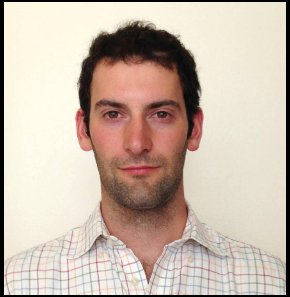
Sample Amplification:

Increasing Dataset Size even when Learning is Impossible

Brian Axelrod

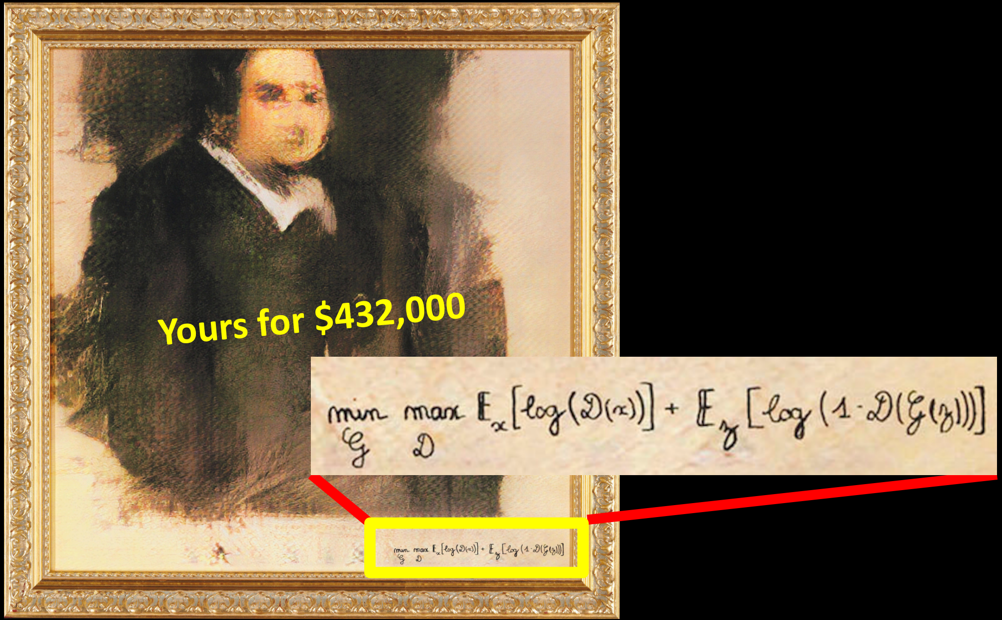
Shivam Garg

Vatsal Sharan



Greg Valiant

Stanford University



What does it mean that a GAN made this image?

(Does it mean that GANs "know" the distribution of renaissance portraits?)

When can you make more data?

Could you generate new samples from a distribution, without even ``*learning*'' it?

New Problem: Sample Amplification

Amplifier

Input: *n* i.i.d. samples from D

Output: m > n “samples”

Input: *m* samples,
distribution *D*

Verifier

Output: ACCEPT or REJECT

Promise: If input is *m* i.i.d. draws from D, then w. prob > ¾, must ACCEPT.

Verifier: 1. Knows *D* 2. Is computationally unbounded 3. Does not know training set

Sample Amplification

Definition: A class of distributions *C* admits (*n*,*m*)-amplification, if there is an *(n,m)* Amplifer s.t. **for all** D ∈ C, any Verifier will ACCEPT with prob > 2/3.

Amplifier

Input: *n* i.i.d. samples from D

Output: m > n "samples"

Input: *m* datapoints, distribution *D*

Verifier

Output: ACCEPT or REJECT

Promise: If input is *m* i.i.d. draws from D, then w. prob > ¾, must ACCEPT.

Verifier: knows *D*, is computationally unbounded

Sample Amplification

Definition: A class of distributions *C* admits (*n*,*m*)-amplification, if there is an *(n,m)* Amplifer s.t. **for all** D ∈ C, any Verifier will ACCEPT with prob > 2/3.

- Every class C admits (n,n)-amplification (why?)
- Verifier does not see Amplifier's *n* input samples. (Otherwise equivalent to *learning*)
- Up to constant factors, equivalent to asking whether Amplifier can output *m* samples, whose T.V. distance to *m* i.i.d. samples from D is small.

Sample Amplification

Definition: A class of distributions *C* admits (*n*,*m*)-amplification, if there is an *(n,m)* Amplifer s.t. **for all** D ∈ C, any Verifier will ACCEPT with prob > 2/3.

Connection to GANs:

Amplifier -> Generator, Verifier -> Discriminator? Not quite..

Similarities in how samples are used and evaluated.

RESULTS

Sample Amplification

Thm 1: Let *C* be class of discrete distributions supported on $\leq k$ elements.
(*n, n* + **n/sqrt(k)**)-amplification is possible (and optimal, to constant factors)

* Nontrivial amplification possible as soon as *n* > sqrt(k).
* *Learning* to nontrivial accuracy requires n=θ(k) samples
* Even with *n* >> *k* can never amplify by arbitrary amount.

Thm 2: Let *C* be class of Gaussians in *d* dimensions, with fixed covariance (e.g. "isotropic"), and **unknown** mean. (*n, n* + **n/sqrt(d)**)-amplification is possible (and optimal, to constant factors)

* Nontrivial amplification possible as soon as *n* > sqrt(d).
* Learning to nontrivial accuracy requires n=θ(d) samples

GAUSSIAN DISTRIBUTION

Thm 2: For Gaussians in **d** dimensions, with fixed covariance, and **unknown** mean:

- *Learning* requires **n = d.**
- Amplification possible starting at **n = sqrt(d)**.

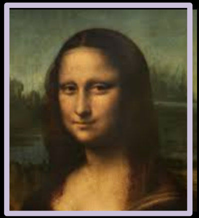
(*n, n* + **n/sqrt(d)**)-amplification is possible (and optimal, to constant factors)

Algorithm:

1) Draw $x_{n+1} \ldots x_m$ using empirical mean u* of input samples.
2) For each input sample x_i “decorrelate” it from u*.
3) Return $x_{n+1} \ldots x_m$ along with “decorrelated” original samples.

Thm 3: If output ⊃ input samples, require **n > d/ log d** for nontrivial amp.

Intuitively, issue is new “samples” would be too correlated with originals:



IS AMPLIFICATION USEFUL?

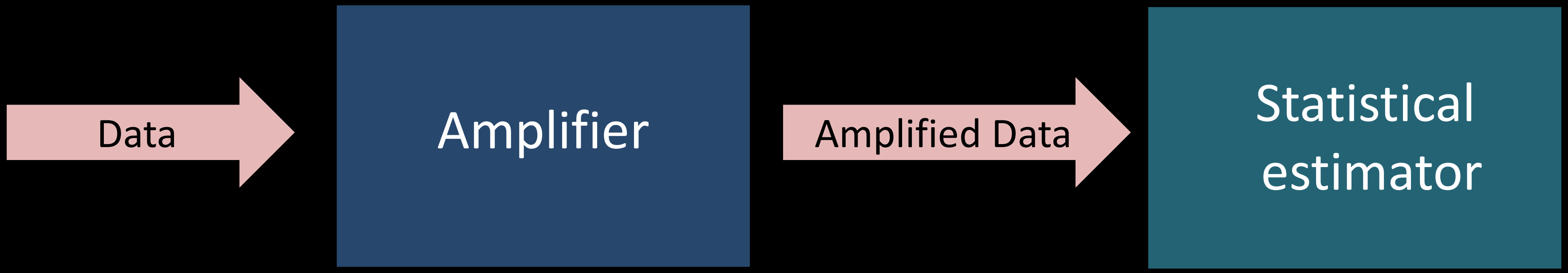

WE'LL MAKE
ALL YOUR DATA
LARGER!

Amplification does not add new information,
but could make original information more easily accessible.

Can widely used statistical tools do better on amplified samples?

Amplification does not add new information,
but could make original information more easily accessible.

Can widely used statistical tools do better on amplified samples?

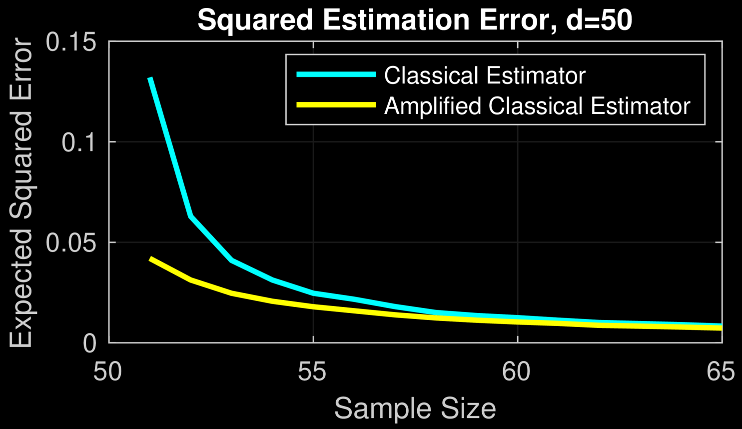


Amplification Maybe Useful?

Given examples $(x, y) \sim D$ estimate error of best linear model

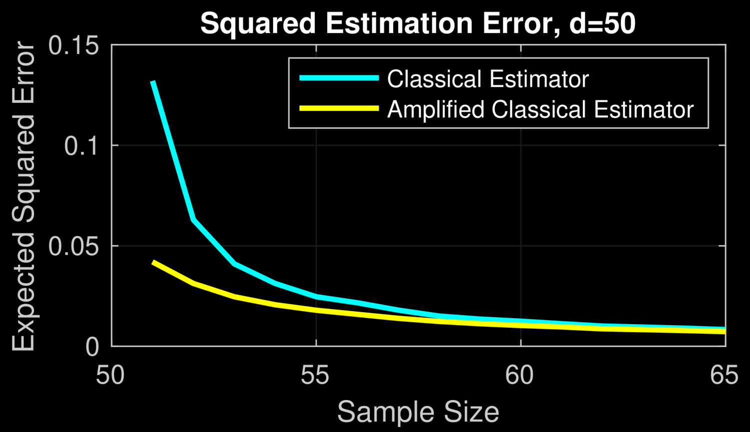
Standard unbiased estimator: Error of least-squares model, scaled down

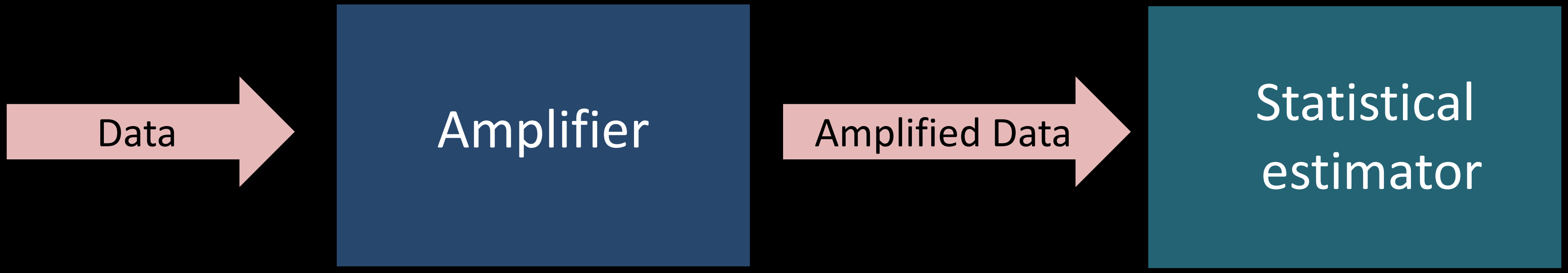
$$x \sim Gaussian(d = 50), y = \theta^T x + Gaussian\ noise$$



Error of classical estimator vs. same estimator on $(n, n+2)$ amplified samples.

Amplification Maybe Useful?





FUTURE DIRECTIONS

What property of a class of distributions determines threshold at which non-trivial amplification is possible?

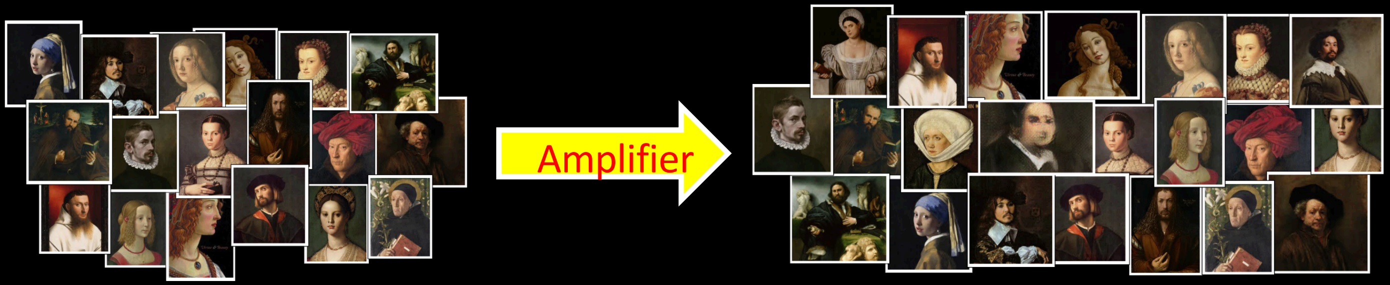
More general amplification schemes?

MORE powerful Verifier?

How much does Verifier need to know about n input samples to preclude amplification without learning?
[How much do we need to know about a GAN's input, to evaluate its output?]

LESS powerful Verifier?

What if Verifier doesn't know D, *only gets sample access?*



THANK YOU!