
MLSS 2020

Aaron Mishkin,
amishkin@cs.ubc.ca
“Stochastic gradient descent (SGD) is today one of the main workhorses for solving large-scale supervised learning and optimization problems.”
—Drori and Shamir [7]
Consensus Says...

...and also Agarwal et al. [1], Assran and Rabbat [2], Assran et al. [3], Bernstein et al. [5], Damaskinos et al. [6], Geffner and Domke [8], Gower et al. [9], Grosse and Salakhudinov [10], Hofmann et al. [11], Kawaguchi and Lu [12], Li et al. [13], Patterson and Gibson [15], Pillaud-Vivien et al. [16], Xu et al. [19], Zhang et al. [20]
**Motivation: Challenges in Optimization for ML**

**Stochastic gradient methods** are the most popular algorithms for fitting ML models,

\[
SGD: \quad w_{k+1} = w_k - \eta_k \nabla f_i (w_k).
\]

But practitioners face major challenges with

- **Speed**: step-size/averaging controls convergence rate.
- **Stability**: hyper-parameters must be tuned carefully.
- **Generalization**: optimizers encode statistical tradeoffs.
Better Optimization via Better Models

Idea: exploit over-parameterization for better optimization.
**Interpolation**

**Loss:** \( f(w) = \frac{1}{n} \sum_{i=1}^{n} f_i(w). \)

**Interpolation** is satisfied for \( f \) if \( \forall w, \)

\[ f(w^*) \leq f(w) \implies f_i(w^*) \leq f_i(w). \]
Constant Step-size SGD

Interpolation and smoothness imply a noise bound,

$$\mathbb{E} \| \nabla f_i(w) \|^2 \leq \rho (f(w) - f(w^*)) .$$

- SGD converges with a constant step-size [4, 17].
- SGD is (nearly) as fast as gradient descent.
- SGD converges to the
  - minimum L$_2$-norm solution for linear regression [18].
  - max-margin solution for logistic regression [14].
  - ??? for deep neural networks.

**Takeaway:** optimization speed and (some) statistical trade-offs.
What about **stability** and **hyper-parameter** tuning?

Is grid-search the best we can do?
Painless SGD
Stochastic Armijo Condition: \( f_i(w_{k+1}) \leq f_i(w_k) - c \eta_k \| \nabla f_i(w_k) \|^2. \)
Theorem 1 (Strongly-Convex). Assuming (a) interpolation, (b) $L_i$-smoothness, (c) convexity of $f_i$’s, and (d) $\mu$ strong-convexity of $f$, SGD with Armijo line-search with $c = 1/2$ in Eq. 1 achieves the rate:

$$\mathbb{E} \left[ \|w_T - w^*\|^2 \right] \leq \max \left\{ \left( 1 - \frac{\bar{\mu}}{L_{\max}} \right), \left( 1 - \bar{\mu} \eta_{\max} \right) \right\}^T \|w_0 - w^*\|^2.$$ 

Theorem 2 (Convex). Assuming (a) interpolation, (b) $L_i$-smoothness and (c) convexity of $f_i$’s, SGD with Armijo line-search for all $c > 1/2$ in Equation 1 and iterate averaging achieves the rate:

$$\mathbb{E} \left[ f(\bar{w}_T) - f(w^*) \right] \leq \frac{c \cdot \max \left\{ \frac{L_{\max}}{2(1-c)}, \frac{1}{\eta_{\max}} \right\}}{(2c - 1)^T} \|w_0 - w^*\|^2.$$ 

Theorem 3 (Non-convex). Assuming (a) the SGC with constant $\rho$ and (b) $L_i$-smoothness of $f_i$’s, SGD with Armijo line-search in Equation 1 with $c = 1 - \frac{L_{\max}}{4\rho L}$ and setting $\eta_{\max} = \frac{2}{\sqrt{5} \rho L}$ achieves the rate:

$$\min_{k=0,\ldots,T-1} \mathbb{E} \|\nabla f(w_k)\|^2 \leq \frac{10\rho L}{T} (f(w_0) - f(w^*)).$$
Painless SGD: Stochastic Armijo in Practice

Classification accuracy for ResNet-34 models trained on MNIST, CIFAR-10, and CIFAR-100.
Thanks for Listening!
**Bonus: Added Cost of Backtracking**

**Backtracking** is low-cost and averages once per-iteration.

### Iteration Costs

<table>
<thead>
<tr>
<th>Experiments</th>
<th>Tuned SGD</th>
<th>SGD + Goldstein</th>
<th>Coin-Betting</th>
<th>AdaBound</th>
<th>Adam</th>
<th>Polyak + Armijo</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CIFAR10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CIFAR100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Experiments</th>
<th>Adam</th>
<th>Polyak + Armijo</th>
<th>Coin-Betting</th>
<th>Nesterov + Armijo</th>
<th>SEG + Lipschitz</th>
</tr>
</thead>
<tbody>
<tr>
<td>mushrooms</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ijcnn</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MF: 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MF: 10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Bonus: Sensitivity to Assumptions

SGD with line-search is **robust**, but can still fail catastrophically.

---

**Bilinear with Interpolation**

- Adam
- Extra-Adam
- SEG + Lipschitz
- SVRE + Restarts

**Bilinear without Interpolation**

- Adam
- Extra-Adam
- SEG + Lipschitz
- SVRE + Restarts


