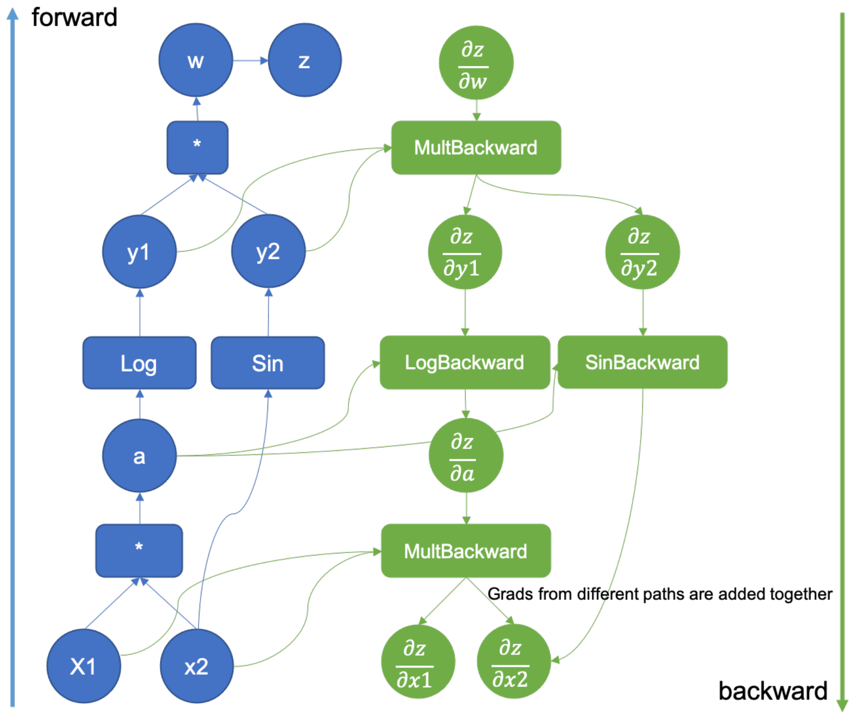
Beyond Worst-Case Analysis in Stochastic Approximation: Moment Estimation Improves Instance Complexity

Jingzhao Zhang (Tsinghua), Hongzhou Lin (Amazon), Subhro Das (IBM),

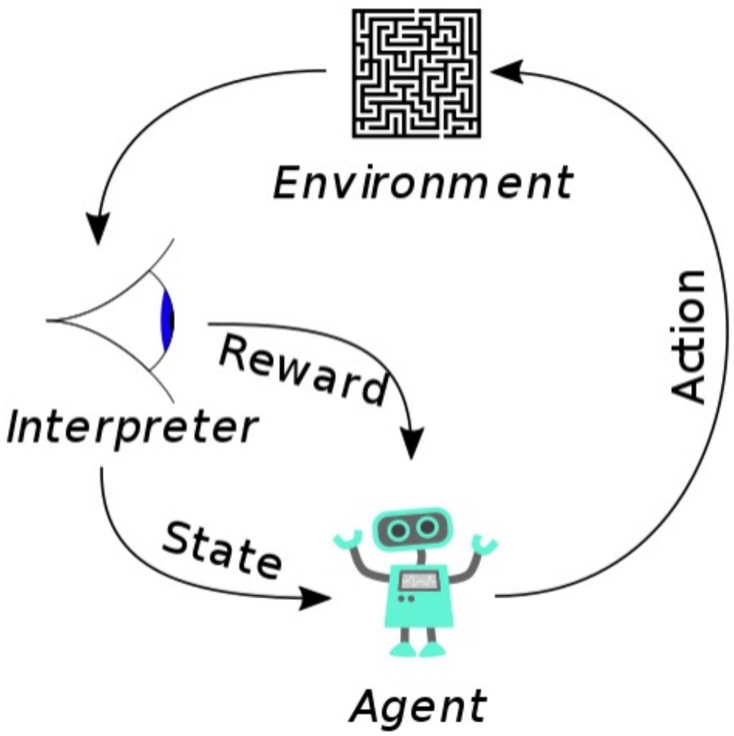
Suvrit Sra (MIT), Ali Jadbabaie (MIT)

ICML 2022

Gradient methods have many applications in modern machine learning



Neural network training



Policy Optimization

Stochastic Approximation: Let's consider a simple smooth convex stochastic approximation problem.

We want to minimize a function f :

$$
\begin{aligned}
&f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle, \\
&\|x_1 - x^*\| \leq R, \\
&\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|,
\end{aligned}
$$

Stochastic approximation assumes exogenous noise model:

$$
\begin{aligned}
&x_{k+1} = x_k - \eta_k g(x_k), \\
&g_k(x) = \nabla f(x) + \xi_k, \\
&\mathbb{E}[\xi_k] = 0, \quad \mathbb{E}[\|\xi_k\|^2] = \sigma_k^2 \leq M^2.
\end{aligned}
$$

This problem seems to be fully solved

- Minimax optimal rates are known:
  - SGD achieves minimax optimal rate:

$$f(x_T) - \min_x f(x) \leq \frac{RM}{\sqrt{T}}$$

  - Similar optimality results were also known in the nonconvex case.
- **However, SGD is often suboptimal in practice.**

[A. S. Nemirovsky and D. B. Yudin. Problem complexity and method efficiency in optimization. Wiley & Sons, 1983]
[Arjevani, Y., Carmon, Y., Duchi, J. C., Foster, D. J., Srebro, N., & Woodworth, B. (2019). Lower bounds for non-convex stochastic optimization. *arXiv preprint arXiv:1912.02365*.]

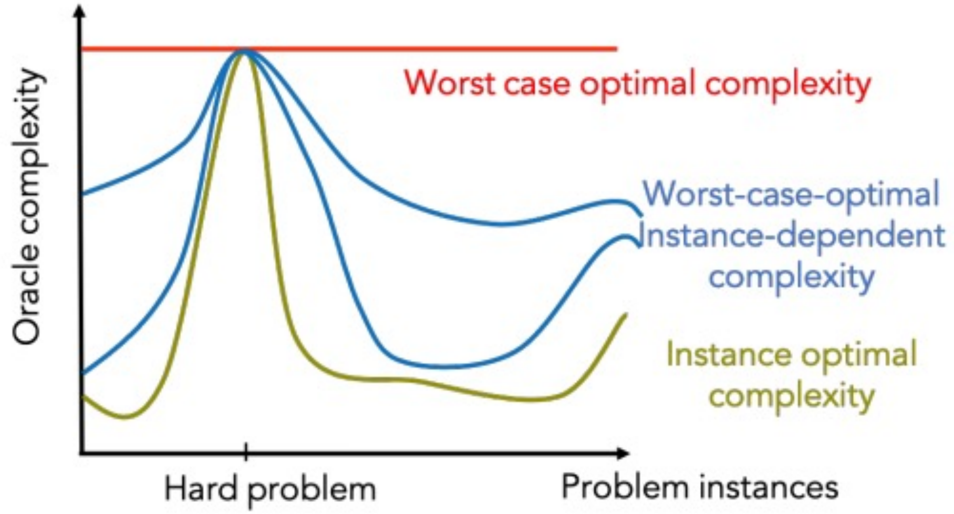
One way to close the gap might be to move beyond the worst case analysis.

Worst case function may not occur.

1. Smooth analysis [Spielman, D. A. 2005.]

2. We may assume a distribution over the problem instances. [Hoare, C. 1962; Pedregosa & Scieur, 2020; Lacotte & Pilanci, 2020; Paquette et al., 2021]

3. We may provide an instance-dependent bound. [Fagin et al., 2003; Afshani et al., 2017, Khamaru et al., 2021; Pananjady & Wainwright, 2020]



One way to close the gap might be to move beyond the worst case analysis.

Worst case function may not occur.

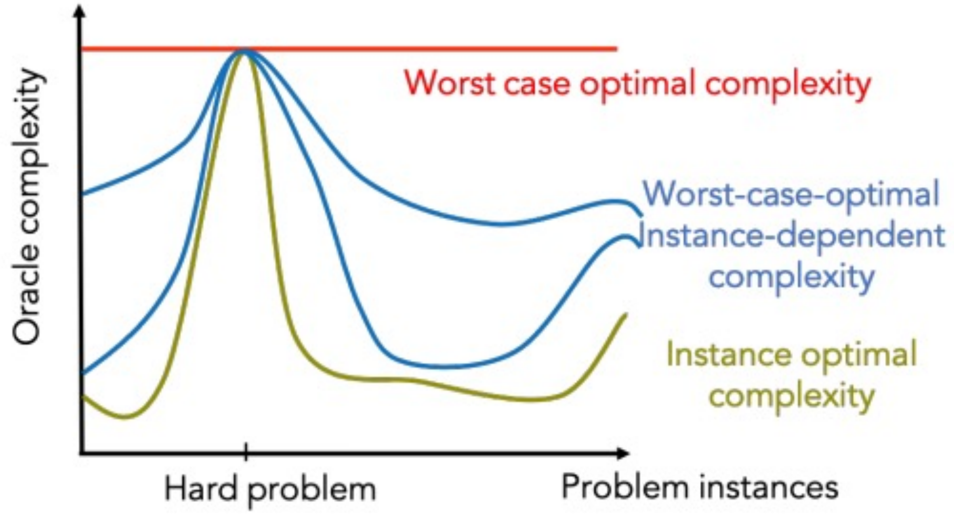
1. Smooth analysis [Spielman, D. A. 2005.]

2. We may assume a distribution over the problem instances. [Hoare, C. 1962; Pedregosa & Scieur, 2020; Lacotte & Pilanci, 2020; Paquette et al., 2021]

3. We may provide an instance-dependent bound. [Fagin et al., 2003; Afshani et al., 2017, Khamaru et al., 2021; Pananjady & Wainwright, 2020]

In our problem: we look for bounds that depend on the iteration-wise noise level

$$\mathbb{E}[\xi_k] = 0, \quad \mathbb{E}[\|\xi_k\|^2] = \sigma_k^2 \leq M^2.$$



From the view of instance-level complexity, SGD is far from optimal.

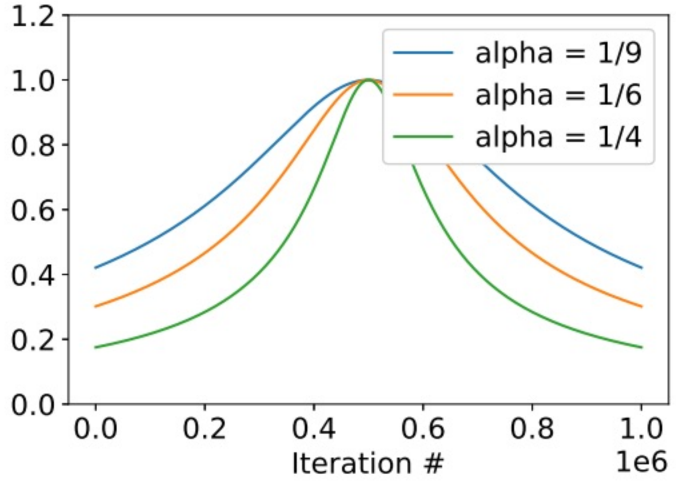
	Worst	Agnostic	Adaptive
Error bound	$\frac{2RM}{\sqrt{T}}$	$(R^2 + \frac{1}{T}\sum_k \sigma_k^2)/\sqrt{T}$	$2R\left(\frac{1}{T}\sum_{k=1}^{T}\sigma_k^2\right)^{1/2}/\sqrt{T}$
η_k	$R/\sqrt{TM^2}$	$1/\sqrt{T}$	$R/\sqrt{\sum_{k=1}^{T}\sigma_t^2}$ or $R/\sqrt{2\sum_{\tau\leq k}\|g_k\|^2}$
Can be achieved via	Fixed step, known R, M	Fixed step, unknown R, M	Fixed step, known R, $\{\sigma_k\}_k$ or Adapt. step, unknown $\{\sigma_k\}_k$

Orabona, Francesco. "A modern introduction to online learning." *arXiv preprint arXiv:1912.13213* (2019).

The gap can not be explained by absolute constants.

Mountain shape noise for different values

$$\sigma_k = \frac{1}{\sqrt{1+T^{2\alpha}\left(\frac{2k}{T}-1\right)^2}}.$$

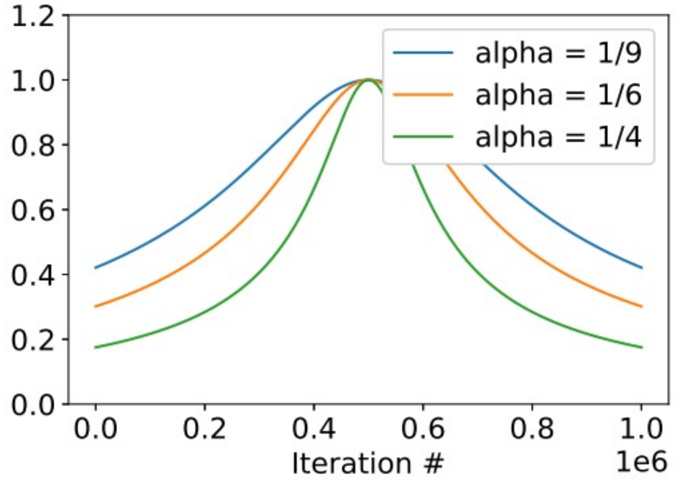


	Worst	Agnostic	Adaptive
Error bound	$T^{-\frac{1}{2}}$	$T^{-\frac{1}{2}}$	$T^{-((1+\alpha)/2)}$

The gap can not be explained by absolute constants.

Mountain shape noise for different values

$$\sigma_k = \frac{1}{\sqrt{1+T^{2\alpha}\left(\frac{2k}{T}-1\right)^2}}.$$



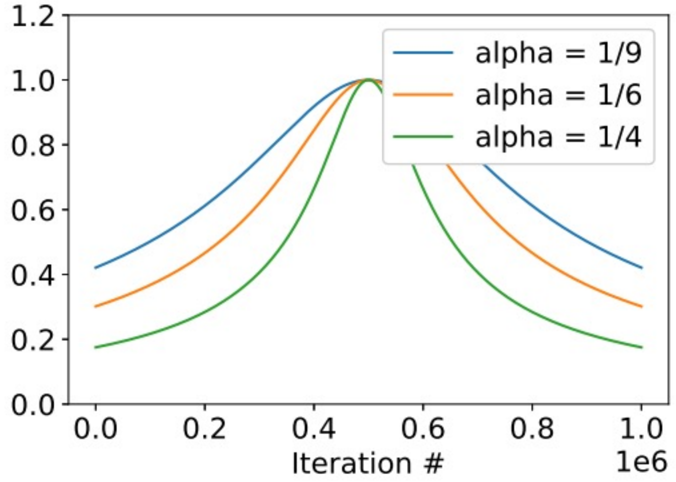
	Worst	Agnostic	Adaptive
Error bound	$T^{-\frac{1}{2}}$	$T^{-\frac{1}{2}}$	$T^{-((1+\alpha)/2)}$

Can we achieve faster convergence from this instance-dependent perspective?

The gap can not be explained by absolute constants.

Mountain shape noise for different values

$$\sigma_k = \frac{1}{\sqrt{1+T^{2\alpha}\left(\frac{2k}{T}-1\right)^2}}.$$



	Worst	Agnostic	Adaptive	Dynamic
Error bound	$T^{-\frac{1}{2}}$	$T^{-\frac{1}{2}}$	$T^{-((1+\alpha)/2)}$	$T^{-((1+2\alpha)/2)}$

Can we achieve faster convergence from this instance-dependent perspective?

From the view of instance-level complexity, SGD is far from optimal.

	Worst	Agnostic	Adaptive	Dynamic
Error bound	$\frac{2RM}{\sqrt{T}}$	$(R^2 + \frac{1}{T}\sum_k \sigma_k^2)/\sqrt{T}$	$2R\left(\frac{1}{T}\sum_{k=1}^{T}\sigma_k^2\right)^{1/2}/\sqrt{T}$	$2\mathrm{R}\left(\frac{1}{T}\sum_{k=1}^{T}\frac{1}{\sigma_k}\right)^{-1}/\sqrt{T}$
η_k	$R/\sqrt{TM^2}$	$1/\sqrt{T}$	$R/\sqrt{\sum_{k=1}^{T}\sigma_t^2}$ or $R/\sqrt{2\sum_{\tau\leq k}\|g_k\|^2}$	$R/(\sigma_k\sqrt{T})$
Can be achieved via	Fixed step, known R, M	Fixed step, unknown R, M	Fixed step, known R, $\{\sigma_k\}_k$ or Adapt. step, unknown $\{\sigma_k\}_k$	Adaptive step, known R, $\{\sigma_k\}_k$

Dynamic error bounds is better but requires knowledge of the noise level.

	Worst	Agnostic	Adaptive	Dynamic
Error bound	$\frac{2RM}{\sqrt{T}}$	$(R^2 + \frac{1}{T}\sum_k \sigma_k^2)/\sqrt{T}$	$2R\left(\frac{1}{T}\sum_{k=1}^{T}\sigma_k^2\right)^{1/2}/\sqrt{T}$	$2\text{R}\left(\frac{1}{T}\sum_{k=1}^{T}\frac{1}{\sigma_k}\right)^{-1}/\sqrt{T}$
η_k	$R/\sqrt{TM^2}$	$1/\sqrt{T}$	$R/\sqrt{\sum_{k=1}^{T}\sigma_t^2}$ or $R/\sqrt{2\sum_{\tau\leq k}\|g_k\|^2}$	$R/(\sigma_k\sqrt{T})$
Can be achieved via	Fixed step, known R, M	Fixed step, unknown R, M	Fixed step, known R, $\{\sigma_k\}_k$ or Adapt. step, unknown $\{\sigma_k\}_k$	Adaptive step, known R, $\{\sigma_k\}_k$

We can achieve this bound with moment estimation under additional regularity conditions.