Reproducibility in Optimization: Theoretical Framework and Limits

Kwangjun Ahn, Prateek Jain, Ziwei Ji, Satyen Kale, Praneeth Netrapalli, Gil I. Shamir

Reproducibility crisis in ML

- With ML models being ubiquitous in practical deployments, *reproducibility* has become critical
- Reproducibility crisis in ML is real!
- **Pineau et al (2021)** discuss reasons (e.g. insufficient docs, insufficient hparam exploration, inaccessible code, etc.) and provide recommendations for reproducibility
- Many papers show that ***same*** model trained with ***same*** data can have ***different*** predictions on the ***same*** test example!
- Reasons: non-convex objective, random init, nondeterminism in training (e.g. shuffling, parallelism, scheduling, hardware differences, round-off errors, etc.)
- Thus, forced to accept some level of irreproducibility despite controlling everything we can

Focus of study: reproducibility in convex optimization

- **Goal**: advance understanding of the **fundamental limits of reproducibility**
- **Specific focus:** *convex optimization*
- Formal definition of reproducibility (informally, deviations in outputs due to errors in basic operations, e.g. gradient computations)
- **Main results:**
  - Lower bounds on irreproducibility of convex optimization procedures in several settings of interest (e.g. smooth, non-smooth, strongly-convex, etc.)
  - Algorithms matching lower bounds
  - Reproducibility in ML settings: finite-sum, and stochastic convex optimization; tight upper and lower bounds

Defining reproducibility

- **Reproducibility:** results of a computation are the same (or largely similar) when the computation is re-run (i.e. same code running on same data)
- Not to be confused with **replicability:** results should be same (or largely similar) when same code is run on *different* data
- Irreproducibility arises when basic ops constituting computation are *inexact*
- Thus, to define reproducibility, need to specify
  - Which basic ops can be inexact, and
  - How to measure differences in results of computation.

Convex function classes

- **Smooth:** convex functions with Lipschitz continuous gradients
- **Non-smooth:** Lipschitz convex functions with potentially non-differentiable points
- **Strongly-convex:** functions *f* such that $f(x) - \frac{\mu}{2}\|x\|^2$ is convex for some $\mu > 0$

Convex optimization procedures

First-order iterative (FOI) convex optimization procedure for function class *F*

- uses ***init*** op which generates an initial point,
- uses ***gradient*** ops, which yield gradients at any given query points

Convex optimization procedures

First-order iterative (FOI) convex optimization procedure for function class *F*

- uses ***init*** op which generates an initial point,
- uses ***gradient*** ops, which yield gradients at any given query points
- constructs iterates as follows:

$$\boldsymbol{x}_t = \boldsymbol{x}_0 - \sum_{i=0}^{t-1} \lambda_i^{(t)} g(\boldsymbol{x}_i) \quad \text{for some } \lambda_i^{(t)},\ i = 0, \ldots, t-1$$

$\boldsymbol{X}_0$ is output of init op, and $g(.)$ is the gradient op

- outputs $\boldsymbol{x}_T$ for some $T > 0$
- Step sizes $\lambda_i^{(t)}$ may be adaptively chosen

Inexact ops in convex optimization

- **Inexact initialization op:** produces an initial point x_0 within distance δ from some reference point.
- **Stochastic gradient op:** produces an unbiased stochastic gradient of *f* at any query point, with variance bounded by δ^2.
- **Non-stochastic inexact gradient op:** produces a non-deterministic vector that is within distance δ from the true gradient of *f* at any query point.

(ε, δ)-deviation

- Consider *first-order* convex optimization procedure, *A,* that is built using init and gradient ops (e.g. gradient descent)
- Fix a class of convex functions *F*: e.g. smooth, non-smooth but Lipschitz, strongly-convex, etc.
- To avoid trivialities, insist that *A* outputs an ε suboptimal solution for any *f* in *F*.
- **(ε, δ)-deviation:** if x_f and x'_f are two outputs of two runs of *A* for *f* in *F*, then (ε, δ)-deviation is

$$\sup_{f \in \mathcal{F}} \mathbb{E} \, \|\boldsymbol{x}_f - \boldsymbol{x}'_f\|^2$$

For stochastic settings

Parameter vs loss/prediction reproducibility

- (ε, δ)-deviation is defined in terms of deviation of output params
- In ML settings, can also measure irreproducibility in terms of deviation in loss on a test example, or deviation in predictions on a test example
- In this work, restrict to parameter reproducibility because:
  - For general convex optimization, loss or prediction reproducibility don't make sense
  - In many ML settings, loss or prediction functions are Lipschitz in parameters; hence param reproducibility bounds can be transformed into loss/prediction reproducibility bounds
  - In practice, learned parametric model can be deployed in unknown ways; hence prediction reproducibility may not make sense even in ML settings
  - In practice, may care about multiple metrics; here param reproducibility is more useful
  - When surrogate losses are used, loss reproducibility may not be useful

Summary of results: Slowed-down SGD/GD are optimal

	Stochastic Inexact Gradient Oracle Theorem 1	Non-stochastic Inexact Gradient Oracle Theorem 2	Inexact Initialization Oracle Theorem 3
Smooth	$\Theta(\delta^2/(T\varepsilon^2))$	$\Theta(\delta^2/\varepsilon^2)$	$\Theta(\delta^2)$
Smooth Strongly-Cvx.	$\Theta(\delta^2/T \wedge \varepsilon)$	$\Theta(\delta^2 \wedge \varepsilon)$	$\Theta(e^{-\Omega(T)}\delta^2 \wedge \varepsilon)$
Nonsmooth	$\Theta(1/(T\varepsilon^2))$	$\Theta(1/(T\varepsilon^2) + \delta^2/\varepsilon^2)$	$\Theta(1/(T\varepsilon^2))$
Nonsmooth Strongly-Cvx.	$\Theta(1/T \wedge \varepsilon)$	$\Theta((1/T + \delta^2) \wedge \varepsilon)$	$\Theta(1/T \wedge \varepsilon)$

- Lower bounds are for first-order iterative algorithms like GD. Also allow *adaptivity.*
- Upper bounds are achieved by “slowed-down GD/SGD”: i.e. run with a smaller learning rate for more iterations

Summary of Results: Non-smooth Settings

	Stochastic Inexact Gradient Oracle Theorem 1	Non-stochastic Inexact Gradient Oracle Theorem 2	Inexact Initialization Oracle Theorem 3
Smooth	$\Theta(\delta^2/(T\varepsilon^2))$	$\Theta(\delta^2/\varepsilon^2)$	$\Theta(\delta^2)$
Smooth Strongly-Cvx.	$\Theta(\delta^2/T \wedge \varepsilon)$	$\Theta(\delta^2 \wedge \varepsilon)$	$\Theta(e^{-\Omega(T)}\delta^2 \wedge \varepsilon)$
Nonsmooth	$\Theta(1/(T\varepsilon^2))$	$\Theta(1/(T\varepsilon^2) + \delta^2/\varepsilon^2)$	$\Theta(1/(T\varepsilon^2))$
Nonsmooth Strongly-Cvx.	$\Theta(1/T \wedge \varepsilon)$	$\Theta((1/T + \delta^2) \wedge \varepsilon)$	$\Theta(1/T \wedge \varepsilon)$

Non-smooth settings:

- Slightest errors in ops can lead to drastic deviation!
- Intuitively, at non-differentiable points errors lead to large deviations
- Standard bound of T = $1/\varepsilon^2$ iterations can lead to *maximal* irreproducibility

Summary of Results: Strongly-Convex Settings

	Stochastic Inexact Gradient Oracle Theorem 1	Non-stochastic Inexact Gradient Oracle Theorem 2	Inexact Initialization Oracle Theorem 3
Smooth	$\Theta(\delta^2/(T\varepsilon^2))$	$\Theta(\delta^2/\varepsilon^2)$	$\Theta(\delta^2)$
Smooth Strongly-Cvx.	$\Theta(\delta^2/T \wedge \varepsilon)$	$\Theta(\delta^2 \wedge \varepsilon)$	$\Theta(e^{-\Omega(T)}\delta^2 \wedge \varepsilon)$
Nonsmooth	$\Theta(1/(T\varepsilon^2))$	$\Theta(1/(T\varepsilon^2) + \delta^2/\varepsilon^2)$	$\Theta(1/(T\varepsilon^2))$
Nonsmooth Strongly-Cvx.	$\Theta(1/T \wedge \varepsilon)$	$\Theta((1/T + \delta^2) \wedge \varepsilon)$	$\Theta(1/T \wedge \varepsilon)$

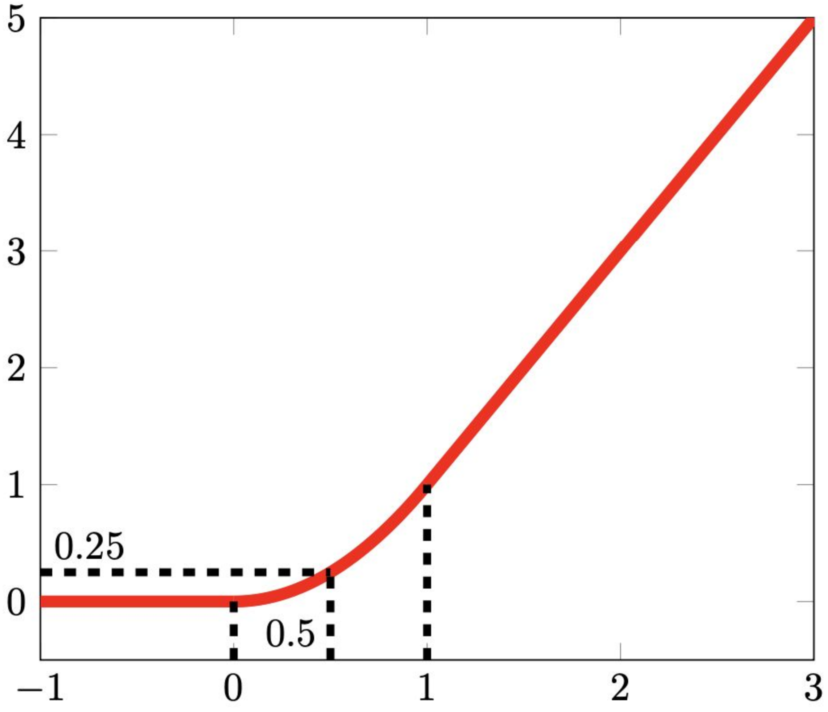
Strongly-convex settings:

- Deviation smaller due to unique minimizer
- ε-accuracy implies ε-deviation automatically

Flavor of proofs: **lower** bound for smooth convex costs with stochastic inexact gradient ops

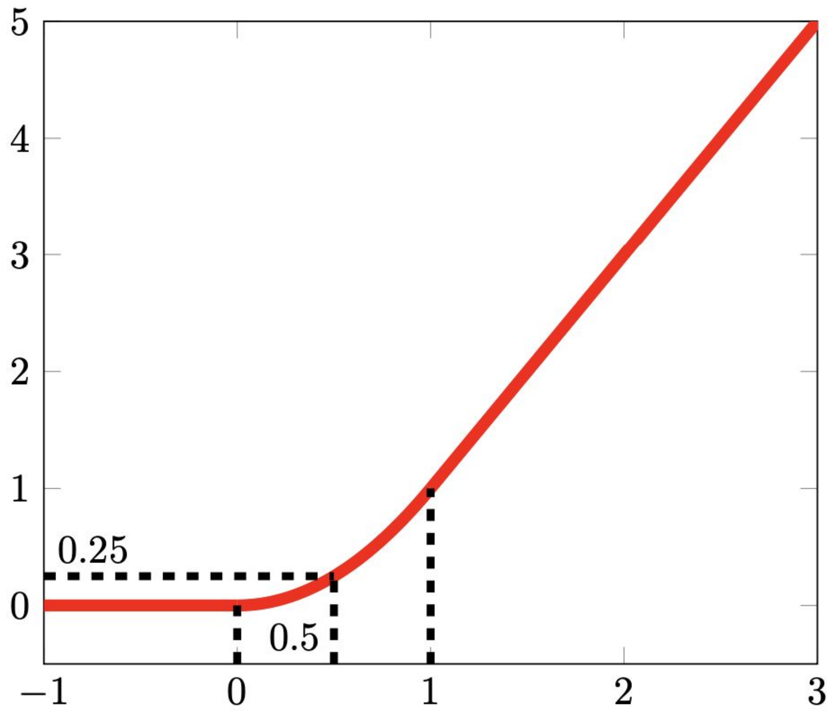
Theorem: Given ε and T, there is a smooth convex function and a δ-bounded stochastic gradient op such that any FOI alg has (ε, δ) deviation at least $\boldsymbol{\delta^2/(\varepsilon^2 T)}$

Flavor of proofs: **lower** bound for smooth convex costs with stochastic inexact gradient ops



$$\mathcal{F}(x) := \begin{cases} x^2 & \text{if } x \in [0,1], \\ 2x-1 & \text{if } x \geq 1, \\ 0 & \text{if } x \leq 0. \end{cases}$$

Flavor of proofs: **lower** bound for smooth convex costs with stochastic inexact gradient ops



Construct T+1 dim function *f* with dummy args $x \in \mathbb{R}^T$ and $y \in \mathbb{R}$ as

$$f(x, y) = 4\epsilon\mathcal{F}(y)$$

If *y* is initialized at 1, to reduce suboptimality gap to ε, *y* must decrease to 0.5

Flavor of proofs: **lower** bound for smooth convex costs

Construct T+1 dim function f with dummy args $x \in \mathbb{R}^T$ and $y \in \mathbb{R}$ as

$$f(x, y) = 4\epsilon \mathcal{F}(y)$$

If *y* is initialized at 1, to reduce suboptimality gap to ε, *y* must decrease to 0.5

Gradient norm at most ε, so for any FOI alg, average step size must be ≈ 1/(εT)

Stochastic gradient op: add ±δ noise to *x* coordinates one by one

When scaled by step sizes, deviation ≈ T x δ^2 x $1/(\varepsilon^2 T^2)$ = $\delta^2/(\varepsilon^2 T)$

Flavor of proofs: **upper** bound for smooth convex costs with stochastic inexact gradient ops

Theorem: Given ε and $T > 1/\varepsilon^2$, SGD with step size $1/(\varepsilon T)$ has (ε, δ) deviation at most $\boldsymbol{\delta^2/(\varepsilon^2 T)}$

Flavor of proofs: **upper** bound for smooth convex costs with stochastic inexact gradient ops

Let x_t and y_t be iterates of SGD and GD resp.:

$$
\begin{aligned}
&\mathbb{E}\left\|\boldsymbol{x}_{t+1}-\boldsymbol{y}_{t+1}\right\|^2=\mathbb{E}\left\|\left(\boldsymbol{x}_t-\eta_t g(\boldsymbol{x}_t)\right)-\left(\boldsymbol{y}_t-\eta_t \nabla f(\boldsymbol{y}_t)\right)\right\|^2 \\
&\quad=\left\|\boldsymbol{x}_t-\boldsymbol{y}_t\right\|^2-2\eta_t \underbrace{\mathbb{E}\left\langle\boldsymbol{x}_t-\boldsymbol{y}_t, g(\boldsymbol{x}_t)-\nabla f(\boldsymbol{y}_t)\right\rangle}_{=\left\langle\boldsymbol{x}_t-\boldsymbol{y}_t, \nabla f(\boldsymbol{x}_t)-\nabla f(\boldsymbol{y}_t)\right\rangle}+\eta_t^2\, \mathbb{E}\left\|g(\boldsymbol{x}_t)-\nabla f(\boldsymbol{y}_t)\right\|^2 \\
&\quad=\left\|\boldsymbol{x}_t-\boldsymbol{y}_t\right\|^2-2\eta_t\left\langle\boldsymbol{x}_t-\boldsymbol{y}_t, \nabla f(\boldsymbol{x}_t)-\nabla f(\boldsymbol{y}_t)\right\rangle+\eta_t^2\, \mathbb{E}\left\|g(\boldsymbol{x}_t)-\nabla f(\boldsymbol{x}_t)\right\|^2 \\
&\qquad+\underbrace{2\eta_t^2\, \mathbb{E}\left\langle g(\boldsymbol{x}_t)-\nabla f(\boldsymbol{x}_t), \nabla f(\boldsymbol{x}_t)-\nabla f(\boldsymbol{y}_t)\right\rangle}_{=0}+\eta_t^2\left\|\nabla f(\boldsymbol{x}_t)-\nabla f(\boldsymbol{y}_t)\right\|^2 \\
&\quad=\left\|\boldsymbol{x}_t-\boldsymbol{y}_t\right\|^2 \underbrace{-2\eta_t\left\langle\boldsymbol{x}_t-\boldsymbol{y}_t, \nabla f(\boldsymbol{x}_t)-\nabla f(\boldsymbol{y}_t)\right\rangle+\eta_t^2\left\|\nabla f(\boldsymbol{x}_t)-\nabla f(\boldsymbol{y}_t)\right\|^2}_{\leq 0} \\
&\qquad+\eta_t^2\, \mathbb{E}\left\|\nabla f(\boldsymbol{x}_t)-g(\boldsymbol{x}_t)\right\|^2 \\
&\quad\leq\left\|\boldsymbol{x}_t-\boldsymbol{y}_t\right\|^2+\eta_t^2 \delta^2 .
\end{aligned}
$$

Flavor of proofs: **upper** bound for smooth convex costs with stochastic inexact gradient ops

Let x_t and y_t be iterates of SGD and GD resp.: $\mathbb{E}\left\|\boldsymbol{x}_T - \boldsymbol{y}_T\right\|^2 \le \delta^2 \sum_t \eta_t^2$

Standard SGD analysis:

$$\mathbb{E}\, f\left(\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_t\right) - f(\boldsymbol{x}_*) \le \frac{L\left\|\boldsymbol{x}_0 - \boldsymbol{x}_*\right\|^2}{2T} + \frac{\left\|\boldsymbol{x}_0 - \boldsymbol{x}_*\right\|^2}{2\eta T} + \frac{\eta\delta^2}{2}$$

Flavor of proofs: **upper** bound for smooth convex costs with stochastic inexact gradient ops

Let x_t and y_t be iterates of SGD and GD resp.: $$\mathbb{E}\left\|\boldsymbol{x}_T - \boldsymbol{y}_T\right\|^2 \leq \delta^2 \sum_t \eta_t^2$$

Standard SGD analysis, with step size 1/(εT), since T > 1/ε^2: $$\mathbb{E}\, f(\bar{\boldsymbol{x}}_T) - f(\boldsymbol{x}_*) \leq O\left(\frac{1}{T} + \varepsilon + \frac{\delta^2}{\varepsilon T}\right) = O(\varepsilon)$$

Deviation bound for step size 1/(εT):

$$\mathbb{E}\left\|\bar{\boldsymbol{x}}_T - \bar{\boldsymbol{y}}_T\right\|^2 \leq \frac{1}{T}\sum_t \mathbb{E}\left\|\boldsymbol{x}_t - \boldsymbol{y}_t\right\|^2$$

Reproducibility in Optimization for ML: Finite-Sum Setting

Finite-Sum setting: optimize $f(\boldsymbol{x}) := \frac{1}{m}\sum_{i=1}^{m} f_i(\boldsymbol{x})$ via inexact gradient ops for ***component*** functions f_i.

Measure of complexity: number of component gradient calls

Lower bound via setting where all f_i are identical: $\Omega\left(1/(T\varepsilon^2) + \delta^2/\varepsilon^2\right)$

"Full batch" gradient descent: $O\left(m/(T\varepsilon^2) + \delta^2/\varepsilon^2\right)$

Main result: SGD (via randomly sampling components) gets optimal deviation!

Variance reduction doesn't help here

Reproducibility in Optimization for ML: Stochastic Convex Opt

Stochastic Convex Opt: optimize $F(\boldsymbol{x}) = \mathbb{E}_{\xi\sim\Xi}\, f(\boldsymbol{x},\xi)$ via access to samples

Given sample $\xi \sim \Xi$ obtain $f(\cdot,\xi)$

I.e. access to $(\boldsymbol{x}, f(\boldsymbol{x},\xi), \nabla f(\boldsymbol{x},\xi), \nabla^2 f(\boldsymbol{x},\xi), \cdots)$ for all ***x***

Assumption: For all ***x***, $\mathbb{E}_{\xi\sim\Xi} \left\|\nabla f(\boldsymbol{x},\xi) - \nabla F(\boldsymbol{x})\right\|^2 \le \delta^2$

For smooth *F*, lower bound for stochastic inexact gradient ops: $\Omega\left(\delta^2/(T\varepsilon^2)\right)$

Main result: Same lower bound holds despite access to more info!

Implication: SGD is optimal in this setting

Follow-Up Work

Optimal Guarantees for Algorithmic Reproducibility and Gradient Complexity in Convex Optimization

Liang Zhang*
ETH Zurich & Max Planck Institute
liang.zhang@inf.ethz.ch

Junchi Yang*
ETH Zurich
junchi.yang@inf.ethz.ch

Amin Karbasi
Yale University & Google Research
amin.karbasi@yale.edu

Niao He
ETH Zurich
niao.he@inf.ethz.ch

Follow-Up Work

	Stochastic Inexact Gradient Oracle Theorem 1	Non-stochastic Inexact Gradient Oracle Theorem 2	Inexact Initialization Oracle Theorem 3
Smooth	$\Theta(\delta^2/(T\varepsilon^2))$	$\Theta(\delta^2/\varepsilon^2)$	$\Theta(\delta^2)$

Need T > 1/ε

Accelerated GD requires only 1/√ε gradient ops for ε accuracy

But: **Attia and Koren (2021)** showed that AGD is unstable

For inexact init, deviation can be as bad as $\Theta(\delta^2 e^{1/\sqrt{\epsilon}})$

Conjecture: acceleration doesn't help for reproducibility

Follow-Up Work

	Stochastic Inexact Gradient Oracle Theorem 1	Non-stochastic Inexact Gradient Oracle Theorem 2	Inexact Initialization Oracle Theorem 3
Smooth	$\Theta(\delta^2/(T\varepsilon^2))$	$\Theta(\delta^2/\varepsilon^2)$	$\Theta(\delta^2)$

Need T > 1/ε

~~**Conjecture:** acceleration doesn't help for reproducibility~~

Zhang et al (2023) give algs using **1/√ε** gradient ops with deviation bounds:

- **$\delta^2/\varepsilon^{2.5}$** for non-stochastic inexact gradient ops
- **δ^2** for inexact init ops (optimal!)

Also extend these results for smooth convex-concave minimax problems

Limitations

- Our results don’t apply to methods like AdaGrad.
- Our lower bounds are only for “span-following” first-order-iterative methods, i.e. each iterate is constructed in the linear span of the previously seen gradients.
- We have info-theoretic lower bounds (i.e. alg can do arbitrary computations with gradients seen) for one setting, believe they hold for all

Further directions

- Extend analysis to non-convex settings
- Prove info-theoretic lower bounds!
- Experimental evidence needed: does slowing-down help in practice?
- Extensions to loss reproducibility or prediction reproducibility that may be more directly relevant to ML
- Extensions to related notions such as *replicability* (parallel work by Impagliazzo et al (2022) has initiated a study)

Thank you!

Questions?