From polarization of belief to Active Learning Theory: a diameter approach

Gauthier Guinet

MIT, December 9, 2020

Motivation:

- **General Polarization phenomena:** *"when different people are exposed to very different sources of information, they are bound to arrive at different conclusions"*

Motivations and Framework

Motivation:

- **General Polarization phenomena:** *"when different people are exposed to very different sources of information, they are bound to arrive at different conclusions"*
  - Big line of work in **Social learning literature** (Bayesian Framework, bounded rationality...)
  - Stochastic models of opinion dynamics (echo chamber, Voter Model...)

Motivations and Framework

Motivation:

- **Our interest:** Yet, individuals exposed to similar information may still end up having substantially different opinions !

Motivation:

- **Our goal:** Under what conditions does polarization of this type arise, and can it be prevented through mild interventions?

Polarization under the lens of Statistical Learning

The objective cost model [Haghtalab et al., 2019]:

- Under realizable distribution D, consistent with $f^\star$, all error-minimizing agents will arrive at hypotheses that are almost in **full agreement** with each other.
- What if we add the notion of **complexity** of hypothesis f, with agents looking for a **balance between accuracy and such complexity** ?

Framework

Notations:

- Distribution $\mathcal{D}$ on $\mathcal{X} \times \{-1, +1\}$, 0-1 loss
- Expected error:

$$\text{err}_{\mathcal{D}}(f) := \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathbb{I}(f(x) \neq y)] = \Pr_{(x,y)\sim\mathcal{D}}[f(x) \neq y]$$

- Empirical Error for sample $\mathcal{S}$:

$$\text{err}_{\mathcal{S}}(f) := \frac{1}{m} \sum_{i=1}^{m} \mathbb{I}\left(f\left(x_i\right) \neq y_i\right)$$

Framework

Notations:

- Disagreement between two hypothesis $f, \tilde{f} \in \mathcal{F}$ (**pseudo-metric**):

$$\Delta_{\mathcal{D}}\left(f, f'\right) := \Pr_{x \sim \mathcal{D} \downarrow \mathcal{X}}\left[f(x) \neq f'(x)\right]$$

- Diameter of any given hypothesis set $\mathcal{H}$:

$$\mathrm{diam}_{\mathcal{D}}(\mathcal{H}) := \sup_{f, f' \in \mathcal{H}} \Delta_{\mathcal{D}}\left(f, f'\right)$$

Framework

Complexity function ϕ:

- "Penalized" type ERM:

$$\text{cost}_{\mathcal{D}}^{\lambda}(f) := \text{err}_{\mathcal{D}}(f) + \lambda\phi(f) \text{ and } \text{cost}_{\mathcal{S}}^{\lambda}(f) := \text{err}_{\mathcal{S}}(f) + \lambda\phi(f)$$

- Stay as general as possible !
    - Penalization or regularization but not only
    - Preferences or prior of agents for certain hypothesis
    - Potentially meta-hypothesis space
    - No structure on $\mathcal{F}$ aside form Δ

Previous results

A quick example:

- **Polarization[Haghtalab et al., 2019]:** There is $\mathcal{F}$ and $\mathcal{D}$ such that for any m and two sets $\mathcal{S}_1, \mathcal{S}_2$ of m i.i.d. samples from $\mathcal{D}$, with probability $\frac{1}{4}$, there exists $f_i \in \operatorname{argmin}_{f \in \mathcal{F}} \operatorname{cost}_{S_i}(f)$ such that $\Delta_{\mathcal{D}}(f_1, f_2) > \frac{1}{6}$.

Previous results

Main result (Informal):

Theorem
For any desired level of disagreement, it's possible to add "small" bias in the distribution $\mathcal{D}$ so that agents learning with "sufficient" samples have disagreement under this threshold.

Main result (Formal Version):

Theorem
For a hypothesis class $\mathcal{F}$ (with finite VC dimension), a realizable distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$, a parameter $\alpha \in [0,1]$ and a maximum level of disagreement $\gamma > 0$, there exists

$$m \in O\left(\gamma^{-4}\alpha^{-2}\left(\mathrm{VCD}(\mathcal{F}) + \ln\left(\frac{1}{\delta}\right)\right)\right)$$

and realizable distribution $\tilde{\mathcal{D}}$, with $\mathcal{TV}(\mathcal{D}, \tilde{\mathcal{D}}) \leq \frac{\alpha}{2}$, such that if two sets S_1 and S_2 of size at least m are sampled from $\widetilde{\mathcal{D}}$, then with probability at least $1-\delta$ any two cost-minimizing hypotheses $f_i \in \mathrm{argmin}_{f\in\mathcal{F}} \mathrm{cost}_{S_i}(f)$ for $i \in \{1,2\}$

1. *have at most γ disagreement over $\mathcal{D}$, i.e., $\Delta_{\mathcal{D}}\left(\widetilde{f_1}, \widetilde{f_2}\right) \leq \gamma$, and*
2. *have a cost that is optimal up to 3α on $\mathcal{D}$, i.e.*

$$\mathrm{cost}_{\mathcal{D}}\left(\widetilde{f_i}\right) \leq \underset{f\in\mathcal{F}}{\mathrm{argmin}}\, \mathrm{cost}_{\mathcal{D}}(f) + 3\alpha$$

Limitations, goals and motivations

Our work:

- **Robustness, Complexity and Learning:** To what extent polarization is robust w.r.t. the complexity functions ? In others words, what is the impact of modifications of the complexity function associated with hypothesis (i.e. *education*) on polarization ?
- **Active Learning and Polarization**: Can we **learn** how to create bias describe above? In particular, what links can be establish with ideas and tools from Active Learning Community?

Few more notations:

- **Rashomon Set [Fisher et al., 2019, Semenova et al., 2020]:**

$$\mathcal{F}_\epsilon^{\mathcal{D}}(\lambda) := \left\{ f \in \mathcal{F} \mid \text{cost}_{\mathcal{D}}^{\lambda}(f) \leq \min_{f' \in \mathcal{F}} \text{cost}_{\mathcal{D}}^{\lambda}\left(f'\right) + \epsilon \right\}$$

- **Core Goal:** What can we say about this set and his diameter in function of λ ?

Few more notations:

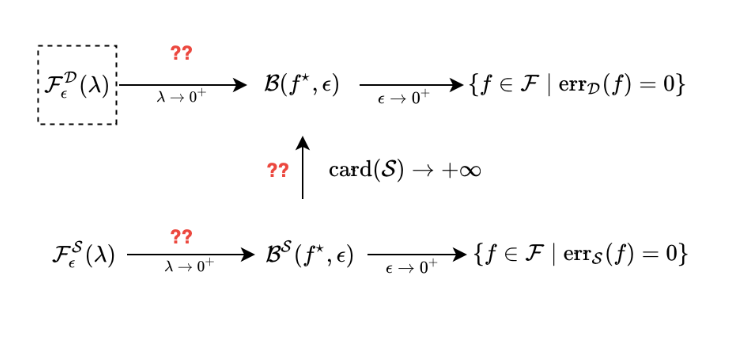
- **Rashomon Set [Fisher et al., 2019, Semenova et al., 2020]:**

$$\mathcal{F}_{\epsilon}^{\mathcal{D}}(\lambda) := \left\{ f \in \mathcal{F} \mid \text{cost}_{\mathcal{D}}^{\lambda}(f) \leq \min_{f' \in \mathcal{F}} \text{cost}_{\mathcal{D}}^{\lambda}\left(f'\right) + \epsilon \right\}$$

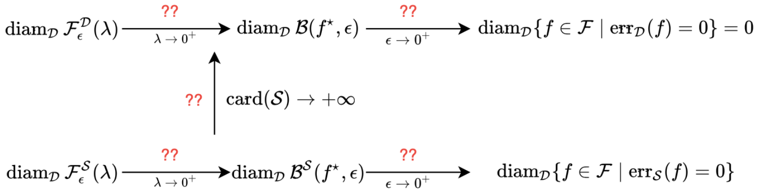
- **ϵ-Ball centered in $f^{\star}$:**

$$\mathcal{B}(f^{\star}, \epsilon) := \left\{ f \in \mathcal{F} \mid \Delta_{\mathcal{D}}(f^{\star}, f) = \text{err}_{\mathcal{D}}(f) - \underbrace{\text{err}\,\mathcal{D}(f^{\star})}_{=0} \leq \epsilon \right\} = \mathcal{F}_{\epsilon}^{\mathcal{D}}(0)$$

Triple convergence phenomena (pointwise vs uniform):



Triple convergence phenomena (pointwise vs uniform):



Hausdorff (pseudo-)distance induced by pseudo metric Δ

$$\begin{aligned} d_{\mathcal{H}}(\epsilon, \lambda) &= d(\mathcal{F}_\epsilon^{\mathcal{D}}(\lambda), \mathcal{B}(f^\star, \epsilon)) \\ &:= \max\Big(\sup_{f \in \mathcal{B}(f^\star, \epsilon)} \inf_{f_\lambda \in \mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)} \Delta(f, f_\lambda), \sup_{f_\lambda \in \mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)} \inf_{f \in \mathcal{B}(f^\star, \epsilon)} \Delta(f, f_\lambda)\Big) \end{aligned}$$

Key properties:

- **Uniform** notion of convergence between set and (thus)

$$|\operatorname{diam}(\mathcal{B}(f^\star, \epsilon)) - \operatorname{diam}(\mathcal{F}_\epsilon^{\mathcal{D}}(\lambda))| \leq 2 d_{\mathcal{H}}(\epsilon, \lambda)$$

Quick intermediate summary

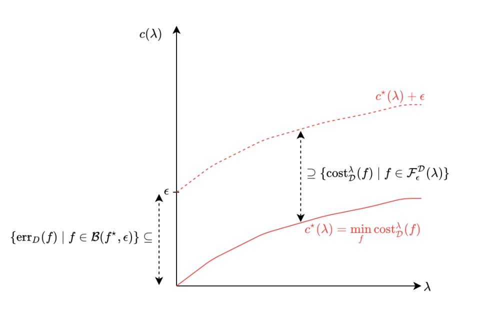
Three step approach:

$$\textbf{Part I:} \quad \sup_{f_\lambda \in \mathcal{F}_\epsilon^\mathcal{D}(\lambda)} \inf_{f \in \mathcal{B}(f^\star, \epsilon)} \Delta(f, f_\lambda)$$

$$\textbf{Part II:} \quad \sup_{f \in \mathcal{B}(f^\star, \epsilon)} \inf_{f_\lambda \in \mathcal{F}_\epsilon^\mathcal{D}(\lambda)} \Delta(f, f_\lambda)$$

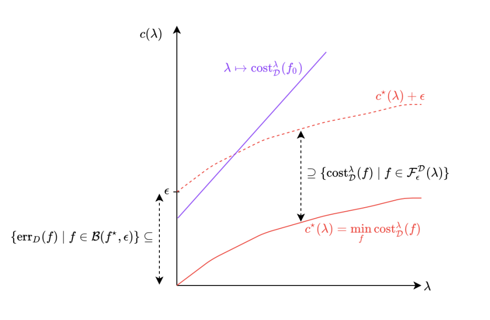
$$\textbf{Part III:} \quad \mathrm{diam}(\mathcal{F}_\epsilon^\mathcal{D}(\lambda))$$

Distance of $\mathcal{F}_\epsilon^{\mathcal{D}}$ to $\mathcal{B}(f^\star, \epsilon)$



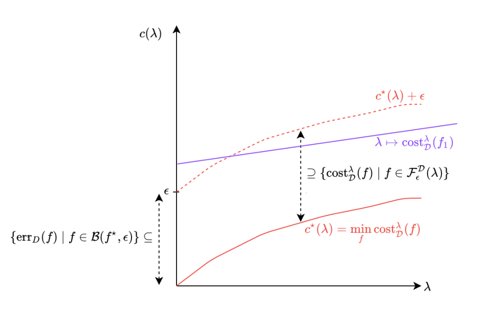
Evolution of Rashomon Set Values in function of λ

Distance of $\mathcal{F}_\epsilon^\mathcal{D}$ to $\mathcal{B}(f^\star, \epsilon)$



Evolution of Rashomon Set Values in function of λ

Distance of $\mathcal{F}_\epsilon^{\mathcal{D}}$ to $\mathcal{B}(f^\star, \epsilon)$



Evolution of Rashomon Set Values in function of λ

Distance of $\mathcal{F}_\epsilon^{\mathcal{D}}$ to $\mathcal{B}(f^\star, \epsilon)$

Difficulties for limits:

- First, the set $\mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)$ might continue to grow when $\lambda \to 0^+$. Thus, considering it at a given time step λ doesn't take into account the fact that it can still increase afterwards.
- Secondly, we need a **uniform parameter** λ_0 associated with the removal of an hypothesis of the class and not a per hypothesis version.

Distance of $\mathcal{F}_{\epsilon}^{\mathcal{D}}$ to $\mathcal{B}(f^{\star}, \epsilon)$

No uniform convergence ?:

Lemma
There exists $\mathcal{F}$ and $\mathcal{D}$ such

$$\forall \lambda > 0, \sup_{f_\lambda \in \mathcal{F}_{\epsilon}^{\mathcal{D}}(\lambda)} \inf_{f \in \mathcal{B}(f^{\star}, \epsilon)} \Delta(f, f_\lambda) > \epsilon \tag{1}$$

Distance of $\mathcal{F}_\epsilon^{\mathcal{D}}$ to $\mathcal{B}(f^\star, \epsilon)$

Upper bound:

Lemma
For a given distribution $\mathcal{D}$, we have:

$$\forall \lambda > 0, \quad \sup_{f_\lambda \in \mathcal{F}_\epsilon^{\mathcal{S}}(\lambda)} \inf_{f \in \mathcal{B}(f^\star, \epsilon)} \Delta(f, f_\lambda) \leq \min(1, \epsilon + c^\star(\lambda) - \Gamma) \leq \min(1, \epsilon + c^\star(\lambda))$$

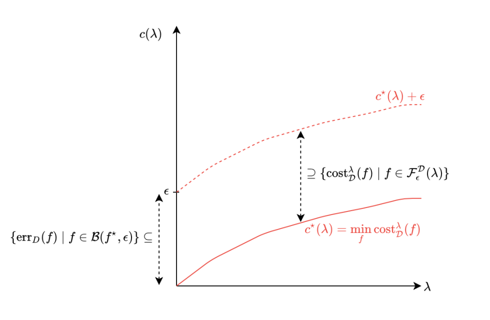
where $\Gamma \geq 0$ is the minimal gradient of an affine function tangent to $\lambda_0 \mapsto c^\star(\lambda_0)$ and going through $c^\star(\lambda) + \epsilon$. Moreover, there exists for all $\lambda > 0$, a distribution $\mathcal{D}$, an $\epsilon > 0$ and an hypothesis space $\mathcal{F}$ where the equality is reached (for a fixed lambda !).

Distance of $\mathcal{F}_\epsilon^\mathcal{D}$ to $\mathcal{B}(f^\star, \epsilon)$

A strong result:

Theorem
If $\{\mathrm{err}_\mathcal{D}(f) \mid f \in \mathcal{F}\}$ *is finite, then there exists* $\lambda_0 > 0$ *such that:*

$$\forall \lambda \leq \lambda_0, \sup_{f_\lambda \in \mathcal{F}_\epsilon^\mathcal{S}(\lambda)} \inf_{f \in \mathcal{B}(f^\star, \epsilon)} \Delta(f, f_\lambda) = 0$$



Evolution of Rashomon Set Values in function of λ

Distance of $\mathcal{B}(f^\star, \epsilon)$ to $\mathcal{F}_\epsilon^{\mathcal{D}}$

A needed distinction between interior and boundary:

Lemma
There exists $\mathcal{F}$ and $\mathcal{D}$ such

$$\forall \lambda > 0, \sup_{f \in \mathcal{B}(f^\star, \epsilon)} \inf_{f_\lambda \in \mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)} \Delta(f, f_\lambda) > \epsilon \quad (2)$$

Distance of $\mathcal{B}(f^\star, \epsilon)$ to $\mathcal{F}_\epsilon^{\mathcal{D}}$

Strong results on convergence:

Theorem
If $\mathcal{F}$ has finite VC dimension, then:

$$\sup_{f \in \mathcal{B}(f^\star, \epsilon)^\circ} \inf_{f_\lambda \in \mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)} \Delta(f, f_\lambda) \underset{\lambda \to 0^+}{\to} 0$$

Distance of $\mathcal{B}(f^\star, \epsilon)$ to $\mathcal{F}_\epsilon^{\mathcal{D}}$

Some other properties under mild assumptions:

Lemma
If there exists $\eta > 0$ *such that* $\forall f \in \mathcal{B}(f^\star, \epsilon)^\circ, \mathrm{err}_{\mathcal{D}}(f) \leq (1-\eta)\epsilon$ *and* ϕ *is bounded on* $\mathcal{B}(f^\star, \epsilon)^\circ$*, then there exists* λ_0 *such that:*

$$\forall \lambda \leq \lambda_0, \sup_{f \in \mathcal{B}(f^\star, \epsilon)^\circ} \inf_{f_\lambda \in \mathcal{F}_\epsilon^{\mathcal{S}}(\lambda)} \Delta(f, f_\lambda) = 0$$

Distance of $\mathcal{B}(f^\star, \epsilon)$ to $\mathcal{F}_\epsilon^{\mathcal{D}}$

The empirical case:

Lemma
If $\mathcal{F}$ has a finite number of patterns on $\mathcal{D}$, then, there exists $\lambda_0 > 0$ such that:

$$\forall \lambda \leq \lambda_0, \sup_{f \in \mathcal{B}(f^\star, \epsilon)^\circ} \inf_{f_\lambda \in \mathcal{F}_\epsilon^{\mathcal{S}}(\lambda)} \Delta(f, f_\lambda) = 0$$

General convergence of diameter

An approximation result:

Theorem
For all $\epsilon > 0$*, there exists* $\lambda_0 > 0$ *such that:*

$$\forall \lambda \leq \lambda_0, d(\mathcal{F}_\epsilon^{\mathcal{D}}(\lambda), \mathcal{B}(f^\star, \epsilon)) \leq 2\epsilon$$

Thus, we have in particular:

$$\forall \lambda \leq \lambda_0, \mathrm{diam}(\mathcal{B}(f^\star, \epsilon)) - 2\epsilon \leq \mathrm{diam}(\mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)) \leq \mathrm{diam}(\mathcal{B}(f^\star, \epsilon)) + 2\epsilon$$

General convergence of diameter

A positive answer:

Corollary
For any $\mathcal{D}$, there exists $\lambda_0 > 0$ such that:

$$\forall \lambda \leq \lambda_0, \mathrm{diam}(\mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)) \leq 6\epsilon$$

A positive answer - II:

Theorem
Under Tsybakov's low-noise assumption [Hanneke, 2011], there exists $\lambda_0 > 0$ *such that, for all* $\epsilon > 0$ *and* $\tilde{\epsilon}$ *arbitrarily close to 0:*

$$\forall \lambda \leq \lambda_0, \mathrm{diam}(\mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)) \leq 2\epsilon + \mu \epsilon^{1/\kappa} + \tilde{\epsilon}$$

A stronger result:

Theorem
If there is a finite number of patterns of $\mathcal{F}$ on $\mathcal{D}$, for all except a finite number of $\epsilon > 0$, there exists $\lambda_0 > 0$ such that:

$$\forall \lambda \leq \lambda_0, \quad d(\mathcal{F}_\epsilon^{\mathcal{D}}(\lambda), \mathcal{B}(f^\star, \epsilon)) = 0$$

From this, we can deduce in particular that:

$$\forall \lambda \leq \lambda_0, \quad \mathrm{diam}(\mathcal{F}_\epsilon^{\mathcal{D}}(\lambda)) = \mathrm{diam}(\mathcal{B}(f^\star, \epsilon))$$

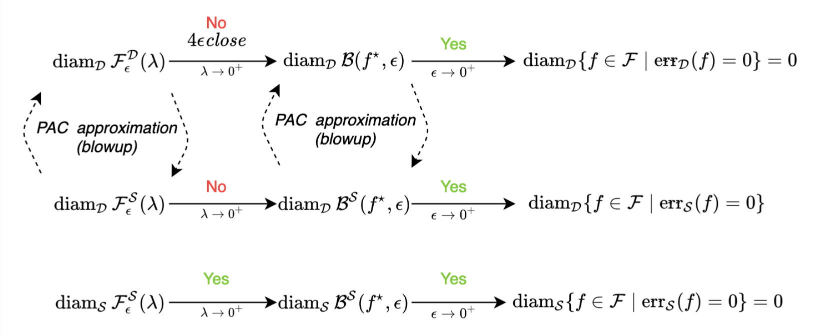
General convergence of diameter

A final result:

Theorem
For any $\epsilon, \delta > 0$*, there exists* $N(\epsilon, \delta)$ *and* $\lambda_0(\epsilon)$ *such that for all set* $\mathcal{S}$ *for size at least* $N(\epsilon, \delta)$*, with probability* $1 - \delta$*:*

$$\forall \lambda \leq \lambda_0, \text{diam}_{\mathcal{D}}(\mathcal{F}_{\epsilon}^{\mathcal{S}}(\lambda)) \leq \text{diam}_{\mathcal{D}}(\mathcal{F}_{2\epsilon}^{\mathcal{D}}(\lambda)) \leq 12\epsilon$$

Robustness, Complexity and Learning- A summary



Global Summary of results

Towards a dynamic reduction of disagreement:

- **Key idea of [Haghtalab et al., 2019]:** Introduce a bias toward an hypothesis f in $\mathcal{F}_\epsilon^{\mathcal{D}}$
- With a distance α allowed between $\mathcal{D}$ and $\tilde{\mathcal{D}}$, guarantees of the form:

$$\mathrm{diam}_{\mathcal{D}}\left(\mathcal{F}_\epsilon^{\tilde{\mathcal{D}}}\right) \in O\left(\frac{\epsilon}{\alpha\,\mathrm{err}_{\mathcal{D}}(f)}\right)$$

where $\tilde{\mathcal{D}} := (1-\alpha)\mathcal{D} + \alpha\mathcal{P}$ and:

$$\mathcal{P} := \mathcal{D} \mid \left\{x \mid \tilde{f}(x) = f^*(x)\right\}$$

Active Learning and Polarization

An active learning idea:

- Subset of hypothesis coherent the labeling of x as y:

$$V_x^y(\mathcal{H}) := \{f \in \mathcal{F} : f(x) = y, h \in \mathcal{H}\}$$

- Suppose a distribution ρ on $\mathcal{F}$ (uniform for instance), or a way of sampling hypothesis.

An active learning idea:

- Construct $\mathcal{E}$, subset of maximal empirical empirical error with ϵ-minimal empirical cost (with sampling guarantees)
- Enforce the modified distribution to say realizable while introducing the bias by picking points $(x^\star, y^\star)$ verifying:

$$\max_{(x,y)\in S} \rho(V_x^y(\mathcal{E}))$$

- Add $(x^\star, y^\star)$ with mass $\alpha_k \mathbb{P}_{\hat{\mathcal{D}}}(x^\star, y^\star)$, where $\hat{\mathcal{D}}$ is a non-parametric estimation of $\mathcal{D}$, and α_k reflect the confidence we have in our estimate.

Some interesting next steps

- Towards action taking context: [Foster et al., 2020]
- Coexistence of different agents: Polarization under the existence of different type of agents (e.g. $\mathcal{F}_1$ vs $\mathcal{F}_2$)

Conclusion

- Thanks for the course !
- Any questions ?

References

Fisher, A., Rudin, C., and Dominici, F. (2019).
All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously.

Foster, D. J., Rakhlin, A., Simchi-Levi, D., and Xu, Y. (2020).
Instance-dependent complexity of contextual bandits and reinforcement learning: A disagreement-based perspective.

Haghtalab, N., Jackson, M., and Procaccia, A. (2019).
Polarization through the lens of learning theory.

Hanneke, S. (2011).
Rates of convergence in active learning.
The Annals of Statistics, 39(1):333–361.

Semenova, L., Rudin, C., and Parr, R. (2020).
A study in rashomon curves and volumes: A new perspective on generalization and model simplicity in machine learning.