Step-Size Stability in Stochastic Optimization: A Theoretical Perspective

Fabian Schaipp [1] Robert M. Gower [2] Adrien Taylor [1]

[1]Inria, ENS, PSL Research University, Paris [2]CCM, Flatiron Institute, New York

TLDR

Adaptive methods need **much less tuning** for the learning rate than SGD. We introduce a single quantity, the **stability index** δ_t, whose scaling in the step size explains this, and **prove** that these methods are always at least as stable as SGD.

Problem setup and motivation

We consider the problem

$$\min_{x\in\mathbb{R}^d} f(x), \qquad f(x) := \mathbb{E}_s\big[f(x,s)\big].$$

Assumption (poster-only): every $x \mapsto f(x,s)$ is convex (can be extended to weakly convex).

Let $g_t \in \partial f(x_t, s_t)$ be a subgradient and $\alpha > 0$ a step size. Then

$$x_{t+1} = x_t - \alpha g_t. \tag{SGD}$$

- The step size α is the **most important hyperparameter** to tune.
- While **empirically** it is well documented that adaptive step-size methods (see below) need less tuning for α, [Schaipp et al., 2024, Islamov et al., 2025], existing theory does not focus on stability.
- **Goal:** a theoretical framework that *quantifies and explains* this stability gap.

Four stochastic methods

Let $g_t \in \partial f(x_t, s_t)$ be a subgradient and $\alpha > 0$ a user-specified step size.

$$\texttt{SPS}: \; x_{t+1} = x_t - \tau_t g_t, \quad \tau_t = \min\Big\{\alpha, \; \frac{f(x_t,s_t)-C_{s_t}}{\|g_t\|^2}\Big\}$$

$$\texttt{NGN}: \; x_{t+1} = x_t - \gamma_t g_t, \quad \gamma_t = \frac{\alpha}{1+\frac{\alpha}{2f(x_t,s_t)}\|g_t\|^2}$$

$$\texttt{SPP}: \; x_{t+1} = \arg\min_{y\in\mathbb{R}^d} f(y,s_t) + \frac{1}{2\alpha}\|y-x_t\|^2$$

Here $C_{s_t} \le \inf f(\cdot, s_t)$ is a (known) lower bound; in machine learning, we can usually set $C_{s_t} = 0$ (non-negative losses).

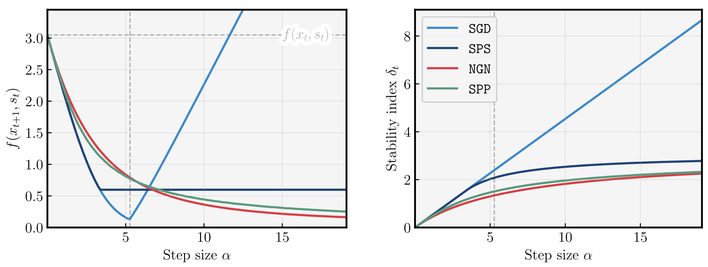

Figure 1. Toy example. **Left:** next-iterate loss vs. step size α. **Right:** the stability index δ_t as a function of α. For large α, stable loss values coincide with a benign ($\approx$ sub-linear) growth of δ_t.

How to quantify stability?

All four methods can be analyzed in a joint framework (stochastic proximal point on a *model* of the loss, see below). Within this framework, we show that a single quantity derived from that model — the **stability index** — governs how suboptimality degrades as α grows.

Model-based framework

Given a model $f_{x_t}(\cdot, s_t)$ of $f(\cdot, s_t)$ around x_t, the update of model-based stochastic proximal point [Asi and Duchi, 2019, Davis and Drusvyatskiy, 2019] is

$$x_{t+1} = \arg\min_{y\in\mathbb{R}^d} f_{x_t}(y, s_t) + \tfrac{1}{2\alpha_t}\|y - x_t\|^2. \tag{1}$$

SGD $\to$ *linear* model,	SPS $\to$ *truncated* model,
NGN $\to$ *square-root* model,	SPP $\to$ *exact* model.

The stability index is defined as follows:

$$\delta_t := f(x_t, s_t) - f_{x_t}(x_{t+1}, s_t) - \tfrac{1}{2\alpha_t}\|x_{t+1} - x_t\|^2. \tag{2}$$

Stability steers suboptimality

Convex case. Let $x_\star$ be a solution and $D := \|x_1 - x_\star\|$. Then,

$$\mathbb{E}[f(x_T) - f(x_\star)] \;\le\; \underbrace{\frac{D^2}{2\sum_t \alpha_t}}_{\text{bias}} + \underbrace{\frac{\sum_t \alpha_t \,\mathbb{E}[\delta_t]}{\sum_t \alpha_t}}_{\text{driven by stability}} + V_T =: \Omega_T,$$

where $V_T := \sum_{k=1}^{T-1} \frac{\alpha_k}{\sum_{t=k+1}^T \alpha_t}\Big(\frac{1}{\sum_{t=k}^T \alpha_t}\sum_{t=k}^T \alpha_t \mathbb{E}[\delta_t]\Big)$.

Takeaway

The stability index directly impacts the suboptimality bound. Slower scaling of δ_t with $\alpha \to +\infty$ corresponds to higher robustness (see Fig. 2).

Illustration. Fixing $\alpha_t = \alpha$ and assuming $\Delta := \mathbb{E}[\delta_t] \propto \alpha^\nu$, the bound becomes $\Omega_T \approx \frac{D^2}{2\alpha T} + \alpha^\nu\,[\,1 + \ln T\,]$.

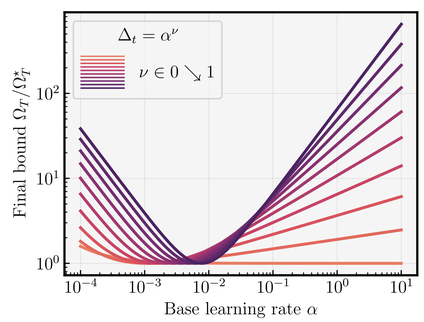

Figure 2. Scaling of stability index with α affects the bound Ω_T.

Stability index for each method

We derive δ_t for each of the methods:

$$\delta_t^{\texttt{SGD}} = \tfrac{\alpha_t}{2}\|g_t\|^2 \qquad (\text{linear in } \alpha)$$

$$\delta_t^{\texttt{SPS}} = \tau_t\big(1 - \tfrac{\tau_t}{2\alpha_t}\big)\|g_t\|^2 \le \min\big\{\alpha_t\|g_t\|^2,\; f(x_t,s_t) - C_{s_t}\big\} \qquad (\le \delta_t^{\texttt{SGD}})$$

$$\delta_t^{\texttt{NGN}} = \tfrac{\gamma_t}{2}\|g_t\|^2 \qquad (\le \delta_t^{\texttt{SGD}})$$

$$\delta_t^{\texttt{SPP}} \le \min\big\{\tfrac{\alpha_t}{2}\|g_t\|^2,\; f(x_t,s_t) - \inf f(\cdot, s_t)\big\} \qquad (\le \delta_t^{\texttt{SGD}})$$

Main result

$\delta_t^{\texttt{SPS}} \le \delta_t^{\texttt{SGD}}$ and $\delta_t^{\texttt{NGN}} \le \delta_t^{\texttt{SGD}}$ and $\delta_t^{\texttt{SPP}} \le \delta_t^{\texttt{SGD}}$ **always**. For SPS/SPP the robustness gains are determined by **interpolation** ($\mathbb{E}_s[f(x_\star) - \inf f(\cdot,s)]$) and the **lower-bound estimation** error ($\mathbb{E}_s[\inf f(\cdot,s) - C_s]$).

Beyond convexity

The same role for δ_t holds in the weakly convex setting: we can bound the Moreau-envelope (a measure for near-stationarity) by a bias term and a stability term with $\mathbb{E}[\delta_t]$.

Experiments

The bound Ω_T qualitatively predicts the **range of well-performing learning rates**, even on non-convex tasks where the theory does not strictly apply.

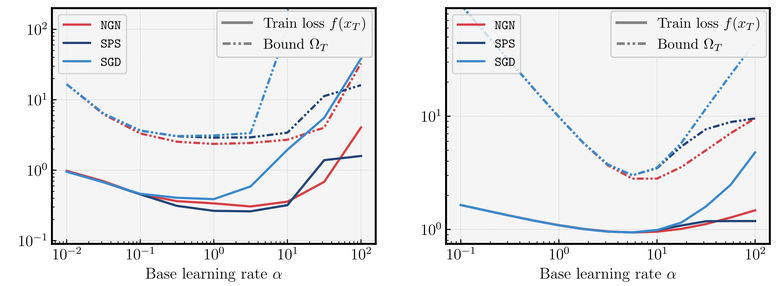

Figure 3. **(Left)** ResNet20 on CIFAR10. SPS and NGN are stable up to $\alpha \approx 10$. SGD degrades for $\alpha \gtrsim 1$; the bound (*dashed*) tracks this. **(Right)** Logistic regression on vowel dataset.

References

Hilal Asi and John C. Duchi. Stochastic (approximate) proximal point methods: convergence, optimality, and adaptivity. *SIAM Journal on Optimization*, 2019.

Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 2019.

Rustem Islamov, Niccolò Ajroldi, Antonio Orvieto, and Aurelien Lucchi. Enhancing optimizer stability: Momentum adaptation of the NGN step-size. In *NeurIPS*, 2025.

Nicolas Loizou, Sharan Vaswani, Issam Hadj Laradji, and Simon Lacoste-Julien. Stochastic Polyak step-size for SGD: An adaptive learning rate for fast convergence. In *AISTATS*, 2021.

Antonio Orvieto and Lin Xiao. An adaptive stochastic gradient method with non-negative Gauss-Newton stepsizes. 2024.

Fabian Schaipp, Ruben Ohana, Michael Eickenberg, Aaron Defazio, and Robert M. Gower. MoMo: Momentum models for adaptive learning rates. In *ICML*, 2024.

Link to paper