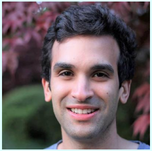
Interpolation, Growth Conditions, and Stochastic Gradient Descent

Aaron Mishkin,
mishkin@stanford.edu





"Stochastic gradient descent (SGD) is today one of the main workhorses for solving large-scale supervised learning and optimization problems."
—Drori and Shamir [2019]

. . . and also Agarwal et al. [2017], Assran and Rabbat [2020], Assran et al. [2018], Bernstein et al. [2018], Damaskinos et al. [2019], Geffner and Domke [2019], Gower et al. [2019], Grosse and Salakhudinov [2015], Hofmann et al. [2015], Kawaguchi and Lu [2020], Li et al. [2019], Patterson and Gibson [2017], Pillaud-Vivien et al. [2018], Xu et al. [2017], Zhang et al. [2016]

Challenges in Optimization for ML

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

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

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.

Challenges in Optimization for ML

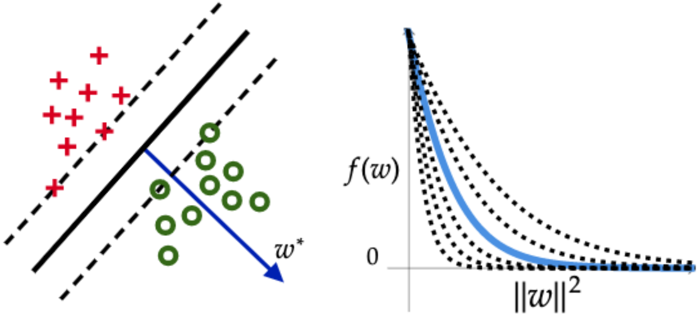
Stochastic gradient methods are the most popular algorithms for fitting ML models,

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

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.

- Intuitively, gradient noise goes to 0 if all data are fit exactly.
- No need for decreasing step-sizes or averaging.

Assumptions for Optimization

Goal: Minimize $f : \mathbb{R}^d \to \mathbb{R}$, where

Assumptions for Optimization

Goal: Minimize $f : \mathbb{R}^d \to \mathbb{R}$, where

- f is **lower-bounded**: $\exists\, w^* \in \mathbb{R}^d$ such that

$$f(w^*) \leq f(w) \qquad \forall w \in \mathbb{R}^d,$$

Assumptions for Optimization

Goal: Minimize $f : \mathbb{R}^d \to \mathbb{R}$, where

- f is **lower-bounded**: $\exists\, w^* \in \mathbb{R}^d$ such that

$$f(w^*) \le f(w) \qquad \forall w \in \mathbb{R}^d,$$

- f is L-**smooth**: $w \mapsto \nabla f(w)$ is L-Lipschitz,

$$\|\nabla f(w) - \nabla f(u)\|_2 \le L\|w - u\|_2 \qquad \forall w, u \in \mathbb{R}^d,$$

Assumptions for Optimization

Goal: Minimize $f : \mathbb{R}^d \to \mathbb{R}$, where

- f is **lower-bounded**: $\exists\, w^* \in \mathbb{R}^d$ such that

$$f(w^*) \le f(w) \qquad \forall w \in \mathbb{R}^d,$$

- f is L-**smooth**: $w \mapsto \nabla f(w)$ is L-Lipschitz,

$$\|\nabla f(w) - \nabla f(u)\|_2 \le L\|w - u\|_2 \qquad \forall w, u \in \mathbb{R}^d,$$

- (Sometimes) f is μ-**strongly-convex**: $\exists\, \mu > 0$ such that,

$$f(u) \ge f(w) + \langle \nabla f(w), u - w\rangle + \frac{\mu}{2}\|u - w\|_2^2 \quad \forall w, u \in \mathbb{R}^d.$$

Interpolation and Growth Conditions

Stochastic First-Order Oracles

Stochastic Oracles:

1. At each iteration k, query oracle $\mathcal{O}$ for stochastic estimates

$$f(w_k, z_k) \quad \text{and} \quad \nabla f(w_k, z_k).$$

Stochastic First-Order Oracles

Stochastic Oracles:

1. At each iteration k, query oracle $\mathcal{O}$ for stochastic estimates

$$f(w_k, z_k) \quad \text{and} \quad \nabla f(w_k, z_k).$$

2. $f(w_k, \cdot)$ is a deterministic function of random variable z_k.

Stochastic First-Order Oracles

Stochastic Oracles:

1. At each iteration k, query oracle $\mathcal{O}$ for stochastic estimates

$$f(w_k, z_k) \quad \text{and} \quad \nabla f(w_k, z_k).$$

2. $f(w_k, \cdot)$ is a deterministic function of random variable z_k.
3. $\mathcal{O}$ is **unbiased**, meaning

$$\mathbb{E}_{z_k}\left[f(w_k, z_k)\right] = f(w_k) \quad \text{and} \quad \mathbb{E}_{z_k}\left[\nabla f(w_k, z_k)\right] = \nabla f(w_k).$$

Stochastic First-Order Oracles

Stochastic Oracles:

1. At each iteration k, query oracle $\mathcal{O}$ for stochastic estimates

$$f(w_k, z_k) \quad \text{and} \quad \nabla f(w_k, z_k).$$

2. $f(w_k, \cdot)$ is a deterministic function of random variable z_k.
3. $\mathcal{O}$ is **unbiased**, meaning

$$\mathbb{E}_{z_k}\left[f(w_k, z_k)\right] = f(w_k) \quad \text{and} \quad \mathbb{E}_{z_k}\left[\nabla f(w_k, z_k)\right] = \nabla f(w_k).$$

4. $\mathcal{O}$ is **individually-smooth**, meaning $f(\cdot, z_k)$ is L_{max}-smooth,

$$\|\nabla f(w, z_k) - \nabla f(u, z_k)\|_2 \leq L_{\text{max}} \|w - u\|_2 \quad \forall w, u \in \mathbb{R}^d,$$

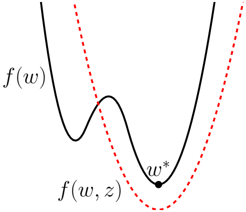
almost surely.

Interpolation as a Property of Oracles

Interpolation is a **local property** of the oracle

Interpolation as a Property of Oracles

Interpolation is a **local property** of the oracle



Definition (Interpolation: Minimizers)

$$w' \in \arg\min f \implies w' \in \arg\min f(\cdot, z_k)$$

Interpolation as a Property of Oracles

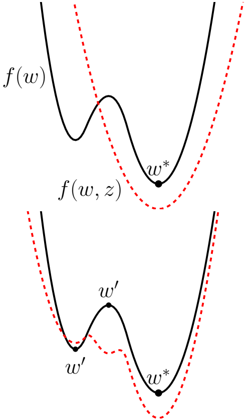
Interpolation is a **local property** of the oracle

Definition (Interpolation: Minimizers)

$$w' \in \arg\min f \implies w' \in \arg\min f(\cdot, z_k)$$

Definition (Interpolation: Stationary Points)

$$\nabla f(w') = 0 \implies \nabla f(w', z_k) \stackrel{\text{a.s.}}{=} 0.$$



Interpolation as a Property of Oracles

Interpolation is a **local property** of the oracle

Definition (Interpolation: Minimizers)

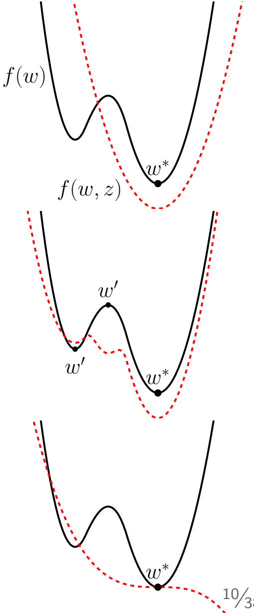
$$w' \in \arg\min f \implies w' \in \arg\min f(\cdot, z_k)$$

Definition (Interpolation: Stationary Points)

$$\nabla f(w') = 0 \implies \nabla f(w', z_k) \overset{\text{a.s.}}{=} 0.$$

Definition (Interpolation: Mixed)

$$w' \in \arg\min f \implies \nabla f(w', z_k) \overset{\text{a.s.}}{=} 0.$$



Growth Conditions: Strong and Weak Growth

Growth conditions control **global behavior** of oracles.

Growth Conditions: Strong and Weak Growth

Growth conditions control **global behavior** of oracles.

Strong Growth+Noise : $\mathbb{E}\left[\|\nabla f(w, z_k)\|^2\right] \leq \rho \|\nabla f(w)\|^2 + \sigma^2.$

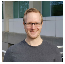
- Does not imply interpolation.

Growth Conditions: Strong and Weak Growth

Growth conditions control **global behavior** of oracles.

Strong Growth+Noise : $\mathbb{E}\left[\|\nabla f(w, z_k)\|^2\right] \leq \rho \|\nabla f(w)\|^2 + \sigma^2.$

- Does not imply interpolation.




Strong Growth : $\mathbb{E}\left[\|\nabla f(w, z_k)\|^2\right] \leq \rho \|\nabla f(w)\|^2.$

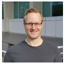
- Implies **stationary-point** interpolation.
- Originally proposed by Solodov [1998], Tseng [1998].

Growth Conditions: Strong and Weak Growth

Growth conditions control **global behavior** of oracles.

Strong Growth+Noise : $\mathbb{E}\left[\|\nabla f(w, z_k)\|^2\right] \leq \rho \|\nabla f(w)\|^2 + \sigma^2.$

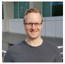
- Does not imply interpolation.




Strong Growth : $\mathbb{E}\left[\|\nabla f(w, z_k)\|^2\right] \leq \rho \|\nabla f(w)\|^2.$

- Implies **stationary-point** interpolation.
- Originally proposed by Solodov [1998], Tseng [1998].





Weak Growth : $\mathbb{E}\left[\|\nabla f(w, z_k)\|^2\right] \leq \alpha \left(f(w) - f(w^*)\right).$

- Implies **mixed** interpolation.

Growth Conditions: Interpolation + Smoothness

Smoothness relates local and global behavior.

Growth Conditions: Interpolation + Smoothness

Smoothness relates local and global behavior.

Lemma (Interpolation and Weak Growth)

If minimizer interpolation holds, then weak growth also holds with

$$\alpha \leq \frac{L_{max}}{L}.$$

Lemma (Interpolation and Strong Growth)

Assume f is μ strongly-convex. If minimizer interpolation holds, then strong growth also holds with

$$\rho \leq \frac{L_{max}}{\mu}.$$

Stochastic Gradient Descent

Fixed Step-size SGD

Fixed Step-Size SGD

0. Choose an initial point $w_0 \in \mathbb{R}^d$.

1. For each iteration $k \geq 0$:
    1.1 Query $\mathcal{O}$ for $\nabla f(w_k, z_k)$.
    1.2 Update input as

$$w_{k+1} = w_k - \eta \nabla f(w_k, z_k).$$

Fixed Step-size SGD

Prior work for SGD under growth conditions or interpolation:

- Convergence under strong growth [Cevher and Vu, 2019, Schmidt and Le Roux, 2013, Solodov, 1998, Tseng, 1998].
- Convergence under weak growth [Vaswani et al., 2019a].
- Convergence under interpolation [Bassily et al., 2018].

Fixed Step-size SGD

Prior work for SGD under growth conditions or interpolation:

- Convergence under strong growth [Cevher and Vu, 2019, Schmidt and Le Roux, 2013, Solodov, 1998, Tseng, 1998].
- Convergence under weak growth [Vaswani et al., 2019a].
- Convergence under interpolation [Bassily et al., 2018].

We provide many improved results:

- **Bigger** step-sizes and **faster** rates for convex and strongly-convex objectives.
- **Almost-sure** convergence under weak/strong growth.
- **Trade-offs** between growth conditions and interpolation.

Convergence for Fixed Step-size SGD

Theorem (Convex + Interpolation)

Assume f is convex and minimizer interpolation holds. Then SGD with $\eta = \frac{1}{L_{max}}$ converges as

$$\mathbb{E}\left[f(\bar{w}_K)\right] - f(w^*) \leq \frac{L_{max}}{2\,K}\|w_0 - w^*\|^2.$$

Convergence for Fixed Step-size SGD

Theorem (Convex + Interpolation)

Assume f is convex and minimizer interpolation holds. Then SGD with $\eta = \frac{1}{L_{max}}$ converges as

$$\mathbb{E}\left[f(\bar{w}_K)\right] - f(w^*) \leq \frac{L_{max}}{2\,K}\|w_0 - w^*\|^2.$$

Comments:

- Improves over worst-case rate with weak growth.

Convergence for Fixed Step-size SGD

Theorem (Convex + Interpolation)

Assume f is convex and minimizer interpolation holds. Then SGD with $\eta = \frac{1}{L_{max}}$ converges as

$$\mathbb{E}\left[f(\bar{w}_K)\right] - f(w^*) \leq \frac{L_{max}}{2\,K}\|w_0 - w^*\|^2.$$

Comments:

- Improves over worst-case rate with weak growth.
- If $L_{\max} = L$, then guarantee is tight with deterministic GD!

Convergence for Fixed Step-size SGD

Theorem (Convex + Interpolation)

Assume f is convex and minimizer interpolation holds. Then SGD with $\eta = \frac{1}{L_{max}}$ converges as

$$\mathbb{E}\left[f(\bar{w}_K)\right] - f(w^*) \leq \frac{L_{max}}{2\,K}\|w_0 - w^*\|^2.$$

Comments:

- Improves over worst-case rate with weak growth.
- If $L_{\max} = L$, then guarantee is tight with deterministic GD!
- Otherwise, stochasticity worsens conditioning of problem.

Weakness of Fixed Step-size SGD

Problem: these rates rely on using the optimal step-size, which depends on L_{max}, α, or ρ.

Weakness of Fixed Step-size SGD

Problem: these rates rely on using the optimal step-size, which depends on $L_{\max}$, α, or ρ.

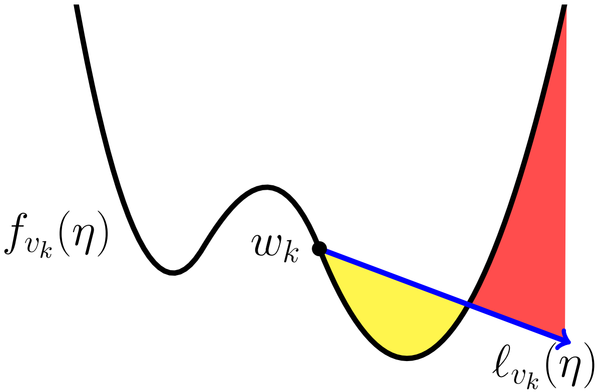
Is **grid-search** really the best way to pick η?

```python

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)

```

SGD: the Armijo Line-search

The **Armijo line-search** is a classic solution to step-size selection.

$$f(\underbrace{w_k - \eta_k \nabla f(w_k)}_{w_{k+1}}) \leq f(w_k) - c \cdot \eta_k \|\nabla f(w_k)\|^2.$$



SGD with Armijo Line-search: Procedure

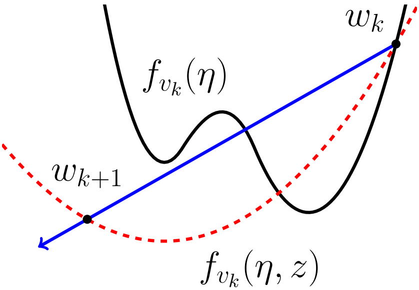
SGD with Armijo Line-Search

0. Choose an initial point $w_0 \in \mathbb{R}^d$.
1. For each iteration k:
   1.1 Query $\mathcal{O}$ for $f(w_k, z_k)$, $\nabla f(w_k, z_k)$.
   1.2 Set $w_{k+1} \leftarrow w_k - \eta_k \nabla f(w_k, z_k)$.
   1.3 Backtrack (decrease η_k) until

$$f(w_{k+1}, z_k) \leq f(w_k, z_k) - c \cdot \eta_k \|\nabla f(w_k, z_k)\|^2.$$

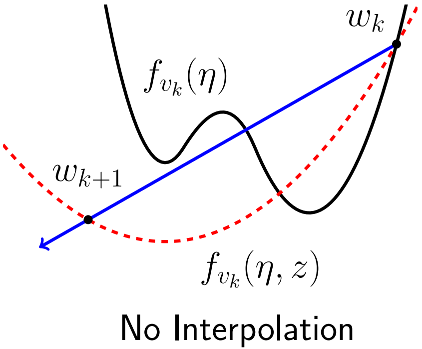
Note: Evaluates Armijo condition on $f(\cdot, z_k)$ instead of f and needs direct access to $f(\cdot, z_k)$ to backtrack.

SGD with Armijo Line-search: Visualization



No Interpolation

SGD with Armijo Line-search: Visualization



SGD with Armijo Line-search: Key Lemma

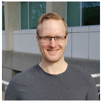
Lemma (Step-size Bound)

Assume minimizer interpolation holds.

Then the **maximal** *step-size satisfying the stochastic Armijo condition satisfies the following:*

$$\frac{2(1-c)}{L_{max}} \leq \eta_{max} \leq \frac{f(w_k, z_k) - f(w^*, z_k)}{c\|\nabla f(w_k, z_k)\|^2}.$$



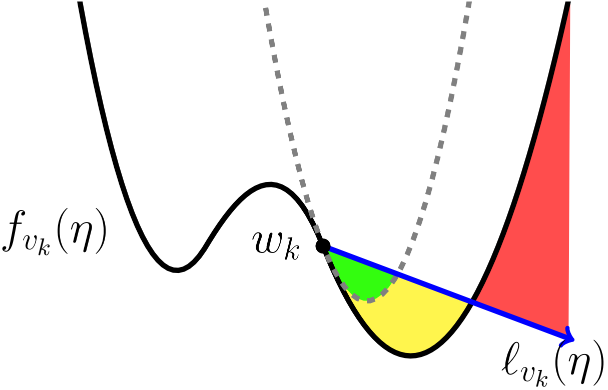




Comments:

- Mirrors classic result in deterministic optimization.
- Easy to relax to a backtracking line-search.

SGD with Armijo Line-Search: Lemma Geometry

$$\frac{2(1-c)}{L_{\mathsf{max}}} \leq \eta_{\mathsf{max}} \leq \frac{f(w_k, z_k) - f(w^*, z_k)}{c\|\nabla f(w_k, z_k)\|^2}.$$



SGD with Armijo Line-search: Convergence

Theorem (Convex + Interpolation)

Assume minimizer interpolation holds and $f(\cdot, z)$ is convex. Then SGD with the Armijo line-search and $c = 1/2$ converges as

$$\mathbb{E}\left[f(\bar{w}_K)\right] - f(w^*) \leq \frac{L_{max}}{2\,K}\|w_0 - w^*\|^2.$$

SGD with Armijo Line-search: Convergence

Theorem (Convex + Interpolation)

Assume minimizer interpolation holds and $f(\cdot, z)$ *is convex. Then SGD with the Armijo line-search and* $c = 1/2$ *converges as*

$$\mathbb{E}\left[f(\bar{w}_K)\right] - f(w^*) \leq \frac{L_{max}}{2\,K} \|w_0 - w^*\|^2.$$

Comments:

- Improves constants in original result [Vaswani et al., 2019b] — line-search is just as fast as the best constant step-size!

SGD with Armijo Line-search: Convergence

Theorem (Convex + Interpolation)

Assume minimizer interpolation holds and $f(\cdot, z)$ is convex. Then SGD with the Armijo line-search and $c = 1/2$ converges as

$$\mathbb{E}\left[f(\bar{w}_K)\right] - f(w^*) \leq \frac{L_{max}}{2\,K}\|w_0 - w^*\|^2.$$

Comments:

- Improves constants in original result [Vaswani et al., 2019b] — line-search is just as fast as the best constant step-size!
- Using the Armijo line-search is (nearly) parameter-free and recovers the deterministic rate when $L_{\max} = L$.

SGD with Armijo Line-search: Convergence

Theorem (Convex + Interpolation)

Assume minimizer interpolation holds and $f(\cdot, z)$ *is convex. Then SGD with the Armijo line-search and* $c = 1/2$ *converges as*

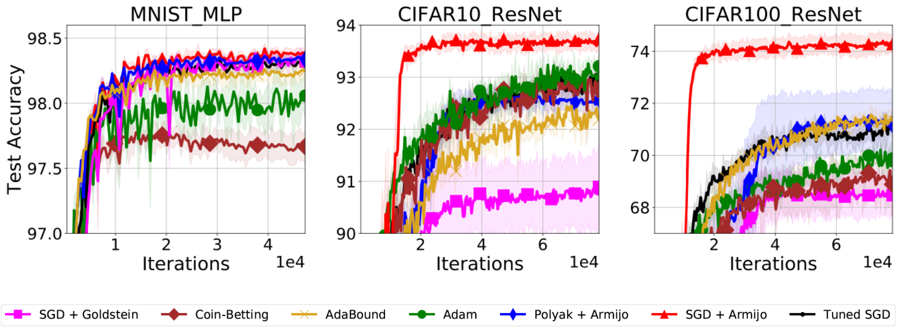
$$\mathbb{E}\left[f(\bar{w}_K)\right] - f(w^*) \leq \frac{L_{max}}{2\,K}\|w_0 - w^*\|^2.$$

Comments:

- Improves constants in original result [Vaswani et al., 2019b] — line-search is just as fast as the best constant step-size!
- Using the Armijo line-search is (nearly) parameter-free and recovers the deterministic rate when $L_{\text{max}} = L$.
- **Strongly-convex** f: we improve rate from $\bar{\mu}$ to μ.

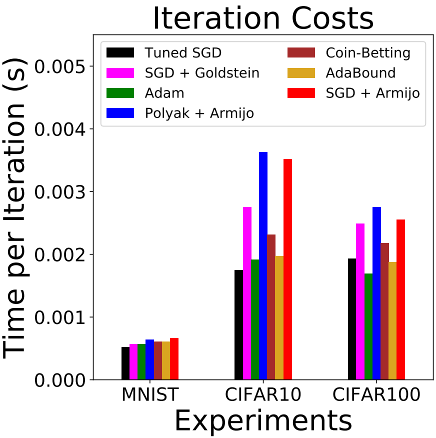
Painless SGD: Stochastic Armijo in Practice

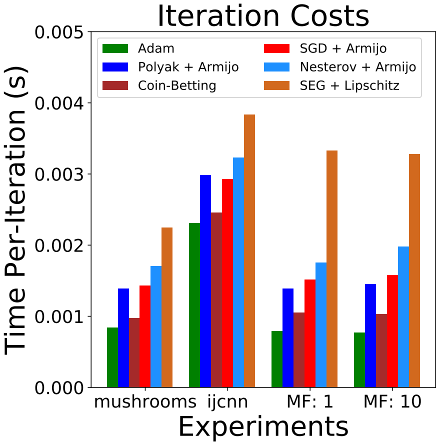
Classification accuracy for ResNet-34 models trained on MNIST, CIFAR-10, and CIFAR-100.



Painless SGD: Added Cost of Backtracking

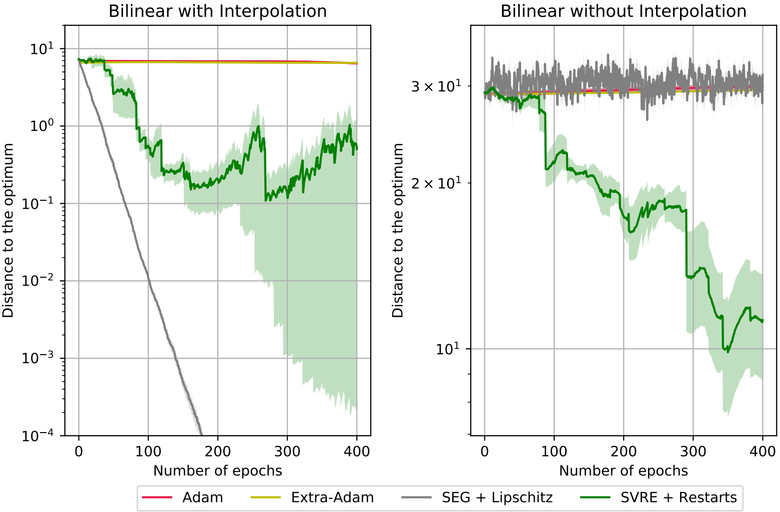
Backtracking is low-cost and averages once per-iteration.





Painless SGD: Sensitivity to Assumptions

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



Acceleration

Stochastic Acceleration

SGD can be accelerated when minimizer interpolation holds:

- Liu and Belkin [2020] modify Nesterov's method and analyze convergence for strongly-convex functions.
  - Requires an additional parameter for the fast rate.

Stochastic Acceleration

SGD can be accelerated when minimizer interpolation holds:

- Liu and Belkin [2020] modify Nesterov's method and analyze convergence for strongly-convex functions.
  - Requires an additional parameter for the fast rate.
- Vaswani et al. [2019a] analyze Nesterov's method under strong growth.
  - Shrinks the step-size and proves a slower rate.

Stochastic Acceleration

SGD can be accelerated when minimizer interpolation holds:

- Liu and Belkin [2020] modify Nesterov's method and analyze convergence for strongly-convex functions.
  - Requires an additional parameter for the fast rate.
- Vaswani et al. [2019a] analyze Nesterov's method under strong growth.
  - Shrinks the step-size and proves a slower rate.

We follow Vaswani et al. [2019a] and close the convergence gap.

Stochastic Acceleration: Main Result

Strong growth implies a modified **descent lemma**,

$$\mathbb{E}_{z_k}[f(w_{k+1})] - f(w_k) \leq \eta \left(1 - \frac{\rho L \eta}{2}\right) \|\nabla f(w_k)\|_2^2.$$

Stochastic Acceleration: Main Result

Strong growth implies a modified **descent lemma**,

$$\mathbb{E}_{z_k}[f(w_{k+1})] - f(w_k) \leq \eta \left(1 - \frac{\rho L \eta}{2}\right) \|\nabla f(w_k)\|_2^2.$$

Analysis proceeds via estimating sequences [Nesterov et al., 2018]!

Stochastic Acceleration: Main Result

Strong growth implies a modified **descent lemma**,

$$\mathbb{E}_{z_k}[f(w_{k+1})] - f(w_k) \leq \eta \left(1 - \frac{\rho L \eta}{2}\right) \|\nabla f(w_k)\|_2^2.$$

Analysis proceeds via estimating sequences [Nesterov et al., 2018]!

Theorem (Acceleration)

Assume f is strongly convex and strong growth holds. Then stochastic acceleration with step-size $\eta = 1/\rho L$ converges as

$$\mathbb{E}\left[f(w_K)\right] - f(w^*) \leq \left(1 - \sqrt{\frac{\mu}{\rho L}}\right)^K \left(f(w_0) - f(w^*) + \frac{\mu}{2}\delta_0\right).$$

Stochastic Acceleration: Main Result

Strong growth implies a modified **descent lemma**,

$$\mathbb{E}_{z_k}[f(w_{k+1})] - f(w_k) \leq \eta \left(1 - \frac{\rho L \eta}{2}\right) \|\nabla f(w_k)\|_2^2.$$

Analysis proceeds via estimating sequences [Nesterov et al., 2018]!

Theorem (Acceleration)

Assume f is strongly convex and strong growth holds. Then stochastic acceleration with step-size $\eta = 1/\rho L$ converges as

$$\mathbb{E}\left[f(w_K)\right] - f(w^*) \leq \left(1 - \sqrt{\frac{\mu}{\rho L}}\right)^K \left(f(w_0) - f(w^*) + \frac{\mu}{2}\delta_0\right).$$

- Improves dependence from ρ to $\sqrt{\rho}$

Stochastic Acceleration: Main Result

Strong growth implies a modified **descent lemma**,

$$\mathbb{E}_{z_k}[f(w_{k+1})] - f(w_k) \leq \eta \left(1 - \frac{\rho L \eta}{2}\right) \|\nabla f(w_k)\|_2^2.$$

Analysis proceeds via estimating sequences [Nesterov et al., 2018]!

Theorem (Acceleration)

Assume f is strongly convex and strong growth holds. Then stochastic acceleration with step-size $\eta = 1/\rho L$ converges as

$$\mathbb{E}\left[f(w_K)\right] - f(w^*) \leq \left(1 - \sqrt{\frac{\mu}{\rho L}}\right)^K \left(f(w_0) - f(w^*) + \frac{\mu}{2}\delta_0\right).$$

- Improves dependence from ρ to $\sqrt{\rho}$
  - Recall: $\sqrt{\rho} = \sqrt{\kappa_{\text{max}}} = \sqrt{L_{\text{max}}/\mu}$ in the worst case.

Takeaways

- **Interpolation**: the oracle model extends interpolation to general stochastic optimization problems.

Takeaways

- **Interpolation**: the oracle model extends interpolation to general stochastic optimization problems.
- **Growth Conditions**: “smooth” oracles satisfying interpolation are well-behaved globally.

Takeaways

- **Interpolation**: the oracle model extends interpolation to general stochastic optimization problems.
- **Growth Conditions**: "smooth" oracles satisfying interpolation are well-behaved globally.
- **SGD**: improved rates show SGD under interpolation is tight with the deterministic setting.

Takeaways

- **Interpolation**: the oracle model extends interpolation to general stochastic optimization problems.
- **Growth Conditions**: "smooth" oracles satisfying interpolation are well-behaved globally.
- **SGD**: improved rates show SGD under interpolation is tight with the deterministic setting.
- **Line-Search**: the Armijo line-search yields fast, parameter-free optimization under interpolation.

Takeaways

- **Interpolation**: the oracle model extends interpolation to general stochastic optimization problems.
- **Growth Conditions**: "smooth" oracles satisfying interpolation are well-behaved globally.
- **SGD**: improved rates show SGD under interpolation is tight with the deterministic setting.
- **Line-Search**: the Armijo line-search yields fast, parameter-free optimization under interpolation.
- **Acceleration**: stochastic acceleration is possible with a penalty of only $\sqrt{\rho}$.

Thanks for Listening!

References I

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.

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

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

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

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.

Volkan Cevher and Bang Công Vu. On the linear convergence of the stochastic gradient method with constant step-size. *Optim. Lett.*, 13(5):1177–1187, 2019.

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.

References III

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

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

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.

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.

References IV

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.

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.

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.

Chaoyue Liu and Mikhail Belkin. Accelerating SGD with momentum for over-parameterized learning. In *8th International Conference on Learning Representations, ICLR 2020*. OpenReview.net, 2020.

Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

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

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.

Mark Schmidt and Nicolas Le Roux. Fast convergence of stochastic gradient descent under a strong growth condition. *arXiv preprint arXiv:1308.6370*, 2013.

References VI

Mikhail V. Solodov. Incremental gradient algorithms with stepsizes bounded away from zero. *Comp. Opt. and Appl.*, 11(1):23–35, 1998.

Paul Tseng. An incremental gradient(-projection) method with momentum term and adaptive stepsize rule. *SIAM Journal on Optimization*, 8(2):506–531, 1998.

Sharan Vaswani, Francis Bach, and Mark W. 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, AISTATS 2019*, volume 89 of *Proceedings of Machine Learning Research*, pages 1195–1204. PMLR, 2019a.

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 *Advances in Neural Information Processing Systems 32: NeurIPS 2019*, pages 3727–3740, 2019b.

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.

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