Non-Uniform Stochastic Average Gradient Method for Training Conditional Random Fields: Supplementary Material

Mark Schmidt, Reza Babanezhad, Mohamed Osama Ahemd,
Aaron Defazio, Ann Clifton, Anoop Sarkar

Abstract

In this supplementary material we provide the proofs of both parts of the the propositions as well as extended experimental results.

Proof of Part (a) of Proposition 1

In this section we consider the minimization problem

$$\min_x f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x), \tag{1}$$

where each f_i' is L-Lipschitz continuous and each f_i is μ-strongly-convex. We will define Algorithm 1 by the sequences $\{x^k\}$, $\{\nu_k\}$, and $\{\phi_j^k\}$ given by

$$\begin{aligned}
\nu_k &= \frac{1}{np_j}[f'_{j_k}(x^k) - f'_{j_k}(\phi_j^k)] + \frac{1}{n}\sum_{i=1}^n f'_i(\phi_i^k),\\
x^{k+1} &= x^k - \frac{1}{\eta}\nu_k,\\
\phi_j^{k+1} &= \begin{cases} f'_{j_k}(x^k) & \text{if } j = j_k,\\ \phi_j^k & \text{otherwise,}\end{cases}
\end{aligned}$$

where $j_k = j$ with probability p_j. In this section we'll use the convention that $x = x^k$, that $\phi_j = \phi_j^k$, and that x^* is the minimizer of f. We first show that ν_k is an unbiased gradient estimator and derive a bound on its variance.

Lemma 1. *We have $\mathbb{E}[\nu_k] = f'(x^k)$ and subsequently*

$$\mathbb{E}\|\nu_k\|^2 \leq 2\mathbb{E}\|\frac{1}{np_j}[f'_j(x) - f'_j(x^*)]\|^2 + 2\mathbb{E}\|\frac{1}{np_j}[f'_j(\phi_j) - f'_j(x^*)]\|^2.$$

Proof. We have

$$\begin{aligned}
\mathbb{E}[\nu_k] &= \sum_{j=1}^n p_{j=1}\left[\frac{1}{np_j}[f'_j(x) - f'_{j_k}(\phi_j)] + \frac{1}{n}\sum_{i=1}^n f'_i(\phi_i)\right]\\
&= \sum_{j=1}^n \left[\frac{1}{n}f'_j(x) - \frac{1}{n}f'_j(\phi_j) + \frac{p_j}{n}\sum_{i=1}^n f'_i(\phi_i)\right]\\
&= \frac{1}{n}\sum_{i=1}^n f'_j(x) - \frac{1}{n}\sum_{i=1}^n f'_j(\phi_j) + \sum_{i=1}^n [p_i]\frac{1}{n}\sum_{i=1}^n f'_j(\phi_j)\\
&= \frac{1}{n}\sum_{i=1}^n f'_i(x) = f'(x).
\end{aligned}$$

To show the second part, we use that $\mathbb{E}\|X - \mathbb{E}[X] + Y\|^2 = \mathbb{E}\|X - \mathbb{E}[X]\|^2 + \mathbb{E}\|Y\|^2$ if X and Y are independent, $\mathbb{E}\|X - \mathbb{E}[X]\|^2 \leq \mathbb{E}\|X\|^2$, and $\|x + y\|^2 \leq 2\|x\|^2 + 2\|y\|^2$,

$$
\begin{aligned}
\mathbb{E}\|\nu_k\|^2 &= \mathbb{E}\|\frac{1}{np_j}[f_j'(x) - f_j'(\phi_j)] + \frac{1}{n}\sum_{i=1}^n f_i'(\phi_i)\|^2 \\
&= \mathbb{E}\|\frac{1}{np_j}[f_j'(x) - f_j'(x^*)] - f'(x) + f'(x) - \frac{1}{np_j}[f_j'(\phi_j) - f_j'(x^*)] - \frac{1}{n}\sum_{i=1}^n f_i'(\phi_i))\|^2 \\
&= \mathbb{E}\|\frac{1}{np_j}[f_j'(x) - f_j'(x^*)] - f'(x) - \frac{1}{np_j}[f_j'(\phi_j) - f_j'(x^*)] - \frac{1}{n}\sum_{i=1}^n f_i'(\phi_i))\|^2 + \|f'(x)\|^2 \\
&\leq \mathbb{E}\|\frac{1}{np_j}[f_j'(x) - f_j'(x^*)] - f'(x)\|^2 + 2\mathbb{E}\|\frac{1}{np_j}[f_j'(\phi_j) - f_j'(x^*)] - \frac{1}{n}\sum_{i=1}^n f_i'(\phi_i))\|^2 + \|f'(x)\|^2 \\
&\leq 2\mathbb{E}\|\frac{1}{np_j}[f_j'(x) - f_j'(x^*)]\|^2 - 2\|f'(x)\|^2 + 2\mathbb{E}\|\frac{1}{np_j}[f_j'(\phi_j) - f_j'(x^*)]\|^2 + \|f'(x)\|^2 \\
&\leq 2\mathbb{E}\|\frac{1}{np_j}[f_j'(x) - f_j'(x^*)]\|^2 + 2\mathbb{E}\|\frac{1}{np_j}[f_j'(\phi_j) - f_j'(x^*)]\|^2.
\end{aligned}
$$

□

We will also make use of the inequality

$$
\langle f'(x), x^* - x\rangle \leq -\frac{\mu}{2}\|x - x^*\|^2 - \frac{1}{2Ln}\sum_{i=1}^n \|f_i'(x^*) - f_i'(x)\|^2, \tag{2}
$$

which follows from Defazio et al. [2014, Lemma 1] using that $f'(x^*) = 0$ and the non-positivity of $\frac{L-\mu}{L}[f(x^*) - f(x)]$. We now give the proof of part (a) of Proposition 1, which we state below.

Proposition 1 (a). *If $\eta = \frac{4L+n\mu}{np_m}$ and $p_m = \min_j\{p_j\}$, then Algorithm 1 has*

$$
\mathbb{E}[\|x^k - x^*\|^2] \leq \left(1 - \frac{np_m\mu}{n\mu + 4L}\right)^t \left[\|x^0 - x^*\| + \frac{2p_m}{(4L + n\mu)^2}\sum_i \frac{1}{p_i}\|\nabla f_i(x^0) - \nabla f_i(x^*)\|^2\right],
$$

.

Proof. We denote the Lyapunov function T^k at iteration k by

$$
T^k = \frac{1}{n}\sum_{i=1}^n \frac{1}{np_j}\|f_i'(\phi_i^k) - f_i'(x^*)\|^2 + c\|x^k - x^*\|^2.
$$

We will will show that $\mathbb{E}[T^{k+1}] \leq (1 - \frac{1}{\kappa})T^k$ for some $\kappa < 1$. First, we write the expectation of the first term as

$$
\begin{aligned}
&\mathbb{E}\left[\sum_i \frac{1}{n^2 p_i}\|f_i'(\phi_i) - f_i'(x^*)\|^2\right] \\
&= \mathbb{E}\left[\frac{1}{n^2 p_j}\|f_j'(x) - f_j'(x^*)\|^2\right] + \sum_i \frac{1}{n^2 p_i}\|f_i'(\phi_i) - f_i'(x^*)\|^2 - E\left[\frac{1}{n^2 p_j}\|f_j'(\phi_j) - f_j'(x^*)\|^2\right] \\
&= \frac{1}{n^2}\sum_i \|f_i'(x) - f_i'(x^*)\|^2 + \frac{1}{n^2}\sum_i \left(\frac{1}{p_i} - 1\right)\|f_i'(\phi_i) - f_i'(x^*)\|^2.
\end{aligned} \tag{3}
$$

Next, we simplify the other term of $\mathbb{E}[T^{k+1}]$,

$$
\begin{aligned}
c\mathbb{E}\|x^{k+1} - x^*\|^2 &= c\mathbb{E}\|x - x^* - \frac{1}{\eta}\nu_k\|^2 \\
&= c\|x - x^*\|^2 + \frac{c}{\eta^2}\mathbb{E}\|\nu_k\|^2 + \frac{2c}{\eta}\langle f'(x), x - x^*\rangle
\end{aligned}
$$

We now use Lemma 1 followed by Inequality (2),

$$\begin{aligned}
c\mathbb{E}\|x^{k+1}-x^*\|^2 &\le c\|x-x^*\|^2 + \frac{c}{\eta^2}2\mathbb{E}\|\frac{1}{np_j}[f_j'(x)-f_j'(x^*)]\|^2 + \frac{c}{\eta^2}2\mathbb{E}\|\frac{1}{np_j}[f_j'(\phi_j)-f_j'(x^*)]\|^2 + \frac{2c}{\eta}\langle f'(x), x-x^*\rangle \\
&\le c(1-\frac{\mu}{\eta})\|x-x^*\|^2 + \frac{2c}{\eta^2}\mathbb{E}\|\frac{1}{np_j}(f_j'(x)-f_j'(x^*))\|^2 \\
&+ \frac{2c}{\eta^2}\mathbb{E}\|\frac{1}{np_j}(f_j'(\phi_j)-f_j'(x^*))\|^2 - \frac{c}{n\eta L}\sum_i \|f_i'(x^*)-f_i'(x)\|^2 \\
&= c(1-\frac{\mu}{\eta})\|x-x^*\|^2 + \sum_i(\frac{2c}{n^2\eta^2 p_i} - \frac{c}{n\eta L})\|f_i'(x)-f_i'(x^*)\|^2 + \sum_i(\frac{2c}{n^2\eta^2 p_i})\|f_i'(\phi_i)-f_i'(x^*)\|^2.
\end{aligned}$$

We use this to bound the expected improvement in the Lyapunov function,

$$\begin{aligned}
\mathbb{E}[T^{k+1}] - T^k &= E[T^{k+1}] - \frac{1}{n}\sum_{i=1}^n \frac{1}{np_j}\|f_i'(\phi_i)-f_i'(x^*)\|^2 - c\|x-x^*\|^2 \\
&\le \frac{1}{n^2}\sum_i \|f_i'(x)-f_i'(x^*)\|^2 + \frac{1}{n^2}\sum_i\left(\frac{1}{p_i}-1\right)\|f_i'(\phi_i)-f_i'(x^*)\|^2 && \text{From (3)} \\
&+ c(1-\frac{\mu}{\eta})\|x-x^*\|^2 + \sum_i(\frac{2c}{n^2\eta^2p_i}-\frac{c}{n\eta L})\|f_i'(x)-f_i'(x^*)\|^2 + \sum_i(\frac{2c}{n^2\eta^2p_i})\|f_i'(\phi_i)-f_i'(x^*)\|^2 && \text{From above} \\
&- \frac{1}{n}\sum_{i=1}^n\frac{1}{np_j}\|f_i'(\phi_i)-f_i'(x^*)\|^2 - c\|x-x^*\|^2 && \text{Definition of } T^k \\
&= \frac{1}{n^2}\sum_i\|f_i'(x)-f_i'(x^*)\|^2 - \frac{1}{n^2}\sum_i\|f_i'(\phi_i)-f_i'(x^*)\|^2 \\
&- \frac{c\mu}{\eta}\|x-x^*\|^2 + \sum_i(\frac{2c}{n^2\eta^2p_i}-\frac{c}{n\eta L})\|f_i'(x)-f_i'(x^*)\|^2 + \sum_i(\frac{2c}{n^2\eta^2p_i})\|f_i'(\phi_i)-f_i'(x^*)\|^2 \\
&= -\frac{1}{\kappa}T^k + \left(\frac{1}{\kappa}-\frac{\mu}{\eta}\right)c\|x-x*\|^2 && (*) \\
&+ \sum_i\left(\frac{2c}{n^2\eta^2p_i}+\frac{1}{n^2}-\frac{c}{n\eta L}\right)\|f_i'(x)-f_i'(x^*)\|^2 \\
&+ \sum_i\left(\frac{2c}{n^2\eta^2p_i}-\frac{1}{n^2}+\frac{1}{n^2\kappa p_i}\right)\|f_i'(\phi_i)-f_i'(x^*)\|^2 \\
&\le -\frac{1}{\kappa}T^k + \left(\frac{1}{\kappa}-\frac{\mu}{\eta}\right)\left[c\|x-x*\|^2\right] \\
&+ \left(\frac{2c}{n^2\eta^2p_m}+\frac{1}{n^2}-\frac{c}{n\eta L}\right)\left[\sum_i\|f_i'(x)-f_i'(x^*)\|^2\right] \\
&+ \left(\frac{2c}{n^2\eta^2p_m}-\frac{1}{n^2}+\frac{1}{n^2\kappa p_m}\right)\left[\sum_i\|f_i'(\phi_i)-f_i'(x^*)\|^2\right],
\end{aligned}$$

where in $(*)$ we add and subtract $\frac{1}{\kappa}T^k$ and in the last line we assumed $c \ge 0$ and used $p_i \ge p_m$. The terms in square brackets are positive, and if we can choose the constants $\{c, \kappa, \eta\}$ to make the round brackets non-positive, we have

$$\mathbb{E}[T^{k+1}] \le \left(1-\frac{1}{\kappa}\right)T^k.$$

For the first expression, choosing $\kappa = \frac{\eta}{\mu}$ makes it zero. We can make the third expression zero under this choice of κ by choosing $c = \frac{\eta^2 p_m}{2} - \frac{\mu\eta}{2}$. This follows because

$$\frac{2c}{n^2\eta^2p_m} - \frac{1}{n^2} + \frac{1}{n^2\kappa p_m} = \frac{2c}{n^2\eta^2p_m} - \frac{1}{n^2} + \frac{\mu}{n^2\eta p_m} = 0,$$

is equivalent to

$$\frac{2c}{n^2\eta^2 p_m} = \frac{1}{n^2} - \frac{\mu}{n^2\eta p_m} \Leftrightarrow c = \frac{\eta^2 p_m}{2} - \frac{\mu\eta}{2}.$$

For the second expression, note that with our choice of c we have

$$\frac{2c}{n^2\eta^2 p_m} + \frac{1}{n^2} - \frac{c}{n\eta L} = \frac{1}{n^2} - \frac{\mu}{n^2\eta p_m} + \frac{1}{n^2} - \frac{\frac{\eta^2 p_m}{2} - \frac{\mu\eta}{2}}{n\eta L},$$

which (multiplying by n) is negative if we have

$$\frac{2}{n} + \frac{\mu}{2L} \leq \frac{\mu}{n\eta p_m} + \frac{\eta p_m}{2L}.$$

Ignoring the last term, we can choose

$$\eta = \frac{4L + n\mu}{np_m}.$$

We will also require that $c \geq 0$ to complete the proof, but this follows because $\eta \geq \frac{\mu}{p_m}$. By using that

$$c\mathbb{E}[\|x^{k+1} - x^*\|^2] \leq \mathbb{E}[T^{k+1}] \leq \left(1 - \frac{1}{\kappa}\right) T^k = \left(1 - \frac{\mu}{\eta}\right) T^k$$

and chaining the expectations while using the definition of η we obtain

$$\begin{aligned}
\mathbb{E}[\|x^k - x^*\|^2] &\leq \left(1 - \frac{\mu}{\eta}\right)^k \frac{T^0}{c} \\
&= \left(1 - \frac{np_m\mu}{n\mu + 4L}\right)^k \left[\|x^0 - x^*\|^2 + \frac{1}{cn}\sum_{i=1}^n \frac{1}{np_j}\|f_i'(\phi_i^0) - f_i'(x^*)\|^2\right].
\end{aligned}$$

To get the final expression, use that

$$\frac{1}{cn^2} = \frac{2}{n^2(\eta^2 p_m - \mu\eta)} \leq \frac{2}{n^2\eta^2 p_m} = \frac{2n^2 p_m^2}{n^2 p_m(4L + n\mu)^2} = \frac{2p_m}{(4L + n\mu)^2}.$$

□

Proof of Part (b) of Proposition 1

In this section we consider the minimization problem

$$\min_x f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x), \tag{4}$$

where each f_i' is L_i-Lipschitz continuous and f is μ-strongly-convex. We will define Algorithm 2, a variant of SAGA, by the sequences $\{x^k\}$, $\{\nu_k\}$, and $\{\phi_j^k\}$ given by

$$\begin{aligned}
\nu_k &= \frac{\bar{L}}{L_i}[f_{j_k}'(x^k) - f_{j_k}'(\phi_j^k)] + \frac{1}{n}\sum_{i=1}^n f_i'(\phi_i^k), \\
x^{k+1} &= x^k - \gamma\nu_k, \\
\phi_j^{k+1} &= \begin{cases} f_{r_k}'(x^k) & \text{if } j = r_k, \\ \phi_j^k & \text{otherwise,} \end{cases}
\end{aligned}$$

where $j_k = j$ with probability $\frac{L_i}{\sum_{j=1}^n L_j}$ and r_k is picked uniformly at random. This is identical to Algorithm 1, except it uses a specific choice of the p_j and the memory ϕ_j is updated based on a different random sample that is sampled uniformly. This algorithm maintains the key property that the expected step is a gradient step, $\mathbb{E}[\nu_k] = f'(x^k)$.

From our assumptions about f and the f_i, we have [Nesterov, 2004, see Chapter 2].

$$f_i(x) \geq f_i(y) + \left\langle f_i'(y), x-y \right\rangle + \frac{1}{2L}\left\|f_i'(x) - f_i'(y)\right\|^2, \tag{5}$$

and

$$f(x) \geq f(y) + \left\langle f'(y), x-y \right\rangle + \frac{\mu}{2}\left\|x-y\right\|^2. \tag{6}$$

We use these to derive several useful inequalities that we will use in the analysis. Adding the former times $\frac{1}{2n}$ for all i to the latter times $\frac{1}{2}$ for $y = x^*$ gives the inequality

$$\left\langle f'(x), x^* - x \right\rangle \leq f(x^*) - f(x) - \frac{\mu}{4}\left\|x^* - x\right\|^2 - \frac{1}{4n}\sum_i \frac{1}{L_i}\left\|f_i'(x^*) - f_i'(x)\right\|^2. \tag{7}$$

Also by applying (5) with $y = x^*$ and $x = \phi_i$, for each f_i and summing, we have that for all ϕ_i and x^*:

$$\frac{1}{n}\sum_i \frac{1}{L_i}\left\|f_i'(\phi_i) - f_i'(x^*)\right\|^2 \leq \frac{2}{n}\sum_i \left[f_i(\phi_i) - f(x^*) - \left\langle f_i'(x^*), \phi_i - x^* \right\rangle\right]. \tag{8}$$

Further, by both minimizing sides of (6) we obtain

$$-\left\|f'(x)\right\|^2 \leq -2\mu\left[f(x) - f(x^*)\right]. \tag{9}$$

We next derive a bound on the variance of the gradient estimate.

Lemma 2. *It holds that for any ϕ_i that with x^{k+1} and x^k as given by Algorithm 2 we have*

$$\begin{aligned}\mathbb{E}\left\|x^{k+1} - x^k\right\|^2 \quad \leq \quad & 2\gamma^2\frac{\bar{L}}{n}\sum_i \frac{1}{L_i}\left\|f_j'(\phi_j^k) - f_j'(x^*)\right\|^2 \\ & + 2\gamma^2\frac{\bar{L}}{n}\sum_i \frac{1}{L_i}\left\|f_j'(x^k) - f_j'(x^*)\right\|^2 - \gamma^2\left\|f'(x^k)\right\|^2.\end{aligned}$$

Proof. We again follow the SAGA argument closely here

$$\begin{aligned} & \mathbb{E}\left\|x^{k+1} - x^k\right\|^2 \\ = \quad & \gamma^2\mathbb{E}\left\|\frac{\bar{L}}{L_j}\left[f_j'(\phi_j^k) - f_j'(x^k)\right] - \frac{1}{n}\sum_{i=1}^n f_i'(\phi_i^k)\right\|^2 \\ = \quad & \gamma^2\mathbb{E}\left\|\frac{\bar{L}}{L_j}\left[f_j'(\phi_j^k) - f_j'(x^*)\right] - \frac{1}{n}\sum_{i=1}^n f_i'(\phi_i^k) - \frac{\bar{L}}{L_j}\left[f_j'(x^k) - f_j'(x^*)\right] - f'(x^k)\right\|^2 \\ & + \gamma^2\left\|f'(x^k)\right\|^2 \\ \leq \quad & 2\gamma^2\mathbb{E}\left\|\frac{\bar{L}}{L_j}\left[f_j'(\phi_j^k) - f_j'(x^*)\right] - \frac{1}{n}\sum_{i=1}^n f_i'(\phi_i^k)\right\|^2 \\ & + 2\gamma^2\mathbb{E}\left\|\frac{\bar{L}}{L_j}\left[f_j'(x^k) - f_j'(x^*)\right] - f'(x^k)\right\|^2 + \gamma^2\left\|f'(x^k)\right\|^2 \\ \leq \quad & 2\gamma^2\mathbb{E}\left\|\frac{\bar{L}}{L_j}\left[f_j'(\phi_j^k) - f_j'(x^*)\right]\right\|^2 \\ & + 2\gamma^2\mathbb{E}\left\|\frac{\bar{L}}{L_j}\left[f_j'(x^k) - f_j'(x^*)\right]\right\|^2 - \gamma^2\left\|f'(x^k)\right\|^2. \end{aligned}$$

We can expand those expectations as follows $\frac{1}{n}\sum_i L_i = \bar{L}$:

$$\begin{aligned}\mathbb{E}\left\|\frac{\bar{L}}{L_i}\left[f_j'(\phi_j^k) - f_j'(x^*)\right]\right\|^2 \quad = \quad & \frac{1}{n\bar{L}}\sum_i L_i\left\|\frac{\bar{L}}{L_i}\left[f_j'(\phi_j^k) - f_j'(x^*)\right]\right\|^2 \\ = \quad & \frac{\bar{L}}{n}\sum_i \frac{1}{L_i}\left\|\left[f_j'(\phi_j^k) - f_j'(x^*)\right]\right\|^2,\end{aligned}$$

and similarly for $\mathbb{E}\left\|\frac{\bar{L}}{L_i}\left[f_j'(x^k)-f_j'(x^*)\right]\right\|^2$. $\square$

We now give the proof of part (b) of Proposition 1, which we state below.

Proposition 1 (b)**.** *If $\gamma=\frac{1}{4L}$, then Algorithm 2 has*

$$E\left[\left\|x^k-x^*\right\|^2\right]\leq\left(1-\min\left\{\frac{1}{3n},\frac{\mu}{8\bar{L}}\right\}\right)^k\left[\left\|x^k-x^*\right\|^2+\frac{n}{2\bar{L}}\left(f(x^0)-f(x^*)\right)\right].$$

Proof. We define the Lyapunov function as

$$T^k=\frac{1}{n}\sum_i f_i(\phi_i^k)-f(x^*)-\frac{1}{n}\sum_i\left\langle f_i'(x^*),\phi_i^k-x^*\right\rangle+c\left\|x^k-x^*\right\|^2.$$

The expectations of the first terms in T^{k+1} are straightforward to simplify:

$$\begin{aligned}\mathbb{E}\left[\frac{1}{n}\sum_i f_i(\phi_i^{k+1})\right] &= \frac{1}{n}f(x^k)+\left(1-\frac{1}{n}\right)\frac{1}{n}\sum_i f_i(\phi_i^k),\\ \mathbb{E}\left[-\frac{1}{n}\sum_i\left\langle f_i'(x^*),\phi_i^{k+1}-x^*\right\rangle\right] &= -\left(1-\frac{1}{n}\right)\frac{1}{n}\sum_i\left\langle f_i'(x^*),\phi_i^k-x^*\right\rangle.\end{aligned}$$

Note that these terms make use of the uniformly sampled $\phi_r^{k+1}=x^k$ value. For the change in the last term of T^k we expand the quadratic and apply $\mathbb{E}[x^{k+1}]=x^k-\gamma f'(x^k)$ to simplify the inner product term:

$$\begin{aligned}&c\mathbb{E}\left\|x^{k+1}-x^*\right\|^2\\ =&c\mathbb{E}\left\|x^k-x^*+x^{k+1}-x^k\right\|^2\\ =&c\left\|x^k-x^*\right\|^2+2c\mathbb{E}\left[\left\langle x^{k+1}-x^k,x^k-x^*\right\rangle\right]+c\mathbb{E}\left\|x^{k+1}-x^k\right\|^2\\ =&c\left\|x^k-x^*\right\|^2-2c\gamma\left\langle f'(x^k),x^k-x^*\right\rangle+c\mathbb{E}\left\|x^{k+1}-x^k\right\|^2.\end{aligned}$$

We now apply Lemma 2 to bound the error term $c\mathbb{E}\left\|x^{k+1}-x^k\right\|^2$, giving:

$$\begin{aligned}&c\mathbb{E}\left\|x^{k+1}-x^*\right\|^2\\ \leq\;& c\left\|x^k-x^*\right\|^2-c\gamma^2\left\|f'(x^k)\right\|^2\\ &-2c\gamma\left\langle f'(x^k),x^k-x^*\right\rangle\\ &+2c\gamma^2\frac{\bar{L}}{n}\sum_i\frac{1}{L_i}\left\|f_i'(\phi_i^k)-f_i'(x^*)\right\|^2+2c\gamma^2\frac{\bar{L}}{n}\sum_i\frac{1}{L_i}\left\|f_i'(x^k)-f_i'(x^*)\right\|^2.\end{aligned}$$

Now we bound $-2c\gamma\left\langle f'(x),x-x^*\right\rangle$ with (7) and then apply (8) to bound $\mathbb{E}\left\|f_j'(\phi_j)-f_j'(x^*)\right\|^2$:

$$\begin{aligned}c\mathbb{E}\left\|x^{k+1}-x^*\right\|^2\leq&\left(c-\frac{1}{2}c\gamma\mu\right)\left\|x^k-x^*\right\|^2\\ &+\left(2c\gamma^2\bar{L}-\frac{1}{2}c\gamma\right)\frac{1}{n}\sum_i\frac{1}{L_i}\left\|f_i'(x^k)-f_i'(x^*)\right\|^2-c\gamma^2\left\|f'(x^k)\right\|^2\\ &-2c\gamma\left[f(x^k)-f(x^*)\right]\\ &+\left(4c\gamma^2\bar{L}\right)\frac{1}{n}\sum_i\left[f_i(\phi_i)-f_i(x^*)-\left\langle f_i'(x^*),\phi_i-x^*\right\rangle\right].\end{aligned}$$

We can now combine the bounds we have derived for each term in T, and pull out a fraction $\frac{1}{\kappa}$ of T^k (for any κ at this point). Together with (9) this yields:

$$\begin{aligned}\mathbb{E}[T^{k+1}] - T^k \leq & -\frac{1}{\kappa}T^k + \left(\frac{1}{n} - 2c\gamma - 2c\gamma^2\mu\right)\left[f(x^k) - f(x^*)\right] \\ & + \left(\frac{1}{\kappa} + 4c\gamma^2\bar{L} - \frac{1}{n}\right)\left[\frac{1}{n}\sum_i f_i(\phi_i^k) - f(x^*) - \frac{1}{n}\sum_i \left\langle f_i'(x^*), \phi_i^k - x^*\right\rangle\right] \\ & + \left(\frac{1}{\kappa} - \frac{1}{2}\gamma\mu\right) c \left\|x^k - x^*\right\|^2 + \left(2\gamma\bar{L} - \frac{1}{2}\right) c\gamma\frac{1}{n}\sum_i \frac{1}{L_i}\left\|f_i'(x^k) - f_i'(x^*)\right\|^2. \qquad (10)\end{aligned}$$

Note that the term in square brackets in the second row is positive in light of (8). We now attempt to find constants that satisfy the required relations. We start with naming the constants that we need to be non-positive:

$$\begin{aligned} c_1 &= \frac{1}{n} - 2c\gamma - 2c\gamma^2\mu, \\ c_2 &= \frac{1}{\kappa} + 4c\gamma^2\bar{L} - \frac{1}{n}, \\ c_3 &= \frac{1}{\kappa} - \frac{1}{2}\gamma\mu, \\ c_4 &= 2\gamma\bar{L} - \frac{1}{2}. \end{aligned}$$

Recall that we are using the step size $\gamma = 1/4\bar{L}$, and thus $c_4 = 0$. Setting c_1 to zero gives

$$c = \frac{1}{2\gamma(1-\gamma\mu)n},$$

which is positive since $\gamma\mu < 1$. Now we look at the restriction that $c_2 \leq 0$ places on κ:

$$\begin{aligned} \frac{1}{\kappa} + 4c\gamma^2\bar{L} - \frac{1}{n} &= \frac{1}{\kappa} + \frac{4\gamma\bar{L}}{2(1-\gamma\mu)n} - \frac{1}{n} \\ &= \frac{1}{\kappa} + \frac{1}{2(1-\gamma\mu)n} - \frac{1}{n} \\ &= \frac{1}{\kappa} + \frac{1}{2(1-\mu/4\bar{L})n} - \frac{1}{n} \\ &\leq \frac{1}{\kappa} + \frac{1}{2(1-\bar{L}/4\bar{L})n} - \frac{1}{n} \\ &= \frac{1}{\kappa} + \frac{2}{3n} - \frac{1}{n} \\ &= \frac{1}{\kappa} - \frac{1}{3n}, \end{aligned}$$

$$\therefore \frac{1}{\kappa} \leq \frac{1}{3n}.$$

We also have the restriction from $c_3 = \frac{1}{\kappa} - \frac{1}{2}\gamma\mu$ of

$$\frac{1}{\kappa} \leq \frac{\mu}{8\bar{L}},$$

therefore we can take

$$\frac{1}{\kappa} = \min\left\{\frac{1}{3n}, \frac{\mu}{8\bar{L}}\right\}.$$

Note that $c\left\|x^k - x^*\right\|^2 \leq T^k$, and therefore by chaining expectations and plugging in constants we get:

$$E\left[\left\|x^k - x^*\right\|^2\right] \leq \left(1 - \min\left\{\frac{1}{3n}, \frac{\mu}{8\bar{L}}\right\}\right)^k \left[\left\|x^k - x^*\right\|^2 + \frac{n}{2\bar{L}}\left(f(x^0) - f(x^*)\right)\right].$$

$\square$

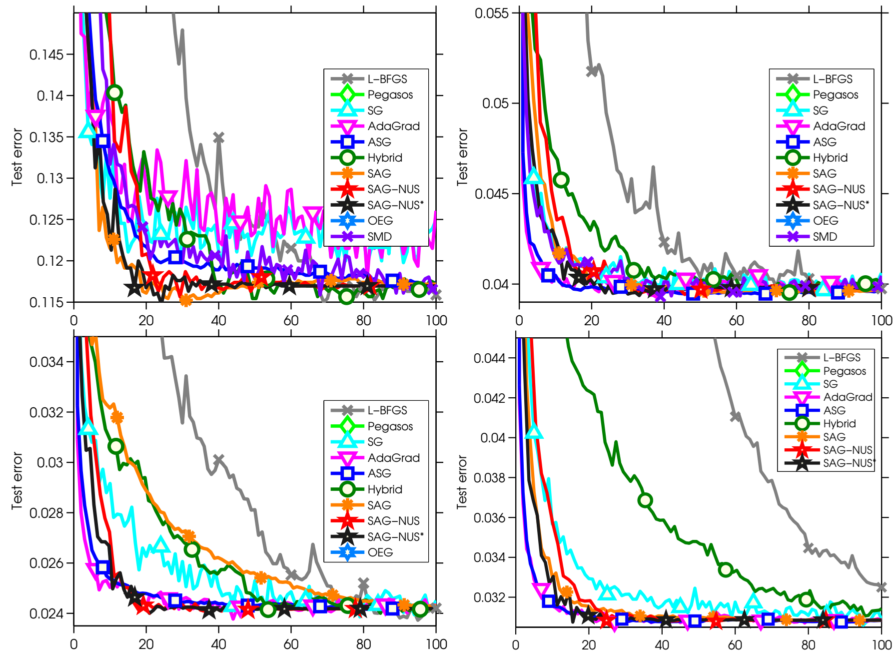

Figure 1: Test error against effective number of passes for different deterministic, stochastic, and semi-stochastic optimization strategies. Top-left: OCR, Top-right: CoNLL-2000, bottom-left: CoNLL-2002, bottom-right: POS-WSJ.

Test Error Plots for All Methods

In the main paper we only plotted test error for a subset of the methods. In Figure 1 we plot the test error of all methods considered in Figure 1 of the main paper. Note that Pegasos and OEG do not appear on the plot (despite being in the legend) because their values exceed the maximum plotted values. In these plots we see that the SAG-NUS methods perform similarly to the best among the optimally-tuned stochastic gradient methods in terms of test error, despite the lack of tuning required to apply these methods.

Runtime Plots

In the main paper we plot the performance against the effective number of passes as an implementation-independent way of comparing the different algorithms. In all cases except OEG and SMD, we implemented a C version of the method and also compared the running times of our different implementations. This ties the results to the hardware used to perform the experiments, and thus says little about the runtime in different hardware settings, but does show the practical performance of the methods in this particular setting. We plot the training objective against runtime in Figure 2 and the testing objective in Figure 3. In general, the runtime plots show the exact same trends as the plots against the effective number of passes. However, we note several small differences:

- *AdaGrad* performs slightly worse in terms of runtime, and was always worse than the basic *SG* method. This seems to be due to the extra square root operators needed to implement the method.
- *Hybrid* performs slightly worse in terms of runtime, although it was still faster than the *L-BFGS* method. This seems to be due to the higher cost of applying the L-BFGS update when the batch size is small.
- *SAG* performs slightly worse in terms of runtime, though it remains among the other top performing methods *Hybrid* and *ASG*. This seems to be due to the higher cost of the memory update associated

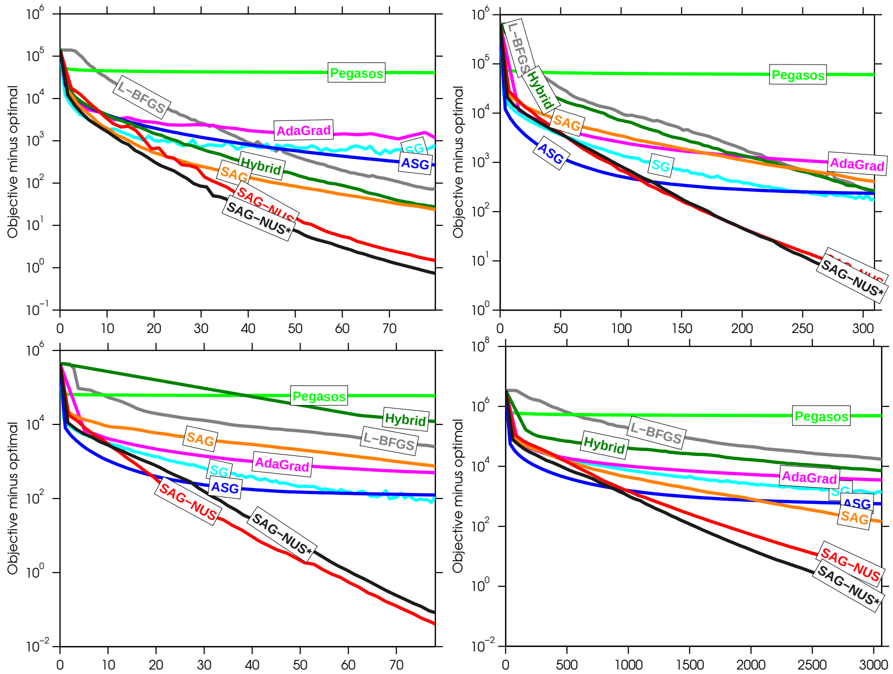

Figure 2: Objective minus optimal objective value against time for different deterministic, stochastic, and semi-stochastic optimization strategies. Top-left: OCR, Top-right: CoNLL-2000, bottom-left: CoNLL-2002, bottom-right: POS-WSJ.

with the algorithm.

- Although both *SAG-NUS* methods still dominate all other methods by a substantial margin, the performance of the new *SAG-NUS** and the existing *SAG-NUS* is much closer in terms of runtime. This seems to be because, although the *SAG-NUS* method does much more backtracking than *SAG-NUS**, these backtracking steps are much cheaper because they only require the forward pass of the forward-backward algorithm. If we compared these two algorithms under more complicated inference schemes, we would expect the advantage of *SAG-NUS** to appear in the runtime, too.

References

A. Defazio, F. Bach, and S. Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in Neural Information Processing Systems*, 2014.

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

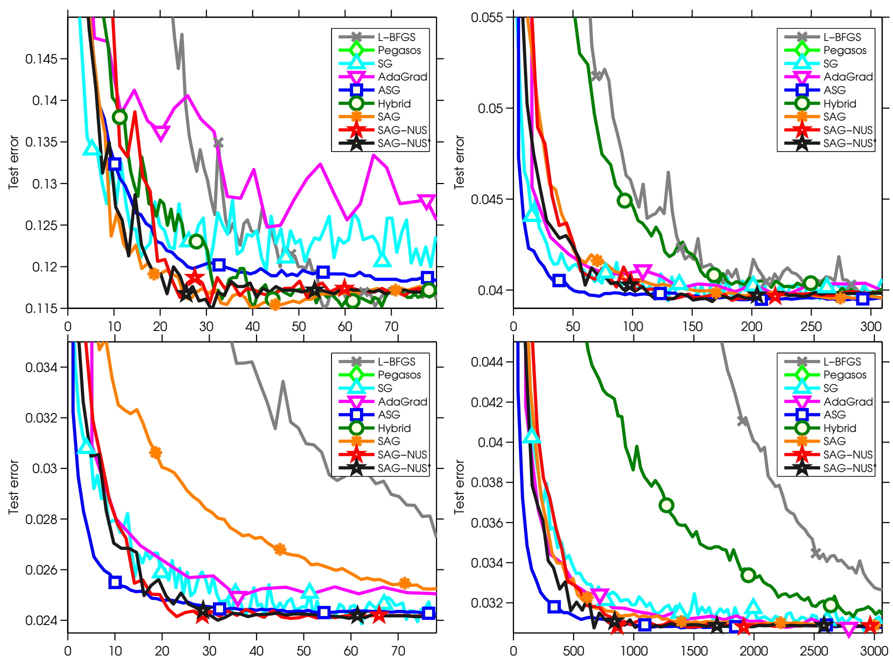

Figure 3: Test error against time for different deterministic, stochastic, and semi-stochastic optimization strategies. Top-left: OCR, Top-right: CoNLL-2000, bottom-left: CoNLL-2002, bottom-right: POS-WSJ.