Optimization with Less Tuning

1 Different Targets

- Parameter-free, as noted by Orabona [2023], originally refers to the algorithms which achieve $\tilde{\mathcal{O}}\left(\frac{\|\mathbf{w}_1-\mathbf{w}_*\|}{\sqrt{T}}\right)$ in convex cases with bounded stochastic subgradients. Similar concepts are also employed in other settings. However, existing results cannot achieve real parameter-free (i.e. need no parameters literally) with tractable efficient algorithms.
- Tuning-free, considered by Khaled and Jin [2024], refers to algorithms that can perform well without exact estimations of constants like L. The results are mainly theoretical and basically lack experimental support. Nonconvex results are considered in this case.
- Schedule-free, considered by Defazio et al. [2024]. This is actually not a rigorous theoretical word but refers to algorithms that can perform well without hand-tuning schedules but do tune constants. This is more empirical, as we still lack rigorous theoretical proof of the benefits of (non-adaptive) schedules.
- Problem-parameter-free, refers to algorithms that can converge without prior knowledge of problem-dependent constants (e.g. smoothness L). There has been a long line of work on it[Malitsky and Mishchenko, 2023, Faw et al., 2022, Zhou et al., 2024]. Nonconvex results are also considered in this case, AdaGrad can be proven to satisfy this requirement.

2 Existing results

We summarize the major existing effort toward optimization with less tuning. Generally, besides Defazio et al. [2024], the existing results basically employ the SGD-like update

$$\mathbf{w}_{t+1}=\mathbf{w}_t-\eta_t\mathbf{g}_t \quad \text{or} \quad \mathbf{w}_{t+1}=\Pi_{\mathcal{W}}\left(\mathbf{w}_t-\eta_t\mathbf{g}_t\right), \tag{1}$$

where $\mathbf{g}_t$ is the stochastic gradient obtained at $\mathbf{w}_t$ and $\pi_{\mathcal{W}}$ indicates projection onto $\mathcal{W}$. The existing results mainly focus on how to adaptively set the step size η_t with less effect of tuning.

Let us define $r_t \triangleq \|\mathbf{w}_t-\mathbf{w}_0\|_2$ and $\bar{r}_t \triangleq \max_{k\le t} r_k \bigvee r_\epsilon$, where r_ϵ is a selected small constant.

2.1 DOG

The DOG algorithm is proposed by Ivgi et al. [2023] (Yair Carmon). It uses SGD-like update (1) with

$$\eta_{\mathrm{DOG},t}=\frac{\bar{r}_t}{\sqrt{\sum_{k=0}^{t}\|\mathbf{g}_k\|_2^2}}$$

Its convergence results include follows.

- Convex with bounded diameter D and universally bounded stochastic gradients $\|\mathbf{g}_t\|\le G$, with probability $1-\delta$,

$$f(\bar{\mathbf{w}}_T)-f^*\le\mathcal{O}\left(\frac{GD}{\sqrt{T}}\log\frac{D}{r_\epsilon}\log\frac{T}{\delta}\right).$$

They further consider a layer-wise version of DOG in the experiments, which performs well.

Remark 1. *The step size used by DOG is actually similar to the best step choice for the scalar version of AdaGrad in convex settings, which is*

$$\eta_{\text{AdaGrad},t} = \frac{\bar{r}_T}{\sqrt{\sum_{k=1}^T \|\mathbf{g}_t\|_2^2}}.$$

2.2 DOWG

The DOWG algorithm is proposed by Khaled et al. [2023] (Chi Jin). It uses SGD-like update (1) with

$$\eta_{\text{DOWG},t} = \frac{\bar{r}_t^2}{\sqrt{\sum_{k=0}^t \bar{r}_k^2 \|\mathbf{g}_k\|_2^2}} \geq \eta_{\text{DOG},t}$$

Its convergence results include follows.

- Deterministic convex G-Lipschitz with bounded diameter D,

$$f(\bar{\mathbf{w}}_T) - f^* \leq \mathcal{O}\left(\frac{GD}{\sqrt{T}} \log \frac{D}{r_\epsilon}\right).$$

- Deterministic convex L-smooth with bounded diameter D,

$$f(\bar{\mathbf{w}}_T) - f^* \leq \mathcal{O}\left(\frac{LD^2}{\sqrt{T}} \log \frac{D}{r_\epsilon}\right).$$

2.3 Schedule-Free SGD

The schedule-free SGD algorithm is proposed by Defazio et al. [2024] (Aaron Defazio).

$$\begin{aligned}
\mathbf{y}_t &= (1-\beta)\mathbf{z}_t + \beta \mathbf{x}_t \\
\mathbf{z}_{t+1} &= \mathbf{z}_t - \eta \nabla f(\mathbf{y}_t; \xi) \\
\mathbf{x}_{t+1} &= (1-c_{t+1})\mathbf{x}_t + c_{t+1}\mathbf{z}_{t+1},
\end{aligned}$$

with $c_{t+1} = 1/(t+1)$ and $\mathbf{z}_1 = \mathbf{x}_1$ as the initial point. In this case, we need to well-tune the hyperparameters η and β. Its convergence results include follows.

- Convex with universally bounded stochastic gradients $\nabla f(\mathbf{y}_t; \xi) \leq G$, let $D = \|\mathbf{x}_1 - \mathbf{x}_*\|$, then for $\eta = \frac{D}{G\sqrt{T}}$ and any $\beta \in [0, 1]$,

$$f(\mathbf{x}_T) - f^* \leq \mathcal{O}\left(\frac{GD}{\sqrt{T}}\right).$$

  Note that this is last iterate convergence, and seems to outperform SGD by logarithmic factors [Shamir and Zhang, 2013].

Besides, Defazio et al. [2024] also proposes schedule-free AdamW using a similar manner as schedule-free SGD, which seems to perform well in experiments.

3 Some Idea

1. We consider a possible coordinate-wise version of Dog, for the j-th coordinate, we use

$$\eta_{t,j} = \frac{\bar{r}_{t,\infty}}{\sqrt{\sum_{k\leq t} \mathbf{g}_{k,j}^2}}, \quad \text{where} \quad r_{t,\infty} \triangleq \|\mathbf{w}_t - \mathbf{w}_0\|_\infty \quad \text{and} \quad \bar{r}_{t,\infty} \triangleq \max_{k\leq t} r_{k,\infty} \bigvee r_\epsilon.$$

   - It might be hard to further incorporate momentum, as well as the exponential moving average version of the preconditioner to let them tuning-free.
   - The stochastic analysis relies largely on universal bounds or subGaussian noise assumptions, it might be hard to obtain convergence in expectation. As shown in Khaled and Jin [2024], it is impossible to obtain tuning-free algorithms with convergence in expectation in nonconvex settings.

2. Consider proving results of schedule-free algorithms in nonconvex cases for adaptive algorithms and also verify them in LLM experiments.

References

Aaron Defazio, Harsh Mehta, Konstantin Mishchenko, Ahmed Khaled, Ashok Cutkosky, et al. The road less scheduled. *arXiv preprint arXiv:2405.15682*, 2024.

Matthew Faw, Isidoros Tziotis, Constantine Caramanis, Aryan Mokhtari, Sanjay Shakkottai, and Rachel Ward. The power of adaptivity in sgd: Self-tuning step sizes with unbounded gradients and affine variance. In *Conference on Learning Theory*, pages 313–355. PMLR, 2022.

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.

Ahmed Khaled and Chi Jin. Tuning-free stochastic optimization. *arXiv preprint arXiv:2402.07793*, 2024.

Ahmed Khaled, Konstantin Mishchenko, and Chi Jin. Dowg unleashed: An efficient universal parameter-free gradient descent method. *Advances in Neural Information Processing Systems*, 36: 6748–6769, 2023.

Yura Malitsky and Konstantin Mishchenko. Adaptive proximal gradient method for convex optimization. *arXiv preprint arXiv:2308.02261*, 2023.

Francesco Orabona. Normalized gradients for all. *arXiv preprint arXiv:2308.05621*, 2023.

Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *International conference on machine learning*, pages 71–79. PMLR, 2013.

Danqing Zhou, Shiqian Ma, and Junfeng Yang. Adabb: Adaptive barzilai-borwein method for convex optimization. *arXiv preprint arXiv:2401.08024*, 2024.