Stab-SGD: Noise-Adaptivity in Smooth Optimization with Stability Ratios

David A. R. Robin
Killian Bakong
Kevin Scaman

Paper link: https://neurips.cc/virtual/2025/poster/118726

INRIA (ARGO) - École Normale Supérieure de Paris
PSL Research University

Smooth Optimization with Stochastic Gradients

Optimize an objective $\mathcal{L} : \mathbb{R}^d \to \mathbb{R}$
Assumption: β-smoothness $\|\nabla\mathcal{L}(y) - \nabla\mathcal{L}(x)\|_2 \leq \beta \|y - x\|_2$.

By a stochastic gradient descent algorithm $\theta : \mathbb{N} \to \mathbb{R}^d$

$$\theta_{t+1} = \theta_t - \eta_t \cdot G_{t+1}$$

with

unbiased gradients	$\mathbb{E}[G_{t+1}] = \nabla\mathcal{L}(\theta_t)$
controlled noise	$\mathbb{E}[\|G_{t+1} - \nabla\mathcal{L}(\theta_t)\|_2^2] \leq \sigma^2 < +\infty$

Typical results[1]:

$$\eta_t = \frac{1}{2}\frac{D_0\sigma^{-1}}{\sqrt{t+1}} \quad \Rightarrow \quad \mathbb{E}\left[\mathcal{L}\left(\bar{\theta}_t\right)\right] - \inf\mathcal{L} \leq 5\, D_0\sigma\,\frac{\log(T+1)}{\sqrt{T+1}}$$

[1] Garrigos & Gower (2023) Thm 5.7 with $\gamma_0 = D_0/(2\sigma)$ if $\sigma > 2\beta D_0$.

Selection of step-size schedule

Under weak convexity with high noise

$$\mathbb{E}\left[\mathcal{L}\left(\bar{\theta}_T\right)\right] - \inf \mathcal{L} \leq \frac{D_0^2}{\eta \cdot T} + 2\eta \cdot \sigma^2$$

Prescription: set $\eta_t = \frac{1}{\sqrt{2}} D_0 \sigma^{-1} / \sqrt{T+1}$ and reach ε in $\mathcal{O}(1/\varepsilon^2)$

Under μ-strong convexity[2]

$$\mathbb{E}\left[\mathcal{L}\left(\theta_T\right)\right] - \inf \mathcal{L} \leq D_0^2 \left(1 - \eta\mu\right)^T + \frac{2\eta}{\mu}\sigma^2$$

Don't average iterates, set $\eta_t = \min\left\{\varepsilon \frac{\mu}{4\sigma^2}, \frac{1}{2\beta}\right\}$ get $\tilde{\mathcal{O}}(1/\varepsilon)$

If the problem is *known* easy → rates are fast.
If it's easy but we don't know → slow rates.

---

[2] Garrigos & Gower (2023) Thm 5.8

Noise-adaptivity

The β-smooth upper bound on loss is

$$\mathcal{L}(\theta_{t+1}) \leq \mathcal{L}(\theta_t) - \eta_t \cdot (\nabla\mathcal{L}(\theta_t) \cdot G_{t+1}) + \frac{\beta}{2}\eta_t^2 \left\|G_{t+1}\right\|_2^2$$

Minimize $\mathbb{E}\left[\,\mathcal{L}(\theta_{t+1}) \mid \theta_t\,\right]$ by setting $\eta_t = \dfrac{1}{\beta}\dfrac{\|\,\mathbb{E}\left[\,G_{t+1} \mid \theta_t\,\right]\|_2^2}{\mathbb{E}\left[\,\|G_{t+1}\|_2^2 \mid \theta_t\,\right]}$

Stab-SGD: use step $\eta_t = \dfrac{1}{\beta}\,\mathrm{Stab}\left(G_{t+1} \mid \theta_t\right)$

For a random variable X with mean $\mu \in \mathbb{R}^d$ and variance $\sigma^2 < \infty$

$$\mathrm{Stab}\left(X\right) = \frac{\|\,\mathbb{E}\left[X\right]\|_2^2}{\mathbb{E}\left[\,\|X\|_2^2\,\right]} = \frac{\|\mu\|_2^2}{\|\mu\|_2^2 + \sigma^2}$$

Convergence Rate of Stab-SGD

Table: Rates under affine variance $\mathbb{V}\left[\, G_{t+1} \,|\, \mathcal{F}_t \,\right] \leq \alpha \|\nabla \mathcal{L}(\theta_t)\|_2^2 + \sigma^2$.

	$\mathbb{E}\left[\, \mathcal{L}(\theta_{T+1}) \,\right] - \mathcal{L}^\star$		$\mathbb{E}\left[\, \frac{1}{T} \sum_t \|\nabla \mathcal{L}(\theta_t)\|_2^2 \,\right]$
	Convex	μ-strongly convex	Non-convex
$\sigma^2 = 0$	$\mathcal{O}\left(T^{-1}\right)$	$\mathcal{O}\left(\exp\left(-\frac{1}{1+\alpha}\frac{\mu}{\beta}T\right)\right)$	$\mathcal{O}\left(T^{-1}\right)$
$\sigma^2 > 0$	$\mathcal{O}\left(T^{-1/3}\right)$	$\mathcal{O}\left(T^{-1}\right)$	$\mathcal{O}\left(T^{-1/2}\right)$

Precise non-asymptotic rates

Convex case: If $\mathcal{L}$ is convex (resp μ-strongly convex),
It holds $\mathbb{E}\left[\,\mathcal{L}(\theta_{T+1})\,\right] \leq (\inf \mathcal{L}) + \varepsilon$ if

$$T \geq \frac{2}{3}\frac{\beta D_0^4 \sigma^2}{\varepsilon^3} + (1+\alpha)\frac{\beta D_0^2}{\varepsilon}$$

$$T \geq \frac{\sigma^2 \beta}{2\mu^2 \varepsilon} + (1+\alpha)\frac{\beta}{\mu}\log\left(\frac{\Delta_0}{\varepsilon}\right)$$

$D_0^2 = \mathbb{E}\left[\,\|\theta_0 - \theta^\star\|_2^2\,\right]$, $\Delta_0 = \mathbb{E}\left[\,\mathcal{L}(\theta_0)\,\right] - (\inf \mathcal{L})$, $\mathcal{L}(\theta^\star) = \inf \mathcal{L}$.

Non-convex case: with $\Delta_0 = \mathbb{E}\left[\,\mathcal{L}(\theta_0)\,\right] - (\inf \mathcal{L})$, for all $T \in \mathbb{N}$

$$\mathbb{E}\left[\frac{1}{T}\sum_{t<T}\|\nabla \mathcal{L}(\theta_t)\|_2^2\right] \leq (1+\alpha)\frac{2\beta\Delta_0}{T} + \sqrt{\frac{2\beta\Delta_0\sigma^2}{T}}$$

Estimating stability ratios

Jackknife estimator of $\mathrm{Stab}(X)$ from iid samples $(X_i \in \mathbb{R}^d)_{i<n}$

$$R_n = \frac{1}{n-1} \frac{\sum_i \sum_{j \neq i} \langle X_i, X_j \rangle}{\sum_i \|X_i\|_2^2}$$

This estimator is consistent $\mathbb{E}\left[(R_n - R_\star)^2\right] \underset{n \to +\infty}{\longrightarrow} 0$

With kurtosis $\kappa = \dfrac{\mathbb{E}\left[\|X\|_2^4\right]}{\mathbb{E}\left[\|X\|_2^2\right]^2}$ and stability ratio $R_\star = \mathrm{Stab}(X)$

$$\mathbb{E}\left[\left(\frac{R_n - R_\star}{R_\star}\right)^2\right] \leq R_\star^{-1} \frac{44 + 4\kappa}{n-1} + R_\star^{-2} \exp\left(-\frac{n}{8\kappa}\right)$$

ResNet-56 experiment on CIFAR-10

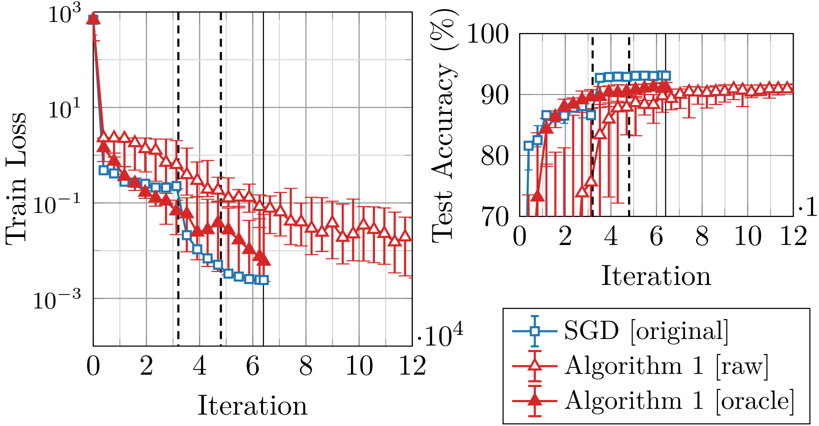

Figure: ResNet-56 on CIFAR-10. Medians and quartiles (20 seeds).

Takeaway: Stability Ratios + Stochastic KŁ integration

- Problem *known easy* vs just *easy* is very different.
  - Noise-adaptive algorithm are possible
  - Shrink step-size by gradient stability ratio

$$\eta_t = \mathrm{Stab}\left(G_{t+1} \mid \theta_t\right)/\beta$$

- Estimating Stability Ratio: $\mathrm{Stab}(X) = \|\,\mathbb{E}[X]\,\|_2^2 / \mathbb{E}\left[\|X\|_2^2\right]$
  - Possible from gradient samples only.
  - Is challenging at low stability.
  - But even simple estimators give good results.
- Convergence Analysis
  - Integration of Kurdyka-Łojasiewicz inequality
  - Pairs well with stochasticity / affine variance (+others).

Stab-SGD: Noise-Adaptivity in Smooth Optimization with Stability Ratios

David A. R. Robin
Killian Bakong
Kevin Scaman

Paper link: https://neurips.cc/virtual/2025/poster/118726

INRIA (ARGO) - École Normale Supérieure de Paris
PSL Research University