The Role of Randomness in Stability

Talk by: Max Hopkins (Princeton-IAS)

Joint with:

Shay Moran
(Technion)

Stability in Learning

- Many standard ML & Optimization methods suffer from *instability*
  - E.g. may find very different local optima over two runs
  - May be heavily sample dependent (e.g. ERM)

Stability in Learning

- Many standard ML & Optimization methods suffer from *instability*
  - E.g. may find very different local optima over two runs
  - May be heavily sample dependent (e.g. ERM)
- Instability leads to critical failures in practice including
  - *Privacy leakage*: data dependence may reveal sensitive info!
  - *Replicability crisis*: instability may lead to non-replicable experiments

Stability in Learning

- Many standard ML & Optimization methods suffer from *instability*
  - E.g. may find very different local optima over two runs
  - May be heavily sample dependent (e.g. ERM)
- Instability leads to critical failures in practice including
  - *Privacy leakage*: data dependence may reveal sensitive info!
  - *Replicability crisis*: instability may lead to non-replicable experiments
- Lead to emergence of powerful theory of stability in learning:

Stability in Learning

- Many standard ML & Optimization methods suffer from *instability*
  - E.g. may find very different local optima over two runs
  - May be heavily sample dependent (e.g. ERM)
- Instability leads to critical failures in practice including
  - *Privacy leakage*: data dependence may reveal sensitive info!
  - *Replicability crisis*: instability may lead to non-replicable experiments
- Lead to emergence of powerful theory of stability in learning:
  1. **Differential privacy** [DMNS'06]: For all neighboring samples S, S':

$$\forall \mathcal{O} : \Pr_r[\mathcal{A}(S;r) \in \mathcal{O}] \leq e^{\varepsilon} \Pr_r[\mathcal{A}(S';r) \in \mathcal{O}] + \delta$$

   - I.e. (randomized) ouptuts of $\mathcal{A}(S)$ and $\mathcal{A}(S')$ are similarly distributed

Stability in Learning

- Many standard ML & Optimization methods suffer from *instability*
  - E.g. may find very different local optima over two runs
  - May be heavily sample dependent (e.g. ERM)
- Instability leads to critical failures in practice including
  - *Privacy leakage*: data dependence may reveal sensitive info!
  - *Replicability crisis*: instability may lead to non-replicable experiments
- Lead to emergence of powerful theory of stability in learning:
  1. **Differential privacy** [DMNS'06]: For all neighboring samples S, S':

$$\forall \mathcal{O} : \Pr_r[\mathcal{A}(S;r) \in \mathcal{O}] \leq e^{\varepsilon} \Pr_r[\mathcal{A}(S';r) \in \mathcal{O}] + \delta$$

     - I.e. (randomized) ouptuts of $\mathcal{A}(S)$ and $\mathcal{A}(S')$ are similarly distributed
  2. **Replicability** [ILPS'22]: For independent samples S, S' and random r:

$$\forall D : \Pr_{S,S' \sim D^n; r \sim \{0,1\}^*}[A(S;r) = A(S';r)] \geq 1 - \rho$$

     - $\mathcal{A}$ replicates over independent samples and *shared* randomness

Stability and Randomness

- Strong stability notions like DP are inherently *randomized*.

Stability and Randomness

- Strong stability notions like DP are inherently *randomized*.
- This is a bit of a problem... clean randomness is expensive!
  - US 2020 Census required (estimated) 90TB of randomness [GL20]
  - Moreover, corrupted/correlated randomness may leak privacy!

Stability and Randomness

- Strong stability notions like DP are inherently *randomized*.
- This is a bit of a problem... clean randomness is expensive!
  - US 2020 Census required (estimated) 90TB of randomness [GL20]
  - Moreover, corrupted/correlated randomness may leak privacy!
- This motivates the study of *randomness complexity* in stability:
  - Can we build randomness-efficient stable algorithms?
  - Can we *characterize* # random bits needed for a given task?

Stability and Randomness

- Strong stability notions like DP are inherently *randomized*.
- This is a bit of a problem... clean randomness is expensive!
  - US 2020 Census required (estimated) 90TB of randomness [GL20]
  - Moreover, corrupted/correlated randomness may leak privacy!
- This motivates the study of *randomness complexity* in stability:
  - Can we build randomness-efficient stable algorithms?
  - Can we *characterize* # random bits needed for a given task?
- Studied for basic estimation tasks in [DPWV'23,CSV'24]
  - Nothing known for more general settings (e.g. classification)

Randomness Complexity and Global Stability

- Fix a learning 'task' $\mathcal{M}$ (e.g. mean estimation, classification...)

Definition (Randomness Complexity)

The *randomness complexity* of $\mathcal{M}$, $\mathbf{C}_{\text{Rep}}$, is the smallest $\ell \in \mathbb{N}$ s.t. $\exists$ a $> 1/2$-replicable algorithm for $\mathcal{M}$ using ℓ random bits:

$$\forall D : \Pr_{S,S' \sim D^n, r \sim \{0,1\}^\ell}[\mathcal{A}(S;r) = \mathcal{A}(S';r)] > \frac{1}{2}$$

Randomness Complexity and Global Stability

- Fix a learning 'task' $\mathcal{M}$ (e.g. mean estimation, classification...)

Definition (Randomness Complexity)

The *randomness complexity* of $\mathcal{M}$, $\mathbf{C}_{\text{Rep}}$, is the smallest $\ell \in \mathbb{N}$ s.t. $\exists$ a $> 1/2$-replicable algorithm for $\mathcal{M}$ using ℓ random bits:

$$\forall D : \Pr_{S,S' \sim D^n, r \sim \{0,1\}^\ell}[\mathcal{A}(S;r) = \mathcal{A}(S';r)] > \frac{1}{2}$$

- We relate C_{Rep} to $\mathcal{M}$'s *global stability*
  - Key notion in study of differentially private learning [BLM'20]
  - Roughly, the best replication probability of any *determinstic* Alg for $\mathcal{M}$

$$\forall D : \Pr_{S,S' \sim D}[\mathcal{A}(S) = \mathcal{A}(S')] \geq \eta$$

Main Result: Characterizing Randomness Complexity

- Define $\mathbf{C}_{\mathsf{Glob}} \coloneqq \log \frac{1}{\eta_{\mathsf{max}}}$

Main Result: Characterizing Randomness Complexity

- Define $\mathbf{C}_{\text{Glob}} := \log \frac{1}{\eta_{\text{max}}}$

Theorem (Characterizing Randomness Complexity)

For any statistical task: $C_{Glob} \leq C_{Rep} \leq C_{Glob} + 1.$

Main Result: Characterizing Randomness Complexity

- Define $\mathbf{C}_{\text{Glob}} := \log \frac{1}{\eta_{\text{max}}}$

Theorem (Characterizing Randomness Complexity)

For any statistical task: $C_{Glob} \leq C_{Rep} \leq C_{Glob} + 1$.

- Further, we can achieve $1 - \rho$ replication in $\lceil C_{\text{Glob}} + \log \frac{1}{\rho} \rceil$ bits!

Main Result: Characterizing Randomness Complexity

- Define $\mathbf{C}_{\text{Glob}} := \log \frac{1}{\eta_{\text{max}}}$

Theorem (Characterizing Randomness Complexity)

For any statistical task: $C_{Glob} \leq C_{Rep} \leq C_{Glob} + 1$.

- Further, we can achieve $1 - \rho$ replication in $\lceil C_{\text{Glob}} + \log \frac{1}{\rho} \rceil$ bits!
- We prove a (weaker) analogous result for Differential Privacy
  - Real statement is more involved
  - *Very* roughly, result of the form:

$$C_{\text{Glob}} - O(1) \leq C_{DP} \leq C_{\text{Glob}} + O(1)$$

Main Result: Characterizing Randomness Complexity

- Define $\mathbf{C}_{\text{Glob}} := \log \frac{1}{\eta_{\max}}$

Theorem (Characterizing Randomness Complexity)

For any statistical task: $C_{Glob} \leq C_{Rep} \leq C_{Glob} + 1$.

- Further, we can achieve $1 - \rho$ replication in $\lceil C_{\text{Glob}} + \log \frac{1}{\rho} \rceil$ bits!
- We prove a (weaker) analogous result for Differential Privacy
  - Real statement is more involved
  - *Very* roughly, result of the form:

$$C_{\text{Glob}} - O(1) \leq C_{DP} \leq C_{\text{Glob}} + O(1)$$

- Real version comes with several caveats
  - Both directions and C_{DP} itself depend on privacy parameters
  - DP $\rightarrow$ Stability requires strong privacy guarantees ($\varepsilon \lesssim 1/\sqrt{n}$)
  - Also depends on #samples & confidence (if not 'user-private')

Application: Stable PAC Learning

- Classical PAC-Learning framework [VC'74,Val'84] consists of:
  - Domain X (e.g. $\mathbb{R}^d$), (binary) hypothesis class H (e.g. halfspaces)
  - **Adversary**: picks unknown distribution D over $X \times \{0,1\}$
  - **Learner:** Given samples from D, output $h \in H$ w/ near-optimal error

Application: Stable PAC Learning

- Classical PAC-Learning framework [VC'74,Val'84] consists of:
  - Domain X (e.g. $\mathbb{R}^d$), (binary) hypothesis class H (e.g. halfspaces)
  - **Adversary**: picks unknown distribution D over $X \times \{0,1\}$
  - **Learner:** Given samples from D, output $h \in H$ w/ near-optimal error

Theorem (Stable PAC Learning)

For any (X, H) with $O(1)$ Littlestone dimension, $\exists$ replicable learner with

1. **Sample Complexity:** $\mathit{poly}(\varepsilon^{-1}, \log(1/\delta))$
2. **Random Bits:** $O\left(\log \frac{1}{\varepsilon}\right)$

Moreover, if $\mathit{Lit}(H) = \infty$, no replicable learner exists.

Application: Stable PAC Learning

- Classical PAC-Learning framework [VC'74,Val'84] consists of:
  - Domain X (e.g. $\mathbb{R}^d$), (binary) hypothesis class H (e.g. halfspaces)
  - **Adversary**: picks unknown distribution D over $X \times \{0,1\}$
  - **Learner:** Given samples from D, output $h \in H$ w/ near-optimal error

Theorem (Stable PAC Learning)

For any (X, H) with $O(1)$ Littlestone dimension, $\exists$ replicable learner with

1. **Sample Complexity:** $poly(\varepsilon^{-1}, \log(1/\delta))$
2. **Random Bits:** $O\left(\log \frac{1}{\varepsilon}\right)$

Moreover, if $Lit(H) = \infty$, no replicable learner exists.

- Implies first (agnostic) stable learners for, e.g.
  - Affine subspaces of $\mathbb{R}^d$
  - Halfspaces with margin

Application: Stable PAC Learning

- Classical PAC-Learning framework [VC'74,Val'84] consists of:
  - Domain X (e.g. $\mathbb{R}^d$), (binary) hypothesis class H (e.g. halfspaces)
  - **Adversary**: picks unknown distribution D over $X \times \{0, 1\}$
  - **Learner:** Given samples from D, output $h \in H$ w/ near-optimal error

Theorem (Stable PAC Learning)

For any (X, H) with $O(1)$ Littlestone dimension, $\exists$ replicable learner with

1. **Sample Complexity:** $poly(\varepsilon^{-1}, \log(1/\delta))$
2. **Random Bits:** $O\left(\log \frac{1}{\varepsilon}\right)$

Moreover, if $Lit(H) = \infty$, no replicable learner exists.

- Implies first (agnostic) stable learners for, e.g.
  - Affine subspaces of $\mathbb{R}^d$
  - Halfspaces with margin
- Also gives first randomness-efficient DP alg for these problems!

Thanks for listening