A. Appendix

A.1. Proofs

Proof of Lemma 1.

Proof. We start with the convergence rate of SAGA established in (Defazio et al., 2014) as

$$\mathbf{E}_{\mathcal{A}}\left[\|\boldsymbol{w}^t - \boldsymbol{w}^*_{\mathcal{S}}\|^2\right] \leq \rho^t_{|\mathcal{S}|}\left[\|\boldsymbol{w}^0 - \boldsymbol{w}^*_{\mathcal{S}}\|^2 + \frac{|\mathcal{S}|}{\mu|\mathcal{S}| + L}\left(\mathcal{R}_{\mathcal{S}}(\boldsymbol{w}^0) - \langle\nabla\mathcal{R}_{\mathcal{S}}(\boldsymbol{w}^*_{\mathcal{S}}), \boldsymbol{w}^0 - \boldsymbol{w}^*_{\mathcal{S}}\rangle - \mathcal{R}^*_{\mathcal{S}}\right)\right]. \tag{12}$$

We then use the L-smoothness assumption of $f_{\boldsymbol{x}}(\boldsymbol{w})$ to relate the suboptimality on the function values to the bound in Eq. (12).

$$\begin{aligned}\mathbf{E}_{\mathcal{A}}\left[|\mathcal{R}_{\mathcal{S}}(\boldsymbol{w}^t) - \mathcal{R}_{\mathcal{S}}(\boldsymbol{w}^*_{\mathcal{S}})|\right] &= \mathbf{E}_{\mathcal{A}}\left[|\mathbf{E}_{\boldsymbol{x}\in\mathcal{S}}\left[f_{\boldsymbol{x}}(\boldsymbol{w}^t)\right] - \mathbf{E}_{\boldsymbol{x}\in\mathcal{S}}\left[f_{\boldsymbol{x}}(\boldsymbol{w}^*_{\mathcal{S}})\right]|\right] \\ &\overset{L-\text{smoothness}}{\leq} L\mathbf{E}_{\mathcal{A}}\left[\|\boldsymbol{w}^t - \boldsymbol{w}^*_{\mathcal{S}}\|^2\right] \\ &\overset{Eq.\ 12}{\leq} \rho^t_{|\mathcal{S}|}C_{\mathcal{S}},\end{aligned}$$

where $C_{\mathcal{S}}$ is the initial suboptimality on the empirical risk defined as:

$$C_{\mathcal{S}} = L\left[\|\boldsymbol{w}^0 - \boldsymbol{w}^*_{\mathcal{S}}\|^2 + \frac{|\mathcal{S}|}{\mu|\mathcal{S}| + L}\left(\mathcal{R}_{\mathcal{S}}(\boldsymbol{w}^0) - \langle\nabla\mathcal{R}_{\mathcal{S}}(\boldsymbol{w}^*_{\mathcal{S}}), \boldsymbol{w}^0 - \boldsymbol{w}^*_{\mathcal{S}}\rangle - \mathcal{R}^*_{\mathcal{S}}\right)\right]$$

Note that this initial error depends on the set $\mathcal{S}$ and its size $|\mathcal{S}|$. In the following Lemma, we propose an upper bound on this initial error that is independent of $\mathcal{S}$ □

Lemma 8. *W.h.p, the initial suboptimality error of sample $\mathcal{S}$ is bounded by:*

$$C_{\mathcal{S}} \leq \xi := \frac{4L}{\mu}\left[\mathcal{R}(\boldsymbol{w}^0) - \mathcal{R}(\boldsymbol{w}^*)\right]$$

Proof. We first use the fact that $\mathcal{R}_{\mathcal{S}}(\boldsymbol{w})$ is μ-strongly convex as well as the optimality of $\boldsymbol{w}^*_{\mathcal{S}}$ to bound $C_{\mathcal{S}}$ as

$$\begin{aligned}C_{\mathcal{S}} &:= L\left(\|\boldsymbol{w}^0 - \boldsymbol{w}^*_{\mathcal{S}}\|^2 + \frac{|\mathcal{S}|}{\mu|\mathcal{S}| + L}\left[\mathcal{R}_{\mathcal{S}}(\boldsymbol{w}^0) - \langle\nabla\mathcal{R}_{\mathcal{S}}(\boldsymbol{w}^*_{\mathcal{S}}), \boldsymbol{w}^0 - \boldsymbol{w}^*_{\mathcal{S}}\rangle - \mathcal{R}_{\mathcal{S}}(\boldsymbol{w}^*_{\mathcal{S}})\right]\right) \\ &\leq \frac{L}{\mu}\left[\mathcal{R}_{\mathcal{S}}(\boldsymbol{w}^0) - \mathcal{R}_{\mathcal{S}}(\boldsymbol{w}^*_{\mathcal{S}})\right] + \frac{|\mathcal{S}|L}{\mu|\mathcal{S}| + L}\left[\mathcal{R}_{\mathcal{S}}(\boldsymbol{w}^0) - \langle\nabla\mathcal{R}_{\mathcal{S}}(\boldsymbol{w}^*_{\mathcal{S}}), \boldsymbol{w}^0 - \boldsymbol{w}^*_{\mathcal{S}}\rangle - \mathcal{R}_{\mathcal{S}}(\boldsymbol{w}^*_{\mathcal{S}})\right] \\ &\leq \frac{L}{\mu}\left[\mathcal{R}_{\mathcal{S}}(\boldsymbol{w}^0) - \mathcal{R}_{\mathcal{S}}(\boldsymbol{w}^*_{\mathcal{S}})\right] + \frac{|\mathcal{S}|L}{\mu|\mathcal{S}| + L}\left[\mathcal{R}_{\mathcal{S}}(\boldsymbol{w}^0) - \mathcal{R}_{\mathcal{S}}(\boldsymbol{w}^*_{\mathcal{S}})\right] \\ &\overset{(L>0)}{\leq} \frac{2L}{\mu}\left[\mathcal{R}_{\mathcal{S}}(\boldsymbol{w}^0) - \mathcal{R}_{\mathcal{S}}(\boldsymbol{w}^*_{\mathcal{S}})\right] \\ &\leq \frac{2L}{\mu}\left[\mathcal{R}_{\mathcal{S}}(\boldsymbol{w}^0) \overset{[1]}{\mp} \mathcal{R}(\boldsymbol{w}^0) \overset{[2]}{\mp} \mathcal{R}(\boldsymbol{w}^*) \overset{[3]}{\mp} \mathcal{R}(\boldsymbol{w}^*_{\mathcal{S}}) - \mathcal{R}_{\mathcal{S}}(\boldsymbol{w}^*_{\mathcal{S}})\right]\end{aligned}$$

We use the generalization bounds in (Vapnik, 1998) to upper bound [1] and [2]. For [3], we used the uniform convergence rate of the ERM that implies (Vapnik, 1998):

$$\mathcal{R}(\boldsymbol{w}^*_{\mathcal{S}}) - \mathcal{R}(\boldsymbol{w}^*) \leq c\sup_{\boldsymbol{w}}|\mathcal{R}_{\mathcal{S}}(\boldsymbol{w}) - \mathcal{R}(\boldsymbol{w})|,$$

where c is a constant. We then get

$$C_{\mathcal{S}} \overset{\text{w.h.p}}{\leq} \frac{2L}{\mu}\left[\mathcal{H}(|\mathcal{S}|) + \mathcal{R}(\boldsymbol{w}^0) - \mathcal{R}(\boldsymbol{w}^*) + c\mathcal{H}(|\mathcal{S}|) + \mathcal{H}(|\mathcal{S}|)\right]. \tag{13}$$

We also make the further assumption that with high probability the initial suboptimality is greater than a constant factor of the statistical accuracy, i.e. $\mathcal{R}(\boldsymbol{w}^0) - \mathcal{R}(\boldsymbol{w}^*) > (2+c)\mathcal{H}(|\mathcal{S}|)$. We can then further upper bound $C_{\mathcal{S}}$ as

$$C_{\mathcal{S}} \leq \frac{4L}{\mu}\left[\mathcal{R}(\boldsymbol{w}^0) - \mathcal{R}(\boldsymbol{w}^*)\right]. \tag{14}$$

□

Lemma 9 (for Proposition 2)**.**

$$V(m) := \frac{D}{m} + Ce^{-\frac{n}{m}}, \textit{ then } \operatorname*{arg\,min}_{0<m\leq n} V(m) = \frac{n}{\log\frac{nC}{D}}$$

Proof.

$$\begin{aligned} \frac{dV}{dm^{-1}} &= D - nCe^{-\frac{n}{m}} \stackrel{!}{=} 0 \\ \iff e^{-\frac{n}{m}} &= \frac{D}{nC} \\ \iff \frac{n}{m} &= \log\frac{nC}{D} \end{aligned}$$

Solving for m, this indeed corresponds to a minimum which can be verified by checking the boundary values $m = n$ and $m \to 0$. □

Lemma 10 (for Theorem 3)**.**

$$\mathbf{E}_{\mathcal{S}|\mathcal{T}}\left[\mathcal{R}_{\mathcal{S}}(\boldsymbol{w}) - \mathcal{R}_{\mathcal{T}}(\boldsymbol{w})\right] \leq \frac{n-m}{n}\left|\mathcal{R}(\boldsymbol{w}) - \mathcal{R}_{\mathcal{T}}(\boldsymbol{w})\right|.$$

Proof.

$$\begin{aligned} \mathbf{E}_{\mathcal{S}|\mathcal{T}}\left[\mathcal{R}_{\mathcal{S}}(\boldsymbol{w}) - \mathcal{R}_{\mathcal{T}}(\boldsymbol{w})\right] &= \mathbf{E}_{\mathcal{S}-\mathcal{T}|\mathcal{T}}\left[\mathcal{R}_{\mathcal{S}}(\boldsymbol{w}) - \mathcal{R}_{\mathcal{T}}(\boldsymbol{w})\right] \\ &= \mathbf{E}_{\mathcal{S}-\mathcal{T}}\left[\frac{1}{n}\left[\sum_{\boldsymbol{x}\in\mathcal{T}} f_{\boldsymbol{x}}(\boldsymbol{w}) + \sum_{\boldsymbol{y}\in\mathcal{S}-\mathcal{T}} f_{\boldsymbol{y}}(\boldsymbol{w})\right] - \frac{1}{m}\sum_{\boldsymbol{x}\in\mathcal{T}} f_{\boldsymbol{x}}(\boldsymbol{w})\right] \\ &= \frac{n-m}{n}\mathbf{E}_{\mathcal{S}-\mathcal{T}}\left[\frac{1}{n-m}\sum_{\boldsymbol{y}\in\mathcal{S}-\mathcal{T}} f_{\boldsymbol{y}}(\boldsymbol{w}) - \mathcal{R}_{\mathcal{T}}(\boldsymbol{w})\right] \\ &= \frac{n-m}{n}\mathbf{E}_{\mathcal{S}-\mathcal{T}}\left[\frac{1}{n-m}\sum_{\boldsymbol{y}\in\mathcal{S}-\mathcal{T}} f_{\boldsymbol{y}}(\boldsymbol{w}) - \mathcal{R}_{\mathcal{T}}(\boldsymbol{w})\right] \\ &= \frac{n-m}{n}\left[\mathbf{E}_{\mathcal{S}-\mathcal{T}}\left[\mathcal{R}_{\mathcal{S}-\mathcal{T}}(\boldsymbol{w})\right] - \mathcal{R}_{\mathcal{T}}(\boldsymbol{w})\right] \\ &= \frac{n-m}{n}\left[\mathcal{R}(\boldsymbol{w}) - \mathcal{R}_{\mathcal{T}}(\boldsymbol{w})\right] \end{aligned}$$

□

A.2. Optimality of the LINEAR Strategy

We here introduce a new notation and chose to represent a sample size schedule by a vector $\boldsymbol{t}^n = \langle t_m \rangle, m < n$ where t_m denotes the number of iterations on sample size m. Note that the total number of iterations up to the sample size n is $T = \sum_{m<n} t_m$. We define n^- as the sample size that we iterate on immediately before sample size n, i.e.

$$n^- = \max\{k < n : t_k > 0\}. \tag{15}$$

We now rewrite the suboptimality bound in terms of the sample size schedule t_n as

$$\begin{aligned} A(\boldsymbol{t}^n) &= \mathbf{E}_{\mathcal{S}}\left[\mathcal{R}_{\mathcal{S}}(\boldsymbol{w}(\boldsymbol{t}^n)) - \mathcal{R}_{\mathcal{S}}(\boldsymbol{w}^*)\right] \\ &= \rho_n^{t_n}\left(A(\boldsymbol{t}^{n^-}) + \frac{n-n^-}{n}\mathcal{H}(n^-)\right), \end{aligned} \tag{16}$$

where the second equality is derived using Lemma 1 and Theorem 3.

One can relate the upper bound $\mathbf{U}(n,n)$ to $A(\boldsymbol{t}^n)$ using the following constrained program:

$$\begin{aligned} \mathbf{U}(n,n) &= \min_{\boldsymbol{t}^n} A(\boldsymbol{t}^n) \\ \text{Subject to } \forall m \leq n &: -t_m \leq 0 \\ \sum_{m\leq n} t_m &= n \end{aligned} \tag{17}$$

In the following we aim at showing that the LINEAR Strategy is the optimal solution of Equation 16. We first prove a Lemma that will be used in the rest of our analysis.

Lemma 11 (Expansion of $A(\boldsymbol{t}^n)$)**.** *if $\mathcal{H}(n) = D/n$, then*

$$A(\boldsymbol{t}^n) := C(\boldsymbol{t}^n) + \sum_{m=m_0+1}^{n} B_m(\boldsymbol{t}^n), \quad \textit{where} \tag{18}$$

$$C(\boldsymbol{t}^n) := \xi \prod_{i=m_0}^{n}\left(\frac{i-1}{i}\right)^{t_i}, \quad B_m(\boldsymbol{t}^n) := \frac{D}{(m-1)m}\prod_{i=m}^{n}\left(\frac{i-1}{i}\right)^{t_i}. \tag{19}$$

Proof. Although one could painstakingly unroll the recursivity in Equation 16, we here provide a simple induction proof. First, one can easily verify that the equation holds for $n = m_0$. For the inductive step, we assume it holds for n^- and prove it holds for all $\{k : n^- < k \leq n\}$. According to the definition of n^-, we have $t_k = 0$ for all $n^- < k < n$, and therefore

$$\rho_k^{t_k} = \prod_{m=n^-+1}^{k} \rho_m^{t_m}. \tag{20}$$

We will also make use of the following equality in our analysis:

$$\frac{k-n^-}{k}\mathcal{H}(n^-) = \mathcal{H}(n^-) - \mathcal{H}(k) \overset{(\mathcal{H}(n)=D/n)}{=} \sum_{m=n^-+1}^{k} \mathcal{H}(m-1) - \mathcal{H}(m). \tag{21}$$

We are now ready to prove the inductive step.

$$A(\boldsymbol{t}^k) \overset{\text{EQ 16}}{=} \rho_k^{t_k} \left(A(\boldsymbol{t}^{n^-}) + \frac{k-n^-}{k} \mathcal{H}(n^-) \right) \tag{22}$$

$$= \rho_k^{t_k} \left(C(\boldsymbol{t}^{n^-}) + \sum_{m=m_0+1}^{n^-} B_m(\boldsymbol{t}^{n^-}) + \frac{k-n^-}{k} \mathcal{H}(n^-) \right) \tag{23}$$

$$\overset{\text{EQ 19, 20}}{=} C(\boldsymbol{t}^k) + \sum_{m=m_0+1}^{n^-} B_m(\boldsymbol{t}^k) + \rho_k^{t_k} \left(\frac{k-n^-}{k} \mathcal{H}(n^-) \right) \tag{24}$$

$$\overset{\text{EQ 21}}{=} C(\boldsymbol{t}^k) + \sum_{m=m_0+1}^{n^-} B_m(\boldsymbol{t}^k) + \rho_k^{t_k} \sum_{m=n^-+1}^{k} \frac{D}{(m-1)m} \tag{25}$$

$$\overset{\text{EQ 20}}{=} C(\boldsymbol{t}^k) + \sum_{m=m_0+1}^{n^-} B_m(\boldsymbol{t}^k) + \sum_{m=n^-+1}^{k} B_m(\boldsymbol{t}^k) \tag{26}$$

$$= C(\boldsymbol{t}^k) + \sum_{m=m_0+1}^{k} B_m(\boldsymbol{t}^k) \tag{27}$$

□

Using the definitions provided in Lemma 11, we investigate the optimality conditions of the optimal sample size strategy. In the following, we simplify our notations and write B_m and C instead of $B_m(\boldsymbol{t}^n)$ and $C(\boldsymbol{t}^n)$.

As a first step in our analysis, we introduce the following equations based on the definitions of B_m and C.

$$B_m = \frac{1}{m(m-1)} \prod_{i \geq m} \left(\frac{i-1}{i} \right)^{t_i} = \frac{m+1}{m-1} \left(\frac{m-1}{m} \right)^{t_m} B_{m+1} \,. \tag{28}$$

$$\prod_{i=m}^{n} \left(\frac{i-1}{i} \right)^{t_i} = \prod_{i=m}^{n} \exp\left(\log\left(\left(\frac{i-1}{i} \right)^{t_i} \right) \right) = \exp\left[\sum_{i=m}^{n} t_i \log\left(1 - \frac{1}{i} \right) \right] . \tag{29}$$

We now compute the derivative of $A(\boldsymbol{t}_*^n)$ as

$$\frac{\partial A(\boldsymbol{t}_*^n)}{\partial t_m} = \log(1 - \frac{1}{m}) \left(C(\boldsymbol{t}^n) + \sum_{k=m_0+1}^{m} B_k(\boldsymbol{t}^n) \right)$$

$$\simeq -\frac{1}{m} \left(C + \sum_{k=m_0+1}^{m} B_k \right) . \tag{30}$$

$C(\boldsymbol{t}^n)$ and $B_m(\boldsymbol{t}^n)$ are *log-convex* (hence *convex*) functions with respect to $\boldsymbol{t}^n$. Since the sum operator preserves convexity (Boyd & Vandenberghe, 2004), $A(\boldsymbol{t}_n)$ is *convex* as well. Let λ_i, ν denote the Lagrangian coefficients associated with the inequality and equality constraints respectively. According the KKT conditions (Boyd & Vandenberghe, 2004) for the the optimal solution, the following inequalities hold:

$$\lambda_m \geq 0 \tag{31}$$

$$-\lambda_m t_m^* = 0 \tag{32}$$

$$\frac{\partial A(\boldsymbol{t}_*^n)}{\partial t_m} - \lambda_m + \nu = 0 \tag{33}$$

According the above condition there are two possible cases for the partial derivative $\frac{\partial A(\boldsymbol{t}_*^n)}{\partial t_m}$:

- For the case of $t_m^* > 0$, the slackness condition 32 implies that $\lambda_m = 0$. Then, according to the condition 33:

$$\begin{aligned} \frac{\partial A(\boldsymbol{t}_*^n)}{\partial t_m} &= -\nu \\ \overset{\text{EQ. 30}}{\Longrightarrow} \frac{1}{m}\left(C + \sum_{k=m_0+1}^{m} B_k\right) &= \nu \end{aligned} \tag{34}$$

- For the case of $t_m^* = 0$, $\lambda_i > 0 (a.)$ holds based on the complementary slackness condition 32.

$$\begin{aligned} \frac{\partial A(\boldsymbol{t}_*^n)}{\partial t_m} &= \lambda_i - \nu \overset{(a.)}{>} -\nu \\ \overset{\text{EQ. 30}}{\Longrightarrow} \frac{1}{m}\left(C + \sum_{k=m_0+1}^{m} B_k\right) &< \nu \end{aligned} \tag{35}$$

In the following two lemmas we use the conditions of optimality derived in Equations 34 and 35 to prove optimality of the LINEAR Strategy. Specifically, we first prove that for the optimal strategy, $t_m > 0$ for $m_0 < m \leq n^-$ and $t_m = 0$ for $m > n^-$. We also prove the optimality of incrementing the sample size by one. In the second lemma, we show that $t_m^* \simeq 2$.

Lemma 12 (Optimality of sample size increment)**.** *For large enough m, a schedule with $t_m = 0$ and $t_{m+1} > 0$ cannot be optimal.*

Proof. Note that by repeated application of Equation (28) we obtain

$$B_{m+1} < B_m < \cdots < B_{m^-+1} \overset{\text{EQ. 34 \& 35}}{<} \nu \tag{36}$$

where optimality conditions *a.* $t_{m^-} > 0$ (EQ.34) and *b.* $t_{m^-+1} = 0$ (EQ.35) yeild the last inequality:

$$B_{m^-+1} = \sum_{k=m_0+1}^{m^-+1} B_k - \sum_{k=m_0+1}^{m^-} B_k \mp C \tag{37}$$

$$\overset{a.}{=} \sum_{k=m_0+1}^{m^-+1} B_k + C - m\nu \tag{38}$$

$$\overset{b.}{<} (m+1)\nu - m\nu = \nu \tag{39}$$

On the other hand, optimality of *a.* $t_{m+1} > 0$ (EQ.34) and *b.* $t_m = 0$ (EQ.35) also imply $B_{m+1} > \nu$ which is in contradiction with the previously established $B_{m+1} < \nu$. Indeed, we have

$$B_{m+1} = \sum_{k=m_0+1}^{m+1} B_k - \sum_{k=m_0+1}^{m} B_k \mp C \tag{40}$$

$$\overset{a.}{=} (m+1)\nu - \sum_{k=m_0+1}^{m} B_k - C \tag{41}$$

$$\overset{b.}{>} (m+1)\nu - m\nu = \nu \tag{42}$$

□

Lemma 13 (Optimality of two iterations)**.** *Consider $\boldsymbol{t}_*^n$ as the minimizer of the optimization problem 17. For sufficiently large $m : m_0 < m \leq n^-$, $t_m^* \simeq 2$.*

Proof. Using Lemma 12, $t_m^* > 0$ holds for $m_0 < m \leq n^-$. We proceed with optimality conditions *a.* $t_m^* > 0$ and *b.* $t_{m-1}^* > 0$ in equation 34.

$$B_m = \sum_{k=m_0+1}^{m} B_k - \sum_{k=m_0+1}^{m-1} B_k \mp C \tag{43}$$

$$\overset{a.}{=} m\nu - \sum_{k=m_0+1}^{m-1} B_k - C \tag{44}$$

$$\overset{b.}{=} m\nu - (m-1)\nu = \nu \tag{45}$$

Consequently, $B_m = B_{m+1} = \nu$. Using Equation 28, one conclude that $t_m^* \simeq 2$:

$$\frac{m-1}{m+1} = \left(\frac{m-1}{m}\right)^{t_m^*} \iff t_m^* = \frac{\log\left(1 - \frac{2}{m+1}\right)}{\log\left(1 - \frac{1}{m}\right)} \simeq \frac{2m}{m+1} \simeq 2\,. \tag{46}$$

□

A.3. Additional Experimental results

A.3.1. Comparison of the two adaptive sample size schemes for dynaSAGA

We here compare the LINEAR and ALTERNATING schemes on the collection of real datasets presented in Table 2 for a regularizer $\lambda = n^{-\frac{1}{2}}$. The results for the empirical and expected risk shown in Figure 6 and Figure 7 show that the ALTERNATING scheme slightly outperforms the LINEAR strategy.

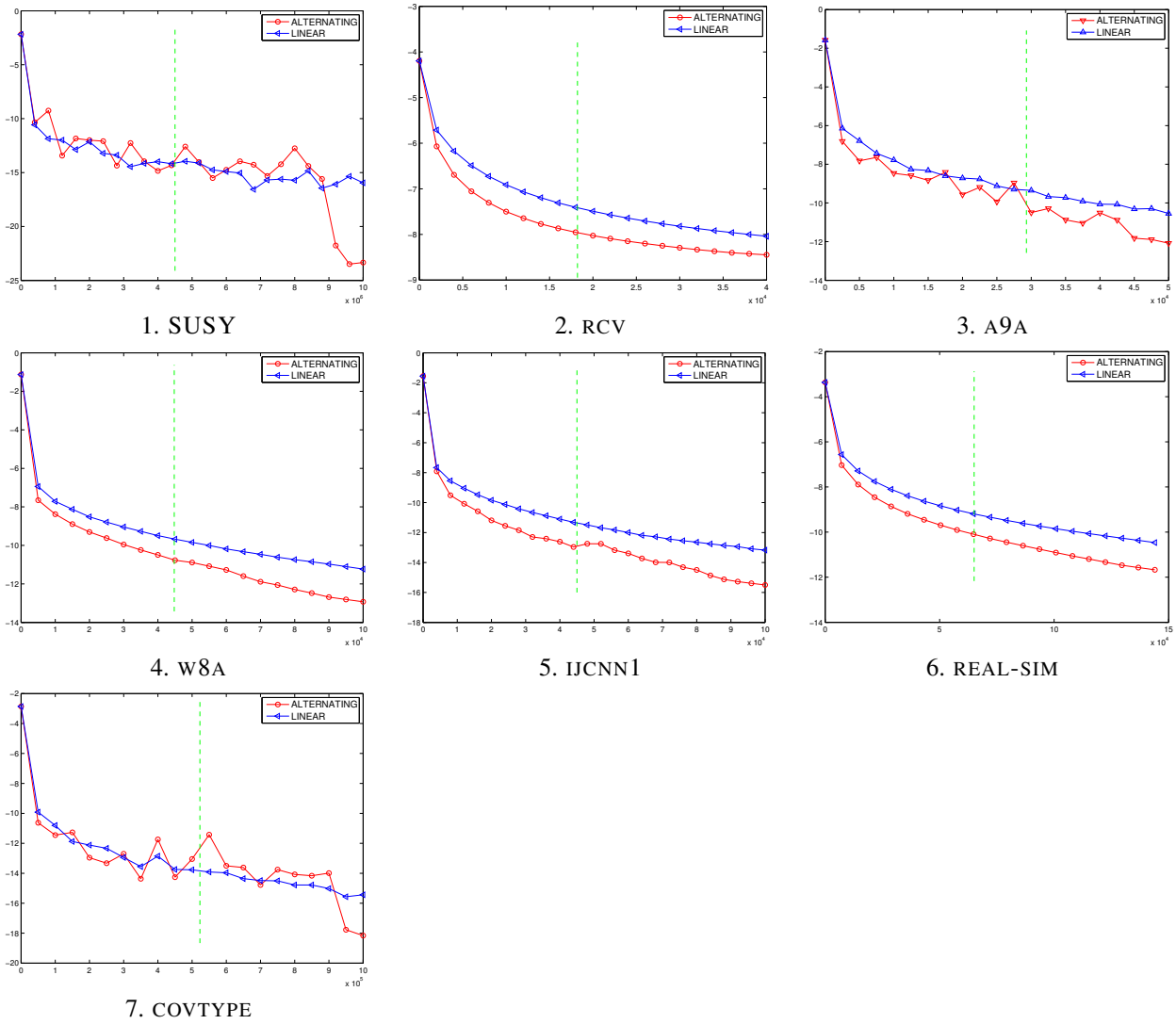
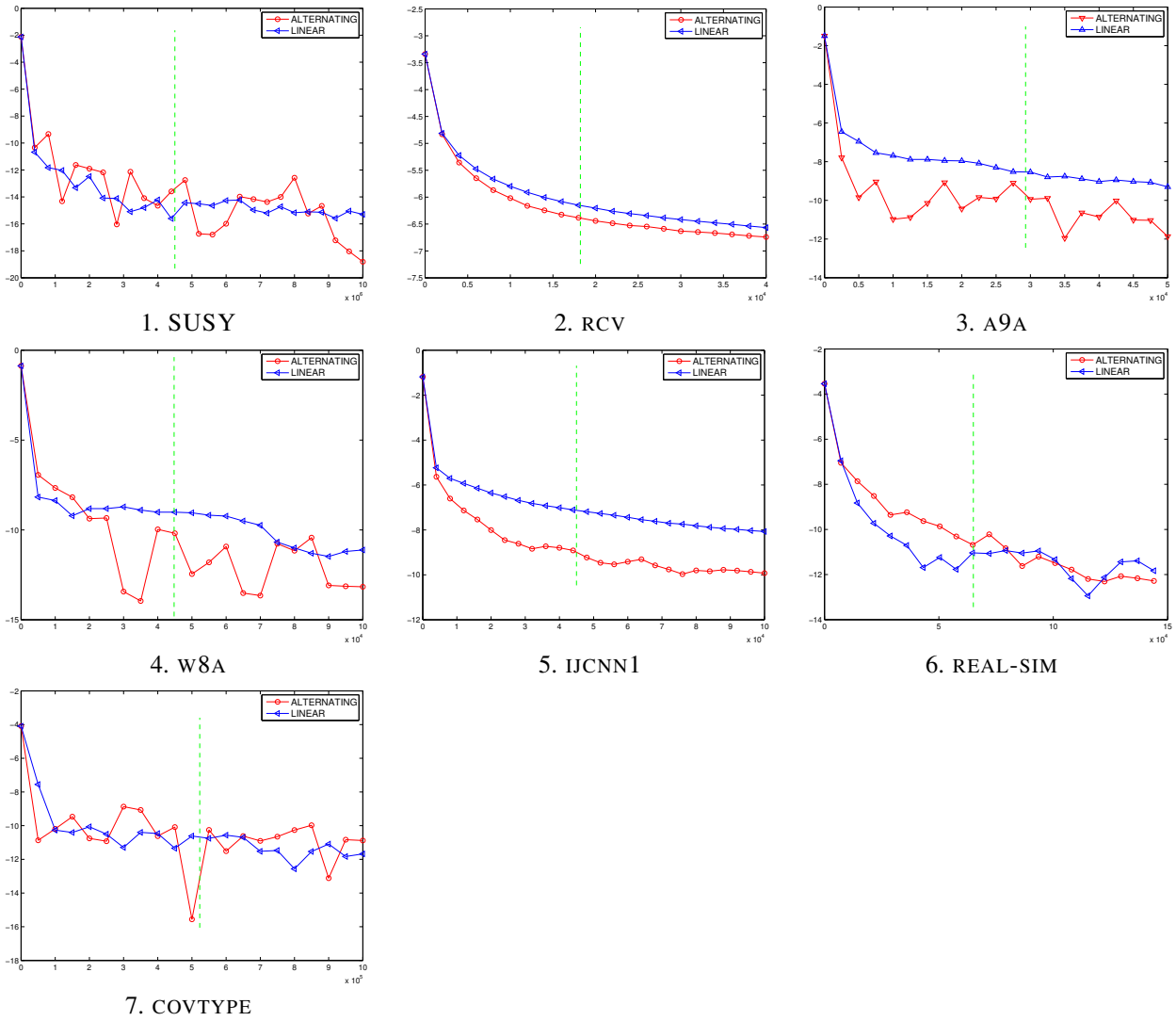


Figure 6. Suboptimality on the empirical risk. The vertical axis shows the suboptimality of the empirical risk, i.e. $\log_2 \mathbf{E}_{10}\left[\mathcal{R}_{\mathcal{T}}(\boldsymbol{w}^t) - \mathcal{R}^*_{\mathcal{T}}\right]$ where the expectation is taken over 10 independent runs. The training set includes 90% of the data. The vertical green dashed line is drawn after exactly one epoch over the data.



1. SUSY 2. RCV 3. A9A

4. W8A 5. IJCNN1 6. REAL-SIM

7. COVTYPE

Figure 7. Suboptimality on the expected risk. The vertical axis shows the suboptimality of the expected risk, i.e. $\log_2 \mathbf{E}_{10}\left[\mathcal{R}_{\mathcal{S}}(\boldsymbol{w}^t) - \mathcal{R}_{\mathcal{S}}(\boldsymbol{w}^*_{\mathcal{T}})\right]$, where $\mathcal{S}$ is a test set which includes 10% of the data and $\boldsymbol{w}^*_{\mathcal{T}}$ is the optimum of the empirical risk on $\mathcal{T}$. The vertical green dashed line is drawn after exactly one epoch over the data.

A.3.2. Effect of the regularizer

We here present additional results for various regularizers of the form $\lambda = \frac{1}{n^p}, p < 1$. In the interest of clarity we only show results on four datasets. We can see a similar trend to the main results presented in the paper for $\lambda = \frac{1}{\sqrt{n}}$ where DynaSAGAshows very fast convergence in terms of both empirical and expected risk. SGD is also very competitive and typically achieves faster convergence than the other baselines, however, its behaviour is not stable throughout all the datasets.

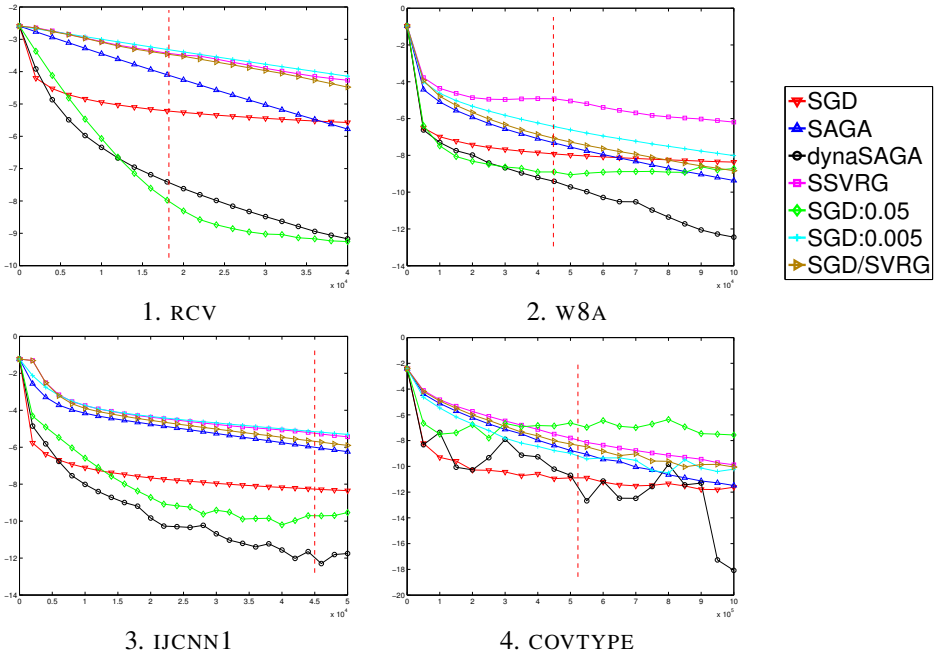


Figure 8. Suboptimality on the empirical risk with regularizer $\lambda = n^{-\frac{2}{3}}$

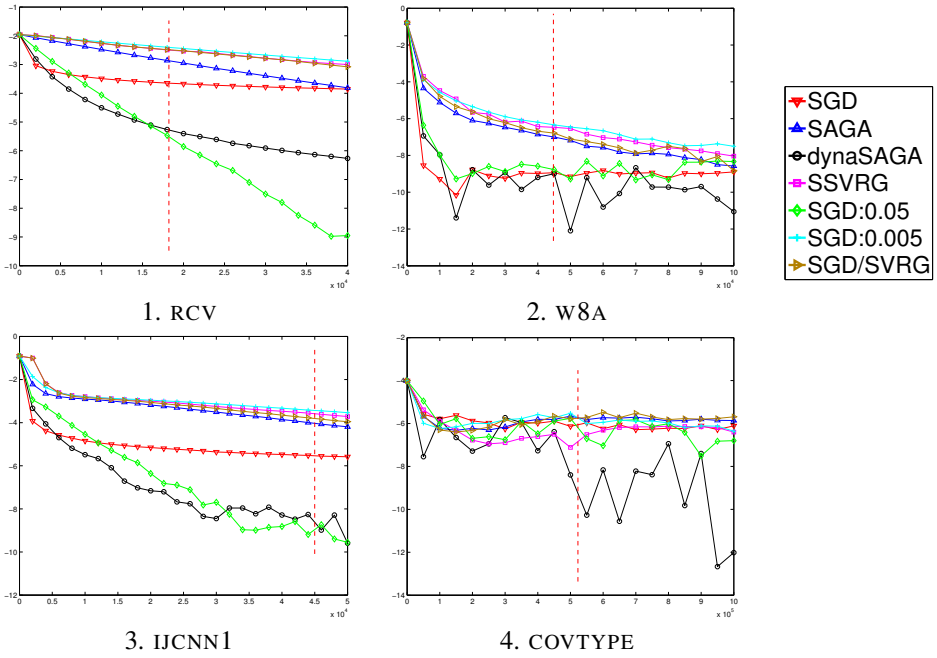


Figure 9. Suboptimality on the expected risk with regularizer $\lambda = n^{-\frac{2}{3}}$

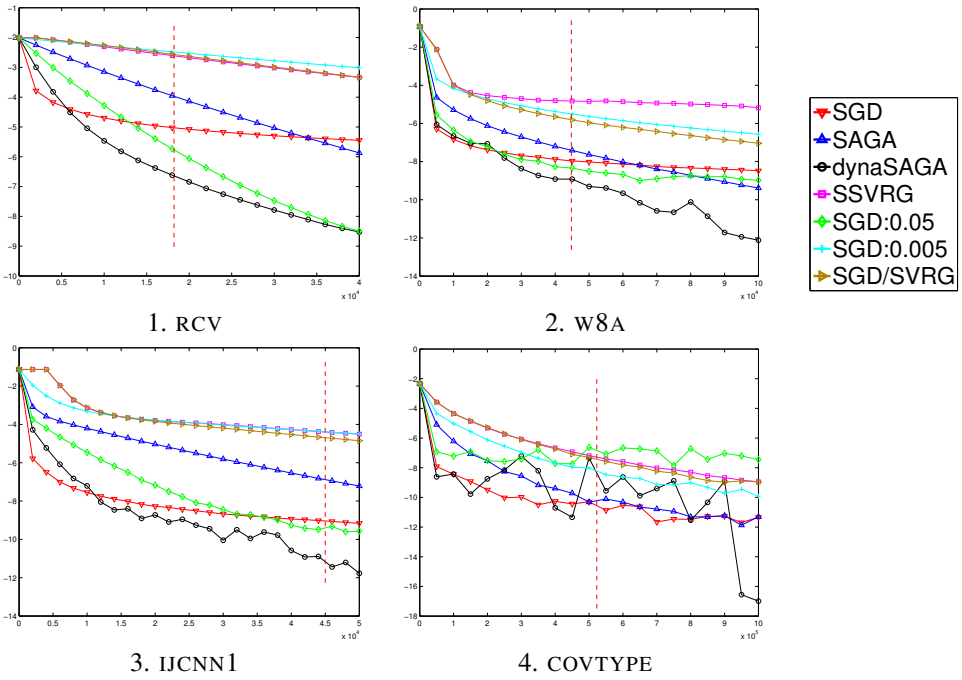


Figure 10. Suboptimality on the empirical risk with regularizer $\lambda = n^{-\frac{3}{4}}$

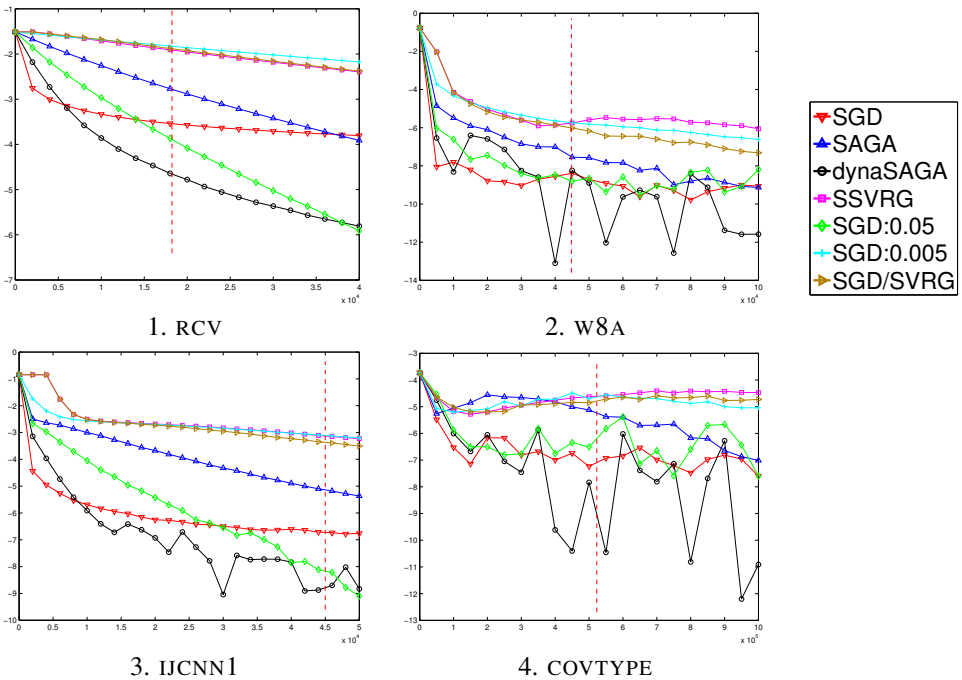


Figure 11. Suboptimality on the expected risk with regularizer $\lambda = n^{-\frac{3}{4}}$

A.4. Details of Experiments

The various parameters of all baselines and DYNASAGA are represented in Table 3.

Table 3. Experimental setting

METHOD	PARAMETER	NOTATION	VALUE
SGD	STEP SIZE	η_t	$\frac{0.1}{0.1+\mu t}$
SAGA	STEP SIZE	η	$\frac{0.3}{L+\mu n}$
SSVRG AND SGD/SVRG	FACTOR FOR INCREASING SAMPLE SIZE	b	3
	A CONSTANT PARAMETER	p	2
	STEP SIZE	η	$\frac{1}{10b^p}$
	INITIAL BATCH SIZE	k_0	κ
	NUMBER OF STEPS ON EACH BATCH SIZE	m	$\frac{\kappa}{\eta}$
SGD:0.05	STEP SIZE	η	0.05
SGD:0.005	STEP SIZE	η	0.005
DYNASAGA	STEP SIZE FOR SAMPLE SIZE m	$\eta(m)$	$\frac{0.3}{L+\mu m}$
	INITIAL BATCH SIZE	k_0	κ
	NUMBER OF ITERATIONS ON SAMPLE SIZE m	$t(m)$	2