On the Finite-Time Complexity and Practical Computation of Approximate Stationarity Concepts of Lipschitz Functions

Lai Tian

Department of Systems Engineering and Engineering Management

The Chinese University of Hong Kong (CUHK)

tianlai@se.cuhk.edu.hk

Joint work w/ Kaiwen Zhou and Anthony Man-Cho So

"Non"-problems are Pervasive

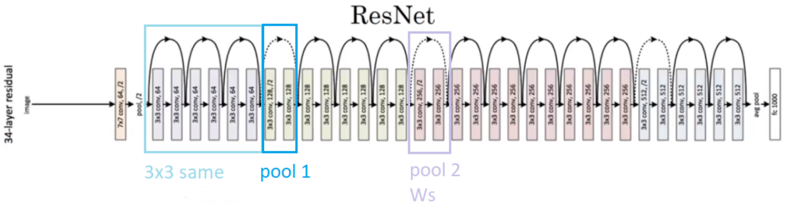


Figure: Modern ReLU Neural Networks.

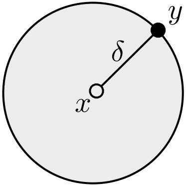
The underlying beast is ferocious!

- In the "non"-setting:
  - subdifferential calculus rule are highly non-trivial;
  - automatic differentiation may be incorrect (Kakade-Lee '18);
  - subgradient flow is pathological (Daniilidis-Drusvyatskiy '20);
  - stationarity concepts are not trivial at all (Li-So-Ma '20).

Main Question:

Can we compute any "stationary point" in a dimension-independent way?

Near-Approximate Stationarity (NAS)



$\mathrm{dist}\Big(0, \partial f(y)\Big) \leq \epsilon$

Definition (Davis-Drusvyatskiy '19, Davis-Grimmer '19)

We say x is an (ϵ, δ)-NAS point of f, if

$$\mathrm{dist}\left(0, \bigcup_{y \in \mathbb{B}_\delta(x)} \partial f(y)\right) \leq \epsilon.$$

Recall $\partial f(x) = \bigcap_{\delta > 0} \bigcup_{y \in \mathbb{B}_\delta(x)} \partial f(y)$.

Finite-Time Analysis: Positive and Negative Results

Recall $f(x)$ is ρ-weakly convex if $f(x) + \frac{\rho}{2}\|x\|^2$ is convex.

Theorem (Davis-Drusvyatskiy '19, Davis-Grimmer '19)

Simple methods compute (ϵ, δ)-NAS points for ρ-weakly convex, L-Lipschitz f, with dimension-independent complexity

$$O\left(\frac{\rho^2 L^2 + \rho L^3}{\epsilon^4} + \frac{\rho L^2 + L^3}{\rho^3 \delta^4}\right).$$

Theorem (T.-So '21)

For any first-order algorithm and finite T, there exist an L-Lip., $\rho(T)$-weakly convex f and an abs. const. $c > 0$, such that, if $0 \le \epsilon, \delta < c$, it cannot compute (ϵ, δ)-NAS points in T steps.

Goldstein Approximate Stationarity (GAS)

Definition (Goldstein '77, Burke-Lewis-Overton '05)

We say x is an (ϵ, δ)-GAS point of f, if

$$\operatorname{dist}\left(0, \operatorname{Conv}\left(\bigcup_{y \in \mathbb{B}_\delta(x)} \partial f(y)\right)\right) \leq \epsilon.$$

Note that (ϵ, δ)-NAS is (ϵ, δ)-GAS but not vice versa.

Remarks.

- Goldstein's conceptual scheme (Goldstein '77)
- existing methods: (Burke-Lewis-Overton '02), (Burke-Lewis-Overton '05), (Kiwiel '07), (Zhang-Lin-Jegelka-Sra-Jadbabaie '20);
- dimension-dependent or use impractical oracle.

Can we have a practical implementation of Goldstein's scheme in a dimension-independent way?

Figure: Allen Abbey Goldstein and Martha Goldstein.

Finite-Time Dimension-Independent Computation

Theorem (T.-So '21, T.-Zhou-So '22)

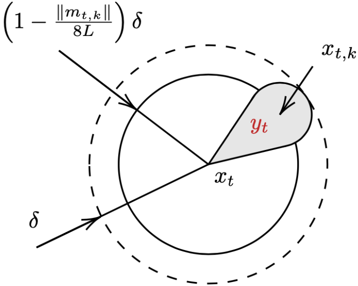
(T.-Zhou-So '22, Algorithm 1) computes an (ϵ, δ)-GAS point with probability at least $1 - \gamma$ using at most

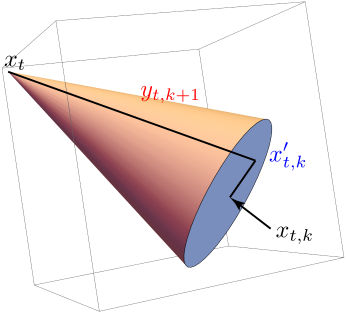
$$\frac{320\Delta L^2}{\epsilon^3 \delta} \log\left(\frac{4\Delta}{\gamma\delta\epsilon}\right) \qquad \text{standard oracle calls.}$$

Remarks.

- using the standard first-order oracle $(f, \nabla f)$;
- only evaluate ∇f at differentiable x;
  - PyTorch/TensorFlow always compute a correct gradient;
- a stochastic version using only ∇f is also available.

New Technique: Random Conic Perturbation



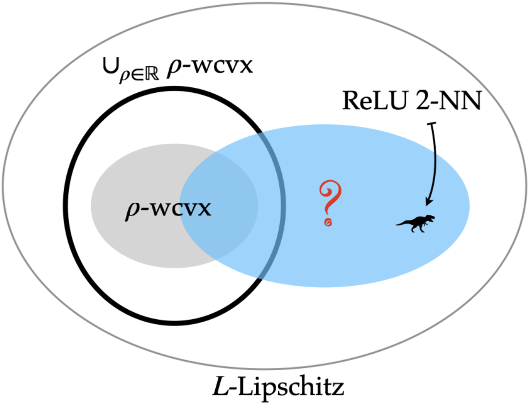


Remarks.

- algorithmically remove the unrealistic subgradient selection oracle of (Zhang-Lin-Jegelka-Sra-Jadbabaie '20);
- exploit the almost everywhere differentiability as guaranteed by Rademacher's theorem.

Compute (ϵ, δ)-NAS for 2-Layer ReLU NN

Given the inapproximability of (T.-So '21), is it still possible?



Yes, if we can catch the dinosaur!

GAS to NAS Reduction

Theorem (T.-Zhou-So '22)

Suppose f is locally Lipschitz and ∂f is (δ, η, κ)-OLC. If x is (ϵ, η)-GAS, then x is also $(\epsilon + \kappa(\delta + \eta), \delta)$-NAS.

Corollary (T.-Zhou-So '22)

We can compute (ϵ, δ)-NAS for 2-Layer ReLU NN in $poly\left(\epsilon^{-1}, \delta^{-1}, L, \kappa(Z), \|Z\|, \log(\gamma^{-1})\right)$ iterations w.p. at least $1 - \gamma$ by PyTorch/TensorFlow, where Z is the data matrix.

- applicable $\forall \#\{\text{hidden units}\}$ (underparameterized regime);
- dimension-independent;
- largely beyond ρ-weakly convexity;
  - ReLU 2-NN is not ρ-weakly convex, $\forall \rho \in \mathbb{R}$;
- many calculus rules and other applications.

main reference:

- L. Tian, K. Zhou, A. M.-C. So. "On the Finite-Time Complexity and Practical Computation of Approximate Stationarity Concepts of Lipschitz Functions," *ICML*, 2022.
- L. Tian, A. M.-C. So. "Computing Goldstein (ϵ, δ)-Stationary Points of Lipschitz Functions in $\widetilde{O}(\epsilon^{-3}\delta^{-1})$ Iterations via Random Conic Perturbation," *arXiv preprint arXiv:2112.09002*, 2021.
- L. Tian, A. M.-C. So. "On the Hardness of Computing Near-Approximate Stationary Points of Clarke Regular Nonsmooth Nonconvex Problems and Certain DC Programs," *ICML BFOM Workshop*, 2021.

Thank You!