Coordinate-Wise Parameter-Free

1 Introduction

To the best of my knowledge, DoG [Ivgi et al., 2023], DoWG, and existing parameter-free results are all for scalar step sizes, which can be non-desirable when training large models. Previously, we have seen the benefits of coordinate-wise step size [Duchi et al., 2011, Liu et al., 2024]. I tried to make DoG coordinate-wise and it seems to work. Let's take the algorithm to be

$$\mathbf{w}_{t+1} = \Pi_{\mathcal{W}}^{\mathbf{\Lambda}_t}\left(\mathbf{w}_t - \eta_t \mathbf{\Lambda}_t^{-1}\mathbf{g}_t\right)$$

where $\mathbf{\Lambda}_t = \mathrm{diag}[\lambda_{t,1}, \dots \lambda_{t,d}]$ and

$$\lambda_{t,j}^2 = \lambda_{t-1,j}^2 + \mathbf{g}_{t,j}^2$$

and we take $\eta_t = \bar{r}_t$ with

$$\bar{r}_t = \max\{\max_{k \le t} r_t, r_\epsilon\} \quad \text{where} \quad r_t \triangleq \|\mathbf{w}_t - \mathbf{w}_0\|_\infty .$$

The settings are similar to that of DoG.

We have the following two key lemmas for obtaining convergence in the deterministic nonsmooth case:

$$\begin{aligned}
\sum_{t=0}^{T-1} \eta_t \langle \mathbf{g}_t, \mathbf{w}_t - \mathbf{w}_* \rangle \le & \sum_{t=0}^{T-1} \left(\|\mathbf{w}_t - \mathbf{w}_*\|_{\mathbf{\Lambda}_t}^2 - \|\mathbf{w}_{t+1} - \mathbf{w}_*\|_{\mathbf{\Lambda}_t}^2 \right) + \sum_{t=0}^{T-1} \eta_t^2 \|\mathbf{g}_t\|_{\mathbf{\Lambda}_t^{-1}}^2 && (1)\\
\le & 2(\bar{d}_T^2 + \bar{r}_T^2)\mathrm{tr}\left(\mathbf{\Lambda}_{T-1}\right) = 2(\bar{d}_T^2 + \bar{r}_T^2) \sum_{j=1}^{d} \sqrt{\sum_{t=0}^{T-1} \mathbf{g}_{t,j}^2} && (2)\\
= & \mathcal{O}\left(D_\infty^2 G_1 \sqrt{T}\right), && (3)
\end{aligned}$$

here we follow DoG to take $\bar{d}_t \triangleq \max_t \|\mathbf{w}_t - \mathbf{w}_*\|_\infty$ and D_∞ denotes the infinite-norm diameter of $\mathcal{W}$ and G_1 is the upper bound on gradient 1-norm. Then based on Lemma 3 of DoG, we can obtain the convergence rate

$$\mathcal{O}\left(\frac{D_\infty G_1}{\sqrt{T}} \log \frac{D_\infty}{r_\epsilon}\right)$$

in the nonsmooth case.

We have convergence results of this coordinate-wise version of DoG:

Theorem 1 (Nonsmooth Convergence)**.** *Assume convex and infinite-norm diameter D_∞ of $\mathcal{W}$, the Coordinate-wise DoG has the following convergence:*

$$\mathbb{E}\left[f(\bar{\mathbf{w}}_T) - f^*\right] \le \mathcal{O}\left(\frac{D_\infty}{T} \sum_{j=1}^{d} \sqrt{\sum_{t=0}^{T-1} \mathbf{g}_{t,j}^2} \log \frac{D_\infty}{r_\epsilon}\right). \quad (4)$$

Or if we assume a coordinate-wise bound bound $\mathbf{G}$ on the subgradient $\mathbf{g}_t$, we have

$$\mathbb{E}\left[f(\bar{\mathbf{w}}_T) - f^*\right] \le \mathcal{O}\left(\frac{D_\infty \|\mathbf{G}\|_1}{\sqrt{T}} \log \frac{D_\infty}{r_\epsilon}\right).$$

Algorithm 1 Coordinate-wise DoG (without projection)

Input: $\mathbf{w}_0 \in \mathbb{R}^d$, $r_\epsilon \in \mathbb{R}$, $\epsilon \in \mathbb{R}$ and batch size $M \in \mathbb{N}$ (Possibly r_ϵ can be chosen by $\alpha(1+\|\mathbf{w}_0\|_\infty)$ with $\alpha \in [10^{-8}, 10^{-6}]$ for language models; ϵ should be small, similar to the ϵ for Adam.)
Initialize $\mathbf{v}_{-1} = \epsilon^2 \mathbf{1}_d$, $\eta_{-1} = r_\epsilon$
for $t = 0$ **to** $T-1$ **do**
  Sample mini-batch $\mathcal{B}_t$ with $|\mathcal{B}_t| \equiv M$ uniformly
  $\mathbf{g}_t = \frac{1}{M}\sum_{\xi\in\mathcal{B}_t} \nabla_{\mathbf{w}} f(\mathbf{w}_t; \xi)$
  $\mathbf{v}_t = \mathbf{v}_{t-1} + (\mathbf{g}_t \odot \mathbf{g}_t)$ ▷ $\odot$ implies coordinate-wise multiplication just like Adam
  $\mathbf{\Lambda}_t = \mathrm{diag}(\sqrt{\mathbf{v}_t})$ ▷ Make the square root of $\mathbf{v}_t$ a diagonal matrix
  $\eta_t = \max\{\eta_{t-1}, \|\mathbf{w}_t - \mathbf{w}_0\|_\infty\}$ ▷ Update step size, need to store $\mathbf{w}_0$ to implement
  $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \mathbf{\Lambda}_t^{-1}\mathbf{g}_t$
end for

Algorithm 2 Coordinate-wise DoG with Momentum

Input: $\mathbf{w}_0 \in \mathbb{R}^d$, $r_\epsilon \in \mathbb{R}$, $\epsilon \in \mathbb{R}$, $\beta \in [0,1]$ (Possibly we can just choose $\beta = 0.9$ to have a try) and batch size $M \in \mathbb{N}$
Initialize $\mathbf{v}_{-1} = \epsilon^2 \mathbf{1}_d$, $\eta_{-1} = r_\epsilon$, $\mathbf{m}_{-1} = 0$
for $t = 0$ **to** $T-1$ **do**
  Sample mini-batch $\mathcal{B}_t$ with $|\mathcal{B}_t| \equiv M$ uniformly
  $\mathbf{g}_t = \frac{1}{M}\sum_{\xi\in\mathcal{B}_t} \nabla_{\mathbf{w}} f(\mathbf{w}_t; \xi)$
  $\mathbf{m}_t = \beta\mathbf{m}_{t-1} + (1-\beta)\mathbf{g}_t$
  $\mathbf{v}_t = \mathbf{v}_{t-1} + (\mathbf{m}_t \odot \mathbf{m}_t)$ ▷ $\odot$ implies coordinate-wise multiplication, here $\mathbf{m}_t$ instead of $\mathbf{g}_t$
  $\mathbf{\Lambda}_t = \mathrm{diag}(\sqrt{\mathbf{v}_t})$ ▷ Make the square root of $\mathbf{v}_t$ a diagonal matrix
  $\eta_t = \max\{\eta_{t-1}, \|\mathbf{w}_t - \mathbf{w}_0\|_\infty\}$ ▷ Update step size, need to store $\mathbf{w}_0$ to implement
  $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \mathbf{\Lambda}_t^{-1}\mathbf{m}_t$
end for

Theorem 2 (Smooth Convergence). *Assume convex, infinite-norm diameter D_∞ of $\mathcal{W}$, and coordinate-wise smoothness $\mathbf{L}$, the Coordinate-wise DoG has the following convergence:*

$$\mathbb{E}\left[f(\bar{\mathbf{w}}_T) - f^*\right] \leq \mathcal{O}\left(\frac{D_\infty^2 \|\mathbf{L}\|_1}{T} \log^2 \frac{D_\infty}{r_\epsilon}\right)$$

These results are generally consistent with the results of AdaGrad compared to SGD. A stochastic version proof should also be applicable if we follow the proof of DoG [Ivgi et al., 2023], which assumes bounded gradients and obtains high-probability results. I am still checking whether convergence in expectation can be applicable.

Possible next steps:

1. stochastic convergence
2. empirical check
3. extensions of the algorithm: exponential moving average, momentum
4. nonconvex

References

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

Maor Ivgi, Oliver Hinder, and Yair Carmon. Dog is sgd's best friend: A parameter-free dynamic step size schedule. In *International Conference on Machine Learning*, pages 14465–14499. PMLR, 2023.

Yuxing Liu, Rui Pan, and Tong Zhang. Large batch analysis for adagrad under anisotropic smoothness. *arXiv preprint arXiv:2406.15244*, 2024.