Investigating the limits of active learning in the Perceptron model

Luca Saglietti

Hugo Cui, Lenka Zdeborova

Outline

- **Pool based** active learning

- Theoretical framework
  Gardner volume and **mutual information**
  Large deviation on the selection process

- Algorithmic implications
  Uncertainty sampling strategies
  Approaching the theoretical bounds with **AMP**

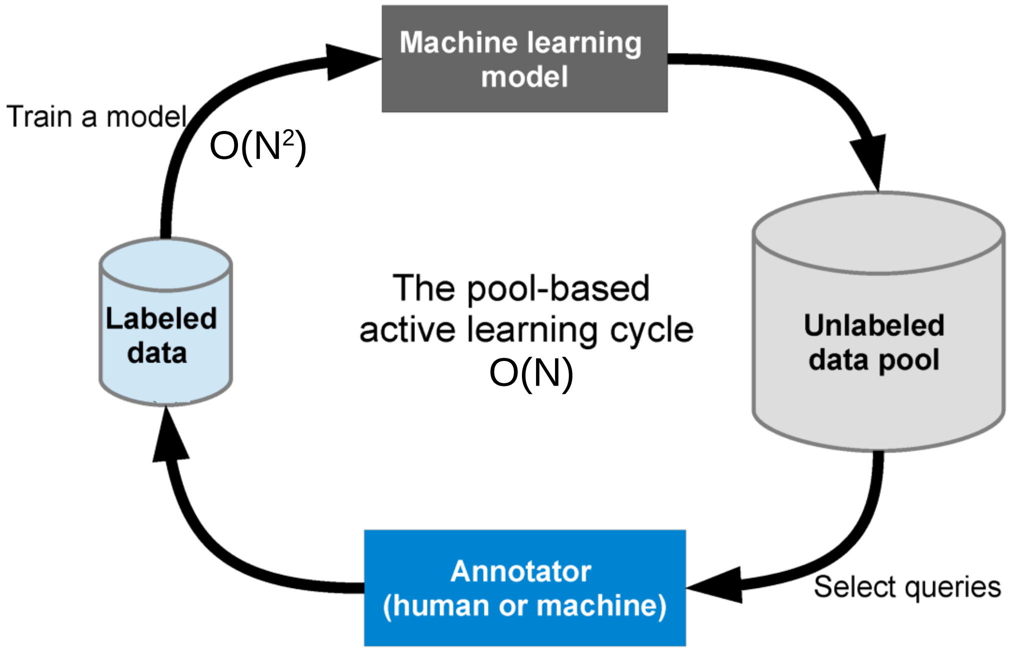
- **Perspectives**

Pool based active learning

Choosing the **most informative** data samples for **labelling** → **best test accuracy**.

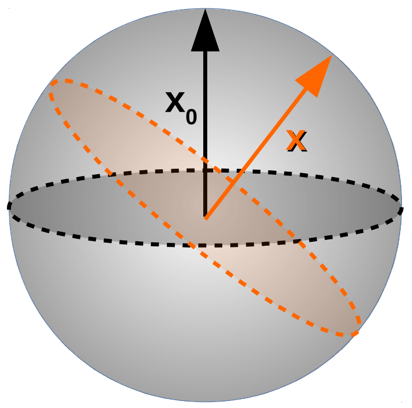
Pool-based active learning: **potential set of samples is large,** but **obtaining the labels is expensive.**
the learner can only query samples that belong to a pre-existing, **fixed pool of samples**.
One is given a certain **budget** → the cardinal of the final training set.

Applications:
- text classification, drug discovery, computational chemistry



$\approx O(N^3)$

Simple learning model



Teacher-**student** Perceptron model (of course!)

Teacher-vector of weights → $\mathbf{x_0}$

Input samples → matrix $\mathbf{F} \in \mathbb{R}^{P \times N}$, $\mathbf{P = \alpha N}$.

Ground truth labels → vector $\mathbf{Y} \in \mathbf{R}^P$ according to $\mathbf{Y = sign(F \cdot x_0)}$.

Student perceptron → $\mathbf{x}$ such that $\mathbf{Y = sign(F \cdot x)}$ on the training set **F**.

Generalization error → distance in weight space between teacher and student functions.

Budget of the student: $\mathbf{0 < n \leq \alpha}$. Select and query the labels of a subset **S** of cardinality $\mathbf{|S| = nN}$, according to some active learning criterion.

NOTE: **F** i.i.d. normal → full set of input data is unstructured and uncorrelated, BUT in the labelled subset **S** non-trivial correlations can appear!

Gardner volume and mutual information

$$\mathcal{I}(\boldsymbol{x}^0;\boldsymbol{Y}|\boldsymbol{F}) = \mathcal{H}(\boldsymbol{Y}|\boldsymbol{F}) - \mathcal{H}(\boldsymbol{Y}|\boldsymbol{F},\boldsymbol{x}^0) = \mathcal{H}(\boldsymbol{Y}|\boldsymbol{F})$$
$$= -\int d\boldsymbol{Y} \int d\boldsymbol{x}^0\, P_X(\boldsymbol{x}^0) P_{\rm out}(\boldsymbol{Y}|\boldsymbol{F}\cdot\boldsymbol{x}^0) \ln \int d\boldsymbol{x} P_X(\boldsymbol{x}) P_{\rm out}(\boldsymbol{Y}|\boldsymbol{F}\cdot\boldsymbol{x})$$

The Gardner volume **v** represents the extent of the version space, i.e. the **entropy of hypotheses** in the model class consistent with the labeled training set. This provides a natural measure of the quality of the student training.

$$\ln v \equiv \frac{1}{N} \mathbb{E}_{\boldsymbol{x}^0,\boldsymbol{Y}} \ln \int d\boldsymbol{x} P_X(\boldsymbol{x}) P_{\rm out}(\boldsymbol{Y}|\boldsymbol{F}\cdot\boldsymbol{x})$$

Large deviations of the selection process

Count the number of possible labelled subsets that induce a given generalization error → Legendre transform.

Introduce a temperature **β** and a chemical potential **Φ** :

$$\mathbb{P}_{\beta,\phi}(\{\sigma_\mu\}) = \left[\int d\boldsymbol{x} P_X(\boldsymbol{x}) \prod_{\mu=1}^{\alpha N} P_{\text{out}}(y^\mu | \boldsymbol{F}^\mu \cdot \boldsymbol{x})^{\sigma_\mu}\right]^\beta e^{\phi \sum_\mu \sigma_\mu}$$

selection variables

Gardner volume

budget

Free entropy :

$$\Phi(\beta,\phi) = \mathbb{E}_{\boldsymbol{F},\boldsymbol{x}^0} \frac{1}{N} \ln \Xi = \mathbb{E}_{\boldsymbol{F},\boldsymbol{x}^0} \frac{1}{N} \ln \sum_{\sigma_\mu} \mathbb{P}_{\beta,\phi}(\{\sigma_\mu\}).$$
$$= \operatorname*{extr}_{v,n} \{\Sigma(n,v) + \beta \ln v + \phi\, n\}.$$

Inverting the Lengendre transform gives us the sought **complexity** :

$$\Sigma(n,v) = \Phi(\beta,\phi) - \beta \ln v - n\phi|_{\partial_\beta \Phi = \ln v, \partial_\phi \Phi = n}.$$

The analysis is completely **agnostic** on how the selection process is achieved

Large deviations of the selection process

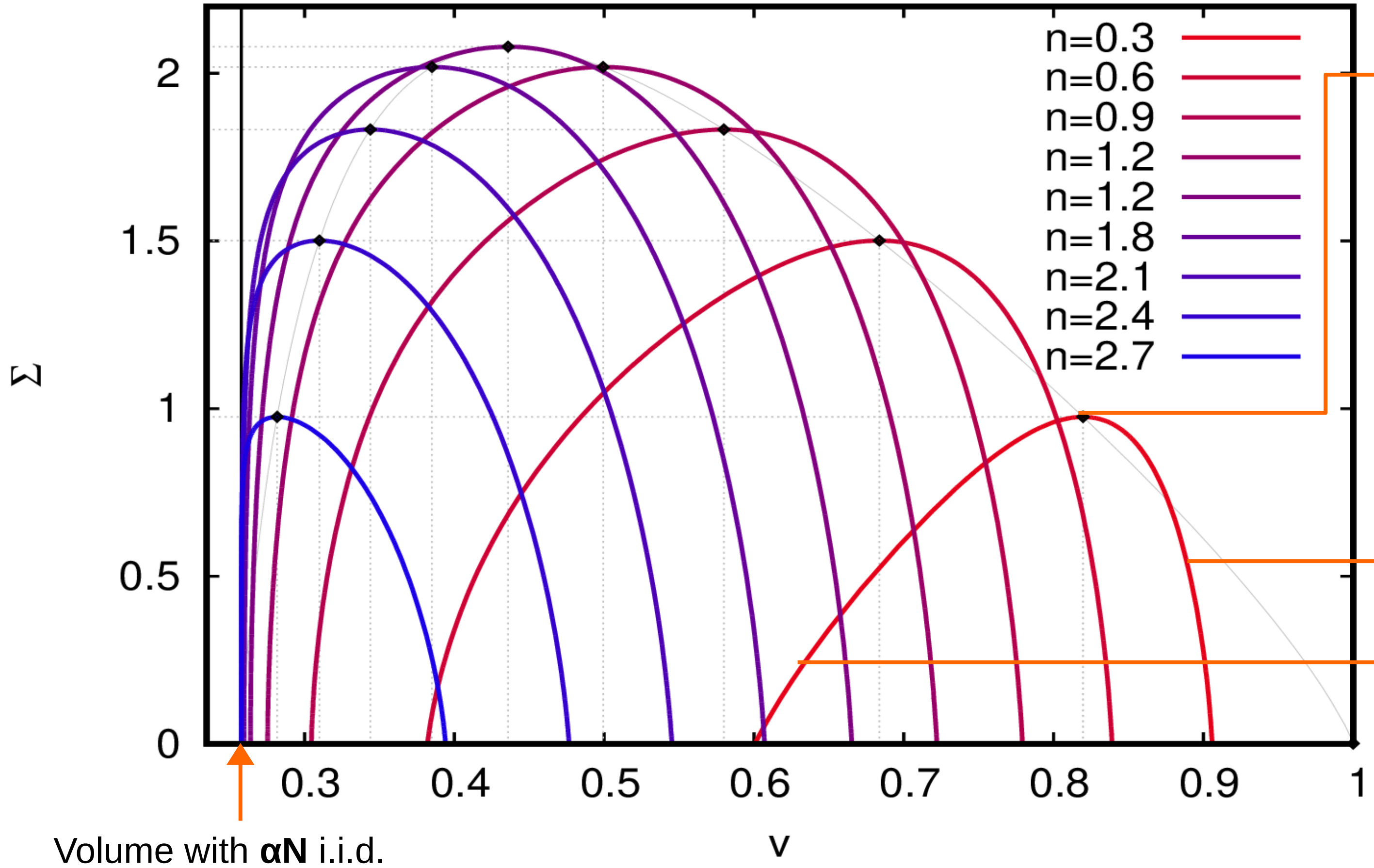
Replica symmetric assumption → order parameters:

- $q = \frac{1}{N}\sum_i \left\langle \langle x_i \rangle_{\boldsymbol{x}|S} \right\rangle_S^2,$ Typical **overlap** between students with **different** labelled subsets
- $Q = \frac{1}{N}\sum_i \left\langle \langle x_i \rangle_{\boldsymbol{x}|S}^2 \right\rangle_S,$ Typical **overlap** between students with **same** labelled subset
- $r = \frac{1}{N}\sum_i \left\langle \langle x_i^2 \rangle_{\boldsymbol{x}|S} \right\rangle_S,$ Typical **norm** of a student
- $m = \frac{1}{N}\sum_i \left\langle \langle x_i x_i^0 \rangle_{\boldsymbol{x}|S} \right\rangle_S,$ Typical **magnetization** of a student

$$\Phi_{\mathrm{RS}}(\beta,\phi) = \underset{\hat{m},\hat{r},\hat{q},\hat{Q},m,r,q,Q}{\mathrm{extr}} \Bigg\{ \frac{\beta}{2} r\hat{r} - \beta m\hat{m} - \frac{\beta(\beta-1)}{2} Q\hat{Q} + \frac{\beta^2}{2} q\hat{q} - \frac{\beta-1}{2}\ln(1+\hat{r}+\hat{Q})$$
$$- \frac{1}{2}\ln(1+\hat{r}-(\beta-1)\hat{Q}+\beta\hat{q}) + \frac{\beta}{2}\frac{\hat{q}+\hat{m}^2}{1+\hat{r}-(\beta-1)\hat{Q}+\beta\hat{q}}$$
$$+ \boxed{2\alpha\int D\eta H\left(-\sqrt{\frac{m^2}{q-m^2}}\eta\right)\ln\left[1+e^{\phi}\int D\zeta H\left(-\frac{1}{\sqrt{r-Q}}(\sqrt{Q-q}\zeta+\sqrt{q}\eta)\right)^{\beta}\right]\Bigg\}}$$

trace over the selection variable

Large deviations: results (α=3)



Volume with **αN** i.i.d. patterns

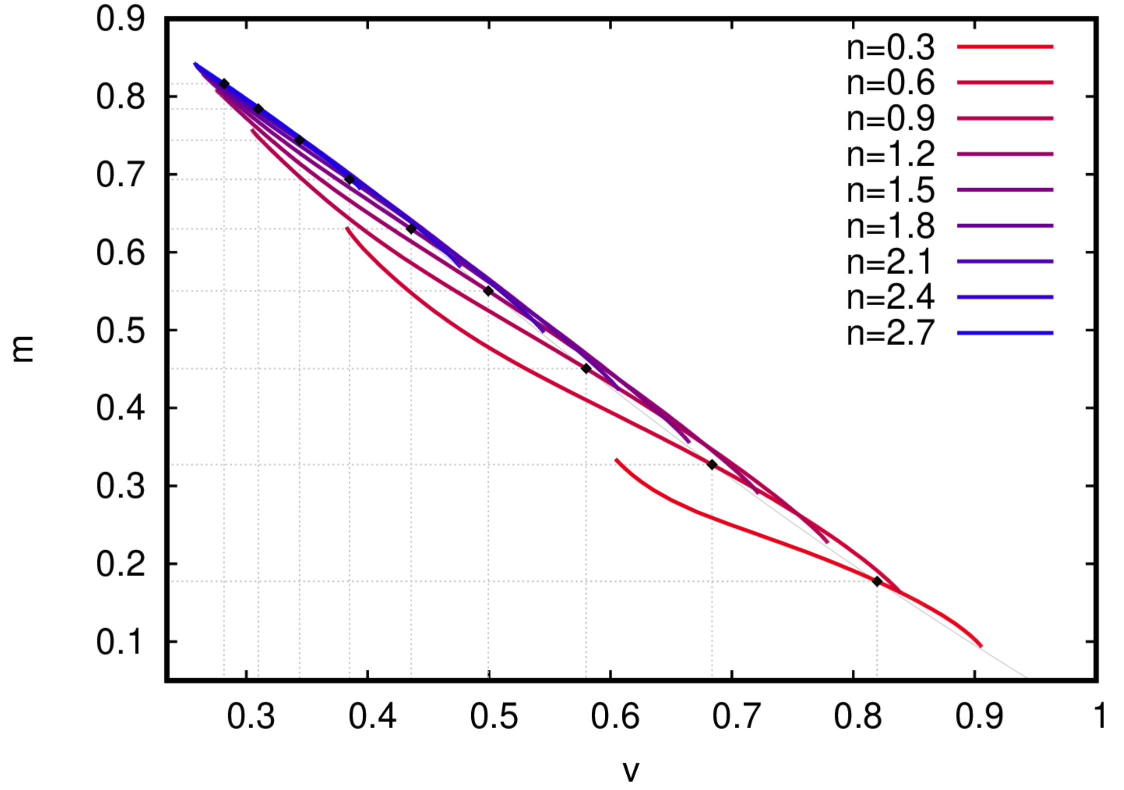
β=0: **typical** Garner volume (**nN** i.i.d. patterns)
→ maximum complexity (binomial distribution)

β>0: atypically **large** Gardner volumes.
→ query worse than random sampling.
Right positive complexity extremum: largest possible volume at budget **n**
→worst generalization.

β<0: atypically **small** Gardner volumes
→ high information content.
The left positive-complexity extremum: smallest possible volume at budget **n**
→best generalization.

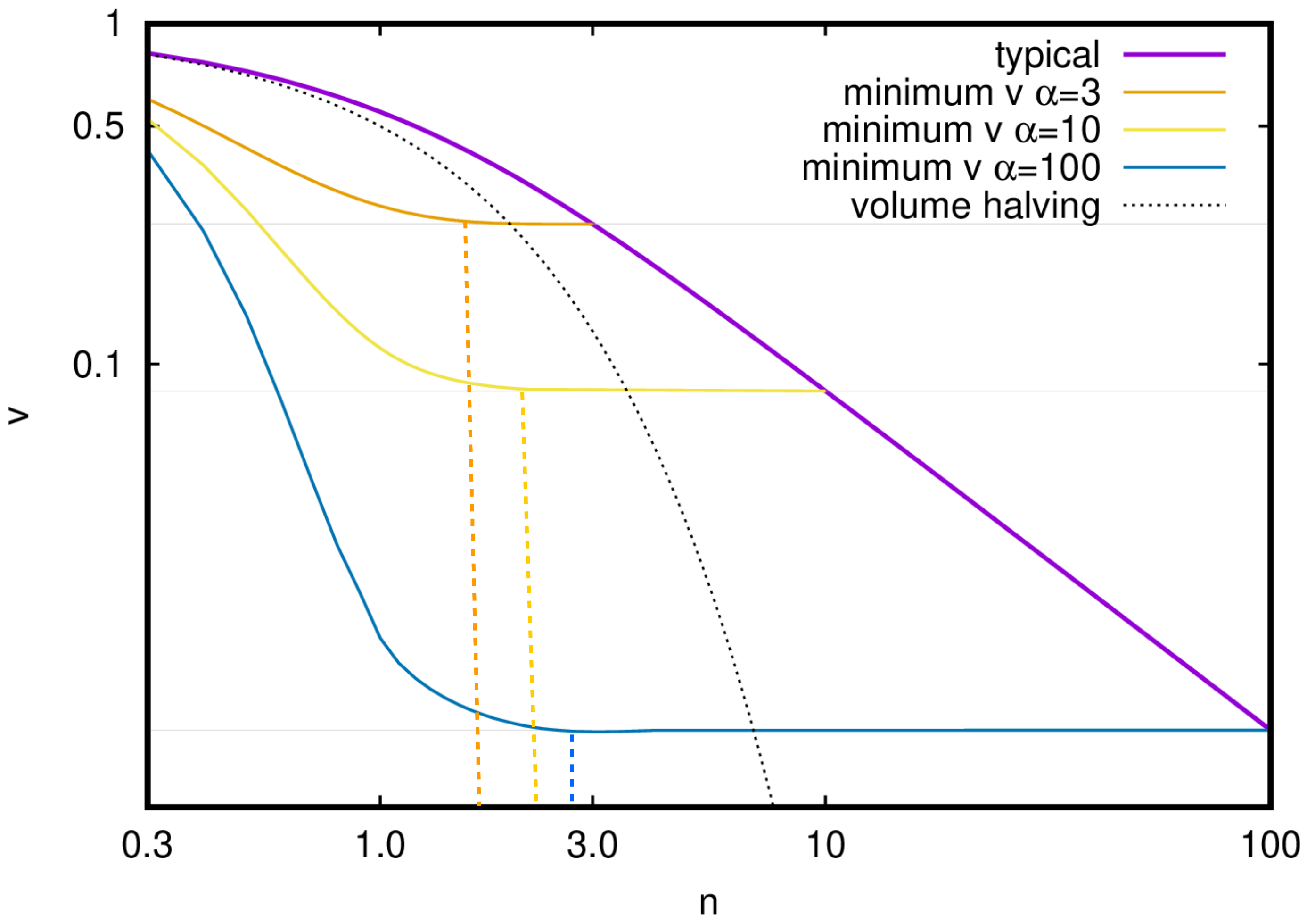
Finding (very) atypical subsets is **HARD**.

Large deviations: results



small Gardner volumes → high magnetizations → low generalization errors

Algorithmic implications



Active learning **lower bound**:

→ all the information contained in the full set of patterns is extracted after querying few samples (**logarithmic fraction**)

→ valid for **any selection strategy**, even for an active learning algorithm that can access additional information on the structure of data (teacher, true labels, etc..)

HOWEVER…

With no prior (or external info) about the generative model the best information gain you can get is **1 bit per pattern**

→ **Volume halving curve**

→ exponential decrease

→ (still not easily achieved)

Uncertainty sampling strategies

When no external prior is available on the data structure, many active learning criteria rely on some form of **label-uncertainty measure**.

→ **Uncertainty sampling**: iteratively selecting and labelling data-points where the **prediction** of the available trained model is the **least confident.**

Active learning CYCLE

1- train model on current labelled subset

2- evaluate model predictions at unlabelled datapoints

3- sort according to confidence (magnitude)

4- query most uncertain samples

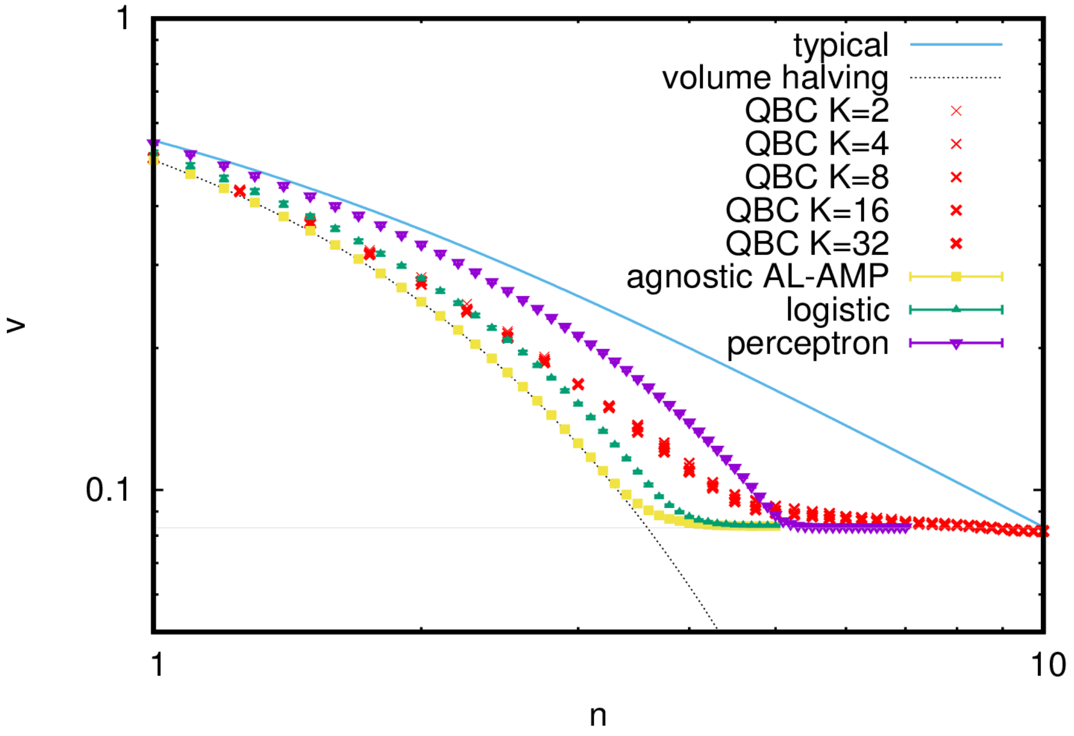
5- repeat ...

→ Let’s **benchmark** some known strategies!

Algorithmic results ($\alpha = 10$)

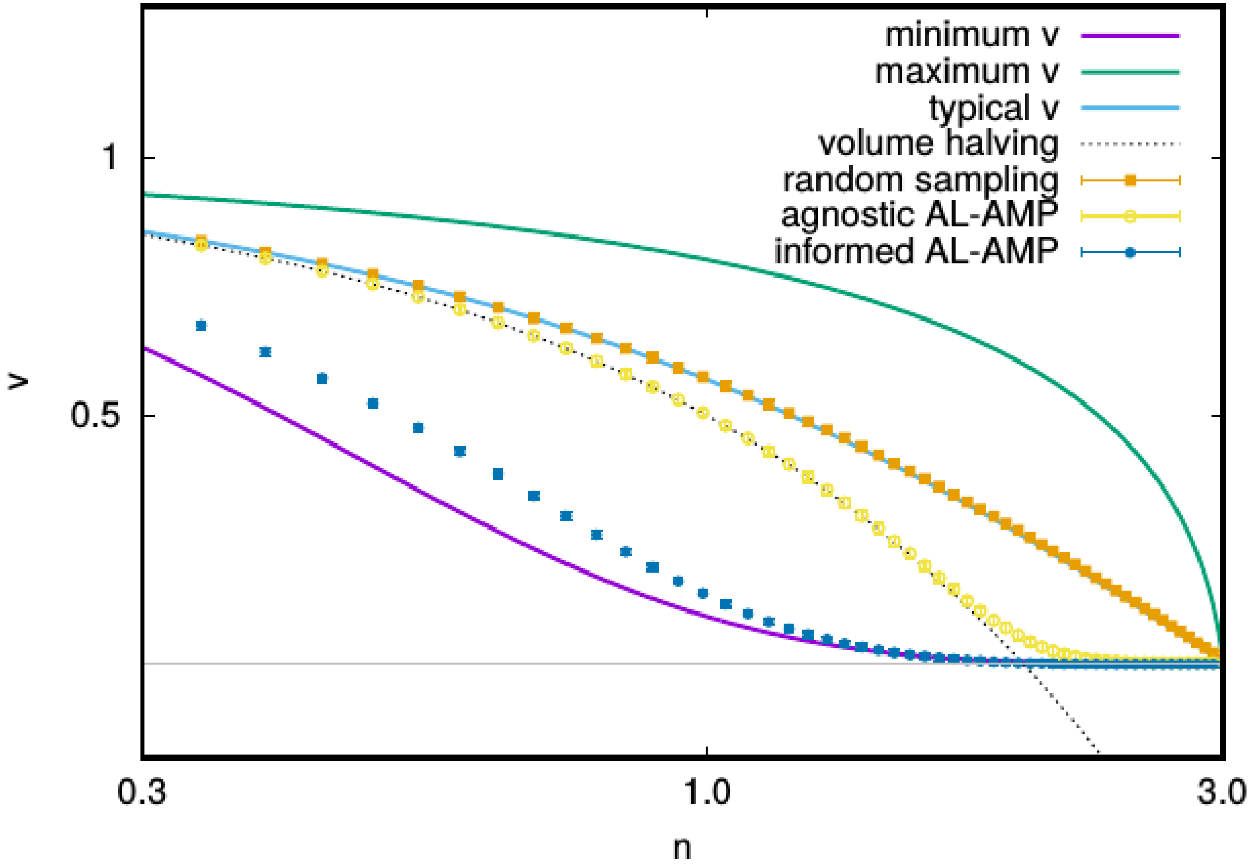
In GLM (i.i.d. Gaussian input data), in high dimension
→ **Approximate Message Passing (AMP)**

At convergence AMP yields an estimator of the **posterior means and variances** → easy to evaluate **model uncertainty** on new data points



Uncertainty sampling strategies		
Heuristic	Required estimates	Sorting criterion
Agnostic AL-AMP	$\hat{\boldsymbol{x}}_{\rm AMP}, \hat{\boldsymbol{\Delta}}_{\rm AMP}$	$\arg\min_\mu \left\| \mathrm{erf}\left(\frac{\boldsymbol{F}'^\mu \hat{\boldsymbol{x}}_{\rm AMP}}{\sqrt{2 (\boldsymbol{F}'^\mu)^2 \hat{\boldsymbol{\Delta}}_{\rm AMP}}} \right) \right\|$
Query by committee	$\{\boldsymbol{x}^k_{\rm SGD}\}_{k=1}^K$	$\arg\min_\mu \left\| \sum_{k=1}^K \mathrm{sign}\left(\boldsymbol{F}'^\mu \cdot \boldsymbol{x}^k_{\rm SGD}\right) \right\|$
Logistic regression	$\boldsymbol{x}_{\rm log}$	$\arg\min_\mu \|\boldsymbol{F}'^\mu \cdot \boldsymbol{x}_{\rm log}\|$
Perceptron learning	$\boldsymbol{x}_{\rm perc}$	$\arg\min_\mu \|\boldsymbol{F}'^\mu \cdot \boldsymbol{x}_{\rm perc}\|$

Algorithmic results ($\alpha = 3$)



Hard to emulate a scenario where the selection algorithm can access **external information** on data structure in our i.i.d. framework!

→ Even in the extreme case where **all the true labels are disclosed to the student**, finding the subset that minimizes the Gardner volume is still a hard problem.

A label-informed AMP algorithm approaches our theoretical bound.

Limits of the approach
and future research

Stability analysis → **1RSB would be needed**

How to study AL in different models (where **volume ≠ mutual info**)?

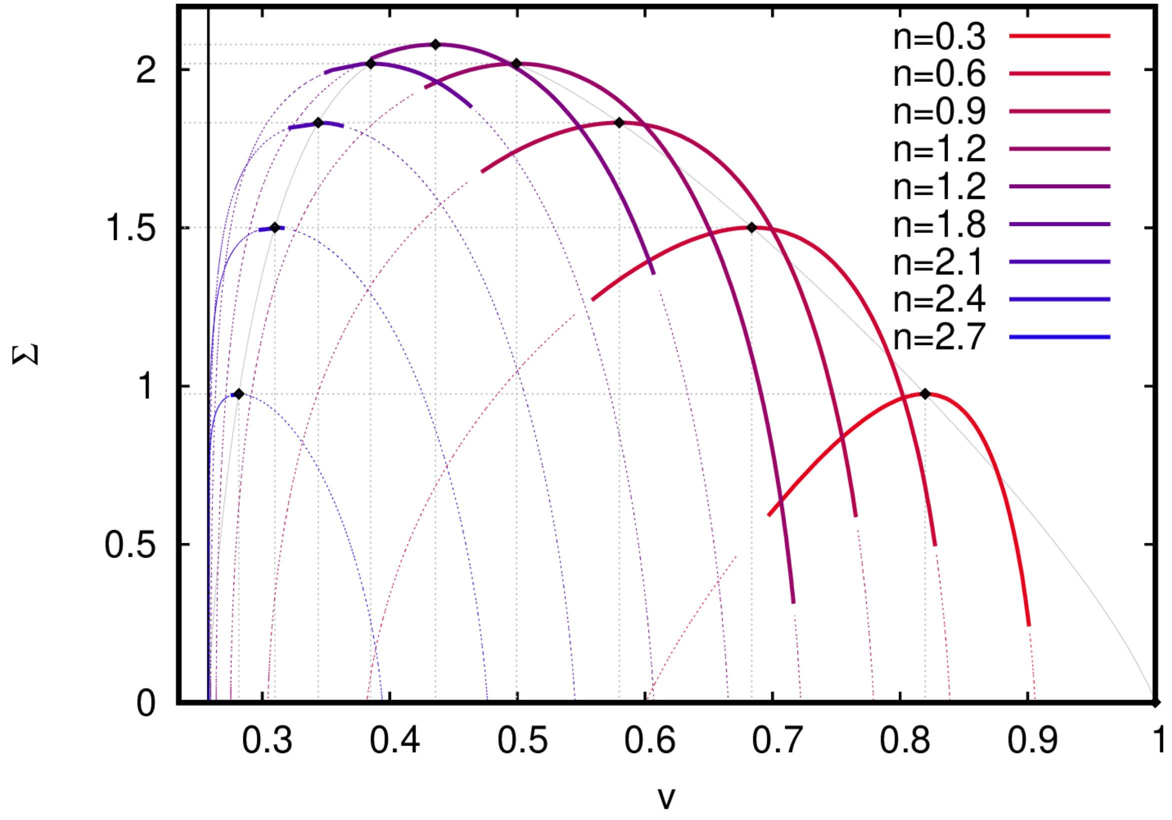
Understand the **convergence issues of AMP** (less constraints → harder)?

Connect with other **label reweighting** strategies (soft labelling, distillation, ...)

THANK YOU FOR YOUR ATTENTION

Stability analysis

Free entropy in the 1RSB ansatz → stability of the RS solution with respect to an **infinitesimal 1RSB perturbation**.



AMP iteration

$$\omega_\mu^t = \sum_i F_i^\mu \hat{x}_i^{t-1} - g_\mu^{t-1} V^{t-1}$$

$$g_\mu^t = \partial_\omega \varphi_\mu^{\mathrm{out},t}$$

$$\Gamma_\mu^t = -\partial_\omega^2 \varphi_\mu^{\mathrm{out},t}$$

$$A^t = c_{\boldsymbol{F}} \sum_\mu \Gamma_\mu^t$$

$$B_i^t = \sum_\mu F_i^\mu g_\mu^t + \hat{x}_i^{t-1} A^t$$

$$\hat{x}_i^t = \partial_B \varphi_i^{\mathrm{in},t}$$

$$\Delta_i^t = \partial_B^2 \varphi_i^{\mathrm{in},t}$$

$$V^t = c_{\boldsymbol{F}} \sum_i \Delta_i^t$$

$$\varphi_\mu^{\mathrm{out},t} = \varphi^{\mathrm{out}}(\omega_\mu^t, V^t, y^\mu)$$

$$\varphi_i^{\mathrm{in},t} = \varphi^{\mathrm{in}}(B_i^t, A^t)$$

Estimating **model uncertainty** → hard problem!

In GLM (i.i.d. Gaussian input data), in high dimension → **Approximate Message Passing (AMP)**

At convergence AMP yields an estimator of the **posterior means and variances**, and a prediction on new data points: