Proof of Convergence

Proof The following proof of DASO's global synchronization method is based heavily on the convergence analysis shown by [43] and will show that the gradients determined with DASO are bounded.

Let $X \subset \mathbb{R}^n$ be a known set, and $f : X \to \mathbb{R}$ a differentiable, convex, L-smooth, and unknown function. Then, the estimator of the stochastic gradient of $f(x)$ is a function $\tilde{g}(x)$ for inputs x determined by the realization of a random variable ζ, such that $\mathbb{E}[\tilde{g}(x;\zeta)] = \nabla f(x:\zeta)$. In the following, ζ is omitted due to space constraints. The stochastic gradient descent (SGD) algorithm updates a model's state at batch $t+1$, x_{t+1}, with the following rule $x_{t+1} = x_t - \eta\tilde{g}(x_t)$, where η is the parametric learning rate. A commonly used variant of SGD in practice is minibatching for computational efficiency reasons. In minibatch SGD, the true stochastic gradient is approximated by averaging across m input items x_i, i.e. $\tilde{G}(x_t) = \frac{1}{m}\sum_{i=1}^{m}\tilde{g}(x_{t,i})$. The model state x_{t+1} for minibatch SGD is

$$x_{t+1} = x_t - \eta\tilde{G}\left(x_t\right) \tag{3}$$

where $\tilde{G}\left(x_t\right)$ is an estimator of $\nabla f\left(x_t\right)$.

Let us now consider, that S subsequent update steps are performed. It is possible to write the model state as:

$$x_{t+S} = x_t - \eta\sum_{i=0}^{S-1}\tilde{G}\left(x_{t+i}\right) \tag{4}$$

One of the primary assumptions in SGD is the Lipschitz-continuous objective gradients. This has the effect that:

$$f\left(x_{t+1}\right) - f\left(x_t\right) \leq -\eta\nabla f\left(x_t\right)^T\mathbb{E}\left[\tilde{g}\left(x_t\right)\right] + \frac{1}{2}\eta^2 L\mathbb{E}\left[\left\|\tilde{g}\left(x_t\right)\right\|_2^2\right] \tag{5}$$

where the Lipschitz constant, L, is greater than zero. Equation (5) implies that the expected decrease in the objective function, $f(x)$, is bounded above by a set quantity, regardless of how the stochastic gradients arrived at x_t [43].

In DASO, the local synchronization step is bound via the same assumptions as minibatch SGD outlined in [43], so long as the iid assumption is upheld. However, the non-standard global synchronization step used in DASO must be shown to be bound under the same principles. DASO's global synchronization is:

$$x_{t+S}^{\mathrm{DASO}} = \frac{2Sx_{l:t+S-1} + \sum_{i=1}^{P}x_{p:t}^{i}}{2S+P} \tag{6}$$

where the l and p subscripts represent the node-local and global model states, S is the number of local update steps before global synchronization, and P is the number of processes.

Similar to Equation (3), this can also be represented via the locally and globally calculated gradients, $\tilde{G}_l\left(x_{l:t}\right)$ and $\tilde{G}_p\left(x_{p:t}\right)$ respectively. The global synchronization function in the gradient representation is as follows:

$$x_{t+S}^{\mathrm{DASO}} = x_t - \alpha\left(2S\sum_{k=0}^{S-1}\tilde{G}_l\left(x_{l:t+k}\right) + \sum_{i=1}^{P}\tilde{G}_p\left(x_{p:t}^{i}\right)\right) \tag{7}$$

where $\alpha = \eta/(2S+P)$. Using this, Equation (3), and the fact that the updates between t and S are local synchronizations which take the form of Equation (4), we find that globally calculated gradients are as follows.

$$\tilde{G}^{\mathrm{DASO}}\left(x_{t+S-1}\right) = P\sum_{\beta=0}^{S-1}\tilde{G}_l\left(x_{l:t+S-\beta}\right) - 2S\tilde{G}_l\left(x_{l:t+S-1}\right) - \sum_{i=1}^{P}\tilde{G}_p\left(x_{p:t}^{i}\right) \tag{8}$$

As all gradient elements in Equation (8) are bound under Equation (5), $\tilde{G}^{\mathrm{DASO}}\left(x_{t+S-1}\right)$ is similarly bounded. □