# Aggregating Data for Optimal Learning



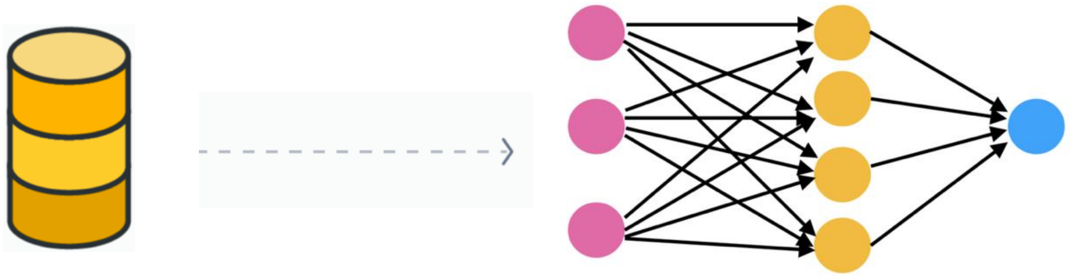

**Sushant Agarwal***[1] Yukti Makhija[2] Rishi Saket[2] Aravindan Raghuveer[2]

[1]Northeastern University [2]Google DeepMind

**UAI 2025**

*Work done during an internship at Google DeepMind

# Supervised Learning



Dataset $\longrightarrow$ Model trained on bags

- Training dataset comprises of $n$ tuples $(x_i, y_i)$.
  - $x_i$ denotes an instance/feature-vector with label $y_i$.
  - Denote the sets by $X, Y$ respectively.
- Train a model to predict labels of unseen instances.

# Learning from Aggregate Labels

Dataset → Split into disjoint bags Only one aggregate bag-label per bag → Model trained on bags

- $X$ is partitioned into disjoint bags $B = \{B_1, B_2, \ldots, B_k\}$.
  - Bag $B_l$ has bag-label $\overline{y_l}$.
- $\overline{y_l}$ is derived from the labels present in $B_l$ via some aggregation function.
- Train a model to predict labels of unseen instances.

# LLP and MIR

We focus on two popular paradigms.

- **Learning from Label Proportions (LLP):** $\overline{y_l}$ is the average of the individual instance labels in $B_l$.
- **Multiple Instance Regression (MIR):** $\overline{y_l}$ is the label of one (undisclosed) instance in $B_l$, chosen uniformly at random.

LLP and MIR formulations are becoming increasingly prevalent.

- Privacy concerns
    - If the bags are of large size, revealing only the aggregate bag-label to the learner provides privacy protection for individual labels.
- Semi-supervised learning
    - One could partition the data into bags, and query an annotator for the label of one of the instances in each bag.

# Problem Statement

- In some cases, bags are fixed.
- In others, these is flexibility in curating the bags.

Main question: Given a lower bound on bag size, what is the ***optimal bagging strategy***, to maximise utility of models trained on these bags?

The minimum bag size constraint is essential, or else the optimal bagging would be the trivial strategy of putting each point in a separate bag.

# Setup

We consider the task of linear regression.

- Assume the existence of an underlying (unknown) $\theta^*$.
    - $y_i = x_i\theta^* + \epsilon_i, \epsilon_i = N(0, \sigma^2)$ .
- Given bags and bag-labels, find estimator $\hat{\theta}$ with maximum utility.
    - Utility defined in terms of closeness to $\theta^*$.
- Train a model on bags by minimizing a given loss function.
    - Instance-level loss
    - Bag-level loss
    - Aggregate-level loss

# Contributions

- **Optimal bagging:** We provide theoretical utility guarantees, and show that in each case, the optimal bagging strategy reduces to finding the optimal $k$-means clustering of the feature vectors or the labels.
- **Differential Privacy**: Apart from the inherent privacy that MIR and LLP offer, we can perturb the labels to obtain formal *label differential-privacy* guarantees, incurring an additional utility error.
- **GLMs:** We extend our results for Linear Regression to Generalized Linear Models (GLMs).
- **Experiments:** We experimentally validate our results on both synthetic and real-world data.

# Loss Functions

- An estimator $\widehat{\theta}$ minimizes **instance-level loss**, if

$$\hat{\theta} := \underset{\theta}{\operatorname{argmin}} \frac{1}{n} \sum_{l=1}^{m} \sum_{i \in B_l} \ell(\overline{y}_l, f_\theta(x_i))$$

- An estimator $\widehat{\theta}$ minimizes **bag-level loss**, if

$$\hat{\theta} := \underset{\theta}{\operatorname{argmin}} \frac{1}{m} \sum_{l=1}^{m} \ell \left( \overline{y}_l, \frac{\sum_{i \in B_l} f_\theta(x_i)}{|B_l|} \right) .$$

- An estimator $\widehat{\theta}$ minimizes **aggregate-level loss**, if

$$\hat{\theta} := \underset{\theta}{\operatorname{argmin}} \frac{1}{m} \sum_{l=1}^{m} \ell \left( \overline{y}_l, f_\theta \left( \frac{\sum_{i \in B_l} x_i}{|B_l|} \right) \right) .$$

# Optimal Bagging

- Intuitively, a bagging provides good utility if the bags are ***homogeneous***, i.e., the feature-vectors and/or labels within a bag are similar.
- By deriving upper bounds on the error, we deduce optimal bagging strategies.

We consider two types of bagging procedures.

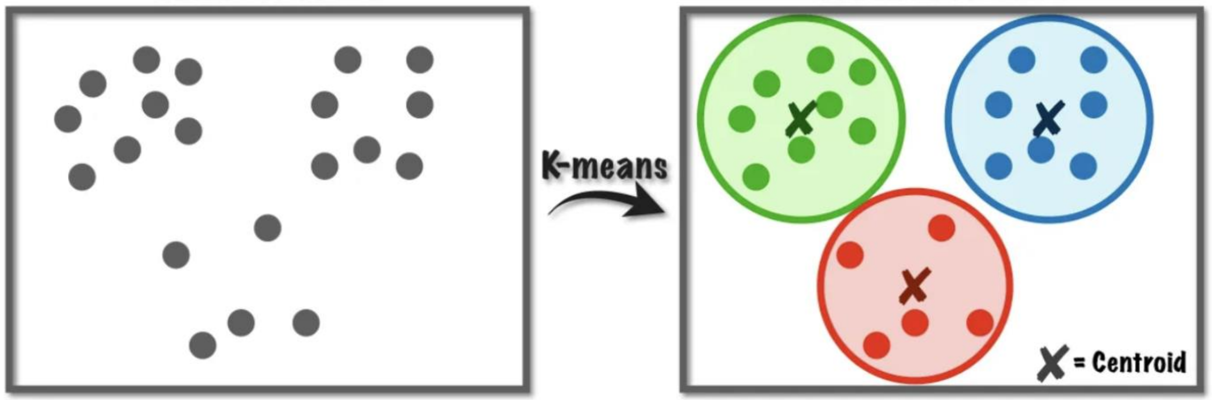
- **Label-dependent bagging:** Individual instance labels are available for the bagging.
- **Label-agnostic bagging:** Individual instance labels are ***not*** available for the bagging.

# Optimal Bagging (label-dependent)

| | Instance-level loss | Bag/Aggregate-level loss |
|---|---|---|
| LLP | 1-dimensional $k$-means clustering of the labels **(Javanmard et al. '24)** | Minimize the condition number of the covariance matrix of each bag's centroid **(Our work)** |
| MIR | 1-dimensional $k$-means clustering of the labels **(Our work)** | Involves both $k$-means clustering of labels, and minimizing the condition number **(Our work)** |

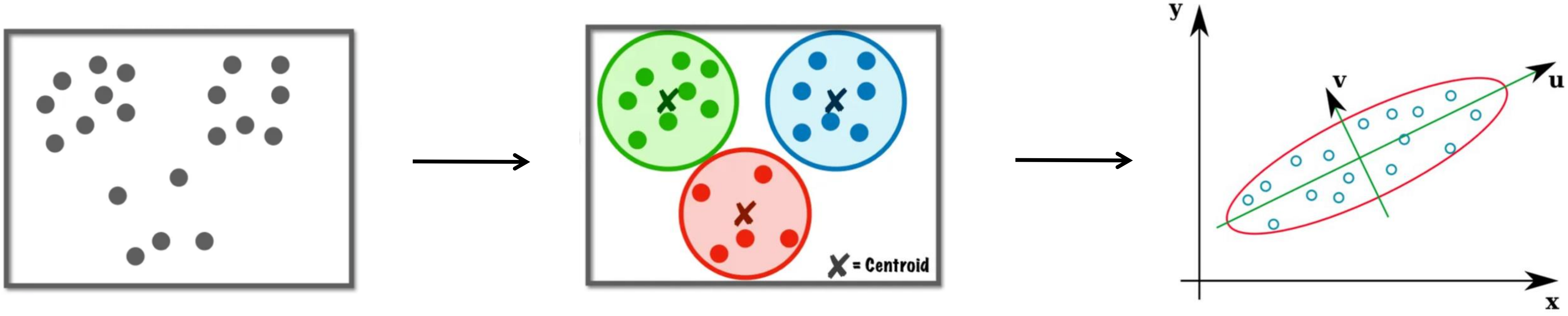
# Optimal Bagging (label-dependent)

$$\underset{\mathbf{S}}{\arg\min} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \left\| \mathbf{x} - \boldsymbol{\mu}_i \right\|^2$$



# Optimal Bagging (label-dependent)

Minimize the condition number of the covariance matrix of each bag's centroid?



# Optimal Bagging (label-agnostic)

- **Instance $\boldsymbol{k}$-means:** We justify that $k$-means of the instances $X$ is a good heuristic for each scenario we consider.
    - $y_i \approx x_i\theta^* \Longrightarrow k$-means of instances is a good heuristic for $k$-means of labels.
    - Maximizing the variance of bag-centroids along a direction $\Longleftrightarrow$ finding an optimal $k$-means clustering of instances projected on that direction.
- **Random bagging:** As a baseline, we also provide a utility analysis of bagging randomly.

# Thanks for listening!

For more details,
check out the paper!

Personal Website
agarwal.sus@northeastern.edu