A Stochastic Gradient Method with an Exponential Convergence Rate for Finite Training Sets

Supplementary Material

Nicolas Le Roux
SIERRA Project-Team
INRIA - ENS
Paris, France
nicolas@le-roux.name

Mark Schmidt
SIERRA Project-Team
INRIA - ENS
Paris, France
mark.schmidt@inria.fr

Francis Bach
SIERRA Project-Team
INRIA - ENS
Paris, France
francis.bach@ens.fr

In this supplementary material, we present

- **A**: a comparison of the convergence rates of primal and dual FG and coordinate-wise methods to the rate of SAG for ℓ_2-regularized least squares in terms of effective passes through the data.
- **B**: additional experimental results on test errors, additional data sets, and searching for the best step-size.
- **C**: proofs of the two propositions.

A Comparison of convergence rates

We consider the ℓ_2-regularized least squares problem

$$\underset{x \in \mathbb{R}^p}{\text{minimize}} \quad g(x) := \frac{\lambda}{2}\|x\|^2 + \frac{1}{2n}\sum_{i=1}^{n}(a_i^T x - b_i)^2,$$

where to apply SG methods and SAG we can use

$$f_i(x) := \frac{\lambda}{2}\|x\|^2 + \frac{1}{2}(a_i^T x - b_i)^2.$$

If we use b to denote a vector containing the values b_i and A to denote a matrix withs rows a_i, we can re-write this problem as

$$\underset{x \in \mathbb{R}^p}{\text{minimize}} \quad \frac{\lambda}{2}\|x\|^2 + \frac{1}{2n}\|Ax - b\|^2.$$

The Fenchel dual of this problem is

$$\underset{y \in \mathbb{R}^n}{\text{minimize}} \quad d(y) := \frac{n}{2}\|y\|^2 + \frac{1}{2\lambda}y^\top AA^\top y + y^\top b.$$

We can obtain the primal variables from the dual variables by the formula $x = (-1/\lambda)A^\top y$. Convergence rates of different primal and dual algorithms are often expressed in terms of the following Lipschitz constants:

$L_g = \lambda + M_\sigma/n$	(Lipschitz constant of g')
$L_g^i = \lambda + M_i$	(Lipschitz constant for all f_i')
$L_g^j = \lambda + M_j/n$	(Lipschitz constant of all g_j')
$L_d = n + M_\sigma/\lambda$	(Lipschitz constant of d')
$L_d^i = n + M_i/\lambda$	(Lipschitz constant of all d_i')

Here, we use M_σ to denote the maximum eigenvalue of $A^\top A$, M_i to denote the maximum squared row-norm $\max_i\{\|a_i\|^2\}$, and M_j to denote the maximum squared column-norm $\max_j\{\sum_{i=1}^n (a_i)_j^2\}$. We use g'_j to refer to element of j of g', and similarly for d'_i. The convergence rates will also depend on the primal and dual strong-convexity constants:

$$\mu_g = \lambda + m_\sigma/n \qquad \text{(Strong-convexity constant of } g\text{)}$$
$$\mu_d = n + m'_\sigma/\lambda \qquad \text{(Strong-convexity constant of } d\text{)}$$

Here, m_σ is the minimum eigenvalue of $A^\top A$, and m'_σ is the minimum eigenvalue of $AA^\top$.

A.1 Full Gradient Methods

Using a similar argument to [1, Theorem 2.1.15], if we use the basic FG method with a step size of $1/L_g$, then $(f(x^k) - f(x^*))$ converges to zero with rate

$$\left(1 - \frac{\mu_g}{L_g}\right)^2 = \left(1 - \frac{\lambda + m_\sigma/n}{\lambda + M_\sigma/n}\right)^2 = \left(1 - \frac{n\lambda + m_\sigma}{n\lambda + M_\sigma}\right)^2 \leq \exp\left(-2\frac{n\lambda + m_\sigma}{n\lambda + M_\sigma}\right).$$

while a larger step-size of $2/(L_g + \mu_g)$ gives a faster rate of

$$\left(1 - \frac{\mu_g + \mu_g}{L_g + \mu_g}\right)^2 = \left(1 - \frac{n\lambda + m_\sigma}{n\lambda + (M_\sigma + m_\sigma)/2}\right)^2 \leq \exp\left(-2\frac{n\lambda_+ m_\sigma}{n\lambda + (M_\sigma + m_\sigma)/2}\right),$$

where the speed improvement is determined by the size of m_σ.

If we use the basic FG method on the dual problem with a step size of $1/L_d$, then $(d(x^k) - d(x^*))$ converges to zero with rate

$$\left(1 - \frac{\mu_d}{L_d}\right)^2 = \left(1 - \frac{n + m'_\sigma/\lambda}{n + M_\sigma/\lambda}\right)^2 = \left(1 - \frac{n\lambda + m'_\sigma}{n\lambda + M_\sigma}\right)^2 \leq \exp\left(-2\frac{n\lambda + m'_\sigma}{n\lambda + M_\sigma}\right).$$

and with a step-size of $2/(L_d + \mu_d)$ the rate is

$$\left(1 - \frac{\mu_d + \mu_d}{L_d + \mu_d}\right)^2 = \left(1 - \frac{n\lambda + m'_\sigma}{n\lambda + (M_\sigma + m'_\sigma)/2}\right)^2 \leq \exp\left(-2\frac{n\lambda + m'_\sigma}{n\lambda + (M_\sigma + m'_\sigma)/2}\right).$$

Thus, whether we can solve the primal or dual method faster depends on m_σ and m'_σ. In the over-determined case where A has independent columns, a primal method should be preferred. In the under-determined case where A has independent rows, we can solve the dual more efficiently. However, we note that a convergence rate on the dual objective does not necessarily yield a corresponding rate in the primal objective. If A is invertible, or it has neither independent columns nor independent rows, then $m_\sigma = m'_\sigma = 0$ and there is no difference between the primal and dual rates.

The AFG method achieves a faster rate. Applied to the primal with a step-size of $1/L_g$ it has a rate of [1, Theorem 2.2.2]

$$\left(1 - \sqrt{\frac{\mu_g}{L_g}}\right) = \left(1 - \sqrt{\frac{\lambda + m_\sigma/n}{\lambda + M_\sigma/n}}\right) = \left(1 - \sqrt{\frac{n\lambda + m_\sigma}{n\lambda + M_\sigma}}\right) \leq \exp\left(-\sqrt{\frac{n\lambda + m_\sigma}{n\lambda + M_\sigma}}\right),$$

and applied to the dual with a step-size of $1/L_d$ is has a rate of

$$\left(1 - \sqrt{\frac{\mu_d}{L_d}}\right) = \left(1 - \sqrt{\frac{n + m'_\sigma\lambda}{n + M_\sigma/\lambda}}\right) = \left(1 - \sqrt{\frac{n\lambda + m'_\sigma}{n\lambda + M_\sigma}}\right) \leq \exp\left(-\sqrt{\frac{n\lambda + m'_\sigma}{n\lambda + M_\sigma}}\right).$$

A.2 Coordinate-Descent Methods

The cost of applying one iteration of an FG method is $O(np)$. For this same cost we could apply p iterations of a coordinate descent method to the primal, assuming that selecting the coordinate to update has a cost of $O(1)$. If we select coordinates uniformly at random, then the convergence rate for p iterations of coordinate descent with a step-size of $1/L_g^j$ is [2, Theorem 2]

$$\left(1 - \frac{\mu_g}{pL_g^j}\right)^p = \left(1 - \frac{\lambda + m_\sigma/n}{p(\lambda + M_j/n)}\right)^p = \left(1 - \frac{n\lambda + m_\sigma}{p(n\lambda + M_j)}\right)^p \leq \exp\left(-\frac{n\lambda + m_\sigma}{n\lambda + M_j}\right).$$

Here, we see that applying a coordinate-descent method can be much more efficient than an FG method if $M_j << M_\sigma$. This can happen, for example, when the number of variables p is much larger than the number of examples n. Further, it is possible for coordinate descent to be faster than the AFG method if the difference between M_σ and M_j is sufficiently large.

For the $O(np)$ cost of one iteration of the FG method, we could alternately perform n iterations of coordinate descent on the dual problem. With a step size of $1/L_d^i$ this would obtain a rate on the dual objective of

$$\left(1 - \frac{\mu_d}{nL_d^i}\right)^n = \left(1 - \frac{n + m'_\sigma/\lambda}{n(n + M_i/\lambda)}\right)^n = \left(1 - \frac{n\lambda + m'_\sigma}{n(n\lambda + M_i)}\right)^n \leq \exp\left(-\frac{n\lambda + m'_\sigma}{n\lambda + M_i}\right),$$

which will be faster than the dual FG method if $M_i << M_\sigma$. This can happen, for example, when the number of examples n is much larger than the number of variables p. The difference between the primal and dual coordinate methods depends on M_i compared to M_j and m_σ compared to m'_σ.

A.3 Stochastic Average Gradient

For the $O(np)$ cost of one iteration of the FG method, we can perform n iterations of SAG. With a step size of $1/2nL_g$, performing n iterations of the SAG algorithm has a rate of

$$\left(1 - \frac{\mu_g}{8nL_g^i}\right)^n = \left(1 - \frac{\lambda + m_\sigma/n}{8n(\lambda + M_i)}\right)^n = \left(1 - \frac{n\lambda + m_\sigma}{8n(n\lambda + nM_i)}\right)^n \leq \exp\left(-\frac{1}{8}\frac{n\lambda + m_\sigma}{n\lambda + nM_i}\right),$$

This is most similar to the rate obtained with the dual coordinate descent method, but is likely to be slower because of the n term scaling M_i. However, the difference will be decreased for over-determined problems when $m_\sigma >> m'_\sigma$.

Under the condition $n \geqslant 8L_g^i/\mu_g = 8(\lambda + M_i)/(\lambda + m_\sigma/n)$, with a step size of $1/2n\mu_g$ performing n iterations of the SAG algorithm has a rate of

$$\left(1 - \frac{1}{8n}\right)^n = \left(1 - \frac{n\lambda}{8n(n\lambda)}\right)^n \leq \exp\left(-\frac{1}{8}\right).$$

Note that depending on the constants this may or may not not be faster than coordinate descent methods. However, if we consider the typical case where $m_\sigma = m'_\sigma = 0$ with $M_i = O(p)$ and $M_j = O(n)$, then if we have $n = 8(\lambda + M_i)/\lambda$ we obtain

$$\left(1 - \frac{1}{8n}\right)^n = \left(1 - \frac{\lambda}{64(\lambda + M_i)}\right)^n = \left(1 - \frac{n\lambda}{64n(\lambda + M_i)}\right)^n \leq \exp\left(-\frac{1}{64}\frac{n\lambda}{\lambda + M_i}\right),$$

Despite the constant of 64 (which is likely to be highly sub-optimal), from these rates we see that SAG outperforms coordinate descent methods when n is sufficiently large.

B Additional experimental results

B.1 Test errors

In Figure 1, we report test errors (after thresholding the predictors at zero to obtain a binary label). Here, we see that SAG is typically among the fastest methods to reach the final test error, though in the case of the *covertype* data the ESG method achieves a lower test-error despite its poor optimization performance.

B.2 Additional data sets

In Figure 2, we report results for the *quantum* ($n = 50000$, $p = 78$) data set obtained from the KDD Cup 2004 website[1], and on the *sido* data set ($n = 12678$, $p = 4932$) obtained from the Causality Workbench website.[2] We make the following observations from these results:

[1] `http://osmot.cs.cornell.edu/kddcup`

[2] `http://www.causality.inf.ethz.ch/home.php`

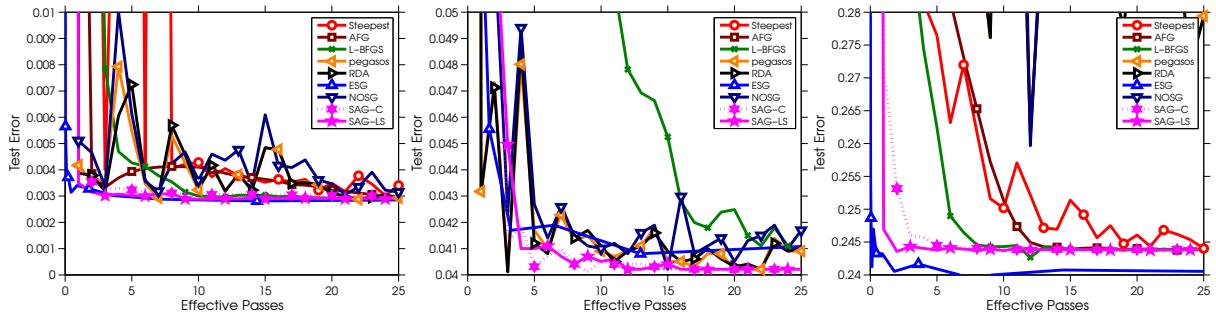

Figure 1: Test errors. From left to right, the *protein*, *rcv1* and *covertype* data sets.

quantum : On this data set the two variants of SAG perform dramatically better than the best competing methods in terms of optimizing the objective, and also outperform the competing methods in terms of reaching the optimal test loss. On this data set, the SG methods (pegasos and RDA) again perform poorly.

sido : The SAG method, particularly with the line-search, is again the most effective method at decreasing the objective function. However, based on the test loss, a regularization parameter of $\lambda = 1/n$ appears to be small and the method is over-fitting. The SG methods again performed poorly on this data set.

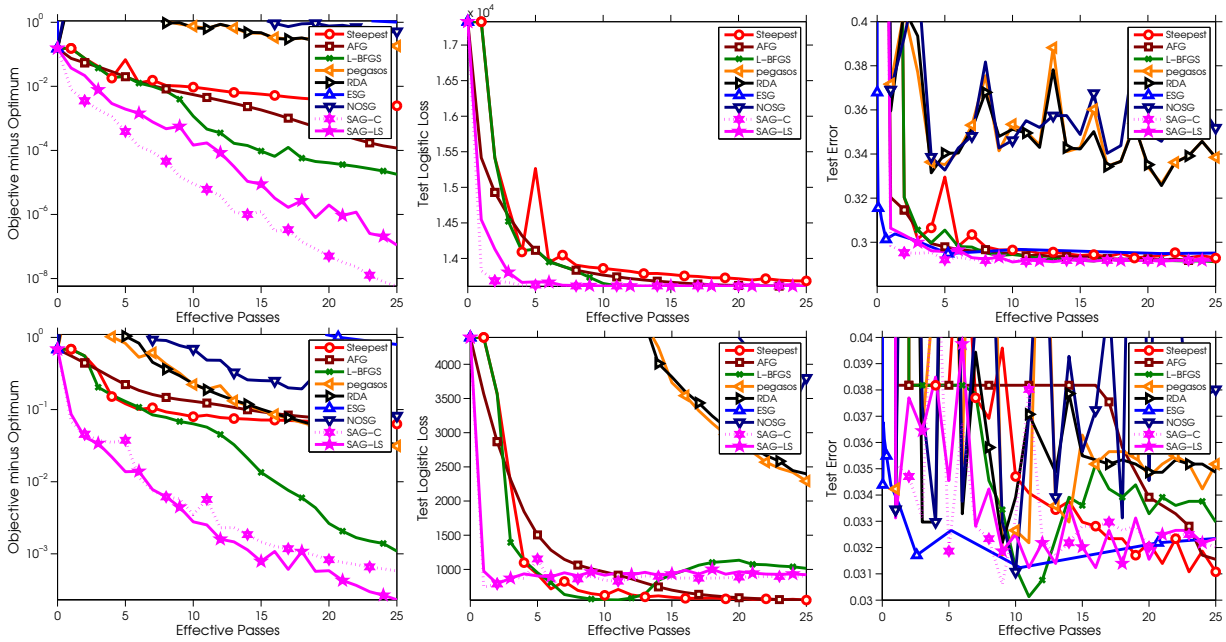

Figure 2: *quantum* and *sido* (bottom) data sets. From left to right: training loss, testing loss, and testing error.

B.3 Searching for best step-sizes

In this series of experiments, we sought to test whether SG methods with a very carefully chosen step size would be competitive with the SAG iterations. In particular, we compared the following variety of basic FG and SG methods.

1. **FG**: The full gradient method described by iteration (3) of the main paper.
2. **AFG**: The accelerated full gradient method of Nesterov, where iterations of (3) are interleaved with an extrapolation step.

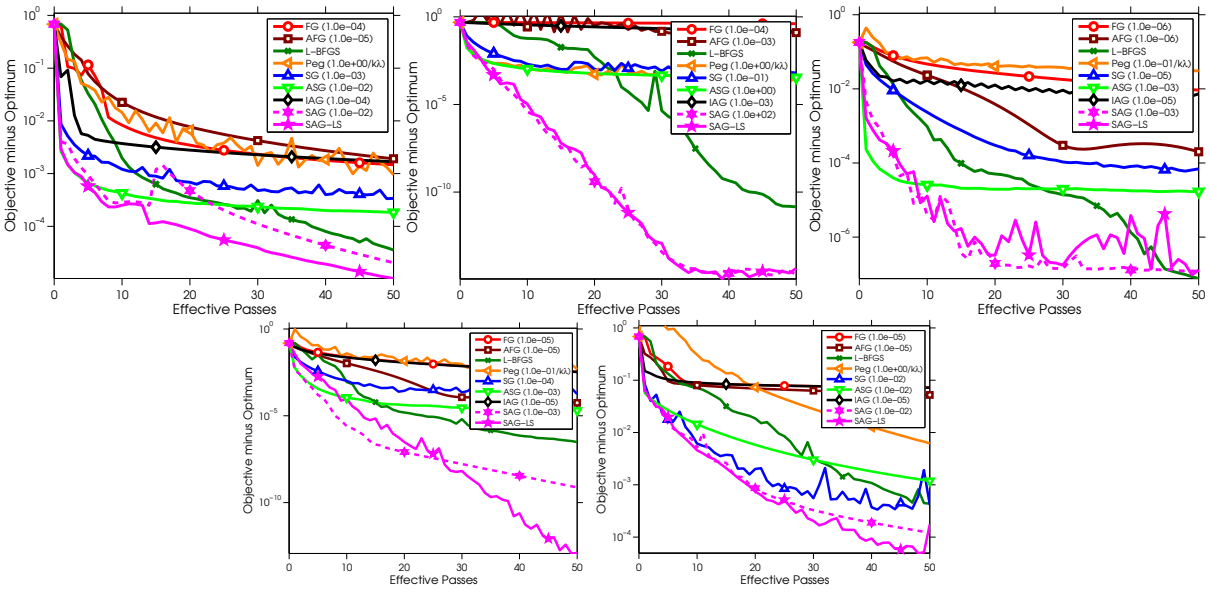

Figure 3: Comparison of optimization strategies that choose the best step-size in hindsight. In the top row are the *protein* (left), *rcv1* (center), and *covertype* (right) data sets. In the bottom row are the *quantum* and *sido* data sets.

3. **peg**: The pegasos algorithm of [3], but where we multiply the step size by a constant.
4. **SG**: The stochastic gradient method described by iteration (4) of the main paper, where we use a constant step-size.
5. **ASG**: The stochastic gradient method described by iteration (4) of the main paper, where we use a constant step size and average the iterates.[3]
6. **IAG**: The incremental aggregated gradient method of [4] described by iteration (5) in the main paper but with a cyclic choice of i_k.
7. **SAG**: The proposed stochastic average gradient method described by iteration (5) in the main paper.

For all of the above methods, we optimized over the full data set and we chose the step size that gave the best performance among powers of 10. We compare these methods to each other and to the L-BFGS and the SAG-LS algorithms from the main paper in Figure 3, which also shows the selected step sizes. In this experiment we see that using a constant or nearly constant step size within SG methods and using averaging tends to perform much better than the basic SG method implemented by pegasos. This makes sense because SG methods with a constant step size have a linear convergence rate when far from the solution. However, the performance is still typically not comparable to that of the SAG iterations, which achieve a linear convergence rate even when close to the solution. We note that the performance and step sizes of the FG and IAG methods are quite similar, while the SAG method chooses much larger step sizes and has much better performance. Finally, the proposed line-search seems to perform as well or better than choosing the optimal fixed step-size in hind sight.

C Proofs of the propositions

We present here the proofs of Propositions 1 and 2.

[3] We have also compared to a variety of other SG methods, such as SG with momentum, SG with gradient averaging, accelerated SG, and using SG but delaying averaging until after the first effective pass. However, none of these SG methods performed better than the ASG method above so we omit them to keep the plots simple.

Proposition 1 *With a step size of $\alpha_k = \frac{1}{2nL}$, the SAG iterations satisfy for $k \geq 1$ that*

$$\mathbb{E}\left[\|x^k - x^*\|^2\right] \leqslant \left(1 - \frac{\mu}{8Ln}\right)^k \left[3\|x_0 - x^*\|^2 + \frac{9\sigma^2}{4L^2}\right] .$$

Proposition 2 *If $\frac{\mu}{L} \geqslant \frac{8}{n}$, with a step size of $\alpha_k = \frac{1}{2n\mu}$ the SAG iterations satisfy for $k \geqslant n$ that*

$$\mathbb{E}\left[g(x^k) - g(x^*)\right] \leqslant C\left(1 - \frac{1}{8n}\right)^k,$$
$$\text{with } C = \left[\frac{16L}{3n}\|x^0 - x^*\|^2 + \frac{4\sigma^2}{3n\mu}\left(8\log\left(1 + \frac{\mu n}{4L}\right) + 1\right)\right] .$$

C.1 Problem set-up and notations

We use $g = \frac{1}{n}\sum_{i=1}^n f_i$ to denote a $\mu-$strongly convex function, where the functions $f_i, i = 1, \ldots, n$ are convex functions from $\mathbb{R}^p$ to $\mathbb{R}$ with L-Lipschitz continuous gradients. Let us denote by x^* the unique minimizer of g.

For $k \geqslant 1$, the stochastic average gradient algorithm performs the recursion

$$x^k = x^{k-1} - \frac{\alpha}{n}\sum_{i=1}^n y_i^k,$$

where an i_k is selected in $\{1, \ldots, n\}$ uniformly at random and we set

$$y_i^k = \begin{cases} f_i'(x^{k-1}) & \text{if } i = i_k, \\ y_i^{k-1} & \text{otherwise.} \end{cases}$$

Denoting z_i^k a random variable which takes the value $1 - \frac{1}{n}$ with probability $\frac{1}{n}$ and $-\frac{1}{n}$ otherwise (thus with zero expectation), this is equivalent to

$$\begin{aligned} y_i^k &= \left(1 - \frac{1}{n}\right) y_i^{k-1} + \frac{1}{n} f_i'(x^{k-1}) + z_i^k \left[f_i'(x^{k-1}) - y_i^{k-1}\right] \\ x^k &= x^{k-1} - \frac{\alpha}{n}\sum_{i=1}^n \left[\left(1 - \frac{1}{n}\right) y_i^{k-1} + \frac{1}{n} f_i'(x^{k-1}) + z_i^k \left[f_i'(x^{k-1}) - y_i^{k-1}\right]\right] \\ &= x^{k-1} - \frac{\alpha}{n}\left[\left(1 - \frac{1}{n}\right) e^\top y^{k-1} + g'(x^{k-1}) + (z^k)^\top \left[f'(x^{k-1}) - y^{k-1}\right]\right], \end{aligned}$$

with

$$e = \begin{pmatrix} I \\ \vdots \\ I \end{pmatrix} \in \mathbb{R}^{np \times p}, \qquad f'(x) = \begin{pmatrix} f_1'(x) \\ \vdots \\ f_n'(x) \end{pmatrix} \in \mathbb{R}^{np}, \qquad z^k = \begin{pmatrix} z_1^k I \\ \vdots \\ z_n^k I \end{pmatrix} \in \mathbb{R}^{np \times p}.$$

Using this definition of z^k, we have $\mathbb{E}[(z^k)(z^k)^\top] = \frac{1}{n} I - \frac{1}{n^2} ee^\top$. Note that, for a given k, the variables $z_1^k, \ldots, z_n^k$ are not independent.

We also use the notation

$$\theta^k = \begin{pmatrix} y_1^k \\ \vdots \\ y_n^k \\ x^k \end{pmatrix} \in \mathbb{R}^{(n+1)p}, \qquad \theta^* = \begin{pmatrix} f_1'(x^*) \\ \vdots \\ f_n'(x^*) \\ x^* \end{pmatrix} \in \mathbb{R}^{(n+1)p} .$$

Finally, if M is a $tp \times tp$ matrix and m is a $tp \times p$ matrix, then:

- $\mathrm{diag}(M)$ is the $tp \times p$ matrix being the concatenation of the t $(p \times p)$-blocks on the diagonal of M;
- $\mathrm{Diag}(m)$ is the $tp \times tp$ block-diagonal matrix whose $(p \times p)$-blocks on the diagonal are equal to the $(p \times p)$-blocks of m.

C.2 Outline of the proofs

Each Proposition will be proved in multiple steps.

1. We shall find a Lyapunov function Q from $\mathbb{R}^{(n+1)p}$ to $\mathbb{R}$ such that the sequence $\mathbb{E}Q(\theta^k)$ decreases at a linear rate.
2. We shall prove that $Q(\theta^k)$ dominates $\|x^k - x^*\|^2$ (in the case of Proposition 2) or $g(x^k) - g(x^*)$ (in the case of Proposition 2) by a constant for all k.
3. In the case of Proposition 2, we show how using one pass of stochastic gradient as the initialization provides the desired result.

Throughout the proofs, $\mathcal{F}_k$ will denote the σ-field of information up to (and including time k), i.e., $\mathcal{F}_k$ is the σ-field generated by $z^1, \dots, z^k$.

C.3 Convergence results for stochastic gradient descent

The constant in both our bounds depends on the initialization chosen. While this does not affect the linear convergence of the algorithm, the bound we obtain for the first few passes through the data is the $O(1/k)$ rate one would get using stochastic gradient descent, but with a constant proportional to n. This problem can be alleviated for the second bound by running stochastic gradient descent for a few iterations before running the SAG algorithm. In this section, we provide bounds for the stochastic gradient descent algorithm which will prove useful for the SAG algorithm.

The assumptions made in this section about the functions f_i and the function g are the same as the ones used for SAG. To get initial values for x^0 and y^0, we will do one pass of standard stochastic gradient.

We denote by $\sigma^2 = \frac{1}{n}\sum_{i=1}^n \|f_i'(x^*)\|^2$ the variance of the gradients at the optimum. We will use the following recursion:

$$\tilde{x}^k = \tilde{x}^{k-1} - \gamma_k f_{i_k}'\left(\tilde{x}^{k-1}\right) \ .$$

Denoting $\delta_k = \mathbb{E}\|\tilde{x}^k - x^*\|^2$, we have (following [5])

$$\delta_k \leqslant \delta_{k-1} - 2\gamma_k(1-\gamma_k L)\mathbb{E}\left[g'(\tilde{x}^{k-1})^\top(\tilde{x}^{k-1} - x^*)\right] + 2\gamma_k^2\sigma^2 \ .$$

Indeed, we have

$$\begin{aligned}
\|\tilde{x}^k - x^*\|^2 &= \|\tilde{x}^{k-1} - x^*\|^2 - 2\gamma_k f_{i_k}'(\tilde{x}^{k-1})^\top(\tilde{x}^{k-1} - x^*) + \gamma_k^2\|f_{i_k}'(\tilde{x}^{k-1})\|^2 \\
&\leqslant \|\tilde{x}^{k-1} - x^*\|^2 - 2\gamma_k f_{i_k}'(\tilde{x}^{k-1})^\top(\tilde{x}^{k-1} - x^*) + 2\gamma_k^2\|f_{i_k}'(x^*)\|^2 + 2\gamma_k^2\|f_{i_k}'(\tilde{x}^{k-1}) - f_{i_k}'(x^*)\|^2 \\
&\leqslant \|\tilde{x}^{k-1} - x^*\|^2 - 2\gamma_k f_{i_k}'(\tilde{x}^{k-1})^\top(\tilde{x}^{k-1} - x^*) + 2\gamma_k^2\|f_{i_k}'(x^*)\|^2 \\
&\quad + 2L\gamma_k^2(f_{i_k}'(\tilde{x}^{k-1}) - f_{i_k}'(x^*))^\top(\tilde{x}^{k-1} - x^*) \ .
\end{aligned}$$

By taking expectations, we get

$$\begin{aligned}
\mathbb{E}\left[\|\tilde{x}^k - x^*\|^2 | \mathcal{F}_{k-1}\right] &\leqslant \|\tilde{x}^{k-1} - x^*\|^2 - 2\gamma_k g'(\tilde{x}^{k-1})^\top(\tilde{x}^{k-1} - x^*) + 2\gamma_k^2\sigma^2 + 2L\gamma_k^2 g'(\tilde{x}^{k-1})^\top(\tilde{x}^{k-1} - x^*) \\
\mathbb{E}\left[\|\tilde{x}^k - x^*\|^2\right] &\leqslant \mathbb{E}\left[\|\tilde{x}^{k-1} - x^*\|^2\right] - 2\gamma_k(1-\gamma_k L)\mathbb{E}\left[g'(\tilde{x}^{k-1})^\top(\tilde{x}^{k-1} - x^*)\right] + 2\gamma_k^2\sigma^2
\end{aligned}$$

Thus, if we take

$$\gamma_k = \frac{1}{2L + \frac{\mu}{2}k} \ ,$$

we have $\gamma_k \leqslant 2\gamma_k(1-\gamma_k L)$ and

$$\begin{aligned}
\delta_k &\leqslant \delta_{k-1} - \gamma_k\mathbb{E}\left[g'(\tilde{x}^{k-1})^\top(x^{k-1} - x^*)\right] + 2\gamma_k^2\sigma^2 \\
&\leqslant \delta_{k-1} - \gamma_k\left[\mathbb{E}\left[g(x^{k-1}) - g(x^*)\right] + \frac{\mu}{2}\delta_{k-1}\right] + 2\gamma_k^2\sigma^2 \text{ using the strong convexity of } g \\
\mathbb{E}g(x^{k-1}) - g(x^*) &\leqslant -\frac{1}{\gamma_k}\delta_k + \left(\frac{1}{\gamma_k} - \frac{\mu}{2}\right)\delta_{k-1} + 2\gamma_k\sigma^2 \\
&\leqslant -\left(2L + \frac{\mu}{2}k\right)\delta_k + \left(2L + \frac{\mu}{2}(k-1)\right)\delta_{k-1} + 2\gamma_k\sigma^2 \ .
\end{aligned}$$

Averaging from $i = 0$ to $k-1$ and using the convexity of g, we have

$$\begin{aligned}
\frac{1}{k}\sum_{i=0}^{k-1}\mathbb{E}g(x^{k-1}) - g(x^*) &\leqslant \frac{2L}{k}\delta_0 + \frac{2\sigma^2}{k}\sum_{i=1}^{k}\gamma_i \\
\mathbb{E}g\left(\frac{1}{k}\sum_{i=0}^{k-1}x^i\right) - g(x^*) &\leqslant \frac{2L}{k}\delta_0 + \frac{2\sigma^2}{k}\sum_{i=1}^{k}\gamma_i \\
&\leqslant \frac{2L}{k}\|x^0 - x^*\|^2 + \frac{2\sigma^2}{k}\sum_{i=1}^{k}\frac{1}{2L + \frac{\mu}{2}i} \\
&\leqslant \frac{2L}{k}L\|x^0 - x^*\|^2 + \frac{2\sigma^2}{k}\int_0^k \frac{1}{2L + \frac{\mu}{2}t}dt \\
&\leqslant \frac{2L}{k}\|x^0 - x^*\|^2 + \frac{4\sigma^2}{k\mu}\log\left(1 + \frac{\mu k}{4L}\right) .
\end{aligned}$$

C.4 Important lemma

In both proofs, our Lyapunov function contains a quadratic term $R(\theta^k) = (\theta^k - \theta^*)^\top \begin{pmatrix} A & b \\ b^\top & c \end{pmatrix} (\theta^k - \theta^*)$ for some values of A, b and c. The lemma below computes the value of $R(\theta^k)$ in terms of elements of θ^{k-1}.

Lemma 1 *If $P = \begin{pmatrix} A & b \\ b^\top & c \end{pmatrix}$, for $A \in \mathbb{R}^{np\times np}$, $b \in \mathbb{R}^{np\times p}$ and $c \in \mathbb{R}^{p\times p}$, then*

$$\begin{aligned}
&\mathbb{E}\left[(\theta^k - \theta^*)^\top \begin{pmatrix} A & b \\ b^\top & c \end{pmatrix} (\theta^k - \theta^*)\middle| \mathcal{F}_{k-1}\right] \\
&\quad= (y^{k-1} - f'(x^*))^\top \left[\left(1 - \frac{2}{n}\right)S + \frac{1}{n}\operatorname{Diag}(\operatorname{diag}(S))\right](y^{k-1} - f'(x^*)) \\
&\quad+ \frac{1}{n}(f'(x^{k-1}) - f'(x^*))^\top \operatorname{Diag}(\operatorname{diag}(S))(f'(x^{k-1}) - f'(x^*)) \\
&\quad+ \frac{2}{n}(y^{k-1} - f'(x^*))^\top [S - \operatorname{Diag}(\operatorname{diag}(S))] (f'(x^{k-1}) - f'(x^*)) \\
&\quad+ 2\left(1 - \frac{1}{n}\right)(y^{k-1} - f'(x^*))^\top \left[b - \frac{\alpha}{n}ec\right](x^{k-1} - x^*) \\
&\quad+ \frac{2}{n}(f'(x^{k-1}) - f'(x^*))^\top \left[b - \frac{\alpha}{n}ec\right](x^{k-1} - x^*) \\
&\quad+ (x^{k-1} - x^*)^\top c(x^{k-1} - x^*) ,
\end{aligned}$$

with

$$S = A - \frac{\alpha}{n}be^\top - \frac{\alpha}{n}eb^\top + \frac{\alpha^2}{n^2}ece^\top .$$

Note that for square $n \times n$ matrix, $\operatorname{diag}(M)$ denotes a vector of size n composed of the diagonal of M, while for a vector m of dimension n, $\operatorname{Diag}(m)$ is the $n \times n$ diagonal matrix with m on its diagonal. Thus $\operatorname{Diag}(\operatorname{diag}(M))$ is a diagonal matrix with the diagonal elements of M on its diagonal, and $\operatorname{diag}(\operatorname{Diag}(m)) = m$.

Proof Throughout the proof, we will use the equality $g'(x) = e^\top f'(x)/n$. Moreover, all conditional expectations of linear functions of z^k will be equal to zero.

We have

$$\begin{aligned}
&\mathbb{E}\left[(\theta^k - \theta^*)^\top \begin{pmatrix} A & b \\ b^\top & c \end{pmatrix} (\theta^k - \theta^*)\middle| \mathcal{F}_{k-1}\right] \\
&\quad= E\left[(y^k - f'(x^*))^\top A(y^k - f'(x^*)) + 2(y^k - f'(x^*))^\top b(x^k - x^*) + (x^k - x^*)^\top c(x^k - x^*)|\mathcal{F}_{k-1}\right] .
\end{aligned} \tag{1}$$

The first term (within the expectation) on the right-hand side of Eq. (1) is equal to

$$\begin{aligned}(y^k - f'(x^*))^\top A(y^k - f'(x^*)) = &\left(1 - \frac{1}{n}\right)^2 (y^{k-1} - f'(x^*))^\top A(y^{k-1} - f'(x^*)) \\ &+ \frac{1}{n^2}(f'(x^{k-1}) - f'(x^*))^\top A(f'(x^{k-1}) - f'(x^*)) \\ &+ [\mathrm{Diag}(z^k)(f'(x^{k-1}) - y^{k-1})]^\top A[\mathrm{Diag}(z^k)(f'(x^{k-1}) - y^{k-1})] \\ &+ \frac{2}{n}\left(1 - \frac{1}{n}\right)(y^{k-1} - f'(x^*))^\top A(f'(x^{k-1}) - f'(x^*)) \ .\end{aligned}$$

The only random term (given $\mathcal{F}_{k-1}$) is the third one whose expectation is equal to

$$\begin{aligned}&\mathbb{E}\left[[\mathrm{Diag}(z^k)(f'(x^{k-1}) - y^{k-1})]^\top A[\mathrm{Diag}(z^k)(f'(x^{k-1}) - y^{k-1})]|\mathcal{F}_{k-1}\right] \\ &\quad = \frac{1}{n}(f'(x^{k-1}) - y^{k-1})^\top \left[\mathrm{Diag}(\mathrm{diag}(A)) - \frac{1}{n}A\right](f'(x^{k-1}) - y^{k-1}) \ .\end{aligned}$$

The second term (within the expectation) on the right-hand side of Eq. (1) is equal to

$$\begin{aligned}(y^k - f'(x^*))^\top b(x^k - x^*) = &\left(1 - \frac{1}{n}\right)(y^{k-1} - f'(x^*))^\top b(x^{k-1} - x^*) \\ &+ \frac{1}{n}(f'(x^{k-1}) - f'(x^*))^\top b(x^{k-1} - x^*) \\ &- \frac{\alpha}{n}\left(1 - \frac{1}{n}\right)^2 (y^{k-1} - f'(x^*))^\top be^\top (y^{k-1} - f'(x^*)) \\ &- \frac{\alpha}{n}\frac{1}{n}\left(1 - \frac{1}{n}\right)(f'(x^{k-1}) - f'(x^*))^\top be^\top (y^{k-1} - f'(x^*)) \\ &- \frac{\alpha}{n}\frac{1}{n}\left(1 - \frac{1}{n}\right)(y^{k-1} - f'(x^*))^\top be^\top (f'(x^{k-1}) - f'(x^*)) \\ &- \frac{\alpha}{n}\frac{1}{n^2}(f'(x^{k-1}) - f'(x^*))^\top be^\top (f'(x^{k-1}) - f'(x^*)) \\ &- \frac{\alpha}{n}[\mathrm{Diag}(z^k)(f'(x^{k-1}) - y^{k-1})]^\top b(z^k)^\top \left[(f'(x^{k-1}) - y^{k-1})\right]\end{aligned}$$

The only random term (given $\mathcal{F}_{k-1}$) is the last one whose expectation is equal to

$$\begin{aligned}&\mathbb{E}\left[[\mathrm{Diag}(z^k)(f'(x^{k-1}) - y^{k-1})]^\top b(z^k)^\top \left[(f'(x^{k-1}) - y^{k-1})\right]|\mathcal{F}_{k-1}\right] \\ &\quad = \frac{1}{n}(f'(x^{k-1}) - y^{k-1})^\top \left(\mathrm{Diag}(\mathrm{diag}(be^\top) - \frac{1}{n}be^\top\right)(f'(x^{k-1}) - y^{k-1}) \ .\end{aligned}$$

The last term on the right-hand side of Eq. (1) is equal to

$$\begin{aligned}(x^k - x^*)^\top c(x^k - x^*) = &(x^{k-1} - x^*)^\top c(x^{k-1} - x^*) \\ &+ \frac{\alpha^2}{n^2}\left(1 - \frac{1}{n}\right)^2 (y^{k-1} - f'(x^*))^\top ece^\top (y^{k-1} - f'(x^*)) \\ &+ \frac{\alpha^2}{n^2}\frac{1}{n^2}(f'(x^{k-1}) - f'(x^*))^\top ece^\top (f'(x^{k-1}) - f'(x^*)) \\ &- \frac{2\alpha}{n}\left(1 - \frac{1}{n}\right)(x^{k-1} - x^*)^\top ce^\top (y^{k-1} - f'(x^*)) \\ &- \frac{2\alpha}{n}\frac{1}{n}(x^{k-1} - x^*)^\top ce^\top (f'(x^{k-1}) - f'(x^*)) \\ &+ \frac{2\alpha^2}{n^2}\frac{1}{n}\left(1 - \frac{1}{n}\right)(y^{k-1} - f'(x^*))^\top ece^\top (f'(x^{k-1}) - f'(x^*)) \\ &+ \frac{\alpha^2}{n^2}\left[(z^k)^\top (f'(x^{k-1}) - y^{k-1})\right]^\top c\left[(z^k)^\top (f'(x^{k-1}) - y^{k-1})\right] \ .\end{aligned}$$

The only random term (given $\mathcal{F}_{k-1}$) is the last one whose expectation is equal to

$$\mathbb{E}\left[\left[(z^k)^\top(f'(x^{k-1})-y^{k-1})\right]^\top c\left[(z^k)^\top(f'(x^{k-1})-y^{k-1})\right]|\mathcal{F}_{k-1}\right]$$
$$=\frac{1}{n}(f'(x^{k-1})-y^{k-1})^\top\left[\mathrm{Diag}(\mathrm{diag}(ece^\top))-\frac{1}{n}ece^\top\right](f'(x^{k-1})-y^{k-1})\,.$$

Summing all these terms together, we get the following result:

$$\begin{aligned}
&\mathbb{E}\left[(\theta^k-\theta^*)^\top\begin{pmatrix}A & b\\ b^\top & c\end{pmatrix}(\theta^k-\theta^*)\middle|\mathcal{F}_{k-1}\right]\\
&\quad=\left(1-\frac{1}{n}\right)^2(y^{k-1}-f'(x^*))^\top S(y^{k-1}-f'(x^*))\\
&\quad+\frac{1}{n^2}(f'(x^{k-1})-f'(x^*))^\top S(f'(x^{k-1})-f'(x^*))\\
&\quad+\frac{1}{n}(f'(x^{k-1})-y^{k-1})^\top\left[\mathrm{Diag}(\mathrm{diag}(S))-\frac{1}{n}S\right](f'(x^{k-1})-y^{k-1})\\
&\quad+\frac{2}{n}\left(1-\frac{1}{n}\right)(y^{k-1}-f'(x^*))^\top S(f'(x^{k-1})-f'(x^*))\\
&\quad+2\left(1-\frac{1}{n}\right)(y^{k-1}-f'(x^*))^\top\left[b-\frac{\alpha}{n}ec\right](x^{k-1}-x^*)\\
&\quad+\frac{2}{n}(f'(x^{k-1})-f'(x^*))^\top\left[b-\frac{\alpha}{n}ec\right](x^{k-1}-x^*)\\
&\quad+(x^{k-1}-x^*)^\top c(x^{k-1}-x^*)
\end{aligned}$$

with $S=A-\frac{\alpha}{n}be^\top-\frac{\alpha}{n}eb^\top+\frac{\alpha^2}{n^2}ece^\top=A-bc^{-1}b^\top+(b-\frac{\alpha}{n}ec)c^{-1}(b-\frac{\alpha}{n}ec)^\top$.

Rewriting $f'(x^{k-1})-y^{k-1}=(f'(x^{k-1})-f'(x^*))-(y^{k-1}-f'(x^*))$, we have

$$\begin{aligned}
&f'(x^{k-1})-y^{k-1})^\top\left[\mathrm{Diag}(\mathrm{diag}(S))-\frac{1}{n}S\right](f'(x^{k-1})-y^{k-1})\\
&\quad=(f'(x^{k-1})-f'(x^*))^\top\left[\mathrm{Diag}(\mathrm{diag}(S))-\frac{1}{n}S\right](f'(x^{k-1})-f'(x^*))\\
&\quad+(y^{k-1}-f'(x^*))^\top\left[\mathrm{Diag}(\mathrm{diag}(S))-\frac{1}{n}S\right](y^{k-1}-f'(x^*))\\
&\quad-2(y^{k-1}-f'(x^*))^\top\left[\mathrm{Diag}(\mathrm{diag}(S))-\frac{1}{n}S\right](f'(x^{k-1})-f'(x^*)).
\end{aligned}$$

Hence, the sum may be rewritten as

$$\begin{aligned}
&\mathbb{E}\left[(\theta^k-\theta^*)^\top\begin{pmatrix}A & b\\ b^\top & c\end{pmatrix}(\theta^k-\theta^*)\middle|\mathcal{F}_{k-1}\right]\\
&\quad=(y^{k-1}-f'(x^*))^\top\left[\left(1-\frac{2}{n}\right)S+\frac{1}{n}\mathrm{Diag}(\mathrm{diag}(S))\right](y^{k-1}-f'(x^*))\\
&\quad+\frac{1}{n}(f'(x^{k-1})-f'(x^*))^\top\mathrm{Diag}(\mathrm{diag}(S))(f'(x^{k-1})-f'(x^*))\\
&\quad+\frac{2}{n}(y^{k-1}-f'(x^*))^\top[S-\mathrm{Diag}(\mathrm{diag}(S))](f'(x^{k-1})-f'(x^*))\\
&\quad+2\left(1-\frac{1}{n}\right)(y^{k-1}-f'(x^*))^\top\left[b-\frac{\alpha}{n}ec\right](x^{k-1}-x^*)\\
&\quad+\frac{2}{n}(f'(x^{k-1})-f'(x^*))^\top\left[b-\frac{\alpha}{n}ec\right](x^{k-1}-x^*)\\
&\quad+(x^{k-1}-x^*)^\top c(x^{k-1}-x^*)
\end{aligned}$$

This concludes the proof. ■

C.5 Analysis for $\alpha = \frac{1}{2nL}$

We now prove Proposition 1, providing a bound for the convergence rate of the SAG algorithm in the case of a small step size, $\alpha = \frac{1}{2nL}$.

Proof

Step 1 - Linear convergence of the Lyapunov function

In this case, our Lyapunov function is quadratic, i.e.,

$$Q(\theta^k) = (\theta^k - \theta^*)^\top \begin{pmatrix} A & b \\ b^\top & c \end{pmatrix} (\theta^k - \theta^*) \; .$$

We consider

$$\begin{aligned} A &= 3n\alpha^2 I + \frac{\alpha^2}{n}(\frac{1}{n} - 2)ee^\top \\ b &= -\alpha(1 - \frac{1}{n})e \\ c &= I \\ S &= 3n\alpha^2 I \\ b - \frac{\alpha}{n}ec &= -\alpha e \; . \end{aligned}$$

The goal will be to prove that $\mathbb{E}[Q(\theta^k)|\mathcal{F}_{k-1}] - (1-\delta)Q(\theta^{k-1})$ is negative for some $\delta > 0$. This will be achieved by bounding all the terms by a term depending on $g'(x^{k-1})^\top(x^{k-1} - x^*)$ whose positivity is guaranteed by the convexity of g.

We have, with our definition of A, b and c:

$$\begin{aligned} S - \mathrm{Diag}(\mathrm{diag}(S)) &= 3n\alpha^2 I - 3n\alpha^2 I = 0 \\ e^\top(f'(x^{k-1}) - f'(x^*)) &= n[g'(x^{k-1}) - g'(x^*)] = ng'(x^{k-1}) \; . \end{aligned}$$

This leads to (using the lemma of the previous section):

$$\begin{aligned}
\mathbb{E}[Q(\theta^k)|\mathcal{F}_{k-1}] &= \mathbb{E}\left[(\theta^k-\theta^*)^\top \begin{pmatrix} A & b \\ b^\top & c \end{pmatrix}(\theta^k-\theta^*)\middle|\mathcal{F}_{k-1}\right] \\
&= \left(1-\frac{1}{n}\right)3n\alpha^2(y^{k-1}-f'(x^*))^\top(y^{k-1}-f'(x^*)) \\
&\quad + (x^{k-1}-x^*)^\top(x^{k-1}-x^*) - \frac{2\alpha}{n}(x^{k-1}-x^*)^\top e^\top(f'(x^{k-1})-f'(x^*)) \\
&\quad + 3\alpha^2(f'(x^{k-1})-f'(x^*))^\top(f'(x^{k-1})-f'(x^*)) \\
&\quad - 2\alpha\left(1-\frac{1}{n}\right)(y^{k-1}-f'(x^*))^\top e(x^{k-1}-x^*) \\
&= \left(1-\frac{1}{n}\right)3n\alpha^2(y^{k-1}-f'(x^*))^\top(y^{k-1}-f'(x^*)) \\
&\quad + (x^{k-1}-x^*)^\top(x^{k-1}-x^*) - 2\alpha(x^{k-1}-x^*)^\top g'(x^{k-1}) \\
&\quad + 3\alpha^2(f'(x^{k-1})-f'(x^*))^\top(f'(x^{k-1})-f'(x^*)) \\
&\quad - 2\alpha\left(1-\frac{1}{n}\right)(y^{k-1}-f'(x^*))^\top e(x^{k-1}-x^*) \\
&\leqslant \left(1-\frac{1}{n}\right)3n\alpha^2(y^{k-1}-f'(x^*))^\top(y^{k-1}-f'(x^*)) \\
&\quad + (x^{k-1}-x^*)^\top(x^{k-1}-x^*) - 2\alpha(x^{k-1}-x^*)^\top g'(x^{k-1}) \\
&\quad + 3\alpha^2 nL(x^{k-1}-x^*)^\top g'(x^{k-1}) \\
&\quad - 2\alpha\left(1-\frac{1}{n}\right)(y^{k-1}-f'(x^*))^\top e(x^{k-1}-x^*)\,.
\end{aligned}$$

The third line is obtained using the Lipschitz property of the gradient, that is

$$\begin{aligned}
(f'(x^{k-1})-f'(x^*))^\top(f'(x^{k-1})-f'(x^*)) &= \sum_{i=1}^n \|f_i'(x^{k-1})-f_i'(x^*)\|^2 \\
&\leqslant \sum_{i=1}^n L(f_i'(x^{k-1})-f_i'(x^*))^\top(x^{k-1}-x^*) \\
&= nL(g'(x^{k-1})-g'(x^*))^\top(x^{k-1}-x^*)\,,
\end{aligned}$$

where the inequality in the second line stems from [1, Theorem 2.1.5].
We have

$$\begin{aligned}
(1-\delta)Q(\theta^{k-1}) &= (1-\delta)(\theta^{k-1}-\theta^*)^\top \begin{pmatrix} A & b \\ b^\top & c \end{pmatrix}(\theta^{k-1}-\theta^*) \\
&= (1-\delta)(y^{k-1}-f'(x^*))^\top\left[3n\alpha^2 I + \frac{\alpha^2}{n}\left(\frac{1}{n}-2\right)ee^\top\right](y^{k-1}-f'(x^*)) \\
&\quad + (1-\delta)(x^{k-1}-x^*)^\top(x^{k-1}-x^*) \\
&\quad - 2\alpha(1-\delta)\left(1-\frac{1}{n}\right)(y^{k-1}-f'(x^*))^\top e(x^{k-1}-x^*)\,.
\end{aligned}$$

The difference is then:

$$\begin{aligned}
&\mathbb{E}[Q(\theta^k)|\mathcal{F}_{k-1}] - (1-\delta)Q(\theta^{k-1}) \\
&\quad\leqslant (y^{k-1} - f'(x^*))^\top \left[3n\alpha^2\left(\delta - \frac{1}{n}\right) I + (1-\delta)\frac{\alpha^2}{n}\left(2 - \frac{1}{n}\right) ee^\top\right] (y^{k-1} - f'(x^*)) \\
&\qquad + \delta(x^{k-1} - x^*)^\top (x^{k-1} - x^*) \\
&\qquad - (2\alpha - 3\alpha^2 nL)(x^{k-1} - x^*)^\top g'(x^{k-1}) \\
&\qquad - 2\alpha\delta\left(1 - \frac{1}{n}\right)(y^{k-1} - f'(x^*))^\top e(x^{k-1} - x^*).
\end{aligned}$$

Note that for any symmetric negative definite matrix M and for any vectors s and t we have

$$(s + \frac{1}{2}M^{-1}t)^\top M(s + \frac{1}{2}M^{-1}t) \leqslant 0,$$

and thus that

$$s^\top M s + s^\top t \leqslant -\frac{1}{4}t^\top M^{-1} t\ .$$

Using this fact with

$$\begin{aligned}
M &= \left[3n\alpha^2\left(\delta - \frac{1}{n}\right) I + (1-\delta)\frac{\alpha^2}{n}\left(2 - \frac{1}{n}\right) ee^\top\right] \\
&= \left[3n\alpha^2\left(\delta - \frac{1}{n}\right)\left(I - \frac{ee^\top}{n}\right) + \alpha^2\left(3n\delta - 1 - 2\delta + \frac{\delta - 1}{n}\right)\frac{ee^\top}{n}\right] \\
s &= y^{k-1} - f'(x^*) \\
t &= -2\alpha\delta\left(1 - \frac{1}{n}\right) e(x^{k-1} - x^*)\ ,
\end{aligned}$$

we have

$$\begin{aligned}
&(y^{k-1} - f'(x^*))^\top \left[3n\alpha^2\left(\delta - \frac{1}{n}\right) I + (1-\delta)\frac{\alpha^2}{n}\left(2 - \frac{1}{n}\right) ee^\top\right] (y^{k-1} - f'(x^*)) \\
&\qquad - 2\alpha\delta\left(1 - \frac{1}{n}\right)(y^{k-1} - f'(x^*))^\top e(x^{k-1} - x^*) \\
&\leqslant -\alpha^2\delta^2\left(1 - \frac{1}{n}\right)^2 (x^{k-1} - x^*)^\top e^\top \left[3n\alpha^2\left(\delta - \frac{1}{n}\right)\left(I - \frac{ee^\top}{n}\right)\right. \\
&\qquad\qquad \left. + \alpha^2\left(3n\delta - 1 - 2\delta + \frac{\delta - 1}{n}\right)\frac{ee^\top}{n}\right]^{-1} e(x^{k-1} - x^*) \\
&= -\frac{\alpha^2\delta^2\left(1 - \frac{1}{n}\right)^2 n}{\alpha^2\left[3n\delta - 1 - 2\delta + \frac{\delta-1}{n}\right]}\|x^{k-1} - x^*\|^2 \\
&= -\frac{\delta^2\left(1 - \frac{1}{n}\right)^2 n}{3n\delta - 1 - 2\delta + \frac{\delta-1}{n}}\|x^{k-1} - x^*\|^2\ .
\end{aligned}$$

A sufficient condition for M to be negative definite is to have $\delta \leqslant \frac{1}{3n}$.

The bound then becomes

$$\begin{aligned}
\mathbb{E}[Q(\theta^k)|\mathcal{F}_{k-1}] - (1-\delta)Q(\theta^{k-1}) \leqslant &-(2\alpha - 3\alpha^2 nL)(x^{k-1} - x^*)^\top g'(x^{k-1}) \\
&+ \left(\delta - \frac{\delta^2\left(1 - \frac{1}{n}\right)^2}{\left[3n\delta - 1 - 2\delta + \frac{\delta-1}{n}\right]} n\right)\|x^{k-1} - x^*\|^2\ .
\end{aligned}$$

We now use the strong convexity of g to get the inequality

$$\|x^{k-1} - x^*\|^2 \leqslant \frac{1}{\mu}(x^{k-1} - x^*)^\top g'(x^{k-1})\ .$$

This yields the final bound

$$\mathbb{E}[Q(\theta^k)|\mathcal{F}_{k-1}] - (1-\delta)Q(\theta^{k-1}) \leqslant -\left(2\alpha - 3\alpha^2 nL + \frac{\delta^2\left(1-\frac{1}{n}\right)^2}{\left[3n\delta - 1 - 2\delta + \frac{\delta-1}{n}\right]}\frac{n}{\mu} - \frac{\delta}{\mu}\right)(x^{k-1}-x^*)^\top g'(x^{k-1}).$$

Since we know that $(x^{k-1}-x^*)^\top g'(x^{k-1})$ is positive, due to the convexity of g, we need to prove that $\left(2\alpha - 3\alpha^2 nL + \frac{\delta^2\left(1-\frac{1}{n}\right)^2}{\left[3n\delta - 1 - 2\delta + \frac{\delta-1}{n}\right]}\frac{n}{\mu} - \frac{\delta}{\mu}\right)$ is positive.

Using $\delta = \frac{\mu}{8nL}$ and $\alpha = \frac{1}{2nL}$ gives

$$\begin{aligned}
2\alpha - 3\alpha^2 nL + \frac{\delta^2\left(1-\frac{1}{n}\right)^2}{\left[3n\delta - 1 - 2\delta + \frac{\delta-1}{n}\right]}\frac{n}{\mu} - \frac{\delta}{\mu} &= \frac{1}{nL} - \frac{3}{4nL} - \frac{1}{8nL} - \frac{\delta^2\left(1-\frac{1}{n}\right)^2\frac{n}{\mu}}{1-3n\delta+2\delta+\frac{1-\delta}{n}} \\
&\geqslant \frac{1}{8nL} - \frac{\delta^2\frac{n}{\mu}}{1-3n\delta} \\
&= \frac{1}{8nL} - \frac{\frac{\mu}{64nL^2}}{1-\frac{3\mu}{8L}} \\
&\geqslant \frac{1}{8nL} - \frac{\frac{\mu}{64nL^2}}{1-\frac{3}{8}} \\
&= \frac{1}{8nL} - \frac{\mu}{40nL^2} \\
&= \frac{1}{8nL} - \frac{1}{40nL} \\
&\geqslant 0\,.
\end{aligned}$$

Hence,

$$\mathbb{E}[Q(\theta^k)|\mathcal{F}_{k-1}] - (1-\delta)Q(\theta^{k-1}) \leqslant 0\,.$$

We can then take a full expectation on both sides to obtain:

$$\mathbb{E}Q(\theta^k) - (1-\delta)\mathbb{E}Q(\theta^{k-1}) \leqslant 0\,.$$

Since Q is a non-negative function (we show below that it dominates a non-negative function), this results proves the linear convergence of the sequence $\mathbb{E}Q(\theta^k)$ with rate $1-\delta$. We have

$$\mathbb{E}Q(\theta^k) \leqslant \left(1 - \frac{\mu}{8nL}\right)^k Q(\theta^0)\,.$$

Step 2 - Domination of $\|x^k - x^*\|^2$ by $Q(\theta^k)$

We now need to prove that $Q(\theta^k)$ dominates $\|x^k - x^*\|^2$. If $P - \begin{pmatrix} 0 & 0 \\ 0 & \frac{1}{3}I \end{pmatrix}$ is positive definite, then $Q(\theta^k) \geqslant \frac{1}{3}\|x^k - x^*\|^2$.

We shall use the Schur complement condition for positive definiteness. Since A is positive definite, the other condition to verify is $\frac{2}{3}I - b^\top A^{-1} b \succ 0$.

$$\begin{aligned}
\frac{2}{3}I - \alpha^2\left(1-\frac{1}{n}\right)^2 e^\top\left[\left(3n\alpha^2 + \frac{\alpha^2}{n} - 2\alpha^2\right)\frac{ee^\top}{n}\right]^{-1} e &= \frac{2}{3}I - \frac{n\left(1-\frac{1}{n}\right)^2}{3n+\frac{1}{n}-2}\frac{ee^\top}{n} \\
&\succ \frac{2}{3}I - \frac{n}{3n-2}\frac{ee^\top}{n} \\
&\succ 0 \text{ for } n \geqslant 2\,,
\end{aligned}$$

and so P dominates $\begin{pmatrix} 0 & 0 \\ 0 & \frac{1}{3}I \end{pmatrix}$.

This yields

$$\begin{aligned}\mathbb{E}\|x^k - x^*\|^2 &\leqslant 3\mathbb{E}Q(\theta^k) \\ &\leqslant 3\left(1 - \frac{\mu}{8nL}\right)^k Q(\theta^0)\ .\end{aligned}$$

We have

$$\begin{aligned}Q(\theta^0) &= 3n\alpha^2 \sum_i \|y_i^0 - f_i'(x^*)\|^2 + \frac{(1-2n)\alpha}{n^2}\left\|\sum_i y_i^0\right\|^2 - 2\alpha\left(1 - \frac{1}{n}\right)(x^0 - x^*)^\top\left(\sum_i y_i^0\right) + \|x^0 - x^*\|^2 \\ &= \frac{3}{4nL^2}\sum_i \|y_i^0 - f_i'(x^*)\|^2 + \frac{(1-2n)}{2n^3L}\left\|\sum_i y_i^0\right\|^2 - \frac{n-1}{n^2L}(x^0 - x^*)^\top\left(\sum_i y_i^0\right) + \|x^0 - x^*\|^2\ .\end{aligned}$$

Initializing all the y_i^0 to 0, we get

$$Q(\theta^0) = \frac{3\sigma^2}{4L^2} + \|x^0 - x^*\|^2\ ,$$

and

$$\mathbb{E}\|x^k - x^*\|^2 \leqslant \left(1 - \frac{\mu}{8nL}\right)^k\left(\frac{9\sigma^2}{4L^2} + 3\|x^0 - x^*\|^2\right)\ .$$

■

C.6 Analysis for $\alpha = \frac{1}{2n\mu}$

Step 1 - Linear convergence of the Lyapunov function

We now prove Proposition 2, providing a bound for the convergence rate of the SAG algorithm in the case of a small step size, $\alpha = \frac{1}{2n\mu}$.

We shall use the following Lyapunov function:

$$Q(\theta^k) = 2g\left(x^k + \frac{\alpha}{n}e^\top y^k\right) - 2g(x^*) + (\theta^k - \theta^*)^\top\begin{pmatrix} A & b \\ b^\top & c\end{pmatrix}(\theta^k - \theta^*)\ ,$$

with

$$\begin{aligned}A &= \frac{\eta\alpha}{n}I + \frac{\alpha}{n}(1-2\nu)ee^\top \\ b &= -\nu e \\ c &= 0\ .\end{aligned}$$

This yields

$$\begin{aligned}S &= \frac{\eta\alpha}{n}I + \frac{\alpha}{n}ee^\top \\ \mathrm{Diag}(\mathrm{diag}(S)) &= \frac{(1+\eta)\alpha}{n}I \\ S - \mathrm{Diag}(\mathrm{diag}(S)) &= \frac{\alpha}{n}(ee^\top - I) \\ \left(1 - \frac{2}{n}\right)S + \frac{1}{n}\mathrm{Diag}(\mathrm{diag}(S)) &= \left(1 - \frac{2}{n}\right)\left[\frac{\eta\alpha}{n}I + \frac{\alpha}{n}ee^\top\right] + \frac{1}{n}\frac{(1+\eta)\alpha}{n}I = \left(1 - \frac{2}{n}\right)\frac{\alpha}{n}ee^\top + \left(\eta - \frac{\eta-1}{n}\right)\frac{\alpha}{n}I\ .\end{aligned}$$

We have

$$\begin{aligned}
&\mathbb{E}[Q(\theta^k)|\mathcal{F}_{k-1}] - (1-\delta)Q(\theta^{k-1}) \\
&\quad = 2g(x^{k-1}) - 2g(x^*) - 2(1-\delta)g\left(x^{k-1} + \frac{\alpha}{n}e^\top y^{k-1}\right) + 2(1-\delta)g(x^*) \\
&\qquad + (y^{k-1} - f'(x^*))^\top \left[\left(1 - \frac{2}{n}\right)\frac{\alpha}{n}ee^\top + \left(\eta - \frac{\eta-1}{n}\right)\frac{\alpha}{n}I - (1-\delta)\frac{\eta\alpha}{n}I \right. \\
&\qquad\qquad \left. -(1-\delta)\frac{\alpha}{n}(1-2\nu)ee^\top\right](y^{k-1} - f'(x^*)) \\
&\qquad - \frac{2\nu}{n}(x^{k-1} - x^*)^\top e^\top (f'(x^{k-1}) - f'(x^*)) \\
&\qquad + \frac{(1+\eta)\alpha}{n^2}(f'(x^{k-1}) - f'(x^*))^\top (f'(x^{k-1}) - f'(x^*)) \\
&\qquad + \frac{2\alpha}{n^2}(y^{k-1} - f'(x^*))^\top \left[ee^\top - I\right](f'(x^{k-1}) - f'(x^*)) \\
&\qquad + 2\left(\frac{1}{n} - \delta\right)\nu(y^{k-1} - f'(x^*))^\top e(x^{k-1} - x^*).
\end{aligned}$$

Our goal will now be to express all the quantities in terms of $(x^{k-1} - x^*)^\top g'(x^{k-1})$ whose positivity is guaranteed by the convexity of g.

Using the convexity of g, we have

$$-2(1-\delta)g\left(x^{k-1} + \frac{\alpha}{n}e^\top y^{k-1}\right) \leqslant -2(1-\delta)\left[g(x^{k-1}) + \frac{\alpha}{n}g'(x^{k-1})e^\top y^{k-1}\right] .$$

Using the Lipschitz property of the gradients of f_i, we have

$$\begin{aligned}
(f'(x^{k-1}) - f'(x^*))^\top (f'(x^{k-1}) - f'(x^*)) &= \sum_{i=1}^n \|f_i'(x^{k-1}) - f_i'(x^*)\|^2 \\
&\leqslant \sum_{i=1}^n L(f_i'(x^{k-1}) - f_i'(x^*))^\top (x^{k-1} - x^*) \\
&= nL(g'(x^{k-1}) - g'(x^*))^\top (x^{k-1} - x^*) .
\end{aligned}$$

Using $e^\top[f'(x^{k-1}) - f'(x^*)] = ng'(x^{k-1})$, we have

$$\begin{aligned}
-\frac{2\nu}{n}(x^{k-1} - x^*)^\top e^\top (f'(x^{k-1}) - f'(x^*)) &= -2\nu(x^{k-1} - x^*)^\top g'(x^{k-1}) \\
\frac{2\alpha}{n^2}(y^{k-1} - f'(x^*))^\top ee^\top (f'(x^{k-1}) - f'(x^*)) &= \frac{2\alpha}{n}(y^{k-1} - f'(x^*))^\top eg'(x^{k-1}) .
\end{aligned}$$

Reassembling all the terms together, we get

$$\begin{aligned}
&\mathbb{E}[Q(\theta^k)|\mathcal{F}_{k-1}] - (1-\delta)Q(\theta^{k-1}) \\
&\quad \leqslant 2\delta[g(x^{k-1}) - g(x^*)] + \frac{2\delta\alpha}{n}g'(x^{k-1})e^\top y^{k-1} \\
&\qquad + (y^{k-1} - f'(x^*))^\top \left[\left(1 - \frac{2}{n}\right)\frac{\alpha}{n}ee^\top + \left(\eta - \frac{\eta-1}{n}\right)\frac{\alpha}{n}I - (1-\delta)\frac{\eta\alpha}{n}I - \right. \\
&\qquad\qquad \left. (1-\delta)\frac{\alpha}{n}(1-2\nu)ee^\top\right](y^{k-1} - f'(x^*)) \\
&\qquad - \left(2\nu - \frac{(1+\eta)\alpha L}{n}\right)(x^{k-1} - x^*)^\top g'(x^{k-1}) \\
&\qquad - \frac{2\alpha}{n^2}(y^{k-1} - f'(x^*))^\top \big(f'(x^{k-1}) - f'(x^*)\big) \\
&\qquad + 2\left(\frac{1}{n} - \delta\right)\nu(y^{k-1} - f'(x^*))^\top e(x^{k-1} - x^*).
\end{aligned}$$

Using the convexity of g gives

$$2\delta[g(x^{k-1}) - g(x^*)] \leqslant 2\delta[x^{k-1} - x^*]^\top g'(x^{k-1}) \ ,$$

and, consequently,

$$\begin{aligned}
&\mathbb{E}[Q(\theta^k)|\mathcal{F}_{k-1}] - (1-\delta)Q(\theta^{k-1}) \\
&\quad \leqslant 2\delta[(x^{k-1}) - (x^*)]^\top g'(x^{k-1}) + \frac{2\delta\alpha}{n} g'(x^{k-1}) e^\top y^{k-1} \\
&\qquad + (y^{k-1} - f'(x^*))^\top \left[\left(1 - \frac{2}{n}\right) \frac{\alpha}{n} ee^\top + \left(\eta - \frac{\eta - 1}{n}\right) \frac{\alpha}{n} I \right. \\
&\qquad\qquad \left. -(1-\delta)\frac{\eta\alpha}{n} I - (1-\delta)\frac{\alpha}{n}(1-2\nu) ee^\top \right] (y^{k-1} - f'(x^*)) \\
&\qquad - \left(2\nu - \frac{(1+\eta)\alpha L}{n}\right) (x^{k-1} - x^*)^\top g'(x^{k-1}) \\
&\qquad - \frac{2\alpha}{n^2} (y^{k-1} - f'(x^*))^\top \big(f'(x^{k-1}) - f'(x^*)\big) \\
&\qquad + 2\left(\frac{1}{n} - \delta\right) \nu (y^{k-1} - f'(x^*))^\top e(x^{k-1} - x^*) \ .
\end{aligned}$$

If we regroup all the terms in $[(x^{k-1}) - (x^*)]^\top g'(x^{k-1})$ together, and all the terms in $(y^{k-1} - f'(x^*))^\top$ together, we get

$$\begin{aligned}
&\mathbb{E}[Q(\theta^k)|\mathcal{F}_{k-1}] - (1-\delta)Q(\theta^{k-1}) \\
&\quad \leqslant \frac{\alpha}{n} (y^{k-1} - f'(x^*))^\top \left[\left(\delta\eta - \frac{\eta - 1}{n}\right) I + \left(\delta - \frac{2}{n} + 2\nu(1-\delta)\right) ee^\top \right] (y^{k-1} - f'(x^*)) \\
&\qquad - \left(2\nu - 2\delta - \frac{(1+\eta)\alpha L}{n}\right) (x^{k-1} - x^*)^\top g'(x^{k-1}) \\
&\qquad + 2(y^{k-1} - f'(x^*))^\top \left[-\frac{\alpha}{n^2}(f'(x^{k-1}) - f'(x^*)) + (\frac{1}{n} - \delta)\nu e(x^{k-1} - x^*) + \frac{\delta\alpha}{n} e g'(x^{k-1}) \right] \ .
\end{aligned}$$

Let us rewrite this as

$$\begin{aligned}
&\mathbb{E}[Q(\theta^k)|\mathcal{F}_{k-1}] - (1-\delta)Q(\theta^{k-1}) \\
&\qquad \leqslant (y^{k-1} - f'(x^*))^\top \left(\tau_{y,I} I + \tau_{y,e} \frac{ee^\top}{n}\right) (y^{k-1} - f'(x^*)) \\
&\qquad + \tau_{x,g} (x^{k-1} - x^*)^\top g'(x^{k-1}) \\
&\qquad + (y^{k-1} - f'(x^*))^\top \left[\tau_{y,f}(f'(x^{k-1}) - f'(x^*)) + \tau_{y,x} e(x^{k-1} - x^*) + \tau_{y,g} e g'(x^{k-1})\right]
\end{aligned}$$

with

$$\begin{aligned}
\tau_{y,I} &= \frac{\alpha}{n}\left(\delta\eta - \frac{\eta - 1}{n}\right) \\
\tau_{y,e} &= \alpha\left(\delta - \frac{2}{n} + 2\nu(1-\delta)\right) \\
\tau_{x,g} &= -(2\nu - 2\delta - \frac{(1+\eta)\alpha L}{n}) \\
\tau_{y,f} &= -\frac{2\alpha}{n^2} \\
\tau_{y,x} &= 2\left(\frac{1}{n} - \delta\right)\nu \\
\tau_{y,g} &= \frac{2\delta\alpha}{n} \ .
\end{aligned}$$

Assuming that $\tau_{y,I}$ and $\tau_{y,e}$ are negative, we have by completing the square that

$$
\begin{aligned}
&(y^{k-1} - f'(x^*))^\top \left(\tau_{y,I} I + \tau_{y,e} \frac{ee^\top}{n}\right) (y^{k-1} - f'(x^*)) \\
&\quad + (y^{k-1} - f'(x^*))^\top \left(\tau_{y,f}(f'(x^{k-1}) - f'(x^*)) + \tau_{y,x} e(x^{k-1} - x^*) + \tau_{y,g} e g'(x^{k-1})\right) \\
&\leqslant -\frac{1}{4} \left(\tau_{y,f}(f'(x^{k-1}) - f'(x^*)) + \tau_{y,x} e(x^{k-1} - x^*) + \tau_{y,g} e g'(x^{k-1})\right)^\top \left(\frac{1}{\tau_{y,I}} \left(I - \frac{ee^\top}{n}\right) + \frac{1}{\tau_{y,I} + \tau_{y,e}} \frac{ee^\top}{n}\right) \\
&\quad \left(\tau_{y,f}(f'(x^{k-1}) - f'(x^*)) + \tau_{y,x} e(x^{k-1} - x^*) + \tau_{y,g} e g'(x^{k-1})\right) \\
&= -\frac{1}{4} \frac{\tau_{y,f}^2}{\tau_{y,I}} \|f'(x^{k-1}) - f'(x^*)\|^2 - \frac{1}{4} \tau_{y,f}^2 n \|g'(x^{k-1})\|^2 \left(\frac{1}{\tau_{y,I} + \tau_{y,e}} - \frac{1}{\tau_{y,I}}\right) \\
&\quad - \frac{1}{4} \frac{\tau_{y,x}^2 n}{\tau_{y,I} + \tau_{y,e}} \|x^{k-1} - x^*\|^2 - \frac{1}{4} \frac{\tau_{y,g}^2 n}{\tau_{y,I} + \tau_{y,e}} \|g'(x^{k-1})\|^2 \\
&\quad - \frac{1}{2} \frac{\tau_{y,f} \tau_{y,x} n}{\tau_{y,I} + \tau_{y,e}} (x^{k-1} - x^*)^\top g'(x^{k-1}) - \frac{1}{2} \frac{\tau_{y,f} \tau_{y,g} n}{\tau_{y,I} + \tau_{y,e}} \|g'(x^{k-1})\|^2 - \frac{1}{2} \frac{\tau_{y,g} \tau_{y,x} n}{\tau_{y,I} + \tau_{y,e}} (x^{k-1} - x^*)^\top g'(x^{k-1}) \,,
\end{aligned}
$$

where we used the fact that $(f'(x^{k-1}) - f'(x^*))^\top e = g'(x^{k-1})$. After reorganization of the terms, we obtain

$$
\begin{aligned}
\mathbb{E}[Q(\theta^k)|\mathcal{F}_{k-1}] - (1-\delta) Q(\theta^{k-1}) \leqslant &\left[\tau_{x,g} - \frac{n \tau_{y,x}}{2(\tau_{y,I} + \tau_{y,e})} (\tau_{y,f} + \tau_{y,g})\right] (x^{k-1} - x^*)^\top g'(x^{k-1}) \\
&- \left[\frac{1}{4} \tau_{y,f}^2 n \left(\frac{1}{\tau_{y,I} + \tau_{y,e}} - \frac{1}{\tau_{y,I}}\right) + \frac{1}{4} \frac{\tau_{y,g}^2 n}{\tau_{y,I} + \tau_{y,e}} + \frac{1}{2} \frac{\tau_{y,f} \tau_{y,g} n}{\tau_{y,I} + \tau_{y,e}}\right] \|g'(x^{k-1})\|^2 \\
&- \frac{1}{4} \frac{\tau_{y,f}^2}{\tau_{y,I}} \|f'(x^{k-1}) - f'(x^*)\|^2 - \frac{1}{4} \frac{\tau_{y,x}^2 n}{\tau_{y,I} + \tau_{y,e}} \|x^{k-1} - x^*\|^2 \,.
\end{aligned}
$$

We now use the strong convexity of the function to get the following inequalities:

$$
\begin{aligned}
\|f'(x^{k-1}) - f'(x^*)\|^2 &\leqslant L n (x^{k-1} - x^*)^\top g'(x^{k-1}) \\
\|x^{k-1} - x^*\|^2 &\leqslant \frac{1}{\mu} (x^{k-1} - x^*)^\top g'(x^{k-1}) \,.
\end{aligned}
$$

Finally, we have

$$
\begin{aligned}
&\mathbb{E}[Q(\theta^k)|\mathcal{F}_{k-1}] - (1-\delta) Q(\theta^{k-1}) \\
&\quad \leqslant \left[\tau_{x,g} - \frac{n \tau_{y,x}}{2(\tau_{y,I} + \tau_{y,e})} (\tau_{y,f} + \tau_{y,g}) - \frac{Ln}{4} \frac{\tau_{y,f}^2}{\tau_{y,I}} - \frac{1}{4\mu} \frac{\tau_{y,x}^2 n}{\tau_{y,I} + \tau_{y,e}}\right] (x^{k-1} - x^*)^\top g'(x^{k-1}) \\
&\quad\quad - \left[\frac{1}{4} \tau_{y,f}^2 n \left(\frac{1}{\tau_{y,I} + \tau_{y,e}} - \frac{1}{\tau_{y,I}}\right) + \frac{1}{4} \frac{\tau_{y,g}^2 n}{\tau_{y,I} + \tau_{y,e}} + \frac{1}{2} \frac{\tau_{y,f} \tau_{y,g} n}{\tau_{y,I} + \tau_{y,e}}\right] \|g'(x^{k-1})\|^2 \,.
\end{aligned}
$$

If we choose $\delta = \frac{\widetilde{\delta}}{n}$ with $\widetilde{\delta} \leqslant \frac{1}{2}$, $\nu = \frac{1}{2n}$, $\eta = 2$ and $\alpha = \frac{1}{2n\mu}$, we get

$$\begin{aligned}
\tau_{y,I} &= \frac{1}{2n^2\mu}\left(\frac{2\widetilde{\delta}}{n} - \frac{1}{n}\right) = -\frac{1-2\widetilde{\delta}}{2n^3\mu} \leqslant 0 \\
\tau_{y,e} &= \frac{1}{2n\mu}\left(\frac{\widetilde{\delta}}{n} - \frac{2}{n} + \frac{1}{n}\left(1 - \frac{\widetilde{\delta}}{n}\right)\right) = -\frac{1}{2n^2\mu}\left(1 - \widetilde{\delta} + \frac{\widetilde{\delta}}{n}\right) \leqslant 0 \\
\tau_{x,g} &= -\left(\frac{1}{n} - \frac{2\widetilde{\delta}}{n} - \frac{3L}{2n^2\mu}\right) = \frac{3L}{2n^2\mu} - \frac{1-2\widetilde{\delta}}{n} \\
\tau_{y,f} &= -\frac{1}{n^3\mu} \\
\tau_{y,x} &= \frac{1-\widetilde{\delta}}{n^2} \\
\tau_{y,g} &= \frac{\widetilde{\delta}}{n^3\mu}\,.
\end{aligned}$$

Thus,

$$\begin{aligned}
&\tau_{x,g} - \frac{n\tau_{y,x}}{2(\tau_{y,I}+\tau_{y,e})}(\tau_{y,f}+\tau_{y,g}) - \frac{Ln}{4}\frac{\tau_{y,f}^2}{\tau_{y,I}} - \frac{1}{4\mu}\frac{\tau_{y,x}^2 n}{\tau_{y,I}+\tau_{y,e}} \\
&\quad\leqslant \frac{3L}{2n^2\mu} - \frac{1-2\widetilde{\delta}}{n} - \frac{\frac{1-\widetilde{\delta}}{2n}\frac{2\widetilde{\delta}-1}{n^3\mu}}{\tau_{y,I}+\tau_{y,e}} + \frac{Ln}{4}\frac{\frac{1}{n^6\mu^2}}{\frac{1-2\widetilde{\delta}}{2n^3\mu}} - \frac{1}{4\mu}\frac{\frac{(1-\widetilde{\delta})^2}{n^3}}{\tau_{y,I}+\tau_{y,e}} \\
&\quad= \frac{L}{n^2\mu}\left[\frac{3}{2} + \frac{1}{2(1-2\widetilde{\delta})}\right] - \frac{1-2\widetilde{\delta}}{n} - \frac{1}{\mu n^3(\tau_{y,I}+\tau_{y,e})}\left[\frac{(1-\widetilde{\delta})^2}{4} + \frac{(1-\widetilde{\delta})(2\widetilde{\delta}-1)}{2n}\right] \\
&\quad\leqslant \frac{L}{n^2\mu}\frac{2-3\widetilde{\delta}}{1-2\widetilde{\delta}} - \frac{1-2\widetilde{\delta}}{n} + \frac{1}{\mu n^3\left(\frac{1-2\widetilde{\delta}}{2n^3\mu} + \frac{1}{2n^2\mu}\left(1-\widetilde{\delta}+\frac{\widetilde{\delta}}{n}\right)\right)}\frac{(1-\widetilde{\delta})^2}{4} \\
&\quad= \frac{L}{n^2\mu}\frac{2-3\widetilde{\delta}}{1-2\widetilde{\delta}} - \frac{1-2\widetilde{\delta}}{n} + \frac{(1-\widetilde{\delta})^2}{2-4\widetilde{\delta}+2n-2n\widetilde{\delta}+2\widetilde{\delta}} \\
&\quad= \frac{L}{n^2\mu}\frac{2-3\widetilde{\delta}}{1-2\widetilde{\delta}} - \frac{1-2\widetilde{\delta}}{n} + \frac{1-\widetilde{\delta}}{2(1+n)} \\
&\quad\leqslant \frac{L}{n^2\mu}\frac{1-3\widetilde{\delta}}{1-2\widetilde{\delta}} - \frac{1-2\widetilde{\delta}}{n} + \frac{1-\widetilde{\delta}}{2n} \\
&\quad= \frac{L}{n^2\mu}\frac{2-3\widetilde{\delta}}{1-2\widetilde{\delta}} - \frac{1-3\widetilde{\delta}}{2n}\,.
\end{aligned}$$

This quantity is negative for $\widetilde{\delta} \leqslant \frac{1}{3}$ and $\frac{\mu}{L} \geqslant \frac{4-6\widetilde{\delta}}{n(1-2\widetilde{\delta})(1-3\widetilde{\delta})}$. If we choose $\widetilde{\delta} = \frac{1}{8}$, then it is sufficient to have $\frac{n\mu}{L} \geqslant 8$.

To finish the proof, we need to prove the positivity of the factor of $\|g'(x^{k-1})\|^2$.

$$\begin{aligned}
\frac{1}{4}\tau_{y,f}^2 n\left(\frac{1}{\tau_{y,I}+\tau_{y,e}} - \frac{1}{\tau_{y,I}}\right) + \frac{1}{4}\frac{\tau_{y,g}^2 n}{\tau_{y,I}+\tau_{y,e}} + \frac{1}{2}\frac{\tau_{y,f}\tau_{y,g}n}{\tau_{y,I}+\tau_{y,e}} &= \frac{n}{4}\frac{1}{\tau_{y,I}+\tau_{y,e}}(\tau_{y,f}+\tau_{y,g})^2 - \frac{n}{4}\frac{\tau_{y,f}^2}{\tau_{y,I}} \\
&\geqslant \frac{n}{4}\frac{(\tau_{y,f}+\tau_{y,g})^2}{\tau_{y,I}} - \frac{n}{4}\frac{\tau_{y,f}^2}{\tau_{y,I}} \\
&= \frac{n}{4\tau_{y,I}}\tau_{y,g}(2\tau_{y,f}+\tau_{y,g}) \\
&\geqslant 0\,.
\end{aligned}$$

Then, following the same argument as in the previous section, we have

$$\begin{aligned}\mathbb{E}Q(\theta^k) &\leqslant \left(1-\frac{1}{8n}\right)^k Q(\theta^0) \\ &= \left(1-\frac{1}{8n}\right)^k \left[2(g(x^0)-g(x^*))+\frac{\sigma^2}{n\mu}\right] ,\end{aligned}$$

with $\sigma^2 = \frac{1}{n}\sum_i \|f_i'(x^*)\|^2$ the variance of the gradients at the optimum.

Step 2 - Domination of $g(x^k)-g(x^*)$ by $Q(\theta^k)$

We now need to prove that $Q(\theta^k)$ dominates $g(x^k)-g(x^*)$.

$$\begin{aligned}Q(\theta^k) &= 2g\left(x^k+\frac{\alpha}{n}e^\top y^k\right) - 2g(x^*) + (\theta^k-\theta^*)^\top \begin{pmatrix} A & b \\ b^\top & c\end{pmatrix}(\theta^k-\theta^*) \\ &= 2g\left(x^k+\frac{\alpha}{n}e^\top y^k\right) - 2g(x^*) + \frac{1}{n^2\mu}\sum_i \left\|y_i^k - f_i'(x^*)\right\|^2 + \frac{n-1}{2n^3\mu}\|e^\top y\|^2 - \frac{1}{n}(x^k-x^*)^\top(e^\top y^k) \\ &\geqslant 2g(x^k) + \frac{2\alpha}{n}g'(x^k)^\top(e^\top y^k) - 2g(x^*) \\ &\quad + \frac{1}{n^2\mu}\sum_i \left\|\frac{1}{n}e^\top y^k + y_i^k - \frac{1}{n}e^\top y^k - f_i'(x^*)\right\|^2 + \frac{n-1}{2n^3\mu}\|e^\top y\|^2 - \frac{1}{n}(x^k-x^*)^\top(e^\top y^k) \\ &\quad \text{using the convexity of } g \text{ and the fact that } \sum_i f_i'(x^*) = 0 \\ &= 2g(x^k) - 2g(x^*) + \left(\frac{2\alpha}{n}g'(x^k) - \frac{1}{n}(x^k-x^*)\right)^\top (e^\top y^k) \\ &\quad + \frac{1}{n^3\mu}\|e^\top y^k\|^2 + \frac{1}{n^2\mu}\sum_i \left\|y_i^k - \frac{1}{n}e^\top y^k - f_i'(x^*)\right\|^2 + \frac{n-1}{2n^3\mu}\|e^\top y\|^2 \\ &\geqslant 2g(x^k) - 2g(x^*) + \left(\frac{2\alpha}{n}g'(x^k) - \frac{1}{n}(x^k-x^*)\right)^\top (e^\top y^k) + \frac{n+1}{2n^3\mu}\|e^\top y\|^2 \\ &\quad \text{by dropping some terms.}\end{aligned}$$

The quantity on the right-hand side is minimized for $e^\top y = \frac{n^3\mu}{n+1}\left(\frac{1}{n}(x^k-x^*) - \frac{2\alpha}{n}g'(x^k)\right)$. Hence, we have

$$
\begin{aligned}
Q(\theta^k) &\geqslant 2g(x^k) - 2g(x^*) - \frac{n^3\mu}{2(n+1)} \left\| \frac{1}{n}(x^k - x^*) - \frac{2\alpha}{n} g'(x^k) \right\|^2 \\
&= 2g(x^k) - 2g(x^*) - \frac{n^3\mu}{2(n+1)} \left(\frac{1}{n^2}\|x^k - x^*\|^2 + \frac{4\alpha^2}{n^2}\|g'(x^k)\|^2 - \frac{4\alpha}{n^2}(x^k - x^*)^\top g'(x^k) \right) \\
&\geqslant 2g(x^k) - 2g(x^*) - \frac{n^3\mu}{2(n+1)} \left(\frac{1}{n^2}\|x^k - x^*\|^2 + \frac{4\alpha^2}{n^2}\|g'(x^k)\|^2 \right) \\
&\quad \text{using the convexity of } g \\
&\geqslant 2g(x^k) - 2g(x^*) - \frac{n\mu}{2(n+1)} \left(1 + \frac{L^2}{\mu^2 n^2} \right) \|x^k - x^*\|^2 \\
&\quad \text{using the Lipschitz continuity of } g' \\
&\geqslant 2g(x^k) - 2g(x^*) - \frac{n\mu}{2(n+1)} \frac{65}{64} \|x^k - x^*\|^2 \text{ since } \frac{\mu}{L} \geqslant \frac{8}{n} \\
&\geqslant 2g(x^k) - 2g(x^*) - \frac{n}{(n+1)} \frac{65}{64} (g(x^k) - g(x^*)) \\
&\geqslant \frac{63}{64}(g(x^k) - g(x^*)) \\
&\geqslant \frac{6}{7}(g(x^k) - g(x^*)) \, .
\end{aligned}
$$

We thus get

$$
\begin{aligned}
\mathbb{E}\left[g(x^k) - g(x^*)\right] &\leqslant 2\mathbb{E}Q(\theta^k) \\
&= \left(1 - \frac{1}{8n}\right)^k \left[\frac{7}{3}(g(x^0) - g(x^*)) + \frac{7\sigma^2}{6n\mu} \right] \, .
\end{aligned}
$$

Step 3 - Initialization of x^0 using stochastic gradient descent

During the first few iterations, we obtain the $O(1/k)$ rate obtained using stochastic gradient descent, but with a constant which is proportional to n. To circumvent this problem, we will first do n iterations of stochastic gradient descent to initialize x^0, which will be renamed x^n to truly reflect the number of iterations done.

Using the bound from section C.3, we have

$$
\mathbb{E}g\left(\frac{1}{n} \sum_{i=0}^{n-1} \tilde{x}^i \right) - g(x^*) \leqslant \frac{2L}{n}\|x^0 - x^*\|^2 + \frac{4\sigma^2}{n\mu} \log\left(1 + \frac{\mu n}{4L}\right) \, .
$$

And so, using $x^n = \frac{1}{n}\sum_{i=0}^{n-1} \tilde{x}^i$, we have for $k \geqslant n$

$$
\mathbb{E}\left[g(x^k) - g(x^*)\right] \leqslant \left(1 - \frac{1}{8n}\right)^{k-n} \left[\frac{14L}{3n}\|x^0 - x^*\|^2 + \frac{28\sigma^2}{3n\mu} \log\left(1 + \frac{\mu n}{4L}\right) + \frac{7\sigma^2}{6n\mu} \right] \, .
$$

Since

$$
\left(1 - \frac{1}{8n}\right)^{-n} \leqslant \frac{8}{7} \, ,
$$

we get

$$
\mathbb{E}\left[g(x^k) - g(x^*)\right] \leqslant \left(1 - \frac{1}{8n}\right)^{k} \left[\frac{16L}{3n}\|x^0 - x^*\|^2 + \frac{32\sigma^2}{3n\mu} \log\left(1 + \frac{\mu n}{4L}\right) + \frac{4\sigma^2}{3n\mu} \right] \, .
$$

References

[1] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*. Springer, 2004.

[2] Y. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *CORE Discussion Paper*, 2010.

[3] S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos: Primal estimated sub-gradient solver for svm. *ICML*, 2007.

[4] D. Blatt, A. O. Hero, and H. Gauchman. A convergent incremental gradient method with a constant step size. *SIAM Journal on Optimization*, 18(1):29–51, 2007.

[5] F. Bach and E. Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *NIPS*, 2011.