浙江大學
ZHEJIANG UNIVERSITY

# On Convergence of Adam for Stochastic Optimization under Relaxed Assumptions

Yusu Hong and Junhong Lin

[NeurIPS 2024]
Poster time: Thu 12 Dec 11 a.m. PST — 2 p.m. PST
Poster place: Poster Room - TBD

# Main Contributions

## Relaxed assumptions

- Almost-surely affine variance noise;
- $(L_0, L_q)$-smoothness (generalized smoothness).

## Main results

A $\tilde{\mathcal{O}}\left(1/\sqrt{T}\right)$ rate for Adam to find a stationary point under these assumptions.

# Algorithm

**Algorithm 1** Adam

**Input:** Horizon $T$, $\boldsymbol{x}_1 \in \mathbb{R}^d$, $\beta_1, \beta_2 \in [0, 1)$, $\boldsymbol{m}_0 = \boldsymbol{v}_0 = \mathbf{0}_d$, $\eta, \epsilon > 0$, $\boldsymbol{\epsilon} = \epsilon \mathbf{1}_d$

**for** $s = 1, \cdots, T$ **do**

Draw a new sample $\boldsymbol{z}_s$ and generate $\boldsymbol{g}_s = g(\boldsymbol{x}_s, \boldsymbol{z}_s)$;

$\boldsymbol{m}_s = \beta_1 \boldsymbol{m}_{s-1} + (1 - \beta_1)\boldsymbol{g}_s$;

$\boldsymbol{v}_s = \beta_2 \boldsymbol{v}_{s-1} + (1 - \beta_2)\boldsymbol{g}_s^2$;

$\eta_s = \eta\sqrt{1 - \beta_2^s}/(1 - \beta_1^s)$, $\boldsymbol{\epsilon}_s = \boldsymbol{\epsilon}\sqrt{1 - \beta_2^s}$;

$\boldsymbol{x}_{s+1} = \boldsymbol{x}_s - \eta_s \cdot \boldsymbol{m}_s / \left(\sqrt{\boldsymbol{v}_s} + \boldsymbol{\epsilon}_s\right)$;

**end for**

The two corrective terms in [Kingma and Ba, 2015] are incorporated into $\eta_s$.

[Kingma and Ba, 2015]. Adam: A method for stochastic optimization, ICLR 2015.

# Preliminary

**Unconstrained optimization**

$$\min_{x \in \mathbb{R}^d} f(x) = \mathbb{E}_{\xi \sim \mathcal{P}}[f(x, \xi)].$$

**Standard assumptions**

- Bounded below: $f(x) \geq f^* > 0, \forall x \in \mathbb{R}^d;$
- Unbiased estimator: $\mathbb{E}[g(x) \mid x] = \nabla f(x), \forall x \in \mathbb{R}^d.$

# Almost-surely Affine Variance Noise

$$\|g(x) - \nabla f(x)\|^2 \leq \sigma_0^2 + \sigma_1^2 \|\nabla f(x)\|^2, a.s., \quad \forall x \in \mathbb{R}^d.$$

**Remark:**

- Could be extended to sub-Gaussian type;
- Covering two standard noise types: **bounded noise** and **sub-Gaussian noise;**
- Examples: robust linear regression, multi-layer network with perturbed by noise [Bottou et al., 2018], [Faw et al., 2022];

Bottou et al., 2018. Optimization methods for large-scale machine learning, SIAM Review 2018.
Faw et al., 2022. The Power of Adaptivity in SGD: Self-Tuning Step Sizes with Unbounded Gradients and Affine Variance, COLT 2022.

# $(L_0, L_q)$-Smoothness

$(L_0, L_q)$-smoothness [Zhang et al., 2020]: for $q \in [0, 2)$,

$$\|\nabla f(y) - \nabla f(x)\| \leq (L_0 + L_q \|\nabla f(x)\|^q)\|y - x\|, \forall \|y - x\| \leq 1/L_q.$$

**Remark:**

- $(L_0, L_q)$-smoothness implies $L$-smoothness;
- Practical examples: $x^k, P(x)/Q(x), a^{b^x}$;
- Objective functions in training language models satisfies $(L_0, L_q)$-smoothness.

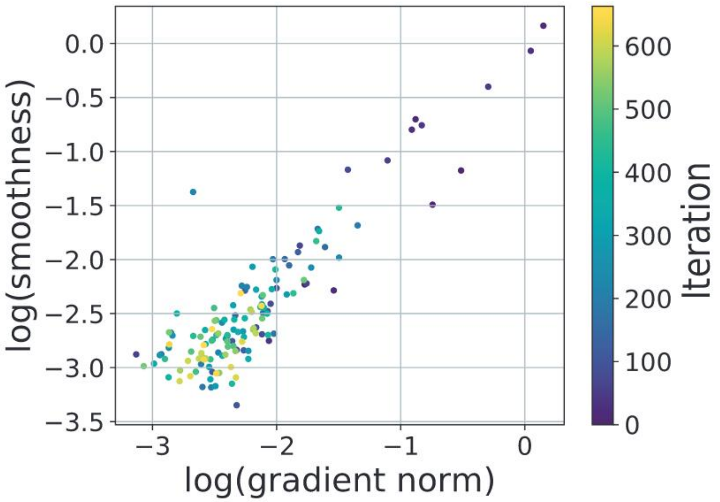


Figure 1 [Zhang et al, 2020]. Gradient norm vs local gradient Lipschitz constant on a log-scale along the training trajectory for AWD-LSTM.

Zhang et al, 2020. Why gradient clipping accelerates training: A theoretical justification for adaptivity, ICLR 2020.

# Probabilistic Convergence Rate (I)

Theorem 1. Let $f$ be $L$-smooth and the almost-surely affine variance noise holds. If

$$0 \leq \beta_1 < \beta_2 < 1, \quad \beta_2 \sim 1 - \mathcal{O}(1/T), \quad \eta, \epsilon \sim \mathcal{O}\left(1/\sqrt{T}\right),$$

then, with probability at least $1 - \delta$,

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla f(x)\|^2 \lesssim \tilde{\mathcal{O}}\left(\frac{1}{\sqrt{T}}\right).$$

Remark: the convergence does not require the information of $L$ to tune $\eta$.

# Probabilistic Convergence Rate (II)

Theorem 2. Let $f$ be $(L_0, L_q)$-smooth and the almost-surely affine variance noise holds. If

$$0 < \beta_1 \leq \beta_2 \leq 1, \quad \beta_2 \sim 1 - \mathcal{O}(1/T), \quad \eta \sim \mathcal{O}\left(\frac{1}{\mathrm{poly}(\log T)\sqrt{T}}\right), \quad \epsilon \sim \mathcal{O}(1/\sqrt{T}),$$

then, with probability at least $1 - \delta$

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla f(x)\|^2 \lesssim \tilde{\mathcal{O}}\left(\frac{1}{\sqrt{T}}\right).$$

Remark: the convergence requires the information of $L_0$, $L_q$ to tune $\eta$.

# Comparison with Existing Works

| | FCT | Noise | Smooth | Conv. Rate | Conv. Type |
|---|---|---|---|---|---|
| [ZRSKK18] | ✗ | Bounded | $L$ | $\frac{1}{T}+\sigma_0^2$ | $\mathbb{E}$ |
| [CLSH19] | ✗ | Bounded | $L$ | $\frac{1}{\sqrt{T}}$ | $\mathbb{E}$ |
| [ZSJZL19] | ✗ | - | $L$ | $\frac{1}{\sqrt{T}}$ | $\mathbb{E}$ |
| [DMU18] | ✗ | - | $L$ | $\frac{1}{\sqrt{T}}$ | $\mathbb{E}$ |
| [SLHS20] | ✗ | Finite Sum Affine | $L$ | - | $\mathbb{E}$ |
| [DBBU20] | ✗ | Bounded | $L$ | $\frac{1}{\sqrt{T}}$ | $\mathbb{E}$ |
| [ZCSSL22] | ✗ | Finite Sum Affine | $L$ | $\frac{1}{\sqrt{T}}$ | $\mathbb{E}$ |
| [LJR23] | ✓ | Sub-Gaussian | $\boldsymbol{(L_0, L_q)}$ | $\frac{1}{\sqrt{T}}$ | **w.h.p.** |
| [WFZZC23] | ✗ | Coordinate-wise Affine | $L$ | $\frac{1}{\sqrt{T}}$ | $\mathbb{E}$ |
| [HL23] | ✓ | Coordinate-wise Affine | $L$ | $\frac{1}{\sqrt{T}}$ | **w.h.p.** |
| Thm.1 | ✓ | Affine | $L$ | $\frac{1}{\sqrt{T}}$ | **w.h.p.** |
| Thm.2 | ✓ | Affine | $\boldsymbol{(L_0, L_q)}$ | $\frac{1}{\sqrt{T}}$ | **w.h.p.** |

- "FCT" refers to "full corrective terms" in vanilla Adam.
- "w.h.p." refers to high probabiltiy convergence.

# Conclusions

- Convergence of Adam on non-convex landscape;
- Almost surely affine variance noise;
- Smoothness and generalized smoothness;
- A $\tilde{\mathcal{O}}\left(1/\sqrt{T}\right)$ rate for Adam to find a stationary point.

# Thank you for listensing!

Poster time: Thu 12 Dec 11 a.m. PST — 2 p.m. PST
Poster place: Poster Room - TBD