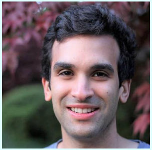
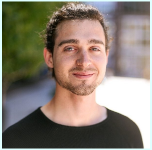
Painless Stochastic Gradient Descent: Interpolation, Line-Search, and Convergence Rates.

MLSS 2020

Aaron Mishkin,
amishkin@cs.ubc.ca

Stochastic Gradient Descent: Workhorse of ML?

"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

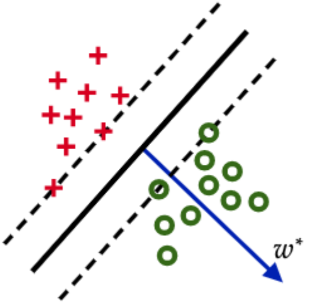
Stochastic gradient methods are the most popular algorithms for fitting ML models,

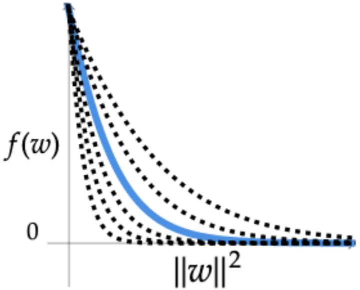
$$\textbf{SGD:} \quad w_{k+1} = w_k - \eta_k \nabla f_i(w_k).$$

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.

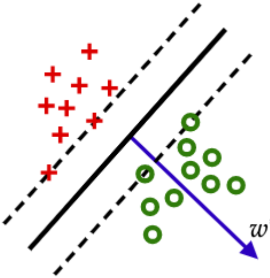
Interpolation

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

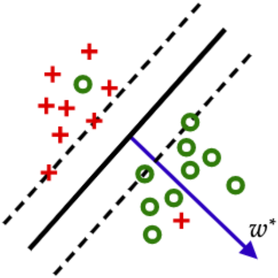
Interpolation is satisfied for f if $\forall w$,

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

Separable



Not Separable



Constant Step-size SGD

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

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

- 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?

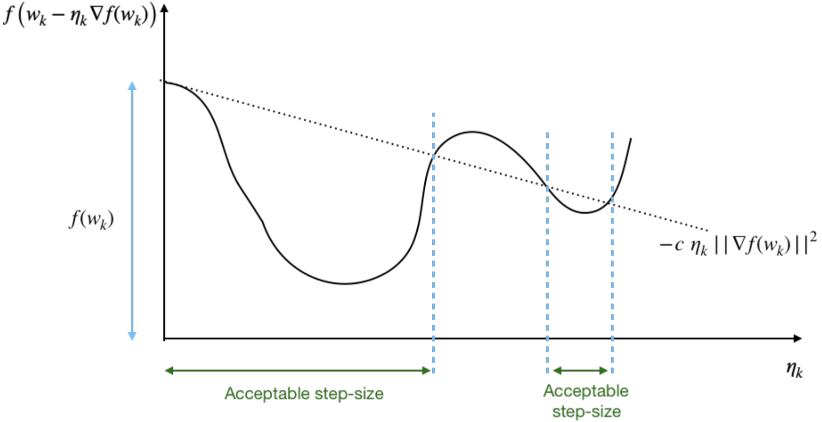
Is grid-search the best we can do?

```
for i, step_size in enumerate(np.logspace(-4,1,12)):
    opt_params["step_size"] = step_size
    results[i] = run_experiment(opt_params, exp_params, data_params, model_fn,
                                objective, error_fn, training_set, test_set)
```

Painless SGD

Painless SGD: Tuning-free SGD via Line-Searches

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



Painless SGD: Stochastic Armijo in Theory

Theorem 1 (Strongly-Convex). *Assuming (a) interpolation, (b) L_i-smoothness, (c) convexity of f_i's, and (d) μ strong-convexity of f, SGD with Armijo line-search with $c = 1/2$ in Eq. 1 achieves the rate:*

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

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:*

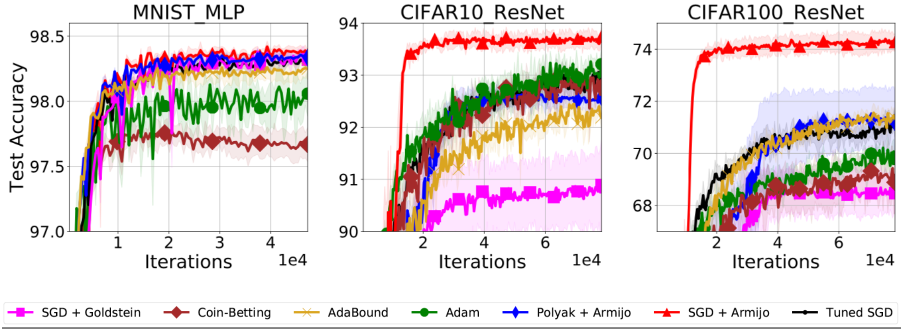
$$\mathbb{E}\left[f(\bar{w}_T) - f(w^*)\right] \leq \frac{c \cdot \max\left\{\frac{L_{max}}{2\,(1-c)}, \frac{1}{\eta_{max}}\right\}}{(2c-1)\,T} \|w_0 - w^*\|^2 .$$

Theorem 3 (Non-convex). *Assuming (a) the SGC with constant ρ 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

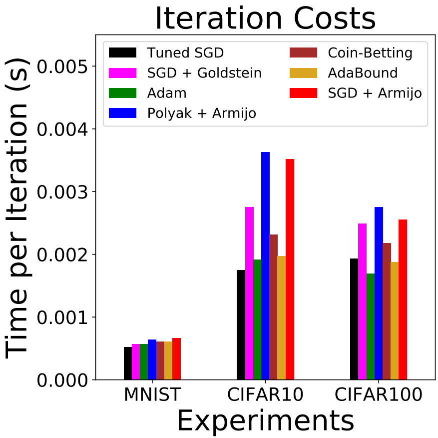
Classification accuracy for ResNet-34 models trained on MNIST, CIFAR-10, and CIFAR-100.

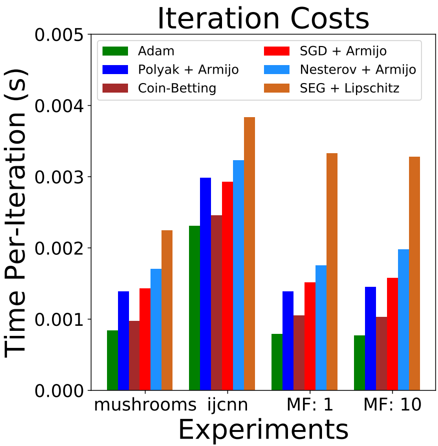


Thanks for Listening!

Bonus: Added Cost of Backtracking

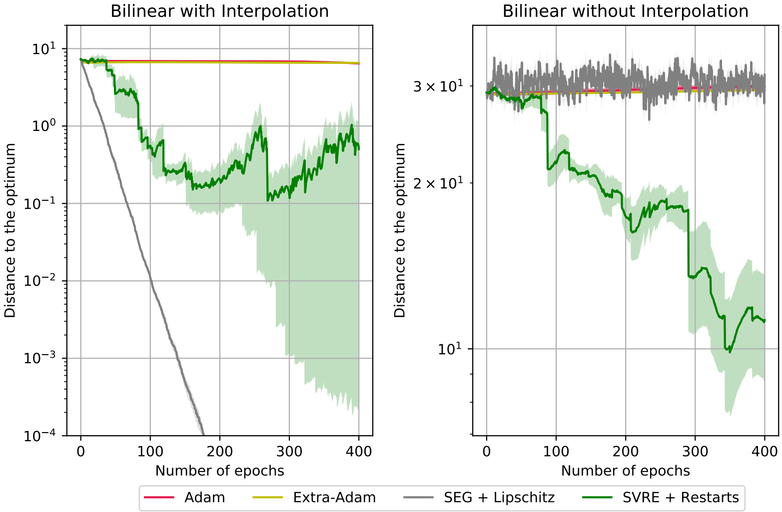
Backtracking is low-cost and averages once per-iteration.





Bonus: Sensitivity to Assumptions

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



References I

[1] Naman Agarwal, Brian Bullins, and Elad Hazan. Second-order stochastic optimization for machine learning in linear time. *The Journal of Machine Learning Research*, 18(1):4148–4187, 2017.

[2] Mahmoud Assran and Michael Rabbat. On the convergence of nesterov's accelerated gradient method in stochastic settings. *arXiv preprint arXiv:2002.12414*, 2020.

[3] Mahmoud Assran, Nicolas Loizou, Nicolas Ballas, and Michael Rabbat. Stochastic gradient push for distributed deep learning. *arXiv preprint arXiv:1811.10792*, 2018.

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

References II

[5] Jeremy Bernstein, Jiawei Zhao, Kamyar Azizzadenesheli, and Anima Anandkumar. signsgd with majority vote is communication efficient and fault tolerant. *arXiv preprint arXiv:1810.05291*, 2018.

[6] Georgios Damaskinos, El Mahdi El Mhamdi, Rachid Guerraoui, Arsany Hany Abdelmessih Guirguis, and Sébastien Louis Alexandre Rouault. Aggregathor: Byzantine machine learning via robust gradient aggregation. In *The Conference on Systems and Machine Learning (SysML), 2019*, number CONF, 2019.

[7] Yoel Drori and Ohad Shamir. The complexity of finding stationary points with stochastic gradient descent. *arXiv preprint arXiv:1910.01845*, 2019.

References III

[8] Tomas Geffner and Justin Domke. A rule for gradient estimator selection, with an application to variational inference. *arXiv preprint arXiv:1911.01894*, 2019.

[9] Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. Sgd: General analysis and improved rates. *arXiv preprint arXiv:1901.09401*, 2019.

[10] Roger Grosse and Ruslan Salakhudinov. Scaling up natural gradient by sparsely factorizing the inverse fisher matrix. In *International Conference on Machine Learning*, pages 2304–2313, 2015.

[11] Thomas Hofmann, Aurelien Lucchi, Simon Lacoste-Julien, and Brian McWilliams. Variance reduced stochastic gradient descent with neighbors. In *Advances in Neural Information Processing Systems*, pages 2305–2313, 2015.

References IV

[12] Kenji Kawaguchi and Haihao Lu. Ordered sgd: A new stochastic optimization framework for empirical risk minimization. In *International Conference on Artificial Intelligence and Statistics*, pages 669–679, 2020.

[13] Liping Li, Wei Xu, Tianyi Chen, Georgios B Giannakis, and Qing Ling. Rsa: Byzantine-robust stochastic aggregation methods for distributed learning from heterogeneous datasets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1544–1551, 2019.

[14] Mor Shpigel Nacson, Nathan Srebro, and Daniel Soudry. Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate. In *AISTATS*, volume 89 of *Proceedings of Machine Learning Research*, pages 3051–3059. PMLR, 2019.

References V

[15] Josh Patterson and Adam Gibson. *Deep learning: A practitioner's approach*. " O'Reilly Media, Inc.", 2017.

[16] Loucas Pillaud-Vivien, Alessandro Rudi, and Francis Bach. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. In *Advances in Neural Information Processing Systems*, pages 8114–8124, 2018.

[17] 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.

References VI

[18] 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.

[19] Peng Xu, Farbod Roosta-Khorasani, and Michael W Mahoney. Second-order optimization for non-convex machine learning: An empirical study. *arXiv preprint arXiv:1708.07827*, 2017.

[20] Jian Zhang, Christopher De Sa, Ioannis Mitliagkas, and Christopher Ré. Parallel sgd: When does averaging help? *arXiv preprint arXiv:1606.07365*, 2016.