Aaron Mishkin

Research Goal: reliable and easy-to-use optimizers for ML.

Challenges in Optimization for ML

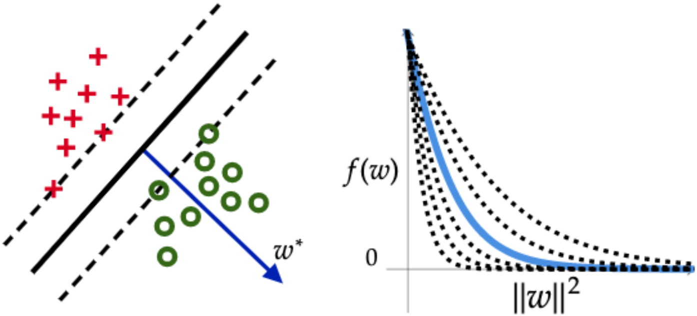
Stochastic gradient methods are the most popular algorithms for fitting ML models,

$$\textbf{SGD:} \quad w^{k+1} = w_k - \eta_k \nabla \tilde{f}(w_k).$$

But practitioners face major challenges with

- **Speed**: step-size decay-schedule controls convergence rate.
- **Stability**: hyper-parameters must be tuned carefully.
- **Generalization**: optimizers encode statistical tradeoffs.

Better Optimization via Better Models



Idea: exploit model properties for better optimization.

Consider minimizing $f(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w)$. We say f satisfies **interpolation** if $\forall w$,

$$f(w^*) \leq f(w) \implies f_i(w^*) \leq f_i(w).$$

First Steps: Constant Step-size SGD

Interpolation and smoothness imply a **noise bound**,

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

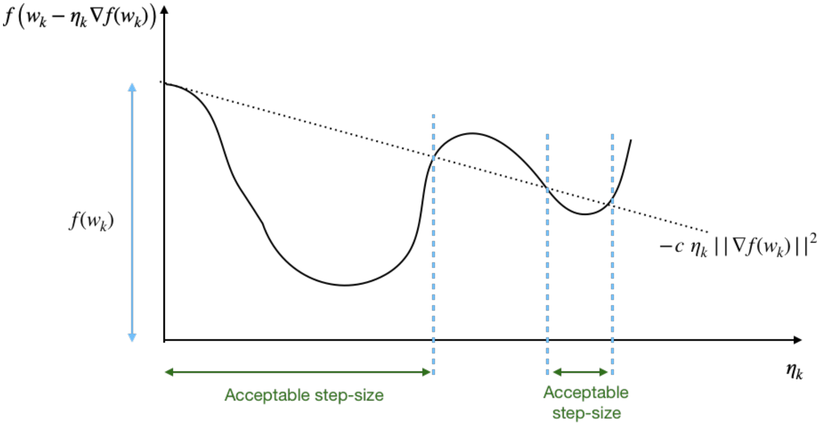
- SGD converges with a **constant step-size** [1, 5].
- SGD is as **fast** as gradient descent.
- SGD converges to the
  - minimum L_2-norm solution for linear regression [7].
  - max-margin solution for logistic regression [4].

Takeaway: optimization speed and (some) statistical trade-offs.

Current Work: Robust Parameter-free SGD

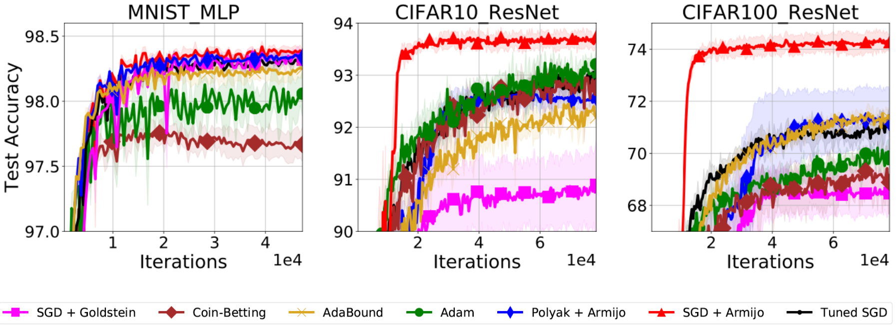
We can even pick η_k using backtracking line-search [6]!

$$\textbf{Armijo Condition}: \; f_i(w_{k+1}) \leq f_i(w_k) - c\,\eta_k \|\nabla f_i(w_k)\|^2.$$



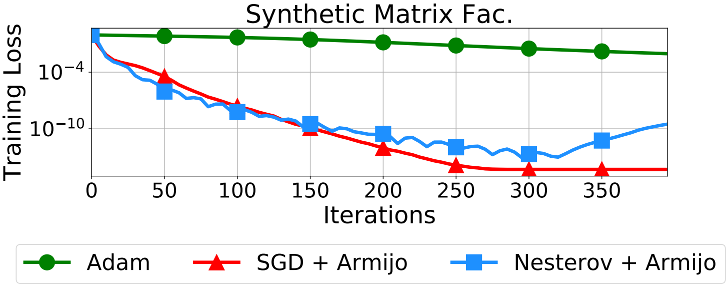
Stochastic Line-Searches in Practice

Classification accuracy for ResNet-34 models trained on MNIST, CIFAR-10, and CIFAR-100.



Questions.

Bonus: Robust Acceleration for SGD



Stochastic acceleration is possible [3, 5], but

- it's **unstable** with the backtracking Armijo line-search; and
- the "acceleration" parameter must be **fine-tuned**.

Potential Solutions:

- more sophisticated line-search (e.g. FISTA [2]).
- stochastic restarts for oscilations.

References I

[1] Raef Bassily, Mikhail Belkin, and Siyuan Ma. On exponential convergence of sgd in non-convex over-parametrized learning. *arXiv preprint arXiv:1811.02564*, 2018.

[2] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sciences*, 2(1):183–202, 2009.

[3] Chaoyue Liu and Mikhail Belkin. Accelerating sgd with momentum for over-parameterized learning. In *ICLR*, 2020.

[4] Mor Shpigel Nacson, Nathan Srebro, and Daniel Soudry. Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate. *arXiv preprint arXiv:1806.01796*, 2018.

References II

[5] Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1195–1204, 2019.

[6] Sharan Vaswani, Aaron Mishkin, Issam H. Laradji, Mark Schmidt, Gauthier Gidel, and Simon Lacoste-Julien. Painless stochastic gradient: Interpolation, line-search, and convergence rates. In *NeurIPS*, pages 3727–3740, 2019.

[7] Ashia C Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro, and Benjamin Recht. The marginal value of adaptive gradient methods in machine learning. In *NeurIPS*, pages 4148–4158, 2017.