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

Shuvomoy Das Gupta

March 29, 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

Problem setup

We are interested in solving the problem

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

where the optimal solution is achieved at $x_\star$. 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 $\|x\| = \sqrt{\langle x \mid y \rangle}$. We let $[1:n] = \{1, 2, \ldots, n\}$ and $z_+ = \max\{z, 0\}$. Also, for notational convenience we denote: $\texttt{sqd}(x) = \|x\|^2$ and $\texttt{ReLU}(z) = \max\{z, 0\}$. 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 $\{1, \ldots, n\}$ is denoted by $\texttt{unif}[1:n]$. One subgradient of the function $f^{[i]}$ evaluated at x is denoted by $\widetilde{\nabla} f(x)$.

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 $t = 0, 1, 2, \ldots, T-1$
sample a function f_i uniformly at random $i \sim \texttt{unif}[1:n]$
set Polyak stepsize $\gamma_t = \begin{cases} \frac{\texttt{ReLU}\left(f^{[i]}(x_t) - f^{[i]}(x_\star)\right)}{\|\widetilde{\nabla} f^{[i]}(x_t)\|^2}, & \text{if } \widetilde{\nabla} f^{[i]}(x_t) \neq 0\| \\ 0, & \text{else,} \end{cases}$
update iterate $x_{t+1} = x_t - \gamma_t \widetilde{\nabla} f^{[i]}(x_t)(x_t)$
end for
3. return x_T

Assumptions

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

- **Star-convexity.**
  $\forall_{i \in [1:n]} f^{[i]}$ star-convex around $x_\star$
  $\overset{\text{def}}{\Leftrightarrow}$
  $\forall_{x \in \operatorname{dom} f}\ f^{[i]}(x) - f^{[i]}(x_\star) \leq \left\langle \widetilde{\nabla} f^{[i]}(x) \mid x - x_\star \right\rangle.$
- **Subgradient-boundedness.**
  $\forall_{i \in [1:N]}\ \forall_{x \in \mathcal{B} = \{y \mid \|y - x_\star\| \leq \|x_0 - x_\star\|\}}\ \forall_{\widetilde{\nabla} f^{[i]}(x) \in \partial f^{[i]}(x)}\ \|\widetilde{\nabla} f^{[i]}(x)\| \leq G.$

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 $i \sim \texttt{unif}[1:N]$. We will come up with an inequality that works for any value of i. We will use the notation $\widetilde{\nabla} f^{[i]}(x_t) \triangleq g_t^{[i]}, \widetilde{\nabla} f(x_t) \triangleq g_t, f^{[i]}(x_t) \triangleq f_t^{[i]}$.

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

$$
\begin{aligned}
&\|x_{t+1} - x_\star\|^2 \\
=&\|x_t - \gamma_t g_t^{[i]} - x_\star\|^2 \\
=&\|(x_t - x_\star) - \gamma_t g_t^{[i]}\|^2 && \triangleright \text{ expand squares} \\
=&\|x_t - x_\star\|^2 + \gamma_t^2 \|g_t^{[i]}\|^2 - 2\gamma_t \left\langle g_t^{[i]} \mid x_t - x_\star \right\rangle
\end{aligned}
$$

/*

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

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

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

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

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

*/

$$\leq \|x_t - x_\star\|^2 + \gamma_t^2 \|g_t^{[i]}\|^2 - 2\gamma_t \left(f_t^{[i]} - f_\star^{[i]}\right) \qquad \triangleright \text{ in this case } \gamma_t = \frac{\mathsf{ReLU}\left(f_t^{[i]} - f_\star^{[i]}\right)}{\|g_t^{[i]}\|^2}$$

$$= \|x_t - x_\star\|^2 + \frac{(\mathsf{sqd} \circ \mathsf{ReLU})\left(f_t^{[i]} - f_\star^{[i]}\right)}{\|g_t^{[i]}\|^4} \|g_t^{[i]}\|^2 - 2\frac{\mathsf{ReLU}\left(f_t^{[i]} - f_\star^{[i]}\right)}{\|g_t^{[i]}\|^2}\left(f_t^{[i]} - f_\star^{[i]}\right)$$

/*

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

*/

$$= \|x_t - x_\star\|^2 + \frac{(\mathsf{sqd} \circ \mathsf{ReLU})\left(f_t^{[i]} - f_\star^{[i]}\right)}{\|g_t^{[i]}\|^2} - 2\frac{(\mathsf{sqd} \circ \mathsf{ReLU})\left(f_t^{[i]} - f_\star^{[i]}\right)}{\|g_t^{[i]}\|^2}$$

$$= \|x_t - x_\star\|^2 - \frac{(\mathsf{sqd} \circ \mathsf{ReLU})\left(f_t^{[i]} - f_\star^{[i]}\right)}{\|g_t^{[i]}\|^2} \ldots . (1)$$

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

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

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

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

From (3), we see that irrespective of the randomness in selecting i, we always have $\|x_{t+1} - x_\star\|^2 \leq \|x_t - x_\star\|^2 \leq \cdots \leq \|x_0 - x_\star\|^2$, 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

$$\|g_t^{[i]}\|^2 \leq G^2$$

$$\Leftrightarrow \frac{1}{\|g_t^{[i]}\|^2} \geq \frac{1}{G^2}$$

$$\Leftrightarrow -\frac{(\mathsf{sqd} \circ \mathsf{ReLU})\left(f_t^{[i]} - f_\star^{[i]}\right)}{\|g_t^{[i]}\|^2} \leq -\frac{(\mathsf{sqd} \circ \mathsf{ReLU})\left(f_t^{[i]} - f_\star^{[i]}\right)}{G^2} \ldots . (4)$$

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

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

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

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

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

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

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

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

$$\begin{aligned}
& (\mathtt{sqd} \circ \mathtt{ReLU})\left(\mathbf{E}\left[\left(f_t^{[i]} - f_\star^{[i]}\right) \mid x_t\right]\right) \le \mathbf{E}\left[(\mathtt{sqd} \circ \mathtt{ReLU})\left(f_t^{[i]} - f_\star^{[i]} \mid x_t\right)\right] \\
\Leftrightarrow & -(\mathtt{sqd} \circ \mathtt{ReLU})\left(\mathbf{E}\left[\left(f_t^{[i]} - f_\star^{[i]}\right) \mid x_t\right]\right) \ge -\mathbf{E}\left[(\mathtt{sqd} \circ \mathtt{ReLU})\left(f_t^{[i]} - f_\star^{[i]} \mid x_t\right)\right] \\
\Leftrightarrow & -\frac{1}{G^2}(\mathtt{sqd} \circ \mathtt{ReLU})\left(\mathbf{E}\left[\left(f_t^{[i]} - f_\star^{[i]}\right) \mid x_t\right]\right) \ge -\frac{1}{G^2}\mathbf{E}\left[(\mathtt{sqd} \circ \mathtt{ReLU})\left(f_t^{[i]} - f_\star^{[i]} \mid x_t\right)\right] \ldots (8)
\end{aligned}$$

Now notice the LHS term in (8):

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

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

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

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

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:

$$
\begin{aligned}
&\mathbf{E}\left[\mathbf{E}\left[\|x_{t+1} - x_\star\|^2 \mid x_t \right] \right] \leq \mathbf{E}\left[\|x_t - x_\star\|^2 - \frac{1}{G^2} \left(f(x_t) - f(x_\star) \right)^2 \right] \\
\Leftrightarrow& \mathbf{E}\left[\|x_{t+1} - x_\star\|^2 \right] \leq \mathbf{E}\left[\|x_t - x_\star\|^2 \right] - \mathbf{E}\left[\frac{1}{G^2} \left(f(x_t) - f(x_\star) \right)^2 \right] \qquad \triangleright \text{ using linearity of expectation on RHS and Adam's law on LHS} \\
\Leftrightarrow& \mathbf{E}\left[\|x_{t+1} - x_\star\|^2 \right] \leq \mathbf{E}\left[\|x_t - x_\star\|^2 \right] - \frac{1}{G^2} \mathbf{E}\left[\left(f(x_t) - f(x_\star) \right)^2 \right] \\
\Leftrightarrow& \frac{1}{G^2} \mathbf{E}\left[\left(f(x_t) - f(x_\star) \right)^2 \right] \leq \mathbf{E}\left[\|x_t - x_\star\|^2 \right] - \mathbf{E}\left[\|x_{t+1} - x_\star\|^2 \right] \ldots . (11)
\end{aligned}
$$

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

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

which yields:

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

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

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

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

Also,

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

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

From (13) and (14), we have

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

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

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

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

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

Thus we have proven that:

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