On Convergence-Diagnostic based Step Sizes for Stochastic Gradient Descent

Aymeric Dieuleveut
CMAP, École Polytechnique, Institut Polytechnique de Paris
Joint work with Scott Pesme and Nicolas Flammarion (EPFL)

10/03/2020 Cirm Luminy - Optimization for Machine Learning
On Convergence-Diagnostic based Step Sizes for Stochastic Gradient Descent

Aymeric Dieuleveut
CMAP, École Polytechnique, Institut Polytechnique de Paris

Joint work with Scott Pesme and Nicolas Flammarion (EPFL)

10/03/2020 Cirm Luminy - Optimization for Machine Learning
1. Feel free to ask any question.

2. Let me ask a few ones first:

   - Who knows about Stochastic Gradient Descent?
   - Who knows the convergence rate for the last iterate instead of the averaged iterate?
   - Who knows about Pflug’s convergence diagnosis?
Why would we still talk about SGD?

Objective function $f: D \to \mathbb{R}$ to minimize

$$\theta_{n+1} = \theta_n - \gamma_{n+1} f'_{n+1}(\theta_n) = \theta_n - \gamma_{n+1} \left( f'(\theta_n) + \xi_{n+1}(\theta_n) \right).$$

What choice for the learning rate $(\gamma_n)_{n \in \mathbb{N}}$?

As often:

- **Theoreticians** (♡) came up with optimal answers (convex setting).
- **Practitioners** do not use them!

*If it works in theory it also works in practice – in theory.*

Why not?

1. Step size in SGD often depends on unknown parameters (esp. $\mu$-strong convexity).
2. May be very sensitive to those parameters.
3. Does not adapt to the noise and function regularity.
A few observations

a) Large learning rates often converge faster at the beginning.
b) But then results in saturation: two phases behavior.
c) Theory suggests to use the Polyak-Ruppert averaged iterate, but the final one might not be that bad.
d) In Deep Learning, common practice is to use a constant learning rate, reduced occasionally.
a) Large learning rates often converge faster at the beginning of the training process.

SGD nearly always results in a Bias (initial condition) - Variance (noise) tradeoff.

A large initial learning rate maximizes the decay of the bias.

![Graph showing the decay of bias and variance over decaying steps.](image)

**Figure 1:** Logistic regression on the Covertype Dataset / Synthetic Dataset
b) Saturation and limit distribution: two phases

- **“Transient phase”** during which the initial conditions are forgotten exponentially fast.
- **"Stationary phase"** where the iterates oscillate around $\theta^*$

![Synthetic logistic dataset d = 2](image)

**Figure 2:** Constant step size SGD (2 dimensionnal) path illustration.

For smooth and strongly convex functions, $\theta_n \xrightarrow{(d)} \pi_\gamma$, “limit distribution”.

$\pi_\gamma$ is a stationary distribution.
c) Polyak–Ruppert averaged iterate vs final one.

Instead of just the final iterate $\theta_n^{(\gamma)}$, we can consider the PR-averaged:

$$
\bar{\theta}_n^{(\gamma)} = \frac{1}{n} \sum_{k=0}^{n-1} \theta_k^{(\gamma)}.
$$

↬ Strongly reduces the impact of the noise.
↬ Slows down the Bias term.

How bad is the last iterate...?

It depends!
<table>
<thead>
<tr>
<th>Final Iterate</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Convex &amp; Smooth</td>
<td></td>
</tr>
<tr>
<td>Strongly convex &amp; Smooth</td>
<td></td>
</tr>
<tr>
<td>No noise (deterministic)</td>
<td></td>
</tr>
<tr>
<td>Finite dimensional quadratic</td>
<td></td>
</tr>
<tr>
<td>Kernel Regression</td>
<td></td>
</tr>
</tbody>
</table>

The Proof by Shamir & Zhang is nice!
c) Polyak-Ruppert averaged iterate vs final one.

<table>
<thead>
<tr>
<th></th>
<th>Final Iterate</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Convex &amp; Smooth</td>
<td>( \frac{\log t}{\nu t} ) ou</td>
<td>( \frac{1}{\nu t} ) ( \log(t) ) opt.</td>
</tr>
<tr>
<td>Strongly convex &amp; Smooth</td>
<td>( \frac{\log t}{\nu t} ) ou</td>
<td>( \frac{1}{\nu t} ) ( \log(t) ) opt.</td>
</tr>
<tr>
<td>No noise (deterministic)</td>
<td>( \checkmark ) opt.</td>
<td></td>
</tr>
<tr>
<td>Finite dimensional quadratic</td>
<td>( \checkmark ) opt.</td>
<td></td>
</tr>
<tr>
<td>Kernel Regression</td>
<td>depends on source condition!</td>
<td>weak case ou on adaptive case bad</td>
</tr>
</tbody>
</table>

The Proof by Shamir & Zhang is nice!
Previous work: with decreasing step-sizes

(Moulines & Bach 2011), smooth + strongly convex

Setting $\gamma_n = \frac{1}{\mu n}$ we get

$$\mathbb{E} \left[ \| \theta_n - \theta^* \|^2 \right] = O \left( \frac{1}{\mu^2 n} \right).$$

(Shamir & Zhang 2012), bounded gradients + strongly convex

Setting $\gamma_n = \frac{1}{\mu n}$ we get

$$\mathbb{E} \left[ f(\theta_n) - f(\theta^*) \right] = O \left( \frac{\log(n)}{\mu n} \right).$$

(Shamir & Zhang 2012), bounded gradients + weakly convex

Setting $\gamma_n = \frac{1}{\sqrt{n}}$ we get

$$\mathbb{E} \left[ f(\theta_n) - f(\theta^*) \right] = O \left( \frac{\log(n)}{\sqrt{n}} \right).$$
d) Deep Learning: training NN

\[ (1 - \text{test\_accuracy}) \]

Figure 3: Typical accuracy curve in deep learning (Cifar10 dataset, Resnet18).
• in the strongly convex case, $\mu$ is often unknown and hard to evaluate.
• a slight misspecification of $\mu$ can lead to arbitrarily slow convergence rates (see Moulines & Bach 2011)
• we would like to make use of the uniform convexity assumption
• ideally we would like a learning rate sequence that adapts to $f$
• these stepsize sequences are not used in practice for deep learning
Natural strategy:

decrease learning rate when no more progress

Hopes: adaptive “restarts” to

- use “maximal step size” as long as useful
- adapt to unknown parameters.

Outline:

1. Convergence properties of SGD with piecewise constant learning rates.
2. Detecting Stationarity: Pflug’s Statistic

“Restart” : nothing to restart, just changing the learning rate!
“Omniscient strategies”. What can we achieve with piecewise constant step sizes?
What rate can you get if you use a large step size for as long as possible and you decrease it when the loss saturates?
Theorem (Needell 2014)

\[ \mathbb{E} \left[ \| \theta_n - \theta^* \|^2 \right] \leq (1 - b \gamma)^n \| \theta_0 - \theta^* \|^2 + c \sigma^2 \gamma + O(\gamma^2), \]

where \( b, c \) depend on \( f \) and \( \sigma^2 = \mathbb{E} \left[ \| \xi(\theta^*) \|^2 \right] \).

**Theoretical procedure:** Let \( p, r \in [0, 1] \). Start with l.r. \( \gamma_0 \), stop at \( \Delta n_1 \):

\[
\mathbb{E} \left[ \| \theta_n - \theta^* \|^2 \right] \leq \left[ 1 - 2\gamma_0 \mu \right]^n \mathbb{E} \left[ \| \theta_0 - \theta^* \|^2 \right] + \left[ \frac{\sigma^2}{\mu} \right] \gamma_0 \Delta n_1 \quad \text{s.t.} \quad p \times ( \ldots )
\]

Set \( \gamma_1 = r \gamma_0 \) and restart from \( \theta_{n_1} = \theta_{\Delta n_1} \):

\[
\mathbb{E} \left[ \| \theta_n - \theta^* \|^2 \right] \leq \left[ 1 - 2\gamma_1 \mu \right]^{(n-n_1)} \mathbb{E} \left[ \| \theta_{n_1} - \theta^* \|^2 \right] + \left[ \frac{\sigma^2}{\mu} \right] \gamma_1 \Delta n_2 \quad \text{s.t.} \quad p \times ( \ldots )
\]

etc.

(Related but slightly different from Hazan Kale 2010, e.g.)
**Theorem (Strongly convex + smooth)**

Following the previous oracle procedure and assuming that $\|\theta_0 - \theta^*\|^2 \leq (p + 1) \frac{\sigma^2}{\mu} \gamma_0$:

\[
\mathbb{E}\left[\|\theta_{n_k} - \theta^*\|^2\right] \leq (p + 1) \frac{\sigma^2}{1 - r} \ln\left((1 + \frac{1}{p}) \frac{1}{\mu r}\right) \frac{1}{\mu^2 n_k}.
\]

\[
\leq O\left(\frac{1}{\mu^2 n_k}\right)
\]

- The upper bound can be optimized over $p$ and $r$
- Purely theoretical result since none of these constants are known.
- The step size sequence produced is piecewise constant and 'imitates' $\gamma_n = 1/\mu n$.

Beyond the Smooth & Strongly convex : uniformly convex functions
Assumptions on $f$

Convexity:

- **Weak convexity**: $f(\theta_1) \geq f(\theta_2) + \langle f'(\theta_2), \theta_1 - \theta_2 \rangle$
- **Strong convexity, $\mu > 0$**: $f(\theta_1) \geq f(\theta_2) + \langle f'(\theta_2), \theta_1 - \theta_2 \rangle + \frac{\mu}{2} \|\theta_1 - \theta_2\|^2$
- **Uniform convexity**: $f$ is uniformly convex with parameters $\mu > 0$, $\rho \in [2, +\infty[$ if:

$$f(\theta_1) \geq f(\theta_2) + \langle f'(\theta_2), \theta_1 - \theta_2 \rangle + \frac{\mu}{\rho} \|\theta_1 - \theta_2\|^\rho$$

Smoothness:

- **($L$-smoothness)** for any $n \in \mathbb{N}$, $f_n$ is $L$-smooth:

$$\|f_n'(\theta_1) - f_n'(\theta_2)\| \leq L \|\theta_1 - \theta_2\| \quad \text{a.s.}$$

- **(Non-smooth, bounded gradients)** bounded gradients framework:

$$\mathbb{E} \left[ \|f_n'(\theta_{n-1})\|^2 \right] \leq G^2$$
Proposition (PDF 2020)

If $f$ is a uniformly convex function with parameter $\rho > 2$ with $G$-bounded gradients then:

$$E \left[ f(\theta_n) - f(\theta^*) \right] \leq C \left( \frac{1}{\gamma n} \right)^{1/\tau} + G^2 \log(n)\gamma$$

Where $\tau = 1 - \frac{2}{\rho} \in [0, 1]$

In the finite horizon framework, this results in:

$$E \left[ f(\theta_n) - f(\theta^*) \right] \leq O \left( \frac{\log N}{N^{1/(1+\tau)}} \right)$$

Notice that $\frac{1}{1+\tau} \in [0.5, 1]$, we have an interpolation between the weakly convex and strongly convex cases.

- Juditsky Nesterov 2014 have a similar rate with a different algorithm
- Roulet et d’Aspremont have the $N^{-1/\tau}$ rate for GD.
Considering the previous upper bound: and following the previous “oracle” procedure (restart when Bias = $p \times \text{Var}$ )

**Theorem (PDF 20)**

$$f(\theta_{n_k}) - f(\theta^*) \leq O\left(\log(n_k)n_k^{-\frac{1}{1+\tau}}\right)$$

As before, the strategy of constant steps with “restart at saturation” gives satisfying rates (as good as the best known strategy for decaying steps)
Numerical simulation in the quadratic case

Figure 4: Oracle constant piece wise SGD
Numerical simulation in the uniformly convex case

Vanilla example: \( f(\theta) = \frac{1}{\rho} \|\theta\|^\rho \) where \( \rho = 2.5 \), rate of \( \sim n^{-0.8} \).

**Figure 5:** Oracle constant piece wise SGD for a uniformly convex function
Oracle procedure has good theoretical guarantees and it adapts to the framework (smoothness, uniform convexity, deterministic).

But:

- Constants are un-known.
- Computing the loss to detect saturation would be very time consuming

Can we detect saturation without having access to the loss values?
Detecting stationarity with statistics. Pflug’s statistic:

\[ S_{n}^{(\gamma)} = \frac{1}{n} \sum_{k=0}^{n-1} \langle f'_{k+1}, f'_{k+2} \rangle \]
Pflug's statistic \( S_n^{(\gamma)} = \frac{1}{n} \sum_{k=0}^{n-1} \langle f'_{k+1}, f'_{k+2} \rangle \)

Pflug's idea:

- During transient phase: \( \mathbb{E} \left[ \langle f'_{n+1}, f'_{n+2} \rangle \right] > 0 \)
- Stationary phase: \( \mathbb{E} \left[ \langle f'_{n+1}, f'_{n+2} \rangle \right] < 0 \)
Algorithm 1 Piecewise constant SGD using Pflug’s statistic

**INPUT:** $\theta_0$, $\gamma_0 > 0$, $n_b > 0$, $r \in [0, 1]$, $N > 0$  
**OUTPUT:** $\theta_N$

$S \leftarrow 0$
$\text{last}_\text{restart} \leftarrow 0$
$\theta_1 \leftarrow \theta_0 - \gamma f'_1(\theta_0)$

**for** $n = 2$ to $N$ **do**

$\theta_n \leftarrow \theta_{n-1} - \gamma_n(\theta_{n-1})$
$S \leftarrow S + \langle f'_n(\theta_{n-1}), f'_{n-1}(\theta_{n-2}) \rangle$

**if** $n > \text{last}_\text{restart} + n_b$ **and** $S < 0$ **then**

$\text{last}_\text{restart} \leftarrow n$
$S \leftarrow 0$
$\gamma \leftarrow r \times \gamma$

**end if**

**end for**

**return** $\theta_N$
Our results

2 main results:

1. Proving that it makes sense
2. Proving that it fails

Why?
1. Convergence:

\[ \Theta_0 \]

\[ \mathbb{P} \left< \Phi_1', \Phi_2' \right> \]

\[ \Theta_0 \sim \Pi \gamma \]

\[ \Theta_i \sim \Pi \gamma \]

2. Wrong intuition:

Wrong intuite: bounce around.

\[ \| \Theta_i - \Theta_{i-1} \| = o(\delta) \]

\[ \| \Theta_0 - \Theta_* \| = o(\sqrt{\delta}) \]
3. positive effect

$\theta < \langle f'(\theta_0), f'(\theta_1) \rangle > 0$

line derivatives!

4. negative effect

$\theta_0 = \theta_+$
$\theta_1 = \gamma \theta_0$

$\langle f'(\theta_0), f'(\theta_1) \rangle$

$\langle f'(\theta_0) + \varepsilon_0, f'(\gamma \varepsilon_0) + \varepsilon_1 \rangle$

$\theta = 0$

$\mathbb{E} \langle \varepsilon_0, \varepsilon_1 \rangle = 0$

$\mathbb{E} \left[ \langle \varepsilon_0, f'(\gamma \varepsilon_0) \rangle \right] < 0$ \implies $\mathbb{E} \langle f'_1, f'_2 \rangle < 0$

Magic!

The negative effect is $2\times$ bigger than the positive one!
Proposition (Pflug 1990), (Chee & Toulis 2018) (PDF 2020)

In the quadratic semi-stochastic setting where \( f(\theta) = \frac{1}{2} \theta^T H \theta \) and i.i.d noise \( \xi_i \) (\( \mathbb{E} [\xi \xi^T] = C \)):

\[
\mathbb{E}_{\pi_\gamma} \left[ \langle f'_1, f'_2 \rangle \right] = \mathbb{E}_{\pi_\gamma} \left[ \langle f'_1(\theta), f'_2(\theta - \gamma f'_1(\theta)) \rangle \right] = -\gamma \text{Tr} \ H C (2I - \gamma H)^{-1} < 0.
\]

1. Proves that asymptotically, under stationary distribution, the inner product is negative on average.
2. The proof in Chee & Toulis (Aistats 18) is incomplete.
3. We also extend the result to a non asymptotic version of the expectation under the restart strategy: if \( \theta_{\text{restart}} \sim \pi_\gamma \) and we restart with a new constant step size \( \gamma_{\text{new}} = r \times \gamma \), . Then:

\[
\mathbb{E}_{\theta_0 \sim \pi_\gamma} \left[ S_n^{(r \gamma)} \right] = \frac{1}{4n} \left( \frac{1}{r} - 1 \right) \text{Tr} \left[ I - (I - r \gamma H)^{2n} \right] C - \frac{1}{2} r \gamma \text{Tr} H C + o_n(\gamma)
\]
We extend the proof to general functions, exhibiting the same balance between the positive and negative parts.

**Theorem (general smooth + strongly convex setting) (PDF 2020)**

For $f$ verifying adequate assumptions:

$$
\mathbb{E}_{\pi_\gamma} \left[ \langle f'_1, f'_2 \rangle \right] = -\frac{1}{2} \gamma \text{Tr} \ f''(\theta^*) \mathcal{C}(\theta^*) + O(\gamma^{3/2}),
$$

where $\mathcal{C}(\theta^*) = \mathbb{E} \left[ \xi(\theta^*)\xi(\theta^*)^T \right]$

**Conclusion:** “it makes sense” the mean of Pflug’s statistic is negative once we have reached the stationary distribution.

**So why does it fail ?**
Implementation of Pflug’s algorithm

Figure 6: Pflug SGD: way to many restarts
Implementation of Pflug’s algorithm

Figure 7: Pflug SGD: way to many restarts
Taking a closer look

- \( \mathbb{E}_{\pi_\gamma}[\langle f'_1, f'_2 \rangle] \propto \gamma \).
- \( \text{Var}\langle f'_1, f'_2 \rangle = C \perp \perp \gamma \).

To detect \( S_n < 0 \) we typically need:

\[
\mathbb{E}[S_n^{(\gamma)}] + \sqrt{\text{Var}(S_n^{(\gamma)})} < 0
\]

\( \Leftrightarrow n > \frac{1}{\gamma^2} \gg n_{opt} = O\left(\frac{1}{\gamma}\right) \)

\[\text{Figure 8: High variance of } \langle f'_k, f'_{k+1} \rangle\]

\[\text{Figure 9: High variance of } S_n.\]
Theorem (Quadratic semi-stochastic framework)

Under symmetry assumptions on the noise, it holds that for all $A > 0$ and $0 \leq \alpha < 2$. Let $n_\gamma = \lfloor A/\gamma^\alpha \rfloor$. It holds that:

$$\mathbb{P}_{\theta_0 \sim \pi_{\gamma/r}} \left( S_{n_\gamma}^{(\gamma)} \leq 0 \right) \xrightarrow{\gamma \to 0} \frac{1}{2}$$

- Therefore no fixed burn-in $n_b$ can solve the variance issue
- We would have to use at least a burn-in scaling as $n_\gamma = \frac{1}{\gamma^2}$, useless since $n_{opt} \propto \frac{1}{\gamma}$.

**Conclusion:** it fails... :(  
(badly... Even mini-batch are not enough... Works if only multiplicative noise but then useless...)
Another heuristic: use
\[ \| \Omega_n \|^2 = \| \theta_n - \theta_0 \|^2. \]
\[ \| \Omega_n \|^2 = \| \eta_n \|^2 + \| \eta_0 \|^2 - 2 \langle \eta_n, \eta_0 \rangle \]

\[ \mathbb{E} [\| \Omega_n \|^2] = \mathbb{E} [\| \eta_n \|^2] + \mathbb{E} [\| \eta_0 \|^2] - 2\eta_0^T (I - \gamma H)^n \eta_0. \]
Figure 10: $\|\theta_n - \theta_0\|^2$ in plain, $\left\| H^{1/2}(\theta_n - \theta^*) \right\|^2$ in dotted
Algorithm 2 Piecewise constant SGD with new diagnosis

**INPUT:** \( \theta_0, \gamma_0 > 0, \ r \in [0, 1], \ N > 0, \ q > 1, \ \text{threshold} \in [0, 1] \)

**OUTPUT:** \( \theta_N \)

\[
\theta_{\text{restart}} \leftarrow \theta_0
\]

\[
\text{for } n = 2 \text{ to } N \text{ do}
\]

\[
\theta_n \leftarrow \theta_{n-1} - \gamma f_n'(\theta_{n-1})
\]

Compute \( \|\Omega_n\|^2 \)

\[
\text{if } \|\Omega_n\|^2 \text{ "has stopped increasing" then}
\]

\[
\gamma \leftarrow r \times \gamma
\]

\[
\theta_{\text{restart}} \leftarrow \theta_n
\]

\[
\text{end if}
\]

\[
\text{end for}
\]

\[
\text{return } \theta_N
\]
Experiments: Least squares
(smooth, strongly convex, synthetic dataset)
Experiments: Logistic regression
(smooth, weakly convex, synthetic dataset)
Experiments: Logistic regression

COVERTYPE dataset

Params: $r = 1/2$, $q = 1.5$, $\text{thresh} = 0.4$

Evolution of the distance-based statistic

Params: $r = 1/4$, $q = 2$, $\text{thresh} = 0.9$

Evolution of the distance-based statistic
Experiments: SVM
(non-smooth, strongly-convex, synthetic dataset)
Experiments: LASSO
(non-smooth, weakly convex, synthetic dataset)
Experiments: Uniformly convex $\rho = 2.5$
Back to the beginning
Training a ResNet18 on Cifar10

Figure 11: Single statistic for whole network
Figure 12: Statistic for each layer (multiple learning rates)
Conclusions

1. Constant step size strategies for SGD restarting “at saturation” result in good convergence rates (in both smooth + strongly convex and uniformly convex settings).

2. Pflug’s strategy for detecting convergence seems sound but cannot work a priori

3. We propose a new statistic based on heuristic arguments, that works well in practice.
Open directions:

1. Theoretical analysis for the “new restart” strategy
2. Restart for the averaged iterate ?
Shameless advertisement

Positions at Polytechnique:

- 2 tenure track assistant professors (Stat & Stat + Energy)
- Postdoc & PhD

Optimization, Learning, Federated Learning, High dimensional statistics.

Figure 13: The place to be
Thank you for listening!
On Convergence-Diagnostic based Step Sizes for Stochastic Gradient Descent

Aymeric Dieuleveut
CMAP, École Polytechnique, Institut Polytechnique de Paris
Joint work with Scott Pesme and Nicolas Flammarion (EPFL)

10/03/2020 Cirm Luminy - Optimization for Machine Learning