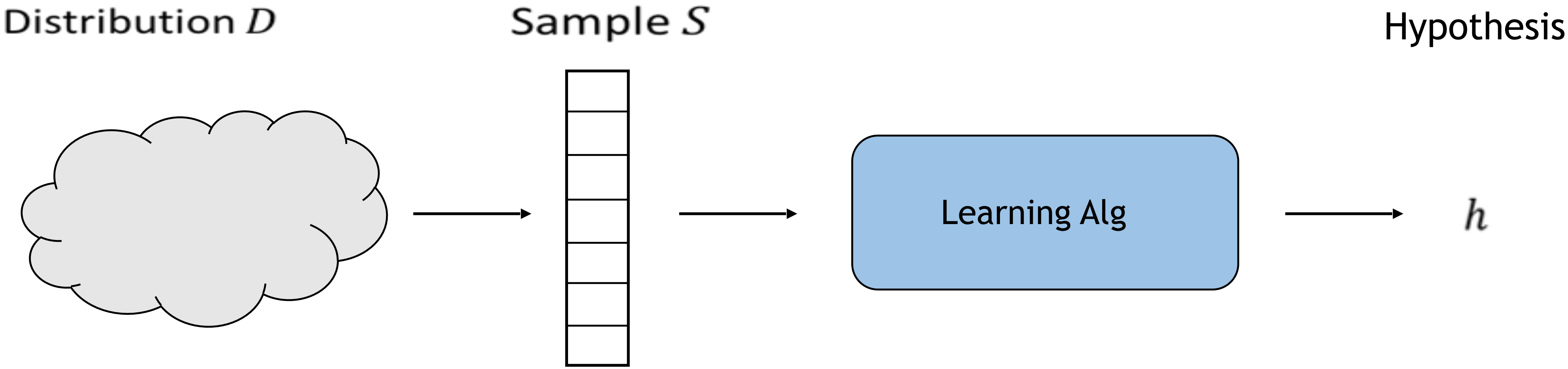
Making Generalization Robust

Katrina Ligett
HUJI & Caltech

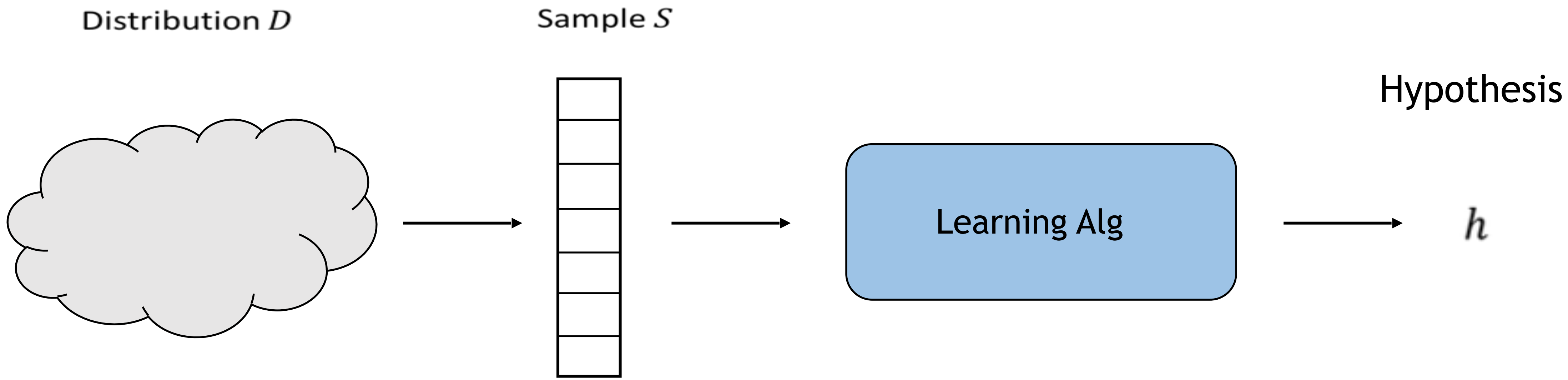
joint with Rachel Cummings, Kobbi Nissim, Aaron Roth, and Steven Wu

A model for science…

A model for science…



- domain: contains all possible examples
- hypothesis: X-> {0,1} labels examples
- learning alg samples labeled examples, returns hypothesis



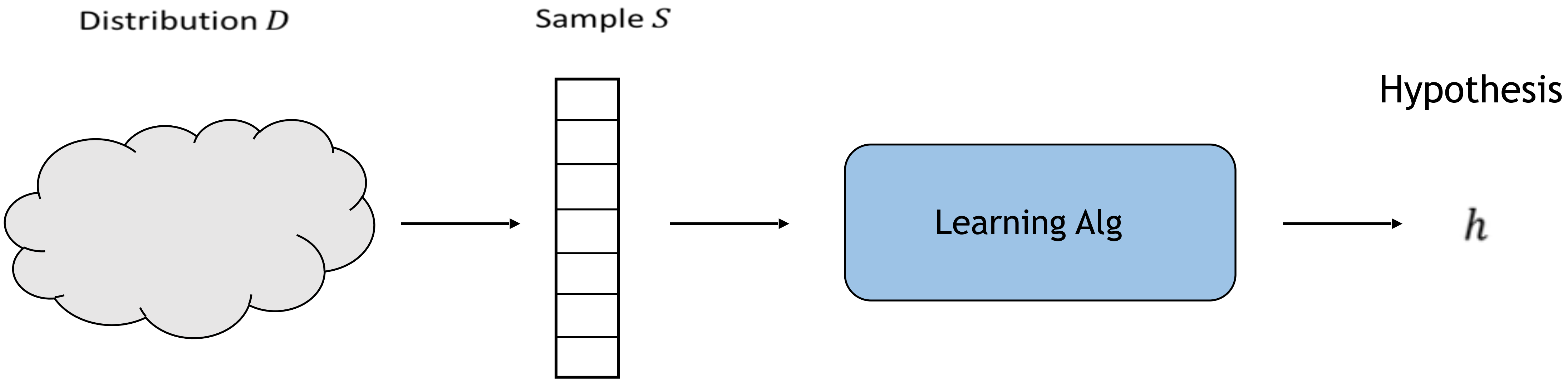
The goal of science:

Find hypothesis that has low true error on the distribution D:

$$\mathrm{err}(h) = \Pr_{x \sim D}[h(x) \neq h^*(x)]$$

Why does science work?

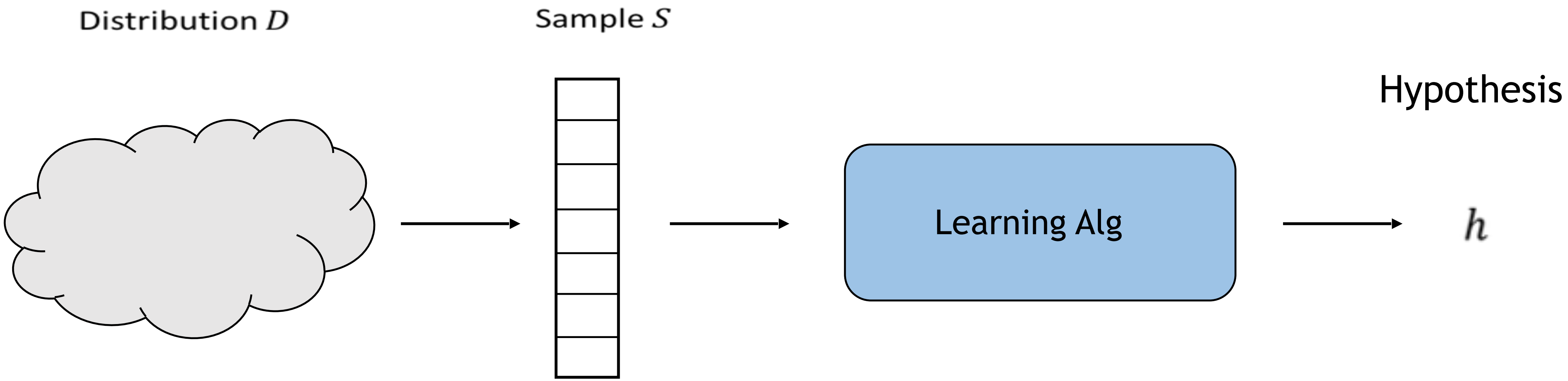
Why does science work?



The goal of science:
Find hypothesis that has low true error on the distribution D:

$$\mathrm{err}(h) = \Pr_{X\sim D}[h(x) \neq h^*(x)]$$

Idea: find hypothesis that has low empirical error on S, plus guarantee that findings on the sample *generalize* to D



Empirical error:

$$err_E(h) = 1/n \sum_{x \in S} \mathbf{1}[h(x) \neq h^*(x)]$$

Generalization: output h s.t.

$$\Pr[|h(S) - h(D)|] \leq \boldsymbol{\alpha}] \geq 1 - \beta$$

THEOREM 6.7 (The Fundamental Theorem of Statistical Learning) *Let $\mathcal{H}$ be a hypothesis class of functions from a domain $\mathcal{X}$ to $\{0,1\}$ and let the loss function be the $0-1$ loss. Then, the following are equivalent:*

1. *$\mathcal{H}$ has the uniform convergence property.*
2. *Any ERM rule is a successful agnostic PAC learner for $\mathcal{H}$.*
3. *$\mathcal{H}$ is agnostic PAC learnable.*
4. *$\mathcal{H}$ is PAC learnable.*
5. *Any ERM rule is a successful PAC learner for $\mathcal{H}$.*
6. *$\mathcal{H}$ has a finite VC-dimension.*

taken from *Understanding Machine Learning*, Shai Shalev-Schwarts and Shai Ben-David

Problem solved!

Problem solved?

Science doesn’t happen in a vacuum.

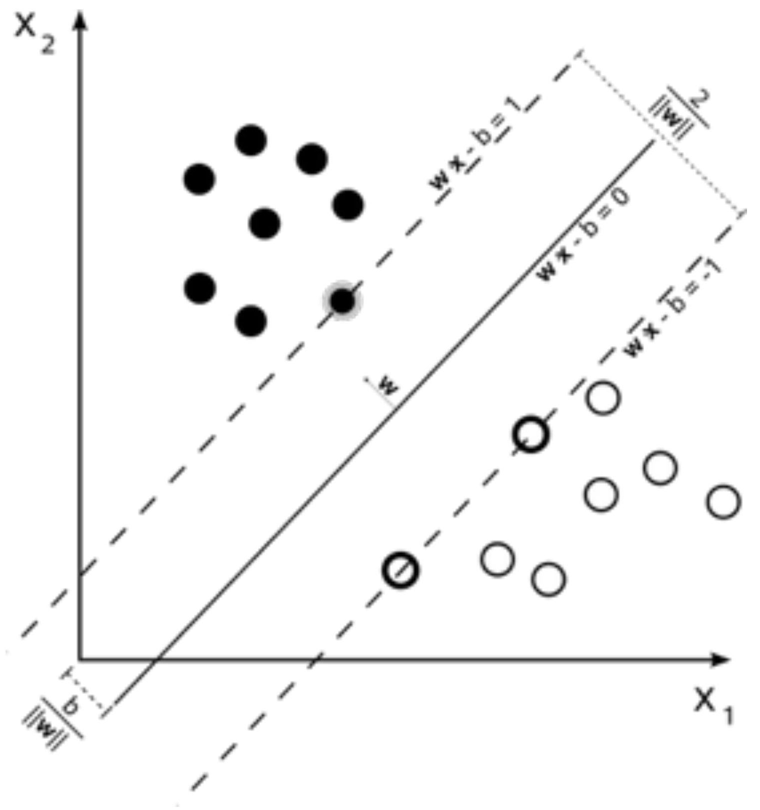
One thing that can go wrong: *post-processing*

Example: $S \in \{0,1\}^n$ generalizing hypothesis h, $|h| = \ell$

$$h' = \langle \underbrace{\quad h \quad}_{\ell} \mid \underbrace{\qquad S \qquad}_{n} \rangle$$ and define $h'(\cdot) = h(\cdot)$

h' generalizes but encodes the entire sample!

- Doesn't have to be explicit or malicious.



- Learning an SVM: Output encodes Support Vectors (sample points)
- This output could be post-processed to obtain a non-generalizing hypothesis: “10% of all data points are x_k”

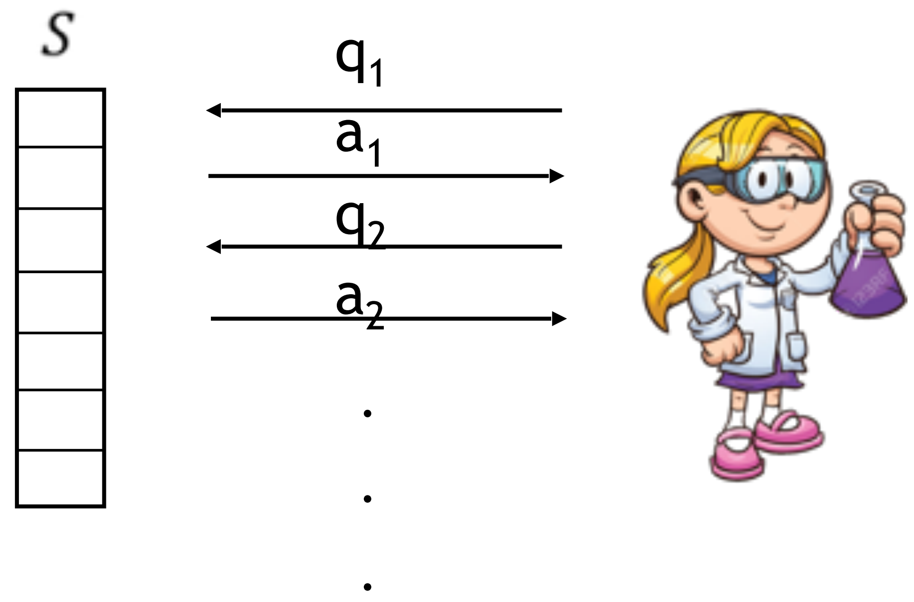
Oh, man. Our approach on this Kaggle competition really failed on the test data. Oh well, let’s try again.

Did you see that paper published by the Smith lab?

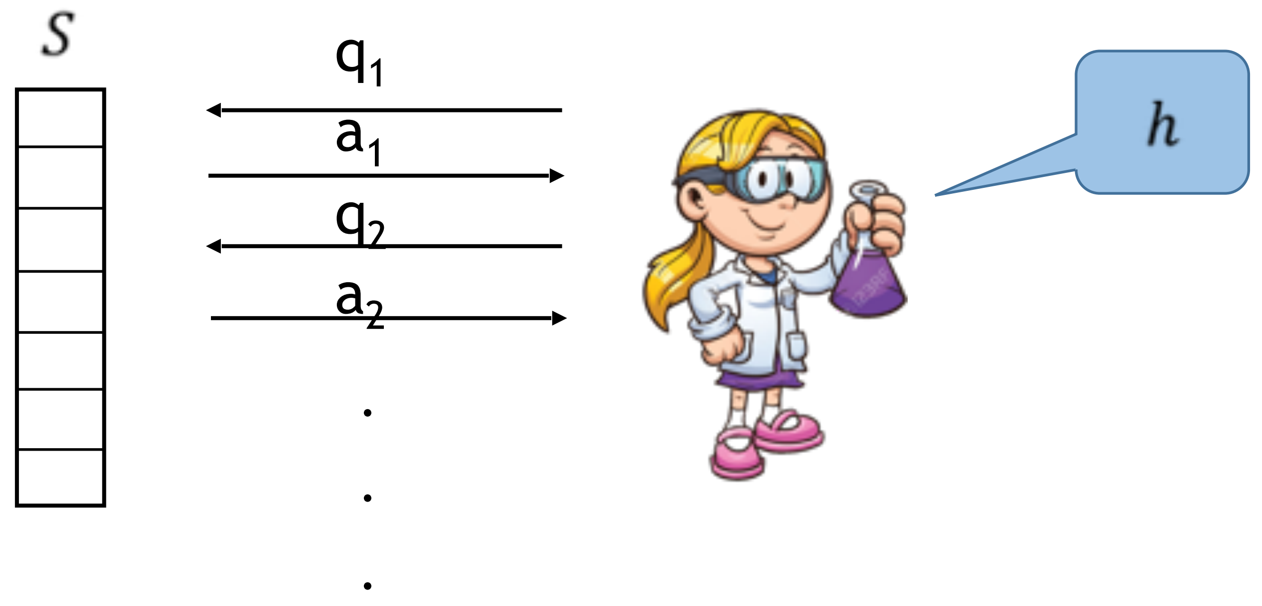
Yeah, I bet they’d see an even bigger effect if they accounted for sunspots!

The journal requires open access to the data—let’s try it and see!

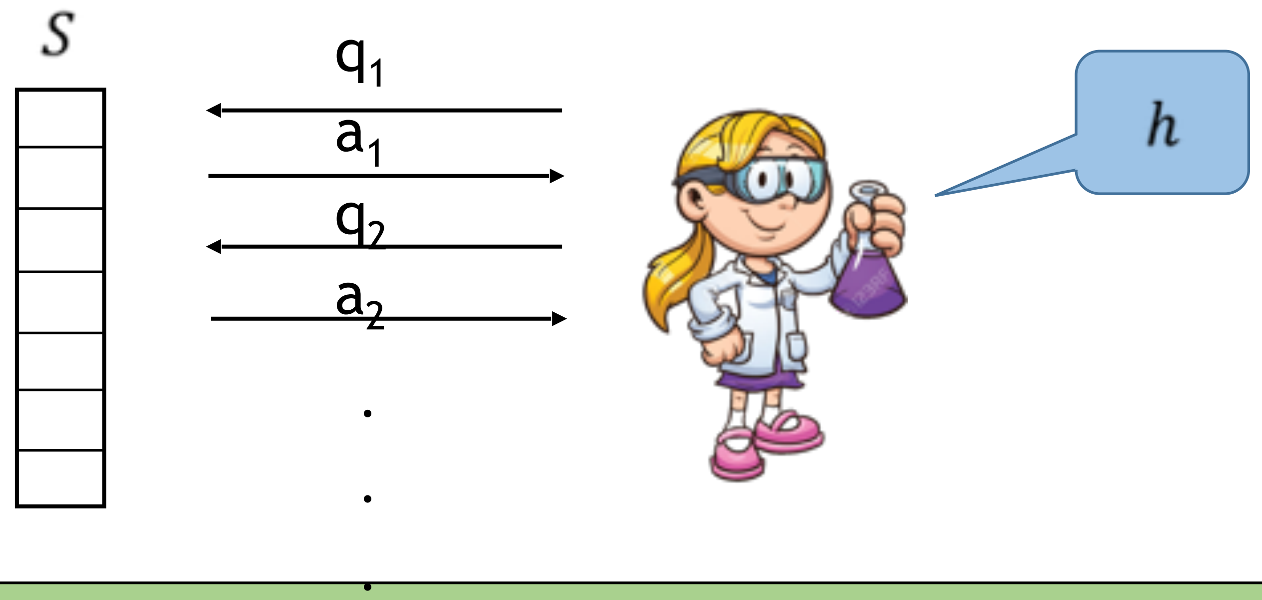
A second big problem: *adaptive composition*



A second big problem: *adaptive composition*

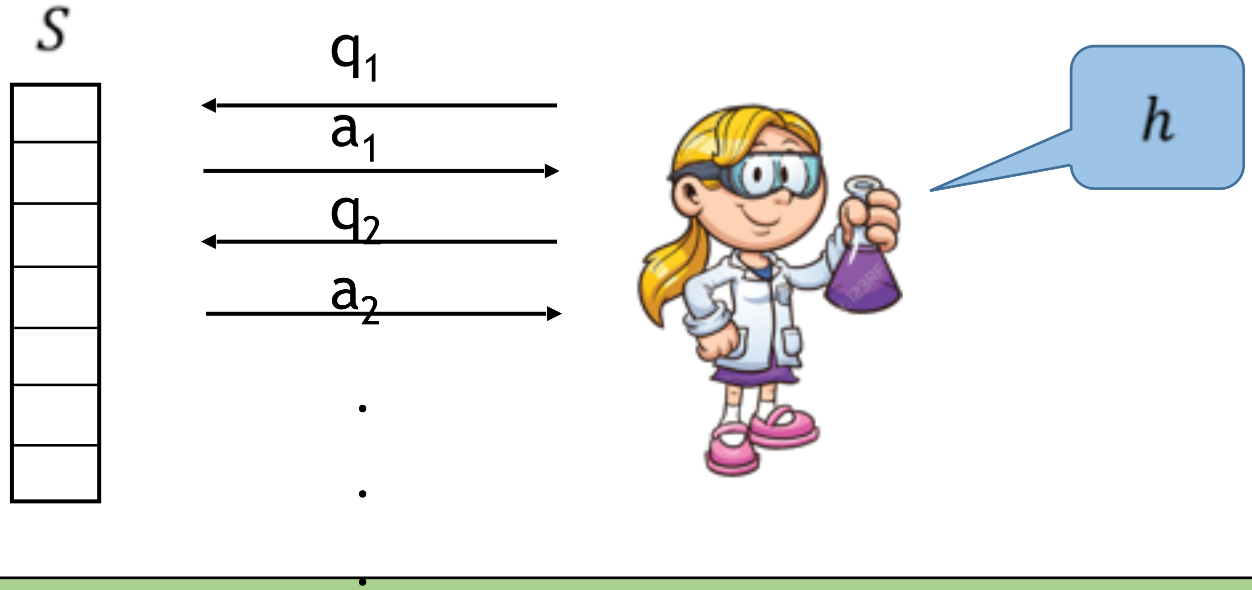


A second big problem: *adaptive composition*



Adaptive composition can cause overfitting!
Generalization guarantees don't "add up"

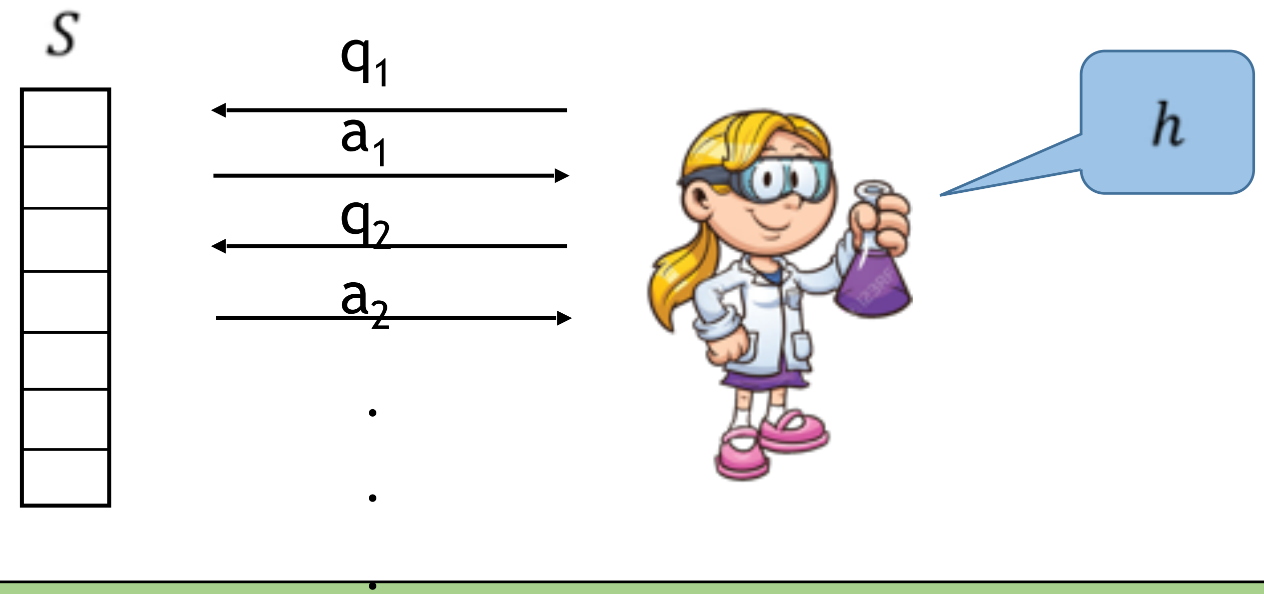
A second big problem: *adaptive composition*



Adaptive composition can cause overfitting!
Generalization guarantees don't "add up"

- Pick parameters; fit model

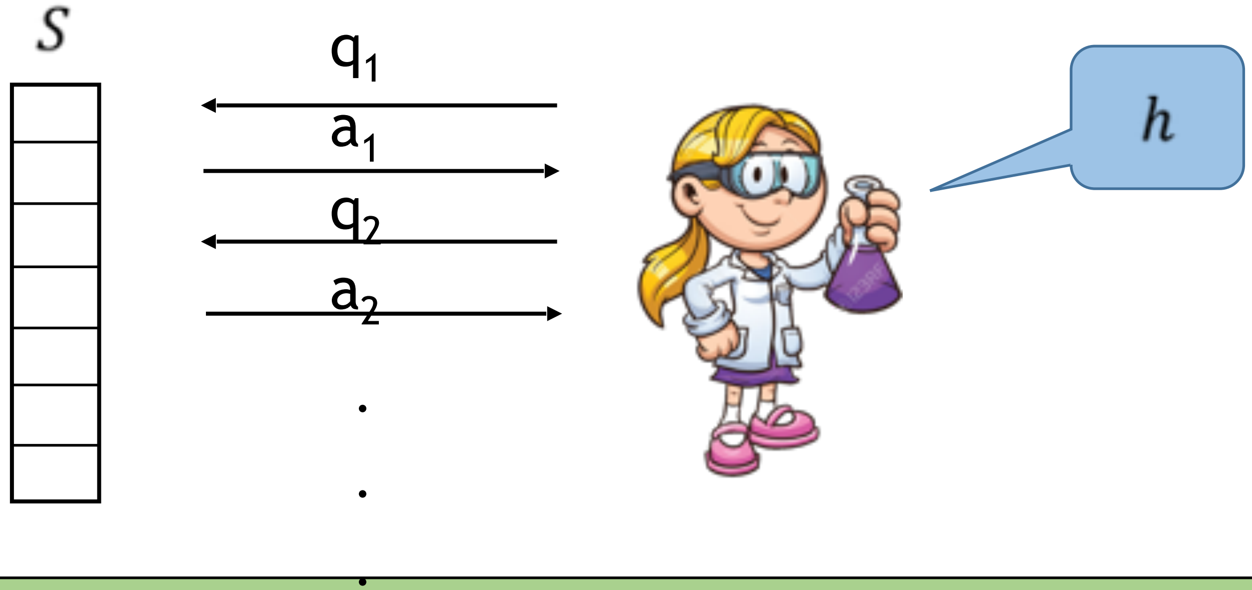
A second big problem: *adaptive composition*



Adaptive composition can cause overfitting!
Generalization guarantees don't "add up"

- Pick parameters; fit model
- ML competitions

A second big problem: *adaptive composition*



Adaptive composition can cause overfitting!
Generalization guarantees don’t “add up”

- Pick parameters; fit model
- ML competitions
- Scientific fields that share one dataset

Some basic questions

- Is it possible to get good learning algorithms that also are robust to post-processing? Adaptive composition?
- How to construct them? Existing algorithms? How much extra data do they need?
- Accuracy + generalization + post-processing-robustness = ?
- Accuracy + generalization + adaptive composition = ?
- What composes with what? How well (how quickly does generalization degrade)? Why?

Notice: generalization doesn't require *correct* hypotheses, just that they *perform the same* on the sample as on the distribution

Generalization alone is easy.
What's interesting: generalization + accuracy.

Generalization + post-processing robustness

- Robust generalization

  "no adversary can use output to find a hypothesis that overfits"

  information-theoretic (could think computational)

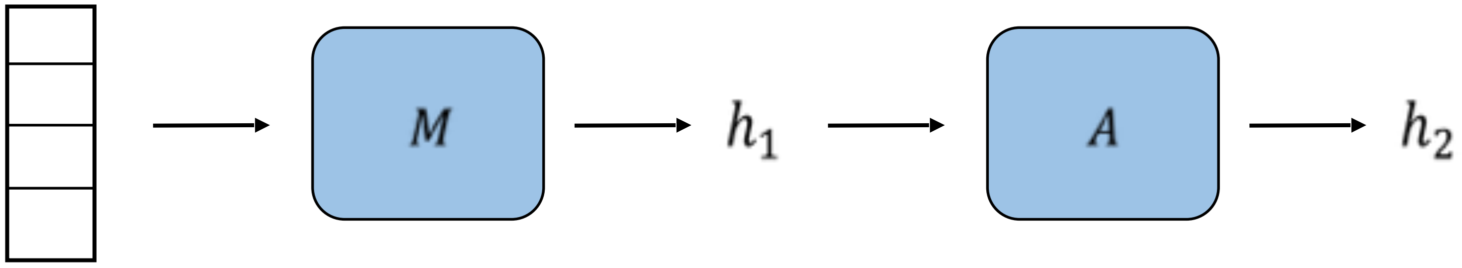
Robust Generalization

Mechanism $M: X^n \to R$ is (α, β)-Robustly Generalizing if

$\forall$ distributions $D \in \Delta X$, $\forall$ adversary A, w.p. $1-\xi$ over $S \sim_{i.i.d.} D^n$,

$$\Pr\left[A\big(M(S)\big) \text{ outputs } h: X \to \{0,1\} \text{ s.t. } |h(S) - h(D)| \le \alpha\right] \ge 1 - \gamma$$

where $\beta = \xi + \gamma$.

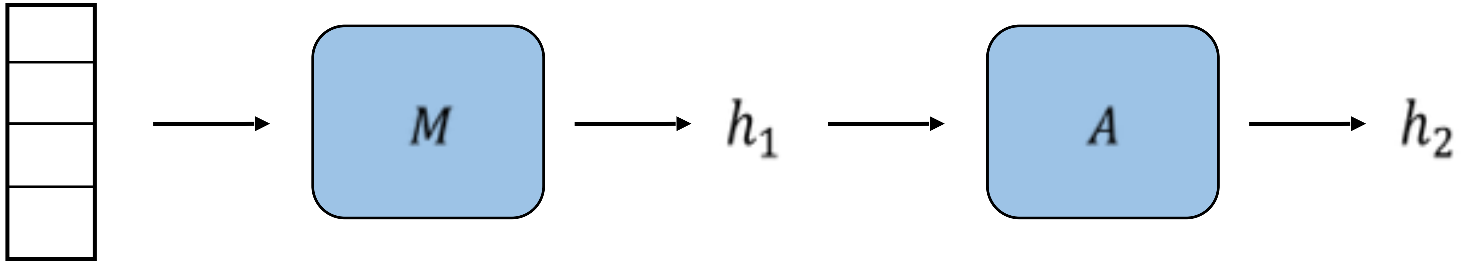


Robust Generalization

Mechanism $M: X^n \to R$ is (α, β)-Robustly Generalizing if
$\forall$ distributions $D \in \Delta X$, $\forall$ adversary A, w.p. $1 - \xi$ over $S \sim_{i.i.d.} D^n$,

$$\Pr\left[A\left(M(S)\right) \text{ outputs } h: X \to \{0,1\} \text{ s.t. } |h(S) - h(D)| \leq \alpha\right] \geq 1 - \gamma$$

where $\beta = \xi + \gamma$.



- Robust to post-processing
- Somewhat robust to adaptive composition (more on this later)

Do Robustly-Generalizing Algs Exist?

Do Robustly-Generalizing Algs Exist?

Yes!

Do Robustly-Generalizing Algs Exist?

Yes!

Do Robustly-Generalizing Algs Exist?

Yes!

- This paper: Compression Schemes -> Robust Generalization

Do Robustly-Generalizing Algs Exist?

Yes!

- This paper: Compression Schemes -> Robust Generalization
- [DFHPRR15a]: Bounded description length -> Robust Generalization

Do Robustly-Generalizing Algs Exist?

Yes!

- This paper: Compression Schemes -> Robust Generalization
- [DFHPRR15a]: Bounded description length -> Robust Generalization
- [BNSSSU16]: Differential privacy -> Robust Generalization

Do Robustly-Generalizing Algs Exist?

Yes!

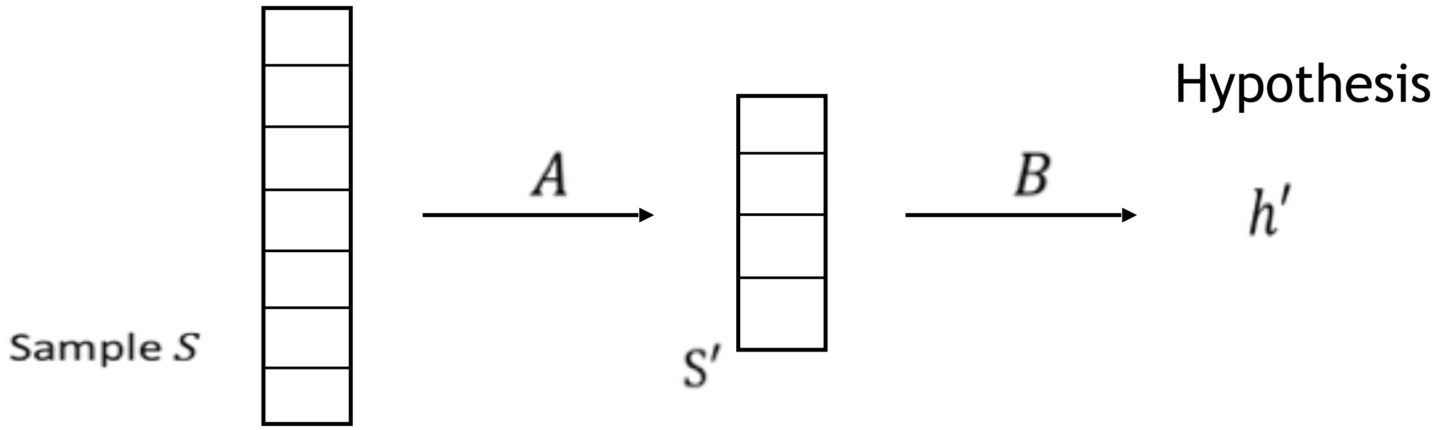
- This paper: Compression Schemes -> Robust Generalization
- [DFHPRR15a]: Bounded description length -> Robust Generalization
- [BNSSSU16]: Differential privacy -> Robust Generalization

Compression schemes

Hypothesis class H has a compression scheme of size k if there exists:

- compression algorithm $A: X^n \to X^k$
- encoding algorithm: $B: X^k \to H$

s.t. $h' = B(A(S))$ is ERM on S, i.e. $err_E(h') \leq err_E(h), \forall h \in H$.

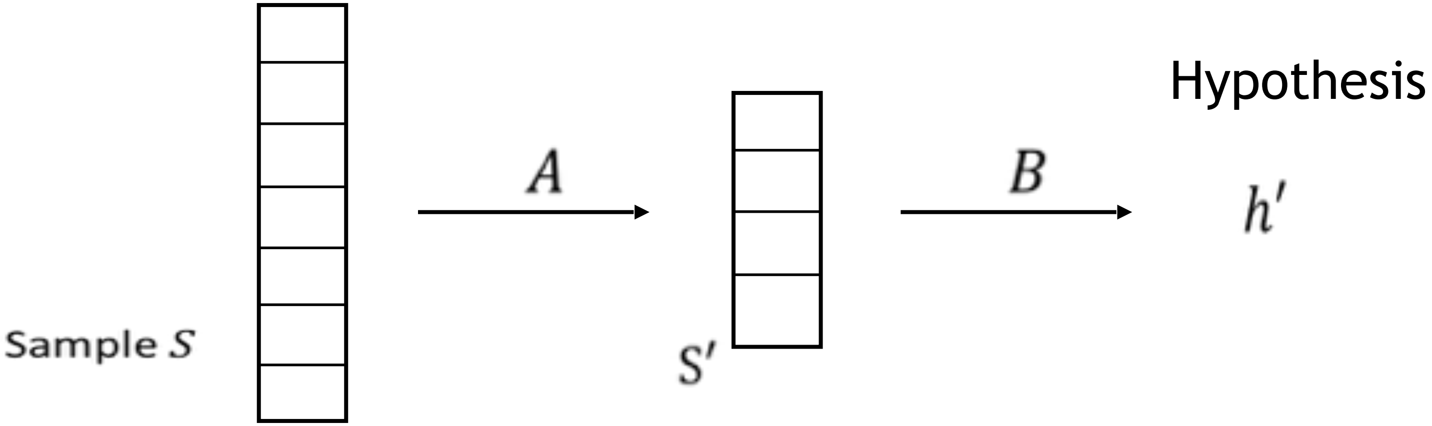


Compression schemes

Hypothesis class H has a compression scheme of size k if there exists:

- compression algorithm $A: X^n \to X^k$
- encoding algorithm: $B: X^k \to H$

s.t. $h' = B(A(S))$ is ERM on S, i.e. $err_E(h') \leq err_E(h), \forall h \in H$.



$L = B \circ A$ is a compression learner

Robust Generalization via compression

Theorem: If class H has a compression scheme of size k, then H is PAC-learnable under RG by a compression learner with

- (α, β)-accuracy
- (ϵ, δ)-RG
- sample complexity $k\ poly(\frac{1}{\alpha}, \frac{1}{\epsilon}, \log\frac{1}{\beta}, \log\frac{1}{\delta})$

Proof idea:

1. Lemma [LW '86]: If class H has a compression scheme of size k, then H is PAC-learnable with (α, β)-accuracy and sample complexity $k\ poly(\frac{1}{\alpha}, \log\frac{1}{\beta})$
2. Lemma: Let A be a compression algorithm then A is (ϵ, δ)-RG for

$$\epsilon = O\left(\sqrt{\frac{k \log(n/\delta)}{n}}\right)$$

What Can be Learned under RG?

Theorem (informal; thanks to Shay Moran): sample complexity of robustly generalizing learning is the *same* up to log factors, as the sample complexity of PAC learning

Do Robustly-Generalizing Algs Exist?

Yes!

- This paper: Compression Schemes -> Robust Generalization
- [DFHPRR15a]: Bounded description length -> Robust Generalization
- [BNSSSU16]: Differential privacy -> Robust Generalization

- Theorem [DFHPRR ‘15]: Let $M: X^n \to R$ s.t. $|R|$ bounded. Then M is (α, β)-RG with $\alpha = \sqrt{\frac{\ln(|R|/\beta)}{2n}}$.

Small description length $\Longrightarrow$ robust generalization

Theorem [BNSSSU ’16]: Let $M: X^n \to R$ be (ϵ, δ)-DP. Then M is $(O(\epsilon), O(\delta/\epsilon))$-RG when $n \geq O(\ln\frac{1}{\delta}/\epsilon^2)$.

Differential privacy $\Longrightarrow$ robust generalization

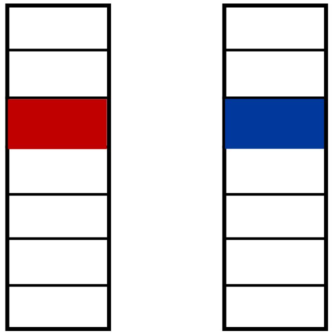
- Adaptive composition *within* each method but not *across*

Differential Privacy [DMNS '06]

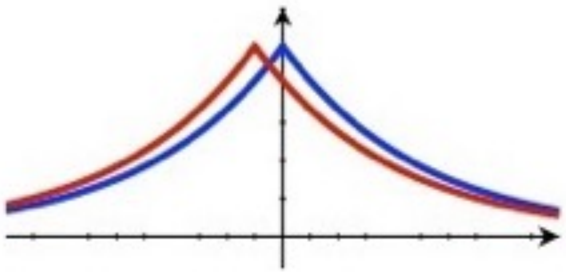
Mechanism $M: X^n \to R$ is (ϵ, δ)-Differentially Private if

$\forall$ pairs of samples S, S' that differ in one element, $\forall\, O \subseteq R$,

$$\Pr[M(S) \in O] \leq \mathrm{e}^{\epsilon} \cdot \Pr[M(S') \in O] + \delta$$



$\Longrightarrow$

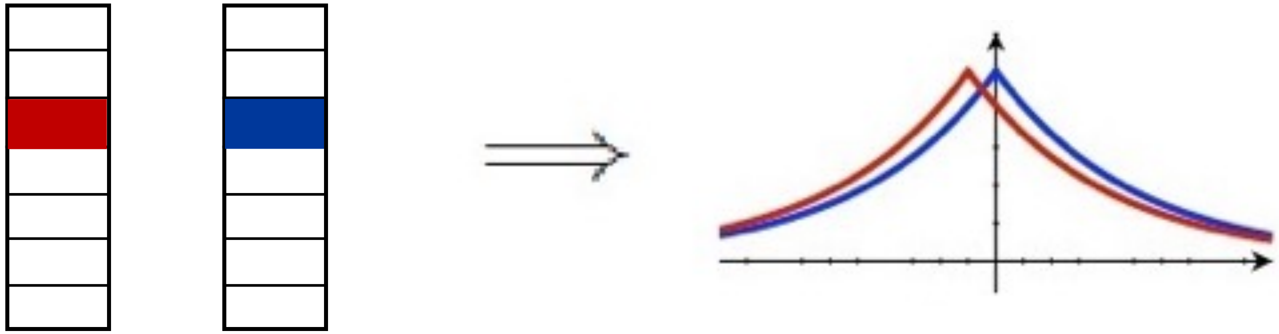


Differential Privacy [DMNS '06]

Mechanism $M: X^n \rightarrow R$ is (ϵ, δ)-Differentially Private if

$\forall$ pairs of samples S, S' that differ in one element, $\forall\, O \subseteq R$,

$$\Pr[M(S) \in O] \leq e^{\epsilon} \cdot \Pr[M(S') \in O] + \delta$$



- Robust to post-processing [DMNS '06] and adaptive composition [DRV '10]
- Necessarily randomized output
- No mention of how samples drawn!

Does DP = RG?

Does DP = RG?

Obvious answer: No, DP algorithms must be randomized and RG can be deterministic.

Is this difference cosmetic?

Theorem (Informal): There exists a learning task than can be solved under RG but not under DP.

Threshold Learning:

$$h_x(y) = \begin{cases} 1 \text{ if } y \le x \\ 0 \text{ if } y > x \end{cases}$$

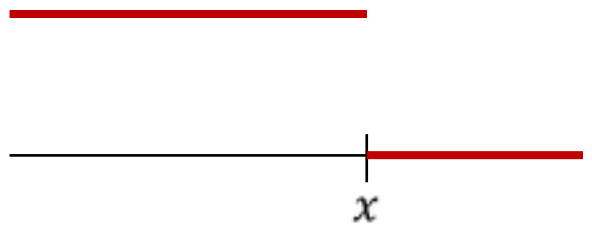
Does DP = RG?

Obvious answer: No, DP algorithms must be randomized and RG can be deterministic.

Is this difference cosmetic?

Theorem (Informal): There exists a learning task than can be solved under RG but not under DP.

Threshold Learning:



$$h_x(y) = \begin{cases} 1 \text{ if } y \leq x \\ 0 \text{ if } y > x \end{cases}$$

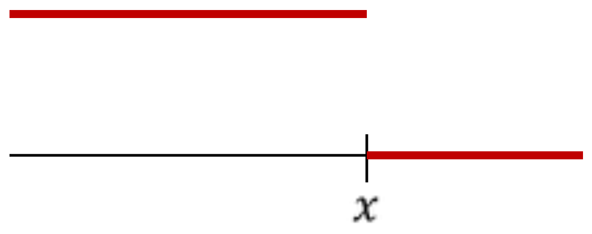
Does DP = RG?

Obvious answer: No, DP algorithms must be randomized and RG can be deterministic.

Is this difference cosmetic?

<u>Theorem (Informal):</u> There exists a learning task than can be solved under RG but not under DP.

Threshold Learning:



$$h_x(y) = \begin{cases} 1 \text{ if } y \le x \\ 0 \text{ if } y > x \end{cases}$$

No "quick fix" to make RG learner satisfy DP

Notions of generalization

- Robust generalization
  "no adversary can use output to find a hypothesis that overfits"
- Differential privacy [DMNS '06]
  "similar samples should have the same output"
- Perfect generalization
  "output reveals nothing about the sample"

Perfect Generalization

Mechanism $M: X^n \to R$ is $(\beta, \epsilon, \delta)$-<u>Perfectly Generalizing</u> if

$\forall$ distributions $D \in \Delta X$, $\exists$ simulator SIM_D, w.p. $1-\beta$ over $S \sim_{i.i.d.} D^n$,

$$\Pr[M(S) \in O] \leq e^{\epsilon} \cdot \Pr[SIM_D \in O] + \delta$$

($SIM_D \approx$ oracle access to the distribution)

PG as a privacy notion

- Differential privacy gives privacy to the individual

Changing one entry in the database shouldn't change the output too much

- Perfect generalization gives privacy to the data provider

(e.g. school, hospital)

Changing the entire sample to something "typical" shouldn't change the output too much

Exponential Mechanism [MT07]

"output an element of the range with probability proportional to exponential of *quality score*"

Let $M: X^n \to R$ be $(\epsilon, 0)$-DP. Define for each $S \in X^n$ and $r \in R$:

$$q(S,r) = \log(\Pr[M(S) = r])$$

Define $M_E: X^n \to R$ as follows

$$\Pr[M_E(S) = r] = \exp(q(S,r))$$

To prove M_E is PG, use SIM_D with output dist.

$$\Pr[SIM_D = r] \propto \exp\left(\boldsymbol{E}_{S \sim_{iid} D^n}[q(S,r)]\right)$$

DP implies PG with worse parameters

<u>Theorem:</u> Let $M: X^n \to R$ be $(\epsilon, 0)$-DP. Then M is

$$(\beta, \sqrt{2n \ln(1/\beta)}\, \epsilon, 0)\text{-PG}.$$

Dependence on n and β asymptotically tight

<u>Proof idea:</u>

1. Every $(\epsilon, 0)$-DP mechanism can be written as Exponential Mechanism
2. Exponential Mechanism satisfies PG

Open: Reduction from (ϵ, δ)-DP to PG

PG implies DP…sort of

PG implies DP…sort of

PG mechanisms are not DP because they can do weird things on a β-fraction of the samples

Example: Output "strange" on one sample, "normal" otherwise

Theorem: Let $M: X^n \to R$ be $(\beta, \epsilon, \delta)$-PG. Define M' on input $S \in X^n$,

1. draw sample $T \in X^n$ i.i.d. from S with replacement
2. output $M(T)$

Then M' is $(4\epsilon, 16\delta + 2\beta)$-DP.

PG implies DP…sort of

PG mechanisms are not DP because they can do weird things on a β-fraction of the samples

Example: Output "strange" on one sample, "normal" otherwise

Theorem: Let $M: X^n \rightarrow R$ be $(\beta, \epsilon, \delta)$-PG. Define M' on input $S \in X^n$,

1. draw sample $T \in X^n$ i.i.d. from S with replacement
2. output $M(T)$

Then M' is $(4\epsilon, 16\delta + 2\beta)$-DP.

Problems that are solvable under PG are also solvable under DP

Notions of generalization

- Robust generalization
  "no adversary can use output to find a hypothesis that overfits"
- Differential privacy [DMNS '06]
  "similar samples should have the same output"
- Perfect generalization
  "output reveals nothing about the sample"

Some basic questions

- Is it possible to get good learning algorithms that also are robust to post-processing? Adaptive composition?
- How to construct them? Existing algorithms? How much extra data do they need?
- Accuracy + generalization + post-processing-robustness = ?
- Accuracy + generalization + adaptive composition = ?
- What composes with what? How well (how quickly does generalization degrade)? Why?

Making Generalization Robust

Katrina Ligett
katrina.ligett@mail.huji.ac.il
HUJI & Caltech

joint with Rachel Cummings, Kobbi Nissim, Aaron Roth, and Steven Wu