_k)-f^*\right]} \\
&=\; \mathbb{E}\left[\|x_k-x_*\|^2\right]-\frac{1}{2L_{\max}}\frac{1}{n}\frac{1}{2n-1}\mathbb{E}\left[f(x_k)-f^*\right]
\end{aligned} \tag{46}$$

Summing up from $k = 0, \ldots, t-1$, and using telescopic cancellation gives

$$\begin{aligned}
\mathbb{E}\left[f(\bar{x}^k) - f^*\right] &\leq \frac{1}{t}\sum_{k=0}^{t-1}\mathbb{E}\left[f(x_k) - f^*\right] \\
&\leq \frac{2L_{\max}n(2n-1)}{t}\sum_{k=0}^{t-1}\left[\mathbb{E}\left[\|x_k - x_*\|^2\right] - \mathbb{E}\left[\|x_{k+1} - x_*\|^2\right]\right] \\
&\leq \frac{2L_{\max}n(2n-1)}{t}\mathbb{E}\left[\|x_k - x_*\|^2\right].
\end{aligned}$$

From Theorem 2 in [17] we have that the PL condition implies the quadratic growth condition, namely

$$f(x) - f(x_*) \geq \frac{\mu}{2}\|x - x_*\|^2.$$

Using the above with $x = x_k$ in (46) gives

$$\mathbb{E}\left[\|x_{k+1} - x_*\|^2\right] \leq \mathbb{E}\left[\|x_k - x_*\|^2\right] - \frac{\mu}{4L_{\max}}\frac{1}{n}\frac{1}{2n-1}\mathbb{E}\left[\|x_k - x_*\|^2\right] \tag{47}$$

which gives the result. ■