

Study notes on “Stochastic Polyak Step-size, a simple step-size tuner with optimal rates” by F. Pedregosa

Shuvomoy Das Gupta

March 28, 2024

Here are my study notes for Fabian Pedregosa’s amazing blog on Stochastic Polyak Step-size; the full citation of Pedregosa’s blog is: *Stochastic Polyak Step-size, a simple step-size tuner with optimal rates*, Fabian Pedregosa, 2023 available at <https://fa.bianp.net/blog/2023/sps/>.

Contents

<i>Problem setup</i>	1
<i>Notation</i>	1
<i>Stochastic Gradient Descent with Polyak Stepsize</i>	1
<i>Assumptions</i>	2
<i>Convergence analysis</i>	2

Problem setup

We are interested in solving the problem

$$p^* = \left(\underset{x \in \mathbb{R}^d}{\text{minimize}} \quad \left\{ f(x) = \frac{1}{n} \sum_{i=1}^n f^{[i]}(x) \right\} \right) \dots (\mathcal{P})$$

where the optimal solution is achieved at x_* . We have the following assumptions regarding the nature of the problem.

Notation

Inner product between vectors x, y is denoted by $\langle x \mid y \rangle$ and Euclidean norm of x is denoted by $\|x\| = \sqrt{\langle x \mid x \rangle}$. We let $[1 : n] = \{1, 2, \dots, n\}$ and $z_+ = \max\{z, 0\}$. Also, for notational convenience we denote: $\text{sqd}(x) = \|x\|^2$ and $\text{ReLU}(z) = \max\{z, 0\}$. Comments are enclosed in `/* this is a comment */`.

Stochastic Gradient Descent with Polyak Stepsize

The algorithm called Stochastic Gradient Descent with Stochastic Polyak Stepsize (SGD-SPS) to solve (\mathcal{P}) is described by Algorithm 1. The uniform distribution with support $\{1, \dots, n\}$ is denoted by $\text{unif}[1 : n]$. One subgradient of the function $f^{[i]}$ evaluated at x is denoted by $\tilde{\nabla} f(x)$.

Algorithm 1 SGD-SPS to solve (\mathcal{P}) **input:** the functions $f^{[i]}$ for $i \in [1 : n]$, iteration limit T **algorithm:****1. initialization:**pick $x_0 \in \mathbb{R}^d$ arbitrarily**2. main iteration:****for** $t = 0, 1, 2, \dots, T - 1$ sample a function f_i uniformly at random $i \sim \text{unif}[1 : n]$ set Polyak stepsize $\gamma_t = \begin{cases} \frac{\text{ReLU}(f^{[i]}(x_t) - f^{[i]}(x_\star))}{\|\tilde{\nabla} f^{[i]}(x_t)\|^2}, & \text{if } \tilde{\nabla} f^{[i]}(x_t) \neq 0 \\ 0, & \text{else,} \end{cases}$ update iterate $x_{t+1} = x_t - \gamma_t \tilde{\nabla} f^{[i]}(x_t)$ **end for****3. return** x_T **Assumptions**

We assume that for all $i, f^{[i]} : \mathbb{R}^d \rightarrow (-\infty, \infty]$ is a nonsmooth, subgradient bounded, and star-convex function, i.e,

- **Star-convexity.**

$$\forall_{i \in [1:n]} f^{[i]} \text{ star-convex around } x_\star$$

$$\stackrel{\text{def}}{\Leftrightarrow}$$

$$\forall_{x \in \text{dom} f} f^{[i]}(x) - f^{[i]}(x_\star) \leq \langle \tilde{\nabla} f^{[i]}(x) \mid x - x_\star \rangle.$$

- **Subgradient-boundedness.**

$$\forall_{i \in [1:N]} \forall_{x \in \mathcal{B} = \{y \mid \|y - x_\star\| \leq \|x_0 - x_\star\|\}} \forall_{\tilde{\nabla} f^{[i]}(x) \in \partial f^{[i]}(x)} \|\tilde{\nabla} f^{[i]}(x)\| \leq G.$$

Convergence analysis

Consider an arbitrary iteration number t and we want to compute iterate x_{t+1} from x_t . Going from x_t to x_{t+1} the randomness lies in the selection of the function f_i by $i \sim \text{unif}[1 : N]$. We will come up with an inequality that works for any value of i . We will use the notation $\tilde{\nabla} f^{[i]}(x_t) \triangleq g_t^{[i]}, \tilde{\nabla} f(x_t) \triangleq g_t, f^{[i]}(x_t) \triangleq f_t^{[i]}$.

Consider the case $\|g_t^{[i]}\| \neq 0$. We have

$$\|x_{t+1} - x_\star\|^2$$

$$\|x_t - \gamma_t g_t^{[i]} - x_\star\|^2$$

$$\|(x_t - x_\star) - \gamma_t g_t^{[i]}\|^2$$

$$\|x_t - x_\star\|^2 + \gamma_t^2 \|g_t^{[i]}\|^2 - 2\gamma_t \langle g_t^{[i]} \mid x_t - x_\star \rangle$$

▷ expand squares

/*

we have $f^{[i]}(x) - f^{[i]}(x_\star) \leq \langle \tilde{\nabla} f^{[i]}(x) \mid x - x_\star \rangle$

$$\stackrel{x := x_t}{\Rightarrow} f^{[i]}(x_t) - f^{[i]}(x_\star) \leq \langle \tilde{\nabla} f^{[i]}(x_t) \mid x_t - x_\star \rangle$$

$$\Leftrightarrow f_t^{[i]} - f_\star^{[i]} \leq \langle g_t^{[i]} \mid x_t - x_\star \rangle$$

$$\Leftrightarrow -\left(f_t^{[i]} - f_\star^{[i]}\right) \geq -\langle g_t^{[i]} \mid x_t - x_\star \rangle$$

$$\therefore -2\gamma_t \langle g_t^{[i]} \mid x_t - x_\star \rangle \leq -2\gamma_t \left(f_t^{[i]} - f_\star^{[i]}\right)$$

*/

$$\begin{aligned} &\leq \|x_t - x_\star\|^2 + \gamma_t^2 \|g_t^{[i]}\|^2 - 2\gamma_t \left(f_t^{[i]} - f_\star^{[i]}\right) \quad \triangleright \text{ in this case } \gamma_t = \frac{\text{ReLU}(f_t^{[i]} - f_\star^{[i]})}{\|g_t^{[i]}\|^2} \\ &= \|x_t - x_\star\|^2 + \frac{(\text{sqd} \circ \text{ReLU})(f_t^{[i]} - f_\star^{[i]})}{\|g_t^{[i]}\|^4} \|g_t^{[i]}\|^2 - 2 \frac{\text{ReLU}(f_t^{[i]} - f_\star^{[i]})}{\|g_t^{[i]}\|^2} \left(f_t^{[i]} - f_\star^{[i]}\right) \end{aligned}$$

/*

$$\text{we have } z \times \text{ReLU}(z) = z \times \max\{z, 0\} = \begin{cases} z^2, & \text{if } z \geq 0 \\ 0, & \text{else} \end{cases} = (\max\{z, 0\})^2 = (\text{sqd} \circ \text{ReLU})(f_t^{[i]} - f_\star^{[i]})$$

*/

$$\begin{aligned} &= \|x_t - x_\star\|^2 + \frac{(\text{sqd} \circ \text{ReLU})(f_t^{[i]} - f_\star^{[i]})}{\|g_t^{[i]}\|^2} - 2 \frac{(\text{sqd} \circ \text{ReLU})(f_t^{[i]} - f_\star^{[i]})}{\|g_t^{[i]}\|^2} \\ &= \|x_t - x_\star\|^2 - \frac{(\text{sqd} \circ \text{ReLU})(f_t^{[i]} - f_\star^{[i]})}{\|g_t^{[i]}\|^2} \dots (1) \end{aligned}$$

Now consider the case $\|g_t^{[i]}\| = 0$, then $x_{t+1} = x_t$ and

$$\|x_{t+1} - x_\star\|^2 = \|x_t - x_\star\|^2 \dots (2)$$

Thus from (1) and (2), we have

$$\|x_{t+1} - x_\star\|^2 \leq \begin{cases} \|x_t - x_\star\|^2 - \frac{(\text{sqd} \circ \text{ReLU})(f_t^{[i]} - f_\star^{[i]})}{\|g_t^{[i]}\|^2}, & \text{with } i \sim \text{unif}[1:n] \text{ and } \|g_t^{[i]}\| \neq 0, \dots (3) \\ \|x_t - x_\star\|^2, & \text{with } i \sim \text{unif}[1:n] \text{ and } \|g_t^{[i]}\| = 0. \end{cases}$$

From (3), we see that irrespective of the randomness in selecting i , we always have $\|x_{t+1} - x_\star\|^2 \leq \|x_t - x_\star\|^2 \leq \dots \leq \|x_0 - x_\star\|^2$, hence we have $x_t \in \mathcal{B} = \{y \mid \|y - x_\star\| \leq \|x_0 - x_\star\|\}$ no matter what. As a result, for the case $\|g_t^{[i]}\| \neq 0$, using the gradient-boundedness assumption we have

$$\begin{aligned} &\|g_t^{[i]}\|^2 \leq G^2 \\ &\Leftrightarrow \frac{1}{\|g_t^{[i]}\|^2} \geq \frac{1}{G^2} \\ &\Leftrightarrow -\frac{(\text{sqd} \circ \text{ReLU})(f_t^{[i]} - f_\star^{[i]})}{\|g_t^{[i]}\|^2} \leq -\frac{(\text{sqd} \circ \text{ReLU})(f_t^{[i]} - f_\star^{[i]})}{G^2} \dots (4) \end{aligned}$$

Next, for the case $\|g_t^{[i]}\| = 0 \Leftrightarrow g_t^{[i]} = 0$, using star-convexity, we have

$$\begin{aligned}
 f^{[i]}(x) - f^{[i]}(x_\star) &\leq \langle \tilde{\nabla} f^{[i]}(x) \mid x - x_\star \rangle \\
 \stackrel{x := x_t}{\Rightarrow} f^{[i]}(x_t) - f^{[i]}(x_\star) &\leq \langle g_t^{[i]} \mid x_t - x_\star \rangle = 0 \\
 \Rightarrow f_t^{[i]} - f_\star^{[i]} &\leq 0 \\
 \Rightarrow \text{ReLU}(f_t^{[i]} - f_\star^{[i]}) &= \max\{f_t^{[i]} - f_\star^{[i]}, 0\} = 0 \\
 \Rightarrow \frac{(\text{sqd} \circ \text{ReLU})(f_t^{[i]} - f_\star^{[i]})}{G^2} &= 0 \dots (5)
 \end{aligned}$$

So, using (4) and (5) in the cases of (3) we get

$$\|x_{t+1} - x_\star\|^2 \leq \|x_t - x_\star\|^2 - \frac{(\text{sqd} \circ \text{ReLU})(f_t^{[i]} - f_\star^{[i]})}{G^2} \text{ with } i \sim \text{unif}[1 : n] \dots (6)$$

Next, on both sides of (6), we take conditional expectation with respect to i given x_t , which we denote by $\mathbf{E}[\cdot \mid x_t] \triangleq \mathbf{E}_{i \sim \text{unif}[1:N]}[\cdot \mid x_t]$, and the resultant inequality is:

$$\begin{aligned}
 &\mathbf{E}[\|x_{t+1} - x_\star\|^2 \mid x_t] \\
 &\leq \mathbf{E}\left[\|x_t - x_\star\|^2 - \frac{(\text{sqd} \circ \text{ReLU})(f_t^{[i]} - f_\star^{[i]})}{G^2} \mid x_t\right] \\
 &= \mathbf{E}[\|x_t - x_\star\|^2 \mid x_t] - \mathbf{E}\left[\frac{(\text{sqd} \circ \text{ReLU})(f_t^{[i]} - f_\star^{[i]})}{G^2} \mid x_t\right] \quad \triangleright \text{using linearity of expectation} \\
 &= \|x_t - x_\star\|^2 - \frac{1}{G^2} \mathbf{E}[(\text{sqd} \circ \text{ReLU})(f_t^{[i]} - f_\star^{[i]}) \mid x_t] \dots (7)
 \end{aligned}$$

\triangleright using "taking out what's known" rule
 $\mathbf{E}[h(X)Y \mid X] = h(X)\mathbf{E}[Y \mid X]$

Recall now Jensen's inequality: if ϕ is a convex function and Z is a random variable, then $\phi(\mathbf{E}[Z]) \leq \mathbf{E}[\phi(Z)]$. Setting $\phi := \text{sqd} \circ \text{ReLU} = \text{sqd}(\text{ReLU}(\cdot))$, which is convex (see Boyd Vandenberghe, Convex Optimization, Figure 3.7) and $Z := [(\text{sqd} \circ \text{ReLU})(f_t^{[i]} - f_\star^{[i]}) \mid x_t]$ we have

$$\begin{aligned}
 &(\text{sqd} \circ \text{ReLU})(\mathbf{E}[(f_t^{[i]} - f_\star^{[i]}) \mid x_t]) \leq \mathbf{E}[(\text{sqd} \circ \text{ReLU})(f_t^{[i]} - f_\star^{[i]}) \mid x_t] \\
 \Leftrightarrow &-(\text{sqd} \circ \text{ReLU})(\mathbf{E}[(f_t^{[i]} - f_\star^{[i]}) \mid x_t]) \geq -\mathbf{E}[(\text{sqd} \circ \text{ReLU})(f_t^{[i]} - f_\star^{[i]}) \mid x_t] \\
 \Leftrightarrow &-\frac{1}{G^2}(\text{sqd} \circ \text{ReLU})(\mathbf{E}[(f_t^{[i]} - f_\star^{[i]}) \mid x_t]) \geq -\frac{1}{G^2} \mathbf{E}[(\text{sqd} \circ \text{ReLU})(f_t^{[i]} - f_\star^{[i]}) \mid x_t] \dots (8)
 \end{aligned}$$

Now notice the LHS term in (8):

$$\mathbf{E}[(f_t^{[i]} - f_\star^{[i]}) \mid x_t]$$

$$\begin{aligned}
&= \mathbf{E}_{i \sim \text{unif}[1:n]} \left[\left(\left(f_t^{[i]} - f_\star^{[i]} \right) \mid x_t \right) \right] \\
&= \left(\left(\frac{1}{n} \sum_{i=1}^n \left(f_t^{[i]} - f_\star^{[i]} \right) \right) \mid x_t \right) \\
&= f(x_t) - f(x_\star),
\end{aligned}$$

where the last term is a random variable in x_t (recall that $\mathbf{E}[Y \mid X]$ is a random variable in X).

From (7), (8), and (9), we have

$$\begin{aligned}
&\mathbf{E} \left[\|x_{t+1} - x_\star\|^2 \mid x_t \right] \\
&\leq \|x_t - x_\star\|^2 - \frac{1}{G^2} (\text{sqd} \circ \text{ReLU}) \left(\mathbf{E} \left[\left(f_t^{[i]} - f_\star^{[i]} \right) \mid x_t \right] \right) \\
&= \|x_t - x_\star\|^2 - \frac{1}{G^2} (\text{sqd} \circ \text{ReLU}) (f(x_t) - f(x_\star)) \\
&= \|x_t - x_\star\|^2 - \frac{1}{G^2} (\max\{f(x_t) - f(x_\star), 0\})^2 \\
&= \|x_t - x_\star\|^2 - \frac{1}{G^2} (f(x_t) - f(x_\star))^2 \dots (10)
\end{aligned}$$

\triangleright as $f(x_t) - f(x_\star) \geq 0$

Now taking expectation with respect to x_t on both sides of (10) and then using Adam's law $\mathbf{E}[\mathbf{E}[Y \mid X]] = \mathbf{E}[Y]$, we get:

$$\begin{aligned}
&\mathbf{E} \left[\mathbf{E} \left[\|x_{t+1} - x_\star\|^2 \mid x_t \right] \right] \leq \mathbf{E} \left[\|x_t - x_\star\|^2 - \frac{1}{G^2} (f(x_t) - f(x_\star))^2 \right] \\
&\Leftrightarrow \mathbf{E} \left[\|x_{t+1} - x_\star\|^2 \right] \leq \mathbf{E} \left[\|x_t - x_\star\|^2 \right] - \mathbf{E} \left[\frac{1}{G^2} (f(x_t) - f(x_\star))^2 \right] \\
&\Leftrightarrow \mathbf{E} \left[\|x_{t+1} - x_\star\|^2 \right] \leq \mathbf{E} \left[\|x_t - x_\star\|^2 \right] - \frac{1}{G^2} \mathbf{E} \left[(f(x_t) - f(x_\star))^2 \right] \\
&\Leftrightarrow \frac{1}{G^2} \mathbf{E} \left[(f(x_t) - f(x_\star))^2 \right] \leq \mathbf{E} \left[\|x_t - x_\star\|^2 \right] - \mathbf{E} \left[\|x_{t+1} - x_\star\|^2 \right] \dots (11)
\end{aligned}$$

\triangleright using linearity of expectation on RHS and Adam's law on LHS

Now, let us do a telescoping sum on (11) for $t = 0, \dots, T$

$$\begin{aligned}
&\frac{1}{G^2} \mathbf{E} \left[(f(x_0) - f(x_\star))^2 \right] \leq \mathbf{E} \left[\|x_0 - x_\star\|^2 \right] - \mathbf{E} \left[\|x_1 - x_\star\|^2 \right] \\
&\frac{1}{G^2} \mathbf{E} \left[(f(x_1) - f(x_\star))^2 \right] \leq \mathbf{E} \left[\|x_1 - x_\star\|^2 \right] - \mathbf{E} \left[\|x_2 - x_\star\|^2 \right] \\
&\frac{1}{G^2} \mathbf{E} \left[(f(x_2) - f(x_\star))^2 \right] \leq \mathbf{E} \left[\|x_2 - x_\star\|^2 \right] - \mathbf{E} \left[\|x_3 - x_\star\|^2 \right] \\
&\vdots \\
&\frac{1}{G^2} \mathbf{E} \left[(f(x_{T-1}) - f(x_\star))^2 \right] \leq \mathbf{E} \left[\|x_{T-1} - x_\star\|^2 \right] - \mathbf{E} \left[\|x_T - x_\star\|^2 \right] \\
&\frac{1}{G^2} \mathbf{E} \left[(f(x_T) - f(x_\star))^2 \right] \leq \mathbf{E} \left[\|x_T - x_\star\|^2 \right] - \mathbf{E} \left[\|x_{T+1} - x_\star\|^2 \right]
\end{aligned}$$

which yields:

$$\frac{1}{G^2} \sum_{k=0}^T \mathbf{E} \left[(f(x_k) - f(x_\star))^2 \right] \leq \mathbf{E} \left[\|x_0 - x_\star\|^2 \right] - \mathbf{E} \left[\|x_{T+1} - x_\star\|^2 \right]$$

$$\begin{aligned}
&= \|x_0 - x_\star\|^2 - \mathbf{E} \left[\|x_{T+1} - x_\star\|^2 \right] &> \text{as } x_0 \text{ is deterministic} \\
&\leq \|x_0 - x_\star\|^2 &> \text{as } -\mathbf{E} \left[\|x_{T+1} - x_\star\|^2 \right] \leq 0 \\
&\Rightarrow \sum_{k=0}^T \mathbf{E} \left[(f(x_k) - f(x_\star))^2 \right] \leq G^2 \|x_0 - x_\star\|^2 \\
&\therefore \frac{1}{T+1} \sum_{k=0}^T \mathbf{E} \left[(f(x_k) - f(x_\star))^2 \right] \leq \frac{G^2 \|x_0 - x_\star\|^2}{T+1} \dots (12)
\end{aligned}$$

Recall now Jensen's inequality again: if ϕ is a convex function and Z is a random variable, then $\phi(\mathbf{E}[Z]) \leq \mathbf{E}[\phi(Z)]$. Setting $\phi := \text{sqd}$ and $Z := f(x_k) - f(x_\star)$ we have

$$\begin{aligned}
&\text{sqd}(\mathbf{E}[f(x_k) - f(x_\star)]) \leq \mathbf{E}[\text{sqd}(f(x_k) - f(x_\star))] \\
&\Rightarrow \min_{k \in [0:T]} (\mathbf{E}[f(x_k) - f(x_\star)])^2 \leq \min_{k \in [0:T]} \mathbf{E} \left[(f(x_k) - f(x_\star))^2 \right] \dots (13)
\end{aligned}$$

Also,

$$\begin{aligned}
&\sum_{k=0}^T \mathbf{E} \left[(f(x_k) - f(x_\star))^2 \right] \geq \sum_{k=0}^T \left(\min_{k \in [0:T]} \mathbf{E} \left[(f(x_k) - f(x_\star))^2 \right] \right) \\
&= \left(\min_{k \in [0:T]} \mathbf{E} \left[(f(x_k) - f(x_\star))^2 \right] \right) \sum_{k=0}^T 1 \\
&= (T+1) \left(\min_{k \in [0:T]} \mathbf{E} \left[(f(x_k) - f(x_\star))^2 \right] \right) \\
&\Rightarrow \frac{1}{T+1} \sum_{k=0}^T \mathbf{E} \left[(f(x_k) - f(x_\star))^2 \right] \geq \min_{k \in [0:T]} \mathbf{E} \left[(f(x_k) - f(x_\star))^2 \right] \dots (14)
\end{aligned}$$

From (13) and (14), we have

$$\min_{k \in [0:T]} (\mathbf{E}[f(x_k) - f(x_\star)])^2 \leq \min_{k \in [0:T]} \mathbf{E} \left[(f(x_k) - f(x_\star))^2 \right] \leq \frac{1}{T+1} \sum_{k=0}^T \mathbf{E} \left[(f(x_k) - f(x_\star))^2 \right] \dots (15)$$

Now, from (15) and (12), we have

$$\min_{k \in [0:T]} (\mathbf{E}[f(x_k) - f(x_\star)])^2 \leq \frac{G^2 \|x_0 - x_\star\|^2}{T+1}.$$

Let the min be achieved at index $\ell \in [0 : T]$, hence using the fact that $\sqrt{\cdot}$ is monotonically increasing on \mathbb{R}_+ (hence would not change direction of inequalities when both sides are nonnegative), we have

$$\begin{aligned}
&(\mathbf{E}[f(x_\ell) - f(x_\star)])^2 \leq \frac{G^2 \|x_0 - x_\star\|^2}{T+1} \\
&\Rightarrow \mathbf{E}[f(x_\ell) - f(x_\star)] \leq \frac{G \|x_0 - x_\star\|}{\sqrt{T+1}}.
\end{aligned}$$

Thus we have proven that:

$$\min_{k \in [0:T]} (\mathbf{E}[f(x_k) - f(x_\star)]) \leq \frac{G \|x_0 - x_\star\|}{\sqrt{T+1}}.$$