Stochastic Polyak Step-size for SGD: An Adaptive Learning Rate for Fast Convergence

Nicolas Loizou[1] Sharan Vaswani[2] Issam Laradji[3] Simon Lacoste-Julien[1]

[1]Mila, UdeM [2] University of Alberta [3] McGill, Element AI

The Problem

$$\min_{x\in\mathbb{R}^d}\left[f(x)=\frac{1}{n}\sum_{i=1}^{n}f_i(x)\right]. \quad (1)$$

We assume f and f_i are smooth functions.
Depending on the model under study, the functions f and f_i can either be strongly-convex, convex, or non-convex.

- $\mathcal{X}^*\subset\mathbb{R}^d$ to be the set of optimal points x^* of (1) ($\mathcal{X}^*\neq\emptyset$)
- f^* : minimum value of f
- For each $i\in\{1,\dots,n\}$: $f_i^*\stackrel{\text{def}}{=}\inf_x f_i(x)$

SGD and the Stochastic Polyak Step-size

$$\textbf{SGD:}\quad x^{k+1}=x^k-\gamma_k\nabla f_i(x^k) \quad (2)$$

Example $i\in[n]$ is chosen uniformly at random and $\gamma_k>0$ is the step-size. For step-size we propose to use the:

Stochastic Polyak Step-size (SPS):

$$\text{SPS:}\quad \gamma_k=\frac{f_i(x^k)-f_i^*}{c\,\|\nabla f_i(x^k)\|^2} \quad (3)$$

and its more conservative variant:

$$\text{SPS}_{\max}:\quad \gamma_k=\min\left\{\frac{f_i(x^k)-f_i^*}{c\|\nabla f_i(x^k)\|^2},\gamma_b\right\} \quad (4)$$

Here $\gamma_b>0$ is a bound that restricts SPS from being very large and is essential to ensure convergence to a small neighborhood around the solution. If $\gamma_b=\infty$ then $\text{SPS}_{\max}$ is equivalent to SPS.

Upper and Lower Bounds of SPS

If functions f_i in problem (1) are μ_i-strongly convex and L_i-smooth, then:

$$\frac{1}{2cL_{\max}}\leq\frac{1}{2cL_i}\leq\gamma_k=\frac{f_i(x^k)-f_i^*}{c\|\nabla f_i(x^k)\|^2}\leq\frac{1}{2c\mu_i}, \quad (5)$$

where $L_{\max}=\max\{L_i\}_{i=1}^n$.

Main Contributions

- We propose a novel adaptive learning rate for SGD: Stochastic Polyak Step-size (SPS), which is a stochastic variant of the classical Polyak step-size (for GD) (Polyak, 1987).
  Attractive choice for typical modern machine learning applications.
- Convergence guarantees of SGD with SPS: Strongly convex, Convex and Non-convex functions.
- Our results require very weak assumptions. In particular, we *do not* assume bounded second moment of the gradients for every x or bounded variance. We rely on the Optimal Objective Difference (see (6)).
- Novel analysis for constant step-size SGD.
- For Over-parametrized models (Interpolation Condition is satisfied), we guarantee: fast convergence to the true solution (like deterministic GD).
- Extensive experimental evaluation.

Main Assumption

Finite optimal objective difference

$$\sigma^2\stackrel{\text{def}}{=}\mathbb{E}_i[f_i(x^*)-f_i^*]=f(x^*)-\mathbb{E}_i[f_i^*]<\infty \quad (6)$$

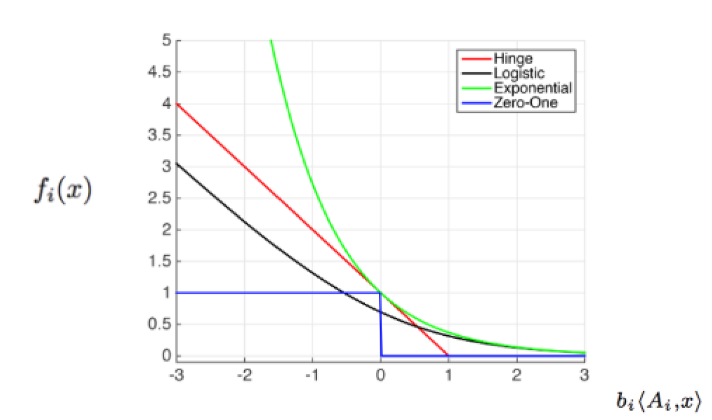
This is a very weak assumption. Moreover when (1) is the training problem of an over-parametrized model, each individual loss function f_i attains its minimum at x^*, and thus $f_i(x^*)-f_i^*=0$. In this *interpolation* setting, it follows that $\sigma=0$.

Comparison to the variance $z^2=\mathbb{E}[\|\nabla f_i(x^*)\|^2]$.
If we assume that all function f_i are μ-strongly convex and L-smooth functions then, $\frac{1}{2L}z^2\leq\sigma^2\leq\frac{1}{2\mu}z^2$.

Evaluating f_i^*

Standard unregularized surrogate loss functions have $f_i^*=0$ (Bartlett et al., 2006). Examples:

- squared loss for regression,
- logistic loss for classification,
- exponential loss (Adaboost algorithm),
- hinge loss (support vector machines)



For the regularized case (e.g. ℓ_2 regularization): f_i^* can be pre-computed in closed form for each i using:

- Lambert W function (Corless et al., 1996).
- or the more general r-Lambert function (Mezo and Baricz, 2017).

Convergence Analysis

Theorem

Let f_i be L_i-smooth convex functions with at least one of them being a strongly convex function. SGD with $\text{SPS}_{\max}$ with $c\geq 1/2$ converges as:

$$\mathbb{E}\|x^k-x^*\|^2\leq(1-\bar{\mu}\alpha)^k\|x^0-x^*\|^2+\frac{2\gamma_b\sigma^2}{\bar{\mu}\alpha}, \quad (7)$$

where $\alpha:=\min\{\frac{1}{2cL_{\max}},\gamma_b\}$, $\bar{\mu}=\mathbb{E}[\mu_i]$ and $L_{\max}=\max\{L_i\}_{i=1}^n$. The best convergence rate and the tightest neighborhood are obtained for $c=1/2$.

Corollaries:

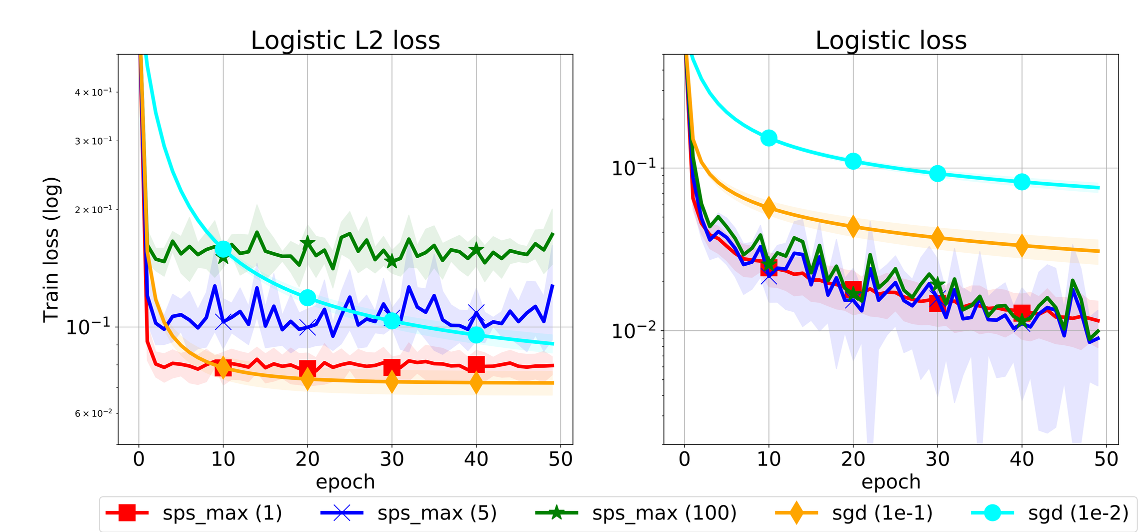
- Assume interpolation ($\sigma=0$). SGD with SPS with $c=1/2$ converges as: $\mathbb{E}\|x^k-x^*\|^2\leq\left(1-\frac{\bar{\mu}}{L_{\max}}\right)^k\|x^0-x^*\|^2$.
- If $\gamma_b\leq\frac{1}{L_{\max}}\Rightarrow$ then method becomes SGD with constant step-size $\gamma\leq\frac{1}{L_{\max}}$ and converges as

$$\mathbb{E}\|x^k-x^*\|^2\leq(1-\bar{\mu}\gamma)^k\|x^0-x^*\|^2+\frac{2\sigma^2}{\bar{\mu}}.$$

Summary of Convergence Analysis Results

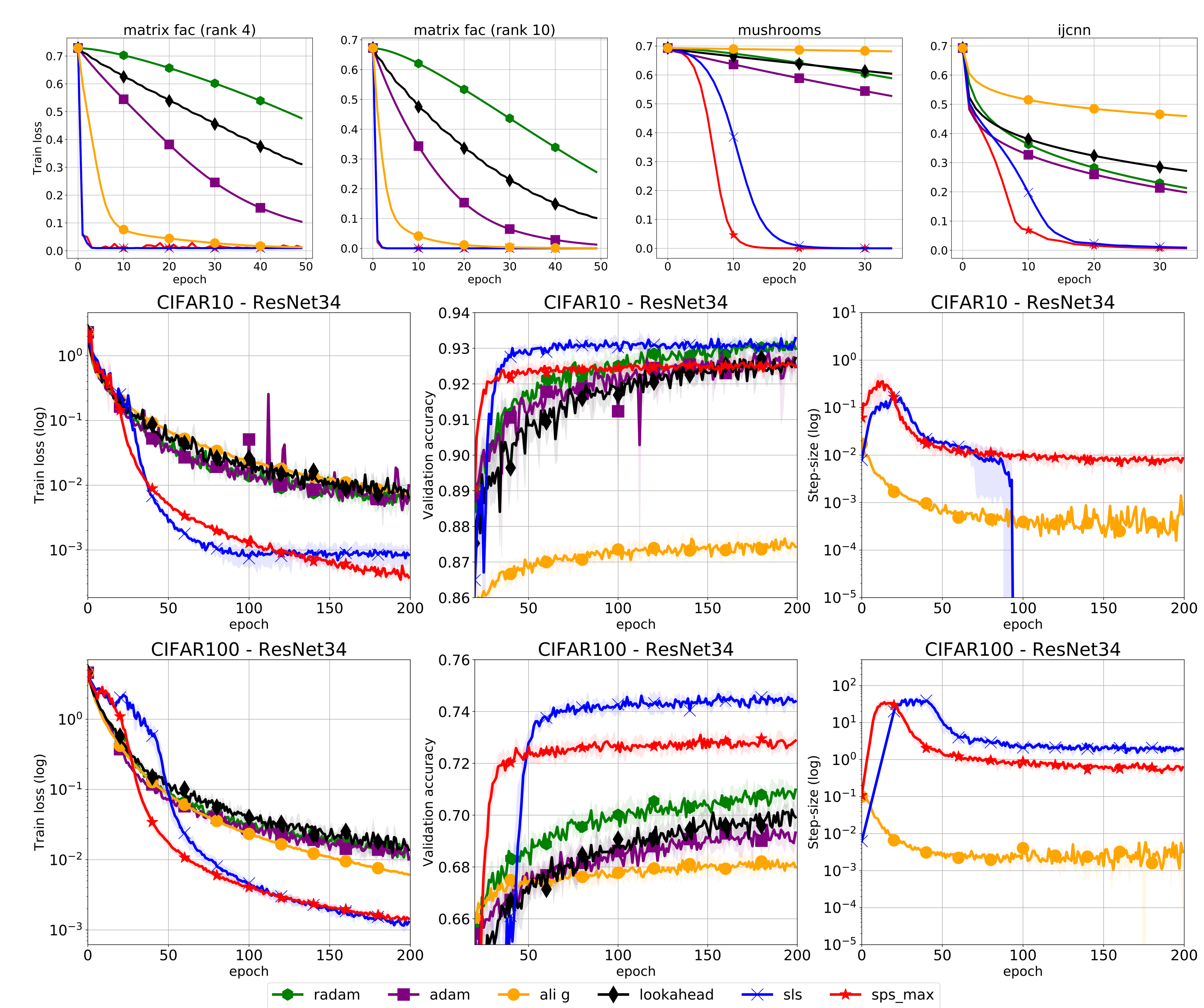
Assumptions	Quantity	Convergence	Neighborhood
Strongly Convex	$\mathbb{E}\|x^k-x^*\|^2$	Linear	$\propto\gamma_b,\sigma^2$
Convex	$\mathbb{E}\left[f(\bar{x}^k)-f(x^*)\right]$	sublinear: $\mathcal{O}(1/k)$	$\propto\gamma_b,\sigma^2$
Polyak-Lojasiewicz (PL)	$\mathbb{E}[f(x^k)-f(x^*)]$	Linear	$\propto\gamma_b,\sigma^2$
Non-Convex $\mathbb{E}[\|\nabla f_i(x)\|^2]\leq\rho\|\nabla f(x)\|^2+\delta$	$\min\mathbb{E}\|\nabla f(x^k)\|^2$	sublinear: $\mathcal{O}(1/k)$	$\propto\gamma_b,\delta$

Experimental Evaluation: Synthetic experiment



Synthetic experiment to benchmark SPS against constant step-size SGD for binary classification using the *(left)* regularized and *(right)* unregularized logistic loss.

Experiments for over-parametrized models



Comparing the performance of optimizers on deep matrix factorization *(top left)* and binary classification using kernels *(top right)* and multi-class classification on CIFAR-10 and CIFAR-100 with ResNet34.