# Aggregating Data for Optimal Learning

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

[1]Northeastern University
[2]Google Deepmind

## Abstract

Multiple Instance Regression (MIR) and Learning from Label Proportions (LLP) are useful learning frameworks, where the training data is partitioned into disjoint sets or *bags*, and only an aggregate label, i.e., *bag-label* for each bag is available to the learner. In the case of MIR, the bag-label is the label of an undisclosed instance from the bag, while in LLP, the bag-label is the mean of the bag's labels. In this paper, we study for various loss functions in MIR and LLP, what is the optimal way to partition the dataset into bags such that the utility for downstream tasks like linear regression is maximized. We theoretically provide utility guarantees, and show that in each case, the optimal bagging strategy (approximately) reduces to finding an optimal clustering of the feature vectors and/or the labels with respect to natural objectives such as $k$-means. We also show that our bagging mechanisms can be made *label-differentially private*, incurring an additional utility error. We then generalize our results to the setting of Generalized Linear Models (GLMs). Finally, we experimentally validate our theoretical results.

## 1 INTRODUCTION

In traditional supervised learning, the training dataset is a set of $n$ tuples of the form $(\mathbf{x}, y)$, where $\mathbf{x}$ is an instance or feature-vector with label $y$ (denote the sets of tuples by $X, Y$ respectively). The objective is to train a model on the training data $(X, Y)$, that predicts the labels of unseen test instances. In this paper, we study the paradigm of *learning from aggregate labels*, in which $X$ is partitioned into $m$ disjoint sets or *bags* of instances $B = \{B_1, \ldots, B_m\}$, and for each bag $B_l$ only one *bag-label* ($\overline{y}_l$) is available to the learner. $\overline{y}_l$ is derived from the instance-labels present in the bag via some aggregation function depending on the scenario. The goal, similar to standard supervised learning, is to train a model that predicts the labels of individual instances. This paradigm of learning from aggregate labels directly generalizes traditional supervised learning, the latter being the special case of unit-sized bags. The two formalizations of our focus are (i) Multiple Instance Regression (MIR), where $\overline{y}_l$ is one of the instance-labels of $B_l$[1], and the instance whose label is chosen as the bag-label is not revealed, and (ii) Learning from Label Proportions (LLP), in which $\overline{y}_l$ is the average of $B_l$'s instance-labels.

The MIR and LLP frameworks are becoming increasingly prevalent, and we briefly discuss two use cases (see Section 1.2 for a more detailed discussion). There are many practical scenarios (eg., medical tests) in which labels are much more private than the features, and we wish to protect the privacy of individual labels from the learner (and any downstream observer of the learners output). In the MIR and LLP setups, if the bags are of large size, revealing only the aggregate bag-label to the learner provides a layer of privacy protection for individual labels. Due to increasing concerns over data privacy, recent regulations on sharing user-level signals across platforms have resulted in aggregation of data, resulting in LLP and MIR formulations for predictive model training on revenue critical advertising datasets (e.g. Apple SKAN and Chrome Privacy Sandbox, see O'Brien et al. [2022]).

In addition to privacy, in many applications, obtaining labeled data is very costly, but unlabeled data is relatively easy to acquire. This is especially relevant as training data is getting increasingly complex, and skilled human annotators are required for data-labeling, leading to semi-supervised learning settings [Van Engelen and Hoos, 2020]. Given a large amount of unlabeled data, and a limited labeling budget, one could partition the data into bags, and query an annotator for the label of one of the instances in each bag. This setting naturally lends itself to the MIR formulation that we study.

[1]We consider the popular case where $\overline{y}_l$ is uniformly random.

In some scenarios, the bags of instances may already be fixed, whereas in other scenarios like semi-supervised learning, there might be flexibility in curating the bags. We study the question of finding the *optimal bagging strategy*, for the purpose of maximising utility of downstream tasks trained on these bags and corresponding bag-labels. We distinguish between baggings based on whether or not labels are available for constructing the bags. We call them (i) label-agnostic bagging, which occur in settings like semi-supervised learning, and (ii) label-dependent bagging, which occur naturally in privacy motivated scenarios.

We consider a regression setting, where instances $\mathbf{x}$ lie in $\mathbb{R}^d$, with labels $y \in \mathbb{R}$. We adopt a standard way to model linear regression, where label $y_i = \mathbf{x}_i^T \theta^* + \gamma_i \,, \gamma_i \sim \mathcal{N}(0, \sigma^2)$, for a fixed underlying model $\theta^*$. Given the bags and corresponding bag-labels, the learner's task is to find an estimator $\hat{\theta}$ with minimal estimation error, by minimizing some loss function. A common loss is *instance-level loss*, that basically assigns the aggregate label of the bag to each point in the bag. An estimator $\hat{\theta}$ minimizes instance-level loss, if

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

where $\ell$ is the squared loss. Another popular loss function is *bag-level loss*, which measures the mismatch between the bag-label and mean of the bag's instance level predictions. An estimator $\hat{\theta}$ minimizes bag-level loss, if

$$\hat{\theta} := \operatorname*{argmin}_{\theta} \frac{1}{m} \sum_{l=1}^{m} \ell\left(\overline{y}_l, \frac{\sum_{i \in B_l} f_\theta(x_i)}{|B_l|}\right) . \tag{2}$$

We also consider *aggregate-level loss*, which penalises the difference between the bag-label and prediction of the mean of the bag instances. An estimator $\hat{\theta}$ minimizes aggregate-level loss, if

$$\hat{\theta} := \operatorname*{argmin}_{\theta} \frac{1}{m} \sum_{l=1}^{m} \ell\left(\overline{y}_l, f_\theta\left(\frac{\sum_{i \in B_l} x_i}{|B_l|}\right)\right) . \tag{3}$$

Given the learning setup (either MIR or LLP, and a loss function), the optimal bagging strategy involves finding the bagging configuration that maximizes the utility of $\hat{\theta}$ trained using the loss function, with utility defined in terms of closeness to $\theta^*$. Note that each bag has size at least $k$ which is a fixed value.

*Remark.* The minimum bag size constraint is essential to define a meaningful problem, otherwise the optimal bagging would be the trivial strategy of putting each point in a separate bag. Smaller bags provide more information about the labels, leading to better utility. However, larger bags are more suitable in cases where MIR and LLP are deployed, such as privacy motivated and semi-supervised learning scenarios.

### 1.1 OUR RESULTS

We briefly summarize our contributions below.

**1) Label-dependent Bagging:** Intuitively, a bagging provides good utility if the bags are *homogeneous*, i.e., the instances and/or instance-labels within a bag are similar. We formalize this intuition below, and study the following learning setups.

*a) MIR, Instance-level loss:* By deriving a sharp upper bound on the estimation error in Theorem 1, we show that finding the optimal bagging reduces to the following $k$-means clustering over the labels,

$$\min_{\mathcal{B}} \sum_{l=1}^{m} \sum_{\tilde{y}_i \in B_l} (\tilde{y}_i - \mu_l)^2, \text{ with } |B_l| \geq k, \forall l \in [m] \tag{4}$$

where $\mu_l$ is the mean of the labels in $B_l$, $\tilde{y}_i$ denotes the expected value of the label of $x_i$, i.e., $\tilde{y}_i := \mathbf{x}_i^T \theta^*$, and $\mathcal{B}$ denotes the set of all baggings of the $n$ samples. This is just a (size-constrained) $k$-means clustering of $\tilde{y}$[2], and intuitively creates bags that are homogeneous w.r.t. labels. The $1d$ clustering problem above can be solved exactly in polynomial time, and turns out to result in a bagging that just sorts the labels in order, and partitions contiguous segments into bags (see Lemma 9).

*b) LLP, Bag-level loss:* By deriving an upper bound on the error in Theorem 2, we show that finding the optimal bagging reduces to the following optimization problem.

$$\min_{\mathcal{B}} \frac{\lambda_{max}(f(X))}{\lambda_{min}(f(X))}, \text{ subject to} |B_l| = k, \forall l \in [m], \tag{5}$$

where $\lambda_{max}/\lambda_{min}$ denote the maximum/minimum eigenvalues of a matrix, and $f(X) = g(X)g(X)^T$, for $g(X) = \left[\left(\frac{\sum_{i \in B_1} x_i}{|B_1|}\right), \ldots, \left(\frac{\sum_{i \in B_m} x_i}{|B_m|}\right)\right]$. Essentially, $f(X)$ is the (sample) covariance matrix of each bag's instance-mean. The optimal bagging strategy involves minimizing the condition number ($\lambda_{max}/\lambda_{min}$ ratio) of $f(X)$, and intuitively creates bags that are homogeneous w.r.t. instances. The above discusses equal sized bags, and we Theorem 7 a corresponding result without the equality constraint.

*c) MIR, Aggregate-level loss:* As seen from the error bound in Theorem 3, the optimal bagging strategy here involves simultaneously minimizing the condition number of $f(X)$, and minimizing the $k$-means clustering objective of $\tilde{y}$, intuitively creating bags that are homogeneous w.r.t. both instances and their labels.

**2) Label-agnostic Bagging:** As seen above, a good bagging has bags that are homogeneous w.r.t. instances and/or labels. A label-agnostic bagging can create baggings that

[2] $\tilde{y}$ is unavailable, but one can instead use $y$ as a proxy, leading to an additional utility error of $n\left(1 - \frac{1}{k}\right)\sigma^2$, see Lemma 10.

are homogeneous w.r.t. instances, but is not able to directly optimize for homogeneity w.r.t. labels. We consider the following 2 label-agnostic bagging strategies.

*a) Instance $k$-means* We justify that the optimal $k$-means clustering of the instances $X$ is an effective label-agnostic bagging strategy for each learning setup we consider. In Instance-MIR, the optimal strategy is a $k$-means clustering of the labels $Y$. We use the fact that $\tilde{Y} = X\theta^*$ to justify that $k$-means of the instances $X$ is a good heuristic for $k$-means of the labels $Y$ (see Section 3.1). In the case of Bag-LLP, the optimal bagging strategy does not involve knowledge of the labels, and minimizes the condition number of the sample covariance matrix of the instance-means of each bag. An eigenvalue of a covariance matrix measures the variance along the corresponding eigenvector. In order to minimize the condition number, we intuitively maximize variance in every direction. We show that maximizing the variance of bag-centroids along a direction is equivalent to finding an optimal $k$-means on $X$ projected on that direction. Hence, we want to reduce the $k$-means objective along every direction, and we justify that $k$-means of $X$ is a good heuristic for the same. For Aggregate-MIR, we must simultaneously minimize the condition number of $f(X)$, and the $k$-means objective over the labels $\tilde{y}$, and $k$-means of $X$ is a good heuristic for both objectives.

*b) Random bagging* We analyse random bagging in Section 3.2. We upper bound the error of random bagging in both Bag-LLP and Aggregate-MIR. Since bounding the condition number term in Equation (5) as a whole is challenging, we provide an upper bound for $\lambda_{\text{max}}$, and a lower bound for $\lambda_{\text{min}}$. As shown in Lemma 2 via an application of Cauchy-Schwarz, aggregating feature vectors does not increase $\lambda_{\text{max}}$. For lower bounding $\lambda_{\text{min}}$ we consider a partitioning strategy where the instances are randomly divided into $2k$-sized *super-bags*. Independently from each super-bag, one $k$-sized bag is is sampled, resulting in a collection of $m/2$ bags which are distributed identically to a random collection of $m/2$ disjoint bags and therefore a lower bound on $\lambda_{\text{min}}$ for these bags is sufficient. Observing that bags are independent in this collection (after fixing the super-bags), we compute $\mu_{\text{min}}$, which is the expected value of $\lambda_{\text{min}}$ for these bags, and use Matrix Chernoff to find a high probability lower bound for $\lambda_{\text{min}}$, as stated in Lemma 3.

**3) Privacy** Apart from the inherent privacy that MIR and LLP offer, we can perturb the labels to obtain formal privacy guarantees in the sense of *label differential privacy*, a popular notion of privacy that measures and prevents the leakage of label information [Chaudhuri and Hsu, 2011]. This incurs an additional utility error, that we formally quantify in Section 4. A larger minimum bag-size $k$ intuitively provides more privacy, and as expected, the error increases with a decrease in $k$.

**4) GLM's** Subsequently, we generalize the previous results for linear regression to the setting of Generalized Linear Model's (GLMs), which includes popular paradigms such as logistic regression. We study both instance-level and aggregate-level losses for MIR under the GLM framework. For Instance-MIR, we derive an upper bound that leads to label $k$-means clustering as the optimal bagging strategy. This result holds across all distributions within the exponential family. For Aggregate-MIR, our objective suggests minimizing the range between the maximum and minimum expected instance-labels within a bag, implying that features with similar expected labels should be grouped together, yielding a clustering-based outcome. This result holds for exponential distributions which have a monotonic first derivative. The detailed analysis is provided in Appendix F.

**5) Experiments** To corroborate our theoretical results, we study the proposed bagging mechanisms through extensive experimentation in Section 5, and demonstrate their effectiveness on each learning setup we consider. We analyse trends obtained by varying various parameters such as the minimum bag size, number of bags, and privacy budget.

### 1.2 RELATED WORK

LLP started with the work of de Freitas and Kück [2005] and has been studied in the context of privacy concerns [Rueping, 2010], lack of supervision due to cost [Chen et al., 2004], or coarse instrumentation [Dery et al., 2017]. While previous works [Quadrianto et al., 2009, Yu et al., 2013, Kotzias et al., 2015, Liu et al., 2019, Scott and Zhang, 2020, Saket et al., 2022] have developed specialized techniques for model training on LLP training data, Yu et al. [2014] defined it in the PAC framework, while Saket [2021, 2022] have shown worst case algorithmic and hardness bounds, and recently Brahmbhatt et al. [2023] gave PAC learning algorithms for Gaussian feature vectors and random bags.

MIR, introduced in Ray and Page [2001], has mostly been studied in applied settings related to remote sensing and image analysis. Popular baseline techniques apply Aggregate-MIR, or Instance-MIR [Wang et al., 2008, Ray and Craven, 2005], whereas several expectation-maximization (EM) based methods have also been proposed [Ray and Page, 2001, Wang et al., 2008, 2012, Wagstaff et al., 2008, Trabelsi and Frigui, 2018]. Recent work of Chauhan et al. [2024] proved bag-to-instance generalization error bounds as well as hardness results for MIR, in the first theoretical exploration of this problem.

Both the above problems, LLP and MIR, have gained renewed interest due to recent restrictions on user data on advertising platforms leading to aggregate conversion labels in reporting systems [O'Brien et al., 2022]. With the goal of preserving the utility of models trained on the aggregate labels, model training techniques for either randomly sampled [Busa-Fekete et al., 2023] or curated bags [Chen et al., 2023, Javanmard et al., 2024] have been proposed.

**Comparison with Javanmard et al. [2024]:** The case of instance-level loss for LLP has been studied in Javanmard et al. [2024], where they show that the optimal bagging strategy reduces to finding the best $k$-means clustering of the labels, very similar to our Instance-MIR objective. This is not very surprising, as LLP and MIR are closely related. Indeed, the expected label of each bag in the MIR setup is exactly the label of the bag in the LLP case. Our focus is on MIR which has not been studied before, and in addition we analyse the popular bag-level loss [Ardehaly and Culotta, 2017] for LLP. They provide an adaptive label-agnostic bagging heuristic, which assumes access to an oracle that provides bag-labels in an online setting. Our work provides label-agnostic bagging algorithm in each case, without assuming access to an online oracle. We provide formal privacy guarantees for each of our methods. They also discuss privacy guarantees for their heuristic algorithm; however, their approach does not provide formal privacy guarantees for label-dependent bagging, which we circumvent by using a private clustering algorithm.

# 2 LABEL-DEPENDENT BAGGING

$X$ has rank $d$, and all expectations henceforth are conditioned on a fixed $X$, unless otherwise stated. The results below provides an upper bound on the error of the estimator $\hat{\theta}$, in terms of a bagging $B$. Most proofs are deferred to Appendix A.

**Theorem 1** (Error Upper Bound, Instance-MIR). *For $\hat{\theta}$ as in* (1)*, for a given bagging $B$,*

$$\mathbb{E}\left[||\hat{\theta} - \theta^*||_2^2\right] \leq C_1 \left( C_2 - \sum_{\ell=1}^{m} \frac{\left(\sum_{i \in B_\ell} \tilde{y}_i\right)^2}{|B_\ell|} \right), \quad (6)$$

*where constants $C_1, C_2$ are independent of $B$.*

In Lemma 9 in Appendix B, we show that finding the optimal $k$-means clustering of the (expected) labels $\tilde{y}$ exactly minimizes $\sum_{\ell=1}^{m} \frac{\left(\sum_{i \in B_\ell} \tilde{y}_i\right)^2}{|B_\ell|}$. Hence, minimizing the bound in (6) over the set of all baggings amounts to the $k$-means optimization problem in (4).

**Theorem 2** (Error Upper Bound, Bag-LLP). *For $\hat{\theta}$ as in* (2)*, for a given bagging $B$ such that $|B_l| = k, \forall l \in [m]$,*

$$\mathbb{E}\left[\|\hat{\theta} - \theta^*\|_2^2\right] \leq \sigma^2 \frac{m}{k} \left( \frac{\lambda_{max}(f(X))}{\lambda_{min}(f(X))} \right)^2. \quad (7)$$

Minimizing the bound in (7) over the set of all baggings amounts to the optimization problem in (5). Theorem 2 is for equal sized bags, and we also show a corresponding result without the equality constraint in Theorem 7.

**Theorem 3** (Error Upper Bound, Aggregate-MIR). *For $\hat{\theta}$ in* (3)*, given a bagging $B$ such that $|B_l| = k, \forall l \in [m]$,*

$$\mathbb{E}\left[\|\hat{\theta} - \theta^*\|_2^2\right] \leq$$
$$C_1 \left( \frac{\lambda_{max}(f(X))}{\lambda_{min}(f(X))} \right)^2 \left( C_2 + \sum_{l=1}^{m} \sum_{\tilde{y}_i \in B_l} (\tilde{y}_i - \mu_l)^2 \right) \quad (8)$$

*where constants $C_1, C_2$ are independent of $B$.*

Minimizing the first term in (8) corresponds to minimizing the condition number of $f(X)$, and minimizing the second term corresponds to finding the optimal $k$-means clustering of $\tilde{y}$. Theorem 3 is for equal sized bags, and we also show a corresponding result without the equality constraint in Theorem 8.

# 3 LABEL-AGNOSTIC BAGGING

## 3.1 INSTANCE $k$-MEANS

We justify that $k$-means of the instances $X$ is an effective label-agnostic bagging heuristic for each setting we consider.

**Instance-MIR** Note that in our setting of linear regression, $\tilde{Y} = X\theta^*$. In other words, $\tilde{Y}$ is just the projection of $X$ along the axis normal to the hyperplane determined by $\theta^*$. Hence, finding an optimal $k$-means clustering of $\tilde{Y}$ is equivalent to minimizing the $k$-means objective of projections along this axis. However, if the labels are not given, this axis is unknown, since $\theta^*$ is unknown. Hence, in order to do a label-agnostic bagging, one must minimize some objective that simultaneously reduces the $k$-means objective along every direction. In Lemma 12, we show that for a given clustering, the $k$-means objective of a dataset is the sum of $k$-means objective of the dataset projected along each coordinate. Given an arbitrary clustering $C$ over $X$ drawn from an isotropic distribution $D$, in expectation the $k$-means clustering objective over $X$ will split equally into $d$ components along each axis (due to symmetry), i.e.,

$$\mathbb{E}[k\text{-means}(C(X_i))] = \frac{1}{d}\mathbb{E}\left[k\text{-means}(C(X))\right], \forall i,$$

where the expectation is over $X$ drawn from $D$. Hence, for isotropic distribution $D$, we would expect that the $k$-means clustering objective along each direction to be roughly equal. Hence, we would also expect that setting $C$ to be the optimal $k$-means clustering over $X$ would simultaneously keep the $k$-means clustering objective low along each direction.

However, the above reasoning holds only for an isotropic distribution. For a non-isotropic distribution, directions with large variance will dominate the $k$-means objective, and therefore directions with small variance might then have a relatively large $k$-means objective. For an isotropic distribution, we avoid the above problem of directions with large variance dominating. However, note that even for a non-isotropic distribution, $\Sigma^{-\frac{1}{2}}X$ is isotropic, where $\Sigma$ is the covariance matrix of the distribution. Essentially, we stretch each direction so that each direction has the same variance. We can now find an optimal $k$-means clustering over $\Sigma^{-\frac{1}{2}}X$. We will then avoid the problem of directions in $X$ with large variance dominating, while also keeping the $k$-means objective along each direction low.

**Bag-LLP** We want to maximize the condition number of $f(X)$. $\lambda_{\max}/\lambda_{\min}$ of a covariance matrix measures the variance along the direction of most/least variance. In Lemma 13, we show that maximizing the variance of bag's instance-centroids along a direction is equivalent to finding an optimal $k$-means on $X$ projected on that direction. We focus on maximizing $\lambda_{\min}$, however, we do not know the corresponding eigenvector. Hence, we must simultaneously reduce the $k$-means objective along every direction, and in the previous section, we justified $k$-means of the instances X is an effective heuristic for this.

**Aggregate-MIR** Note that in order to minimize the error bound, we must simultaneously minimize the condition number of $f(X)$, and the $k$-means objective over the labels $\tilde{Y}$. Earlier, we justified that $k$-means of the instances $X$ is a good heuristic for both objectives.

### 3.2 RANDOM BAGGING

We first state the Matrix Chernoff bound, that we use heavily in this section.

**Lemma 1** (Matrix Chernoff (Corollary 5.2 [Tropp, 2012])). *Consider a finite sequence $\{X_k\}$ of independent, random, self-adjoint matrices that satisfy $X_k \succeq 0$ and $\lambda_{\max}(X_k) \leq R$ almost surely. Compute the minimum and maximum eigenvalues of the sum of expectations, $\mu_{\min} := \lambda_{\min}\left(\sum_k \mathbb{E}X_k\right)$. Then, for $\delta \in [0,1]$*

$$\mathbb{P}\left[\lambda_{\min}\left(\sum\nolimits_k X_k\right) \leq (1-\delta)\mu_{\min}\right] \leq d \cdot \left[\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right]^{\mu_{\min}/R}.$$

**Bag-LLP** We prove the following bound.

**Theorem 4** (Random Bagging Upper Bound, Bag-LLP). *For $\hat{\theta}$ as in (2) and random bagging given by random partitioning into $k$-sized bags,*

$$\mathbb{E}\left[\|\hat{\theta}-\theta^*\|_2^2\right] \leq \frac{16\sigma^2 nk^2}{(1-\delta)^2}\left(\frac{\lambda_{max}(X^TX)}{\lambda_{min}(X^TX)}\right)^2.$$

*w.p. greater than* $1 - d \cdot \left[\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right]^{\frac{\mu_{\min}}{k\beta}}$.

*Proof.* The proof follows from Theorem 2 and Lemmas 2 and 3. □

**Lemma 2** ($\lambda_{max}$ Upper Bound)**.**

$$\lambda_{max}\left(f(X)\right) \leq \lambda_{max}(X^TX).$$

**Lemma 3** ($\lambda_{min}$ Lower Bound)**.**

$$\mathbb{P}\left[\lambda_{min}\left(f(X)\right) > (1-\delta)\frac{\lambda_{min}(X^TX)}{4k^2}\right] \geq 1 - d \cdot \left[\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right]^{\frac{\mu_{\min}}{k\beta}}.$$

*Proof.* Let $X_l$ represent the feature matrices of $B_l$ for $l \in [m]$ We consider the randomized Algorithm 1 which outputs a collection of $m/2$ disjoint bags which are distributed identically to a random subset of $m/2$ disjoint bags, and thus a lower bound for this collection suffices. We have, $\lambda_{min}\left(f(X)\right) = \frac{1}{k^2}\lambda_{min}\left(\sum_{l=1}^m X_l^TX_l\right)$. The feature matrix for bag $B'_l$ sampled using Algorithm 1 can be represented by $X'_l$ for all $l \in [r]$.

> **Algorithm 1: Random Bagging, Bag-LLP**
> **Input:** : Instances $\mathcal{X}$, fixed bag size $k$.
> **Steps:**
> 1. Randomly partition $\mathcal{X}$ into $r$ $2k$-sized *super*-bags, where $r = n/2k$.
>
> $$\mathcal{X} = \cup_{l=1}^r \mathcal{X}_l \text{ and } \mathcal{X}_l \cap \mathcal{X}_{l'} = \phi \text{ for all } l \neq l'$$
>
> 2. For $l = 1, \ldots, r$, a $k$-sized bag $B'_l$ is sampled $u.a.r$ from $\mathcal{X}_l$.
> 3. Output $\mathcal{B}'$ where $\mathcal{B}' = \{B'_l\}_{l \in [r]}$

Figure 1: Algorithm 1: Random Bagging Algorithm, Bag-LLP

$$\frac{1}{k^2}\lambda_{min}\left(\sum_{l=1}^m X_l^TX_l\right) \geq \frac{1}{k^2}\lambda_{min}\left(\sum_{l=1}^r {X'_l}^TX'_l\right) \quad (9)$$

Let $\mu_{min} = \lambda_{min}\left(\sum_{l=1}^r \mathbb{E}\left[{X'_l}^TX'_l\right]\right)/k^2$. We expand ${X'_l}^TX'_l$ and find $\mu_{min}$:

$$\mu_{min} = \frac{1}{k^2}\lambda_{min}\left(\sum_{l=1}^r \mathbb{E}\left[\sum_{x_i,x_j \in B'_l} x_ix_j^T\right]\right)$$
$$= \frac{1}{k^2}\lambda_{min}\left(\sum_{l=1}^r \mathbb{E}\left[\sum_{x_i \in B'_l} x_ix_i^T\right] + \mathbb{E}\left[\sum_{i \neq j} x_ix_j^T\right]\right)$$

In Algorithm 1, $x_i \in \mathcal{X}_l$ get sampled in $B'_l$ with probability $1/2$. Similarly, the probability of sampling the ordered pair

$(x_i, x_j)$ is $2^{2k-2}C_{k-2}/^{2k}C_k = (k-1)/(2k-1)$. Let $\hat{x} = \sum_{x_i \in \mathcal{X}_l} x_i$.

$$\mu_{min} =$$

$$\frac{\lambda_{min}}{k^2}\left(\sum_{l=1}^{r}\sum_{x_i\in\mathcal{X}_l}\frac{1}{2}x_ix_i^T + \sum_{(x_i,x_j)\in\mathcal{X}_l}\frac{k-1}{2k-1}x_ix_j^T\right) =$$

$$\frac{\lambda_{min}}{k^2}\left(\sum_{l=1}^{r}\frac{1}{2}\left(1-\frac{k-1}{2k-1}\right)\sum_{x_i\in\mathcal{X}_l}x_ix_i^T + \frac{k-1}{2(2k-1)}\hat{x}\hat{x}^T\right)$$

$$= \frac{\lambda_{min}}{k^2}\left(\sum_{l=1}^{r}\left(\frac{k}{2(2k-1)}\right)\sum_{x_i\in\mathcal{X}_l}x_ix_i^T + \frac{k-1}{2(2k-1)}\hat{x}\hat{x}^T\right)$$

$$= \frac{\lambda_{min}}{2k^2(2k-1)}\left(kX^TX + (k-1)\sum_{l=1}^{r}\hat{x}\hat{x}^T\right)$$

Since the second term is a summation of $p.s.d$ matrices, we get $\mu_{min} > \lambda_{min}(X^TX)/4k^2$. We assume $\|x\|_2^2 \le \beta$ for all $x \in \mathcal{X}$.

**Lemma 4.** $\lambda_{max}(X_l'^T X_l') \le k\beta$.

Using Lemma 1 and Lemma 4, we get

$$\mathbb{P}\left[\frac{1}{k^2}\lambda_{min}\left(\sum_{l=1}^{m}X_l'^TX_l'\right) \le (1-\delta)\mu_{min}\right] \le d\cdot\left[\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right]^{\frac{\mu_{min}}{k\beta}}$$

Using Equation 9 we get

$$\mathbb{P}\left[\lambda_{min}\left(f(X)\right) > (1-\delta)\frac{\lambda_{min}(X^TX)}{4k^2}\right] \ge 1 - d\cdot\left[\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right]^{\frac{\mu_{\min}}{k\beta}}$$

□

**Aggregate-MIR** We consider a random bagging algorithm similar to the one for Bag-LLP (Algorithm 1) for Aggregate-MIR. The upper bound for Aggregate-MIR (Theorem 3) is product of the label $k$-means objective and the condition number of the bag's instance-centroids. Algorithm 2 takes both these objectives into account. We first sort the instances in increasing order of $\tilde{y}$ and then parition them into contiguous *super-bags* of sizes $2k$. From each super bag, one $k$-sized bag is independently sampled, resulting in a collection of $m/2$ bags. In Theorem 4, we derive an error bound for any arbitrary partitioning of instances into super-bags, and the same bound holds for Algorithm 2. Next, we show that arbitrarily dividing the *super*-bag into two equal sized bags leads to a decrease in the $k$-means objective in Proposition 1.

**Algorithm 2: Random Bagging, Aggregate-MIR**
**Input:** : Instances $\mathcal{X}$, fixed bag size $k$, true labels $\tilde{y}$.
**Steps:**

1. Sort points $\mathcal{X}$ in increasing order of $\tilde{y}$.
2. Partition sorted points into $r$ contiguous *super*-bags of sizes $2k$, where $r = n/2k$.

   $$\mathcal{X} = \cup_{l=1}^{r}\mathcal{X}_l \text{ and } \mathcal{X}_l \cap \mathcal{X}_{l'} = \phi \text{ for all } l \ne l'$$

3. For $l = 1, \ldots, r$, a $k$-sized bag $B_l'$ is sampled $u.a.r$ from $\mathcal{X}_l$.
4. Output $\mathcal{B}'$ where $\mathcal{B}' = \{B_l'\}_{l\in[r]}$

Figure 2: Algorithm 2: Random Bagging Algorithm, Aggregate-MIR

Let $B_l'$ denote a super-bag of size $2k$ for $l \in [r]$ as defined in Algorithm 2. We arbitrarily sample $k$ instances to create a bag $B_l^{(1)}$ and the remaining instances form another bag $B_l^{(2)}$. We know $B_l^{(1)} \cap B_l^{(2)} = \phi$, and $|B_l^{(1)}| = |B_l^{(2)}| = k$.

**Proposition 1** (Optimizing $k$-means in Equation 8)**.** *For super-bags $B_l'$ as defined in Algorithm 2 with arbitrary non-overlapping partitions $B_l^{(1)}$ and $B_l^{(2)}$,*

$$\sum_{l=1}^{r} kmc(\{\tilde{y}_i\}_{i\in B_l'}) \ge \sum_{l=1}^{r} kmc(\{\tilde{y}_i\}_{i\in B_l^{(1)}}) + kmc(\{\tilde{y}_i\}_{i\in B_l^{(2)}})$$

*where kmc($C$) is the $k$-means clustering loss for cluster $C$. kmc($C$) $= \sum_{y_i\in C}(y_i-\mu)^2$, where $\mu$ denotes the mean of cluster $C$.*

We defer the proof to Appendix C.

**Error guarantees of Algorithm 2.** The error for Aggregate MIR, as described in Equation (8) is the the product of the condition number of the bag centroids and a label $k$-means objective. Since analysis of Theorem 4 in Section 3.2 holds for any arbitrary partitioning of instances into *super*-bags, we obtain corresponding bound on the condition number. Proposition 1 shows that the loss of the $k$-bagging will be at most that of the optimal $2k$ clustering.

## 4 PRIVACY

In each of the previous scenarios, the aggregator can modify the bagging procedure to obtain formal label-differential privacy guarantees [Chaudhuri and Hsu, 2011], defined below.

**Definition 1** (Label DP)**.** A randomized algorithm $A$ taking a dataset as an input is $(\varepsilon, \delta)$-label-DP if for two datasets $D$ and $D'$ which differ only on the label of one instance, for any subset $S$ of outputs of $A$,

$$\mathbb{P}[A(D) \in S] \le e^{\varepsilon}\mathbb{P}[A(D') \in S] + \delta.$$

To guarantee label-DP, it is necessary to assume a sensitivity bound on labels, which we achieve by bounding the norm of the labels by a constant $R$. The results below quantifies the additional loss in utility that is incurred due to private bagging in the cases of Instance-MIR, and Bag-LLP. We discuss the corresponding result for Aggregate-MIR in Appendix D, along with the proofs.

**Theorem 5** (Private Error Upper Bound, Instance-MIR). *There exists a bagging $B$ with $|B_l| = k, \forall l \in [m]$, satisfying $(\varepsilon, \delta)$ label-DP, such that for $\hat{\theta}$ in* (1), *we have*

$$\mathbb{E}\left[||\hat{\theta} - \theta^*||_2^2\right] \leq C_1 \left(C_2 + OPT + n\left(1 - \frac{1}{k}\right)\alpha^2 + \frac{d\alpha^2}{k^2}\right),$$

*where $\alpha^2 = \frac{16R^2 \log\left(\frac{1.25}{\delta/2}\right)}{\varepsilon^2}$, $OPT$ is the objective value of the optimal $k$-means clustering over $\tilde{y}$, and constants $C_1, C_2$ are independent of $B$.*

In the label-agnostic setting, one would just need to add noise to the bag-labels. MIR outputs one label at random, hence the sensitivity of the output is $2R$. Due to privacy amplification via subsampling Balle et al. [2018], we add $\mathcal{N}\left(0, \frac{\alpha^2}{k^2}\right)$ noise to the label value to ensure $\left(\frac{\varepsilon}{2}, \frac{\delta}{2}\right)$ label-DP, where $\alpha^2 = \frac{16R^2 \log\left(\frac{1.25}{\delta/2}\right)}{\varepsilon^2}$, leading to an additional error of $\frac{d\alpha^2}{k^2}$. In addition, since the objective here is a label-dependent clustering, we must use a differentially private $k$-means algorithm, leading to additional loss in utility. We show that the simple approach of adding $\mathcal{N}\left(0, \alpha^2\right)$ noise to each label, and then find an optimal clustering over the noise labels, leads to an additional error of $n\left(1 - \frac{1}{k}\right)\alpha^2$. In Appendix D, we discuss how it is possible to achieve better utility, since the above method satisfies the more stringent notion of local-DP, while we only need to satisfy the standard notion of central-DP.

**Theorem 6** (Private Error Upper Bound, Bag-LLP). *There exists a bagging $B$ with $|B_l| = k, \forall l \in [m]$, satisfying $(\varepsilon, \delta)$ label-DP, such that for $\hat{\theta}$ in* (2), *we have*

$$\mathbb{E}\left[||\hat{\theta} - \theta^*||_2^2\right] = OPT\left(\frac{d}{k}\alpha^2 + \sigma^2\frac{m}{k}\right),$$

*where $\alpha^2 = \frac{4R^2 \log\left(\frac{1.25}{\delta}\right)}{\varepsilon^2}$, and $OPT$ is the optimal value of $\left(\frac{\lambda_{max}(f(X))}{\lambda_{min}(f(X))}\right)^2$.*

In this case, the optimal bagging strategy in independent of the labels. Hence, one just needs to add noise to the bag-labels, and not add noise for a private clustering of the labels. LLP outputs the mean of $k$ labels, hence the sensitivity of the output is $\frac{2R}{k}$. We add $\mathcal{N}\left(0, \frac{\alpha^2}{k^2}\right)$ noise to the label value to ensure $(\varepsilon, \delta)$ label-DP, leading to an additional error of $\frac{\alpha^2 m}{k^2}$ over the corresponding non-private bagging mechanism.

| $k$ | Bagging Method | $\|\hat{\theta} - \theta^*\|_2^2$ |
|---|---|---|
| *LLP* | *Bag Loss* | |
| 10 | Instance $k$-means | $0.0082 \pm 0.002$ |
| | Label $k$-means | $0.0458 \pm 0.012$ |
| | Random | $0.0099 \pm 0.002$ |
| 50 | Instance $k$-means | $0.0392 \pm 0.008$ |
| | Label $k$-means | $0.0629 \pm 0.008$ |
| | Random | $0.0423 \pm 0.009$ |
| *MIR* | *Instance Loss* | |
| 10 | Instance $k$-means | $0.0088 \pm 0.002$ |
| | Label $k$-means | $0.0072 \pm 0.002$ |
| | Random | $0.0085 \pm 0.002$ |
| 50 | Instance $k$-means | $0.0388 \pm 0.006$ |
| | Label $k$-means | $0.0404 \pm 0.007$ |
| | Random | $0.0419 \pm 0.006$ |
| *MIR* | *Aggregate Loss* | |
| 10 | Instance $k$-means | $0.0102 \pm 0.002$ |
| | Label $k$-means | $0.0453 \pm 0.008$ |
| | Random | $0.0221 \pm 0.004$ |
| 50 | Instance $k$-means | $0.0437 \pm 0.008$ |
| | Label $k$-means | $0.0601 \pm 0.008$ |
| | Random | $0.0619 \pm 0.012$ |

Table 1: Non-Private Bagging

# 5 EXPERIMENTS

We conduct experiments on synthetically generated data. The synthetic dataset is of the form $(X \in \mathbb{R}^{n \times d}, Y \in \mathbb{R}^n)$ and is generated by first sampling a random ground truth model $\theta^*$ from the standard $d$-dimensional Gaussian distribution, sampling each of the rows of $X$ i.i.d. from the standard $d$-dimensional Gaussian distribution, and then setting $Y = X\theta^* + \gamma$ where each coordinate of $\gamma$ is i.i.d. drawn from $N(0, \sigma^2)$ where $\sigma$ is 0.5. We set $n$ to be $50,000$ and $d$ as 32. We also vary $k$, and use $k = 10, 50$.

We implement 3 bagging mechanisms on each of Instance-MIR, Aggregate-MIR, and Bag-LLP, namely (1) Instance $k$-means, (2) Label $k$-means, and (3) Random bagging. In Table 1, we present the mean and standard deviation of the error, calculated over 15 runs for each experiment. As expected, for Bag-LLP, instance $k$-means performs better than random bagging, which in turn performs better than label $k$-means. For Aggregate-MIR, instance $k$-means consistently performs the best, which is expected, while random bagging overall performs slightly better than label $k$-means. However, for Instance-MIR, all the 3 mechanisms show similar performance. We compute statistical significance of our results using the paired $T$-value test in Appendix E.1.

We also consider the private version of Instance-MIR in Table 2. We set $\delta = 10^{-5}$, and vary $\varepsilon$. For each mechanism, we see that accuracy drops with a decrease in $\varepsilon$. However, the drop is sharper for label $k$-means, which is expected,

| $k$ | Bagging Method | $\varepsilon$ | $\|\hat{\theta} - \theta^*\|_2^2$ |
|---|---|---|---|
| 10 | Instance $k$-means | 0.5 | $0.0621 \pm 0.009$ |
| | | 1.0 | $0.0537 \pm 0.009$ |
| | | 2.0 | $0.0390 \pm 0.008$ |
| | Label $k$-means | 0.5 | $0.0505 \pm 0.005$ |
| | | 1.0 | $0.0362 \pm 0.006$ |
| | | 2.0 | $0.0189 \pm 0.004$ |
| 50 | Instance $k$-means | 0.5 | $0.0656 \pm 0.012$ |
| | | 1.0 | $0.0595 \pm 0.012$ |
| | | 2.0 | $0.0521 \pm 0.009$ |
| | Label $k$-means | 0.5 | $0.0559 \pm 0.008$ |
| | | 1.0 | $0.0480 \pm 0.005$ |
| | | 2.0 | $0.0431 \pm 0.006$ |

Table 2: Private Bagging, Instance-MIR

since unlike feature $k$-means, it is label-dependent, incurring an extra utility error. We also note that that drop in accuracy is sharper for a smaller bag size; this is again expected since the error due to privacy scales with $\frac{1}{k}$.

We also consider non-isotropic distributions. We generate $X$ i.i.d. from $\mathcal{N}(0, \Sigma)$, where $\Sigma$ is determined by sampling $d$ independent values $\{\lambda_1, \cdots, \lambda_d\}$ from a uniform distribution $U(0.1, 10)$ to be the eigenvalues of the $\Sigma$, which is diagonal matrix. We also consider the case where the columns of $\Sigma$ are non-independent. We sample each entry of a Cholesky matrix $M$ of size $d \times d$ from $\mathcal{N}(0, 1)$. We then compute the covariance matrix $M^T M$ and apply a linear transformation to feature vectors $x$ sampled from $\mathcal{N}(0, I)$ using $M$. The resulting set of vectors is non-isotropic with correlated features.

We implement 4 bagging mechanisms on each of Instance-MIR, Aggregate-MIR, and Bag-LLP, namely (1) Instance $k$-means, (2) Label $k$-means, (3) Random bagging, and (4) Scaled Instance $k$-means, that scales the dataset $X$ as $\Sigma^{-\frac{1}{2}} X$ to be isotropic, and then finds an optimal $k$-means clustering on the scaled dataset. In the tables, we present the mean and standard deviation of the error, calculated over 15 runs for each experiment. As expected, in most cases for Bag-LLP (Table 5) and Aggregate-MIR (Table 4), scaled instance $k$-means performs better than instance $k$-means, which in turn performs better than random bagging, which in turn performs better than label $k$-means. However, for Instance-MIR (Table 6), all the mechanisms show similar performance, with label $k$-means showing better performance in many cases. We show the Bag-LLP setup in Table 3. The complete tables (Table 5, Table 4, Table 6) are deferred to the Appendix E.

| $k$ | Bagging Method | $\|\hat{\theta} - \theta^*\|_2^2$ |
|---|---|---|
| | *Independent* | |
| 10 | Scaled Instance $k$-means | $0.008552 \pm 0.00191$ |
| | Instance $k$-means | $0.009739 \pm 0.00201$ |
| | Random | $0.010518 \pm 0.00339$ |
| | Label $k$-means | $0.042496 \pm 0.00626$ |
| 50 | Scaled Instance $k$-means | $0.038586 \pm 0.00784$ |
| | Instance $k$-means | $0.036923 \pm 0.00536$ |
| | Random | $0.039461 \pm 0.00760$ |
| | Label $k$-means | $0.059834 \pm 0.00598$ |
| | *Non-independent* | |
| 10 | Scaled Instance $k$-means | $0.024811 \pm 0.00498$ |
| | Instance $k$-means | $0.032367 \pm 0.00835$ |
| | Random | $0.024585 \pm 0.00755$ |
| | Label $k$-means | $0.052438 \pm 0.00936$ |
| 50 | Scaled Instance $k$-means | $0.049910 \pm 0.00773$ |
| | Instance $k$-means | $0.051425 \pm 0.00895$ |
| | Random | $0.048222 \pm 0.01074$ |
| | Label $k$-means | $0.061918 \pm 0.00820$ |

Table 3: Non-Isotropic Distribution, Bag-LLP

## 6 CONCLUSION

In this paper, we study for various loss functions in MIR and LLP, what is the optimal way to partition the dataset into bags such that the utility for downstream tasks like linear regression is maximized. We theoretically provide utility guarantees, and show that in each case, the optimal bagging strategy (approximately) reduces to finding an optimal $k$-means clustering of the feature vectors or the labels. We also show that our bagging mechanisms can be made label-DP, incurring an additional utility error. We finally generalize our results to the setting of GLMs.

There are several potential directions for future work. While we only considered linear models, it would be interesting to analyse optimal bagging strategies in non-linear models, such as neural networks. One could also consider other popular loss functions for MIR and LLP used in literature. While our work only looked at upper bounds, having corresponding lower bounds would also be interesting.

### References


Ehsan Mohammady Ardehaly and Aron Culotta. Co-training for demographic classification using deep learning from label proportions. In *ICDM*, pages 1017–1024, 2017.

Borja Balle, Gilles Barthe, and Marco Gaboardi. Privacy amplification by subsampling: Tight analyses via couplings and divergences, 2018. URL `https://arxiv.org/abs/1807.01647`.

Björn Bebensee. Local differential privacy: a tutorial, 2019. URL `https://arxiv.org/abs/1907.11908`.

Anand Paresh Brahmbhatt, Rishi Saket, and Aravindan Raghuveer. PAC learning linear thresholds from label proportions. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=5Gw9YkJkFF.

Robert Istvan Busa-Fekete, Heejin Choi, Travis Dick, Claudio Gentile, and Andres Munoz medina. Easy learning from label proportions. *arXiv*, 2023. URL https://arxiv.org/abs/2302.03115.

Kamalika Chaudhuri and Daniel Hsu. Sample complexity bounds for differentially private learning. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 155–186. JMLR Workshop and Conference Proceedings, 2011.

Kushal Chauhan, Rishi Saket, Lorne Applebaum, Ashwinkumar Badanidiyuru, Chandan Giri, and Aravindan Raghuveer. Generalization and learnability in multiple instance regression. In *UAI*, 2024.

L. Chen, Z. Huang, and R. Ramakrishnan. Cost-based labeling of groups of mass spectra. In *Proc. ACM SIGMOD International Conference on Management of Data*, pages 167–178, 2004.

Lin Chen, Thomas Fu, Amin Karbasi, and Vahab Mirrokni. Learning from aggregated data: Curated bags versus random bags. *arXiv*, 2023. URL https://arxiv.org/abs/2305.09557.

N. de Freitas and H. Kück. Learning about individuals from group statistics. In *Proc. UAI*, pages 332–339, 2005.

L. M. Dery, B. Nachman, F. Rubbo, and A. Schwartzman. Weakly supervised classification in high energy physics. *Journal of High Energy Physics*, 2017(5):1–11, 2017.

Adel Javanmard, Matthew Fahrbach, and Vahab Mirrokni. Priorboost: An adaptive algorithm for learning from aggregate responses, 2024. URL https://arxiv.org/abs/2402.04987.

D. Kotzias, M. Denil, N. de Freitas, and P. Smyth. From group to individual labels using deep features. In *Proc. SIGKDD*, pages 597–606, 2015.

J. Liu, B. Wang, Z. Qi, Y. Tian, and Y. Shi. Learning from label proportions with generative adversarial networks. In *Proc. NeurIPS*, pages 7167–7177, 2019.

Zhigang Lu and Hong Shen. Differentially private k-means clustering with convergence guarantee. *IEEE Transactions on Dependable and Secure Computing*, page 1–1, 2020. ISSN 2160-9209. doi: 10.1109/tdsc.2020.3043369. URL http://dx.doi.org/10.1109/TDSC.2020.3043369.

Conor O'Brien, Arvind Thiagarajan, Sourav Das, Rafael Barreto, Chetan Verma, Tim Hsu, James Neufield, and Jonathan J Hunt. Challenges and approaches to privacy preserving post-click conversion prediction. *arXiv preprint arXiv:2201.12666*, 2022.

N. Quadrianto, A. J. Smola, T. S. Caetano, and Q. V. Le. Estimating labels from label proportions. *J. Mach. Learn. Res.*, 10:2349–2374, 2009.

S. Ray and D. Page. Multiple instance regression. In *Proc. ICML*, pages 425–432, 2001.

Soumya Ray and Mark Craven. Supervised versus multiple instance learning: an empirical comparison. In *Proc. ICML*, page 697–704, 2005.

S. Rueping. SVM classifier estimation from group probabilities. In *Proc. ICML*, pages 911–918, 2010.

R. Saket. Learnability of linear thresholds from label proportions. In *Proc. NeurIPS*, 2021. URL https://openreview.net/forum?id=5BnaKeEwuYk.

R. Saket. Algorithms and hardness for learning linear thresholds from label proportions. In *Proc. NeurIPS*, 2022. URL https://openreview.net/forum?id=4LZo68TuF-4.

Rishi Saket, Aravindan Raghuveer, and Balaraman Ravindran. On combining bags to better learn from label proportions. In *AISTATS*, volume 151 of *Proceedings of Machine Learning Research*, pages 5913–5927. PMLR, 2022. URL https://proceedings.mlr.press/v151/saket22a.html.

C. Scott and J. Zhang. Learning from label proportions: A mutual contamination framework. In *Proc. NeurIPS*, 2020.

Thomas Steinke. Composition of differential privacy & privacy amplification by subsampling, 2022. URL https://arxiv.org/abs/2210.00597.

Dong Su, Jianneng Cao, Ninghui Li, Elisa Bertino, and Hongxia Jin. Differentially private $k$-means clustering, 2015. URL https://arxiv.org/abs/1504.05998.

Mohamed Trabelsi and Hichem Frigui. Fuzzy and possibilistic clustering for multiple instance linear regression. In *2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–7, 2018.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12:389–434, 2012.

Jesper E Van Engelen and Holger H Hoos. A survey on semi-supervised learning. *Machine learning*, 109(2):373–440, 2020.

K. L. Wagstaff, T. Lane, and A. Roper. Multiple-instance regression with structured data. In *Workshops Proceedings of the 8th IEEE ICDM*, pages 291–300, 2008.

Z. Wang, V. Radosavljevic, B. Han, Z. Obradovic, and S. Vucetic. *Aerosol Optical Depth Prediction from Satellite Observations by Multiple Instance Regression*, pages 165–176. 2008.

Z. Wang, L. Lan, and S. Vucetic. Mixture model for multiple instance regression and applications in remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 50(6):2226–2237, 2012.

F. X. Yu, D. Liu, S. Kumar, T. Jebara, and S. F. Chang. $\propto$SVM for learning with label proportions. In *Proc. ICML*, volume 28, pages 504–512, 2013.

F. X. Yu, K. Choromanski, S. Kumar, T. Jebara, and S. F. Chang. On learning from label proportions. *CoRR*, abs/1402.5902, 2014. URL `http://arxiv.org/abs/1402.5902`.

# Aggregating Data for Optimal Learning
# (Supplementary Material)

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

[1]Northeastern University
[2]Google Deepmind

## A UTILITY ANALYSIS

### A.1 MIR, INSTANCE-LEVEL LOSS

We denote the uniform distribution by $\Gamma$. Let $\overline{y} = [\overline{y}_1, \ldots, \overline{y}_m]$, where $\overline{y}_l = y_{\Gamma(B_l)}$. We define a random attribution matrix for MIR, $A \in \{0,1\}^{n \times n}$, as follows.

$$A_{(i,j)} = \begin{cases} 1 & \text{if } i \in B_l \text{ and } \overline{y}_l = y_j \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

Note that $\mathbb{E}[A] = S = S^T$ is given by

$$S_{(i,j)} = \begin{cases} \frac{1}{|B_l|} & \text{if } i, j \in B_l \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

The minimizer of (1) is then given by

$$\hat{\theta} = \operatorname*{argmin}_{\theta} \frac{1}{n} \|Ay - X\theta\|_2^2 = (X^T X)^{-1} X^T A \mathbf{y}. \tag{12}$$

We now give a proof sketch for Theorem 1, providing an upper bound for the error of $\hat{\theta}$ (some details are omitted to Appendix A. All the expectations henceforth are over the randomness in $A$ unless otherwise stated.

*Proof.* (of Theorem 1) We begin with the following proposition, and use it to prove the main theorem

**Proposition 2.**

$$\mathbb{E}\left[||\hat{\theta} - \theta^*||_2^2\right] = \mathbb{E}\left[||(X^T X)^{-1} X^T (A - I) X \theta^*||_2^2\right] + \sigma^2 \mathbb{E}\left[||(X^T X)^{-1} X^T A||_F^2\right].$$

*Proof.* (of Proposition 2) By rearranging the terms,

$$\begin{aligned} \hat{\theta} - \theta^* &= (X^T X)^{-1} X^T A \mathbf{y} - \theta^* \\ &= (X^T X)^{-1} X^T A X \theta^* - \theta^* + (X^T X)^{-1} X^T A \gamma \\ &= (X^T X)^{-1} X^T (A - I) X \theta^* + (X^T X)^{-1} X^T A \gamma. \end{aligned}$$

$\gamma$ is independent of $A$ with $\mathbb{E}[\gamma] = 0$, $\mathbb{E}[\gamma\gamma^T] = \sigma^2 I$ and $\mathbb{E}[A] = S$. Using this we get,

$$\begin{aligned}\mathbb{E}\left[||\hat{\theta} - \theta^*||^2\right] &= \mathbb{E}\left[||(X^TX)^{-1}X^T(A-I)X\theta^*||_2^2\right] + \mathbb{E}\left[\mathrm{tr}((X^TX)^{-1}X^TA\gamma\gamma^TA^TX(X^TX)^{-1})\right] \\ &= \mathbb{E}\left[||(X^TX)^{-1}X^T(A-I)X\theta^*||_2^2\right] + \sigma^2\mathbb{E}\left[\mathrm{tr}((X^TX)^{-1}X^TAA^TX(X^TX)^{-1})\right] \\ &= \mathbb{E}\left[||(X^TX)^{-1}X^T(A-I)X\theta^*||_2^2\right] + \sigma^2\mathbb{E}\left[||(X^TX)^{-1}X^TA||_F^2\right]\end{aligned}$$

□

We now upper bound the error in Proposition 2. We simplify the first term.

$$\begin{aligned}\mathbb{E}\left[||(X^TX)^{-1}X^T(A-I)X\theta^*||_2^2\right] &\leq \mathbb{E}\left[||(X^TX)^{-1}X^T||_{op}||(A-I)X\theta^*||_2^2\right] \\ &= ||(X^TX)^{-1}X^T||_{op}^2\mathbb{E}\left[||(A-I)X\theta^*||_2^2\right]\end{aligned}$$

We simplify the RHS above with the following proposition.

**Proposition 3.**

$$\mathbb{E}\left[||(A-I)X\theta^*||_2^2\right] = \left(2||\tilde{y}||_2^2 - 2\sum_{l=1}^{m}\frac{\left(\sum_{i\in B_l}\tilde{y}_i\right)^2}{|B_l|}\right)$$

*Proof.*

$$\begin{aligned}&\mathbb{E}\left[||(A-I)X\theta^*||_2^2\right] \\ &= \mathbb{E}\left[((A-I)X\theta^*)^T(A-I)X\theta^*\right] \\ &= \mathbb{E}\left[\theta^{*T}X^TA^TAX\theta^*\right] - \mathbb{E}\left[\theta^{*T}X^T(A+A^T)X\theta^*\right] + ||X\theta^*||_2^2 \\ &= \mathbb{E}\left[||A\tilde{y}||_2^2\right] - \theta^{*T}X^T(S+S^T)X\theta^* + ||X\theta^*||_2^2 \\ &= \mathbb{E}\left[||AX\theta^*||_2^2\right] - 2\theta^{*T}X^TSX\theta^* + ||\tilde{y}||_2^2\end{aligned}$$

Putting the following two lemmas together, we conclude Proposition 3.

**Lemma 5.** $\mathbb{E}\left[||AX\theta^*||_2^2\right] = ||\tilde{y}||_2^2$.

*Proof.* (of Lemma 5) Let $B(i)$ be the bag containing $x_i$. Note that $AX\theta^* = \left[\tilde{y}_{\Gamma(B(1))}, \ldots, \tilde{y}_{\Gamma(B(n))}\right]^T$

$$\theta^{*T}X^TA^TAX\theta^* = \sum_{i=1}^{i=n}\tilde{y}_{\Gamma(B(i))}^2$$

Then we have

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^{i=n}\tilde{y}_{\Gamma(B(i))}^2\right] &= \sum_{i=1}^{i=n}\left(\sum_{j\in B(i)}\frac{(\tilde{y}_j)^2}{|(B(i))|}\right) \\ &= \sum_{l=1}^{l=m}|B_l|\left(\sum_{j\in B(i)}\frac{(\tilde{y}_j)^2}{|B_l|}\right) \\ &= \sum_{i=1}^{n}(\tilde{y}_i)^2\end{aligned}$$

□

**Lemma 6.** $\theta^{*T}X^TSX\theta^* = \sum_{l=1}^{m}\frac{\left(\sum_{i\in B_l}\tilde{y}_i\right)^2}{|B_l|}$.

*Proof.* (of Lemma 6). Note that $S = M^T M$, where $M \in \mathbb{R}^{m \times n}$ is defined as:

$$M_{(i,j)} = \begin{cases} 1/\sqrt{|B_i|} & \text{if } x_j \in B_i \\ 0 & \text{otherwise.} \end{cases}$$

Thus, $\theta^{*T} X^T S X \theta^* = \theta^{*T} X^T M^T M X \theta^* = ||M\tilde{y}||_2^2$.

$$\begin{aligned} ||M\tilde{y}||_2^2 &= \sum_{l=1}^{m} \left( \sum_{x_i \in B_l} \frac{1}{\sqrt{|B_l|}} \tilde{y}_i \right)^2 \\ &= \sum_{l=1}^{m} \frac{1}{|B_l|} \left( \sum_{x_i \in B_l} \tilde{y}_i \right)^2 \end{aligned}$$

□

□

The following proposition analyses the second term in Proposition 2, and together with Proposition 3 concludes the proof of Theorem 1.

**Proposition 4.**

$$\mathbb{E}\left[||(X^T X)^{-1} X^T A||_F^2\right] \leq d||(X^T X)^{-1} X^T||_{op}^2$$

*Proof.* (of Proposition 4). We use the following inequality:

$$||AB||_F^2 \leq \min\left(||A||_{op}^2 ||B||_F^2, ||B||_{op}^2 ||A||_F^2\right) .$$

$$\mathbb{E}\left[||(X^T X)^{-1} X^T A||_F^2\right] \leq \min\left(\mathbb{E}\left[||(X^T X)^{-1} X^T||_{op}^2 ||A||_F^2\right], \mathbb{E}\left[||(X^T X)^{-1} X^T||_F^2 ||A||_{op}^2\right]\right)$$

We assumed $\text{rank}(X) = d$, hence $||(X^T X)^{-1} X^T||_F \leq \sqrt{d}||(X^T X)^{-1} X^T||_{op}$ .

$$\begin{aligned} \mathbb{E}\left[||(X^T X)^{-1} X^T A||_F^2\right] &\leq \min\left(\mathbb{E}\left[||(X^T X)^{-1} X^T||_{op}^2 ||A||_F^2\right], \mathbb{E}\left[d||(X^T X)^{-1} X^T||_{op}^2 ||A||_{op}^2\right]\right) \\ &= ||(X^T X)^{-1} X^T||_{op}^2 \min\left(\mathbb{E}\left[||A||_F^2\right], d\mathbb{E}\left[||A||_{op}^2\right]\right) \end{aligned}$$

We have $\mathbb{E}\left[||A||_F^2\right] = n$ and $\mathbb{E}\left[||A||_{op}^2\right] = 1$. Also, we are in the setting where $n > d$ to have a well defined regressor. Therefore, we obtain

$$\mathbb{E}\left[||(X^T X)^{-1} X^T A||_F^2\right] \leq d||(X^T X)^{-1} X^T||_{op}^2$$

□

□

### A.2 LLP, BAG-LEVEL LOSS

We define a bagging matrix $S \in \{0, 1\}^{m \times n}$ that encodes the assignment of instances to bags.

$$S_{(l,i)} = \begin{cases} \frac{1}{|B_l|} & \text{if } i \in B_l, \\ 0 & \text{otherwise.} \end{cases} \tag{13}$$

The minimizer of the bag-level loss in matrix form is

$$\hat{\theta} = \underset{\theta}{\text{argmin}} \frac{1}{m} \|S\mathbf{y} - SX\theta\|_2^2. \tag{14}$$

**Theorem 7** (full version of Theorem 2)**.** *For $\hat{\theta}$ as in* (2)*, for a given bagging $B$ with bagging matrix $S$, we have*

$$\mathbb{E}\left[\|\hat{\theta}-\theta^*\|_2^2\right] \leq \sigma^2 \left(\frac{\lambda_{max}((SX)^T SX)}{\lambda_{min}((SX)^T SX)}\right)^2 \left(\sum_{l=1}^{m} \frac{1}{|B_l|}\right)$$

*For equal sized bags of size $k$, this simplifies to*

$$\mathbb{E}\left[\|\hat{\theta}-\theta^*\|_2^2\right] \leq \sigma^2 \frac{m}{k} \left(\frac{\lambda_{max}((SX)^T SX)^{-1}}{\lambda_{min}((SX)^T SX)^{-1}}\right)^2.$$

*Proof.* We start by proving the following lemma

**Lemma 7.**

$$\mathbb{E}\left[\|\hat{\theta}-\theta^*\|_2^2\right] = \sigma^2 \|((SX)^T SX)^{-1}(SX)^T (SS^T)^{1/2}\|_F^2. \tag{15}$$

*Proof.* The minimizer of the bag-level loss in matrix form is

$$\begin{aligned}\hat{\theta} &= \underset{\theta}{\operatorname{argmin}} \frac{1}{m}\|S\mathbf{y} - SX\theta\|_2^2 \\ &= (X^T S^T SX)^{-1} X^T S^T Sy.\end{aligned}$$

By rearranging the terms, we have

$$\begin{aligned}\hat{\theta}-\theta^* &= ((SX)^T SX)^{-1} X^T S^T S\mathbf{y} - \theta^* \\ &= ((SX)^T SX)^{-1} X^T S^T SX\theta^* - \theta^* \\ &+ ((SX)^T SX)^{-1} X^T S^T S\gamma \\ &= ((SX)^T SX)^{-1} X^T S^T S\gamma\end{aligned}$$

Since $\gamma$ is independent of $X$l, with $\mathbb{E}[\gamma] = 0$, and $\mathbb{E}[\gamma\gamma^T] = \sigma^2\mathcal{I}$, we have

$$\mathbb{E}\left[\|\hat{\theta}-\theta^*\|_2^2\right] = \sigma^2 tr(((SX)^T SX)^{-1}(SX)^T SS^T (SX)((SX)^T SX)^{-1})$$

By definition, $SS^T = \text{Diag}(\{\frac{1}{|B_1|}, \frac{1}{|B_2|}, \ldots, \frac{1}{|B_m|}\})$ and the expression simplifies to give:

$$\mathbb{E}\left[\|\hat{\theta}-\theta^*\|_2^2\right] = \sigma^2 \|((SX)^T SX)^{-1}(SX)^T (SS^T)^{1/2}\|_F^2$$

□

Now we upper bound the RHS.

$$\begin{aligned}\mathbb{E}\left[\|\hat{\theta}-\theta^*\|_2^2\right] &= \sigma^2 \|((SX)^T SX)^{-1}(SX)^T (SS^T)^{1/2}\|_F^2 \\ &\leq \sigma^2 \|((SX)^T SX)^{-1}(SX)^T\|_{op}^2 \|(SS^T)^{1/2}\|_F^2 \\ &= \sigma^2 \|((SX)^T SX)^{-1}(SX)^T\|_{op}^2 \left(\sum_{l=1}^{m} \frac{1}{|B_l|}\right) \\ &\leq \sigma^2 \|((SX)^T SX)^{-1}\|_{op}^2 \|(SX)^T\|_{op}^2 \left(\sum_{l=1}^{m} \frac{1}{|B_l|}\right) \\ &\leq \sigma^2 \left(\frac{\lambda_{max}((SX)^T SX)}{\lambda_{min}((SX)^T SX)}\right)^2 \left(\sum_{l=1}^{m} \frac{1}{|B_l|}\right)\end{aligned}$$

□

### A.3 MIR, AGGREGATE-LEVEL LOSS

We define a random attribution matrix $A \in \{0,1\}^{m\times n}$ as follows, to indicate the bag-label of each bag.

$$A_{(l,i)} = \begin{cases} 1 & \text{if } y_i = \Gamma(B_l), \\ 0 & \text{otherwise.} \end{cases} \tag{16}$$

We denote $\mathbb{E}[A] = S$. This turns out to be the same S as (13), and represents the instances in each bag. The minimizer of the aggregate-level loss is

$$\hat{\theta} = \operatorname*{argmin}_{\theta} \frac{1}{m} \|A\mathbf{y} - SX\theta\|_2^2. \tag{17}$$

**Theorem 8** (full version of Theorem 3)**.** *For $\hat{\theta}$ in* (3)*, given a bagging $B$ with bagging matrix $S$,*

$$\mathbb{E}\left[\|\hat{\theta}-\theta^*\|_2^2\right] \le \|((SX)^T SX)^{-1}(SX)^T\|_{op}^2 \left(\sum_{l=1}^{m}\left(\frac{\sum_{i\in B_l}\tilde{y}_i^2}{|B_l|}\right) - \sum_{l=1}^{m}\left(\frac{\sum_{i\in B_l}\tilde{y}_i}{|B_l|}\right)^2 + \sigma^2 n\right)$$

*For equal sized bags, this simplifies to*

$$\mathbb{E}\left[\|\hat{\theta}-\theta^*\|_2^2\right] \le \frac{1}{k}\|((SX)^T SX)^{-1}(SX)^T\|_{op}^2 \left(\sum_{l=1}^{m}\sum_{\tilde{y}_i\in B_l}(\tilde{y}_i-\mu_l)^2 + \sigma^2 nk\right),$$

*Proof.*

$$\begin{aligned}\hat{\theta} &= \operatorname*{argmin}_{\theta} \frac{1}{m}\|A\mathbf{y} - SX\theta\|_2^2 \\ &= (X^T S^T SX)^{-1} X^T S^T A\mathbf{y}.\end{aligned}$$

By rearranging the terms, we have

$$\begin{aligned}\hat{\theta}-\theta^* &= ((SX)^T SX)^{-1}X^T S^T A\mathbf{y} - \theta^* \\ &= ((SX)^T SX)^{-1}X^T S^T AX\theta^* - \theta^* + ((SX)^T SX)^{-1}X^T S^T A\gamma\end{aligned}$$

$\gamma$ is independent of $X$ with $\mathbb{E}[\gamma]=0$ and $\mathbb{E}[\gamma\gamma^T] = \sigma^2\mathcal{I}$. Also, $\mathbb{E}[A]=S$, and $\gamma, A$ are independent. Hence,

$$\begin{aligned}\mathbb{E}\left[\|\hat{\theta}-\theta^*\|_2^2\right] &= \mathbb{E}\left[\|((SX)^T SX)^{-1}(SX)^T AX\theta^* - ((SX)^T SX)^{-1}(SX)^T SX\theta^* + ((SX)^T SX)^{-1}X^T S^T A\gamma\|_2^2\right] \\ &\le \|((SX)^T SX)^{-1}(SX)^T\|_{op}^2 \mathbb{E}[\|(AX\theta^* - SX\theta^*) + A\gamma\|_2^2] \\ &\le \|((SX)^T SX)^{-1}(SX)^T\|_{op}^2 \left(\mathbb{E}[\|AX\theta^* - SX\theta^*\|_2^2] + \mathbb{E}[\|A\gamma\|_2^2]\right) \\ &\le \|((SX)^T SX)^{-1}(SX)^T\|_{op}^2 \left(\mathbb{E}[\|A\tilde{y} - S\tilde{y}\|_2^2] + \mathbb{E}[\|A\gamma\|_2^2]\right)\end{aligned}$$

We now analyse $\mathbb{E}[\|A\tilde{y} - S\tilde{y}\|_2^2]$ in the lemma below.

**Lemma 8.**

$$\mathbb{E}[\|A\tilde{y} - S\tilde{y}\|_2^2] = \sum_{l=1}^{m}\left(\frac{\sum_{i\in B_l}\tilde{y}_i^2}{|B_l|}\right) - \sum_{l=1}^{m}\left(\frac{\sum_{i\in B_l}\tilde{y}_i}{|B_l|}\right)^2$$

*Proof.*

$$
\begin{aligned}
\mathbb{E}[\|A\tilde{y} - S\tilde{y}\|_2^2] &= \mathbb{E}[(A\tilde{y} - S\tilde{y})^T(A\tilde{y} - S\tilde{y})] \\
&= \mathbb{E}[||A\tilde{y}||^2 + ||S\tilde{y}||^2 - 2\tilde{y}^T S^T A\tilde{y}] \\
&= \mathbb{E}[||A\tilde{y}||^2] + \mathbb{E}[||S\tilde{y}||^2] - 2\mathbb{E}[\tilde{y}^T S^T A\tilde{y}] \\
&= \mathbb{E}[||A\tilde{y}||^2] + \mathbb{E}[||S\tilde{y}||^2] - 2\mathbb{E}[\tilde{y}^T S^T S y] \\
&= \mathbb{E}[||A\tilde{y}||^2] + \mathbb{E}[||S\tilde{y}||^2] - 2\mathbb{E}[||S\tilde{y}||^2] \\
&= \mathbb{E}[||A\tilde{y}||^2] - \mathbb{E}[||S\tilde{y}||^2] \\
&= \mathbb{E}[||A\tilde{y}||^2] - ||S\tilde{y}||^2
\end{aligned}
$$

We now analyse $\mathbb{E}[||A\tilde{y}||^2]$

$$
\begin{aligned}
A\tilde{y} &= \left[\tilde{y}_{\Gamma(B_1)}, \ldots, \tilde{y}_{\Gamma(B_m)}\right]^T \\
\implies \tilde{y}^T A^T A\tilde{y} &= \sum_{l=1}^{l=m} \tilde{y}^2_{\Gamma(B_l)}
\end{aligned}
$$

Then we have

$$
\begin{aligned}
\mathbb{E}\left[\tilde{y}^T A^T A\tilde{y}\right] &= \mathbb{E}\left[\sum_{l=1}^{l=m} \tilde{y}^2_{\Gamma(B_l)}\right] \\
&= \sum_{l=1}^{m} \left(\frac{\sum_{i\in B_l} \tilde{y}_i^2}{|B_l|}\right)
\end{aligned}
$$

For equal size bags it simplifies to $\frac{||\tilde{y}||^2}{k}$. We now analyse Term 2 $||S\tilde{y}||^2$

$$
\begin{aligned}
S\tilde{y} &= \left[\frac{\sum_{i\in B_1} \tilde{y}_i}{|B_1|}, \ldots, \frac{\sum_{i\in B_m} \tilde{y}_i}{|B_m|}\right]^T \\
\implies \tilde{y}^T S^T S\tilde{y} &= \sum_{l=1}^{m} \left(\frac{\sum_{i\in B_l} \tilde{y}_i}{|B_l|}\right)^2
\end{aligned}
$$

For equal size bags this simplifies to $\sum_{l=1}^{m} \left(\frac{\sum_{i\in B_l} \tilde{y}_i}{k}\right)^2$. □

It is easy to see that $\mathbb{E}[\|A\gamma\|_2^2] = n\sigma^2$. Combining this with the above lemma, we are done.

□

# B MISSING PROOFS

In this section, we present the missing proofs from the paper, along with some additional results that were briefly mentioned in the main paper.

## B.1 ADDITIONAL RESULTS FROM SECTION 2

Lemma 9 shows that finding the optimal $k$-means clustering of the (expected) labels $\tilde{y}$ exactly maximizes $\sum_{\ell=1}^{m} \frac{\left(\sum_{i\in B_\ell} \tilde{y}_i\right)^2}{|B_\ell|}$. Lemma 10 shows that clustering over $y = \tilde{y} + \gamma$ as a proxy for clustering over $\tilde{y}$ leads to an additional utility error of $\left(1 - \frac{1}{k}\right)\sigma^2 n$. Lemma 11 shows that the $1d$ clustering problem above turns out to result in a bagging that just sorts the labels in order, and partitions contiguous segments into bags.

**Lemma 9** ($k$-means Equivalence)**.** *Maximizing* $\sum_{\ell=1}^{m} \frac{\left(\sum_{i\in B_\ell} \tilde{y}_i\right)^2}{|B_\ell|}$ *corresponds to finding the optimal $k$-means clustering over $\tilde{y}$.*

*Proof.* The $k$-means objective for a bagging $B$ over $\tilde{y}$ is

$$\sum_{l=1}^{m} \sum_{i \in B_l} (\tilde{y}_i - \mu_l)^2 \,,$$

where $\mu_l = \frac{1}{|B_l|} \sum_{i \in B_l} \tilde{y}_i$ is the mean of the entries of $\tilde{y}$ in bag $l$. We expand on the objective below.

$$\begin{aligned}
\sum_{l=1}^{m} \sum_{i \in B_l} (\tilde{y}_i - \mu_l)^2 &= \sum_{l=1}^{m} \sum_{i \in B_l} (\tilde{y}_i^2 + \mu_l^2 - 2\tilde{y}_i \mu_l) \\
&= \sum_{l=1}^{m} \left( \sum_{i \in B_l} \tilde{y}_i^2 + \sum_{i \in B_l} \mu_l^2 - 2 \sum_{i \in B_l} \tilde{y}_i \mu_l \right) \\
&= \sum_{l=1}^{m} \left( \sum_{i \in B_l} \tilde{y}_i^2 + |B_l| \mu_l^2 - 2|B_l| \mu_l^2 \right) \\
&= \sum_{i=1}^{n} \tilde{y}_i^2 - \sum_{l=1}^{m} \left( |B_l| \mu_l^2 \right) \\
&= ||\tilde{y}||_2^2 - \sum_{\ell=1}^{m} \frac{\left( \sum_{i \in B_\ell} \tilde{y}_i \right)^2}{|B_\ell|}
\end{aligned}$$

$||\tilde{y}||_2^2$ is constant, hence minimizing $\sum_{l=1}^{m} \sum_{i \in B_l} (\tilde{y}_i - \mu_l)^2$ is equivalent to maximizing $\sum_{\ell=1}^{m} \frac{\left( \sum_{i \in B_\ell} \tilde{y}_i \right)^2}{|B_\ell|}$. □

**Lemma 10** (Noisy Clustering)**.** *Given $y_i = \tilde{y}_i + \gamma_i$, where $\gamma_i \sim \mathcal{N}(0, \sigma^2)$. Then, given a clustering $B$ over $y$,*

$$\mathbb{E}[k\text{-}means(B(y))] = \mathbb{E}[k\text{-}means(B(\tilde{y}))] + (n - m)\sigma^2$$

*where where k-means$(S(X))$ is the k-means clustering objective of $S$ on $X$. For equal sized bags of size $k$,*

$$\mathbb{E}[k\text{-}means(B(y))] = \mathbb{E}[k\text{-}means(B(\tilde{y}))] + n\left(1 - \frac{1}{k}\right)\sigma^2.$$

*Proof.*

$$
\begin{aligned}
\mathbb{E}[k\text{-means}(B(y))] - \mathbb{E}[k\text{-means}(B(\tilde{y}))] &= \mathbb{E}\left[\sum_{l=1}^{m}\sum_{i\in B_l}(y_i-\mu_l)^2\right] - \mathbb{E}\left[\sum_{l=1}^{m}\sum_{i\in B_l}(\tilde{y}_i-\mu_l)^2\right] \\
&= \mathbb{E}\left[\sum_{l=1}^{m}\sum_{i\in B_l}(y_i-\mu_l)^2 - \sum_{l=1}^{m}\sum_{i\in B_l}(\tilde{y}_i-\mu_l)^2\right] \\
&= \mathbb{E}\left[\sum_{l=1}^{m}\sum_{i\in B_l}\left((y_i-\mu_l)^2-(\tilde{y}_i-\tilde{\mu}_l)^2\right)\right] \\
&= \mathbb{E}\left[\sum_{l=1}^{m}\sum_{i\in B_l}\left((y_i-\tilde{y}_i+\tilde{\mu}_l-\mu_l)(y_i-\mu_l+\tilde{y}_i-\tilde{\mu}_l)\right)\right] \\
&= \mathbb{E}\left[\sum_{l=1}^{m}\sum_{i\in B_l}\left(\left(\gamma_i-\frac{\sum_{i\in B_l}\gamma_i}{|B_l|}\right)\left(2y_i-2\mu_l+\gamma_i-\frac{\sum_{i\in B_l}\gamma_i}{|B_l|}\right)\right)\right] \\
&= \sum_{l=1}^{m}\sum_{i\in B_l}\left(\mathbb{E}\left[\gamma_i^2\right]+\frac{\sum_{i\in B_l}\mathbb{E}\left[\gamma_i^2\right]}{|B_l|^2}-2\frac{\mathbb{E}\left[\gamma_i^2\right]}{|B_l|}\right) \\
&= \sum_{l=1}^{m}\sum_{i\in B_l}\mathbb{E}\left[\gamma_i^2\right]\left(1-\frac{1}{|B_l|}\right) \\
&= \sigma^2\sum_{l=1}^{m}(|B_l|-1) \\
&= \sigma^2(n-m)
\end{aligned}
$$

□

**Lemma 11.** *Sort $\tilde{y}_i$ in non-increasing order as $\tilde{y}_{(1)},\dots,\tilde{y}_{(n)}$. There exists an optimal $k$-means clustering $B^*$ such that $\tilde{y}_{(i)},\tilde{y}_{(j)}\in B_l^* \implies \tilde{y}_{(k)}\in B_l^*, \forall k\in\{i,i+1,\dots,j\}$.*

*Proof.* Follows from Lemma 2.3 in Javanmard et al. [2024]. □

### B.2 ADDITIONAL RESULTS FROM SECTION 3.1

**Lemma 12** ($k$-means Decomposition)**.** *Consider an orthogonal basis $z_1,\dots z_d$. Fix a clustering $S$. We can show the following*

$$
k\text{-}means(S(X)) = \sum_{j=1}^{d} k\text{-}means(S(X_{z_j})),
$$

*where k-means$(S(X))$ is the $k$-means clustering objective of $S$ on $X$, and $X_z$ is the projection of $X$ along $z$.*

*Proof.* Let $X = \{X_1, \dots, X_n\}$.

$$\begin{aligned}
\text{k-means}(S(X)) &= \sum_{l=1}^{m} \sum_{X_i \in S_l} ||X_i - \mu_l||_2^2 \\
&= \sum_{l=1}^{m} \sum_{X_i \in S_l} ||X_i||_2^2 + ||\mu_l||_2^2 - 2X_i^T \mu_l \\
&= \sum_{l=1}^{m} \sum_{X_i \in S_l} \sum_{j=1}^{d} \left( {X_{z_j}}_i^2 + \mu_{l_{z_j}}^2 - 2{X_{z_j}}^T \mu_{l_{z_j}} \right) \\
&= \sum_{j=1}^{d} \sum_{l=1}^{m} \sum_{X_i \in S_l} \left( {X_{z_j}}^T - \mu_{l_{z_j}} \right)^2 \\
&= \sum_{j=1}^{d} \text{k-means}(S(X_{z_j}))
\end{aligned}$$

□

**Lemma 13** ($k$-means-Variance Equivalence)**.** *Consider a direction $z$, and a centred dataset $X$. Given a bagging $S$ over $X$ with $m$ bags of equal size $k$,*

$$\mathit{Var}_z(SX) = \frac{1}{k^2} \left( \mathit{Var}(X_z) - \textit{k-means}(S(X_z)) \right),$$

*Proof.* Say the points are $X_1, \dots, X_n$, and the projections along $z$ are $x_1, \dots, x_n$. Let $\mu = 0$ be the mean of $X$, and $\mu_l$ be the mean of $B_l$. The variance of the $SX$ along $z$ is

$$\begin{aligned}
\text{Var}(SX_z) &= \sum_{l=1}^{m} (\mu_{l_z} - \mu_z)^2 \\
&= \sum_{\ell=1}^{m} \left( \frac{\sum_{i \in B_\ell} x_i}{k} \right)^2 \\
&= \frac{1}{k^2} \left( \sum_{i=1}^{n} x_i^2 - \sum_{\ell=1}^{m} \sum_{i \in B_\ell} (x_i - \mu_{l_z})^2 \right) \\
&= \frac{1}{k^2} \left( \text{Var}(X_z) - \text{k-means}(S(X_z)) \right)
\end{aligned}$$

□

# C RANDOM BAGGING, AGGREGATE-MIR

We propose a random bagging algorithm similar to the one for Bag-LLP (Algorithm 1) for Aggregate-MIR. The upper bound for Aggregate-MIR (Theorem 3) is product of the label $k$-means objective and the condition number. We propose the following algorithm which takes both these objectives into account.

We can analyze the condition number by establishing a lower bound on the minimum eigenvalue of the covariance matrix for the aggregated feature vectors. In Section 3.2, we derive this bound for any fixed partitioning of instances into super-bags, and the same bound holds for Algorithm 2.

Following the analysis in Section 3.2, we get,

$$\mathbb{P}\left[ \lambda_{min}\left( (SX)^T SX \right) > (1-\delta) \frac{\lambda_{min}(X^T X)}{4k^2} \right] \geq 1 - d \cdot \left[ \frac{e^{-\delta}}{(1-\delta)^{1-\delta}} \right]^{\mu_{\min}/k\beta}$$

Let $B_l$ denote a super-bag of size $2k$ for $l \in [r]$. We arbitrarily sample $k$ instances to create a bag $B_l^{(1)}$ and the remaining instances form another bag $B_l^{(2)}$. We know $B_l = B_l^{(1)} \cup B_l^{(2)}$ and $B_l^{(1)} \cap B_l^{(2)} = \phi$. Also, $|B_l^{(1)}| = |B_l^{(2)}| = k$.

**Theorem 9.** *For super-bags $B'_l$ as defined in Algorithm 2 with arbitrary non-overlapping partitions $B_l^{(1)}$ and $B_l^{(2)}$, we have*

$$\sum_{l=1}^{r} \textit{k-means-cluster}(\{\tilde{y}_i\}_{i \in B'_l}) \geq \sum_{l=1}^{r} \textit{k-means-cluster}(\{\tilde{y}_i\}_{i \in B_l^{(1)}}) + \textit{k-means-cluster}(\{\tilde{y}_i\}_{i \in B_l^{(1)}}) \tag{18}$$

*where, $k$-means-cluster($C$) is the $k$-means clustering loss for cluster $C$. This expands to give the following:*

$$\sum_{l=1}^{r} \sum_{i \in B'_l} (\tilde{y}_i - \mu'_l)^2 \geq \sum_{l=1}^{r} \Big( \sum_{j \in B_l^{(1)}} (\tilde{y}_i - \mu_l^{(1)})^2 + \sum_{j \in B_l^{(2)}} (\tilde{y}_i - \mu_l^{(2)})^2 \Big) \tag{19}$$

*where, $\mu$ denotes the respective cluster means.*

*Proof.* We write the $k$-means loss for $B'_l$. Let $\mu'_l = \sum_{j \in B'_l} \tilde{y}_i / 2k$.

$$\begin{aligned}
&\sum_{i \in B'_l} (\tilde{y}_i - \mu'_l)^2 \\
&= \sum_{i \in B'_l} \tilde{y}_i^2 - 2\tilde{y}_i \mu'_l + \mu_l'^2 \\
&= (\sum_{i \in B'_l} \tilde{y}_i^2) - \frac{(\sum_{i \in B'_l} \tilde{y}_i)^2}{k} + \frac{(\sum_{i \in B'_l} \tilde{y}_i)^2}{2k} \\
&= (\sum_{i \in B'_l} \tilde{y}_i^2) + (\frac{1}{4k} - \frac{1}{k})(\sum_{i \in B'_l} \tilde{y}_i)^2 \\
&= (\sum_{i \in B'_l} \tilde{y}_i^2) - \frac{1}{2k}(\sum_{i \in B'_l} \tilde{y}_i)^2
\end{aligned}$$

Next, we write the $k$-means loss for $B_l^{(1)}$. Let $\mu_l^{(1)} = \sum_{j \in B_l^{(1)}} \tilde{y}_i / k$.

$$\begin{aligned}
&\sum_{j \in B_l^{(1)}} (\tilde{y}_i - \mu_l^{(1)})^2 \\
&= \sum_{j \in B_l^{(1)}} \tilde{y}_i^2 - 2\tilde{y}_i \mu_l^{(1)} + {\mu_l^{(1)}}^2 \\
&= (\sum_{j \in B_l^{(1)}} \tilde{y}_i^2) - \frac{2(\sum_{j \in B_l^{(1)}} \tilde{y}_i)^2}{k} + \frac{(\sum_{j \in B_l^{(1)}} \tilde{y}_i)^2}{k} \\
&= (\sum_{j \in B_l^{(1)}} \tilde{y}_i^2) - \frac{1}{k}(\sum_{j \in B_l^{(1)}} \tilde{y}_i)^2
\end{aligned}$$

Similarly, for $B_l^{(2)}$, we get

$$\sum_{j \in B_l^{(2)}} (\tilde{y}_i - \mu_l^{(2)})^2 = (\sum_{j \in B_l^{(2)}} \tilde{y}_i^2) - \frac{1}{k}(\sum_{j \in B_l^{(1)}} \tilde{y}_i)^2$$

We define $\Delta_l = \sum_{i \in B'_l} (\tilde{y}_i - \mu'_l)^2 - \sum_{j \in B_l^{(1)}} (\tilde{y}_i - \mu_l^{(1)})^2 - \sum_{j \in B_l^{(2)}} (\tilde{y}_i - \mu_l^{(2)})^2$.

$$\begin{aligned}
\Delta_l &= \frac{-1}{2k} (\sum_{i \in B'_l} \tilde{y}_i)^2 + \frac{1}{k} [(\sum_{j \in B_l^{(1)}} \tilde{y}_i)^2 + (\sum_{j \in B_l^{(2)}} \tilde{y}_i)^2 + 2 \sum_{i \in B_l^{(1)}} \sum_{j \in B_l^{(2)}} \tilde{y}_i \tilde{y}_j - 2 \sum_{i \in B_l^{(1)}} \sum_{j \in B_l^{(2)}} \tilde{y}_i \tilde{y}_j] \\
&= \frac{-1}{2k} (\sum_{i \in B'_l} \tilde{y}_i)^2 + \frac{1}{k} [(\sum_{j \in B'_l} \tilde{y}_i)^2 - 2 \sum_{i \in B_l^{(1)}} \sum_{j \in B_l^{(2)}} \tilde{y}_i \tilde{y}_j] \\
&= \frac{1}{2k} (\sum_{i \in B'_l} \tilde{y}_i)^2 + \frac{-2}{k} (\sum_{i \in B_l^{(1)}} \sum_{j \in B_l^{(2)}} \tilde{y}_i \tilde{y}_j) \\
&= \frac{1}{2k} [(\sum_{i \in B'_l} \tilde{y}_i)^2 - 4 (\sum_{i \in B_l^{(1)}} \sum_{j \in B_l^{(2)}} \tilde{y}_i \tilde{y}_j)] \\
&= \frac{1}{2k} [(\sum_{j \in B_l^{(1)}} \tilde{y}_i) - (\sum_{j \in B_l^{(2)}} \tilde{y}_i)]^2 \\
&\geq 0
\end{aligned}$$

For any super-bag $B'_l$ for $l \in [r]$, $\Delta_l > 0$. We can now sum over all bags to get the total loss observed after bagging $\Delta = \sum_{l=1}^{r} \Delta \geq 0$.

This implies that the loss incurred by applying the $k$-means objective is higher when the instances are clustered into super-bags of sizes $2k$, compared to our random bagging approach, which creates two non-overlapping bags of sizes $k$ from the super-bags.

□

# D PRIVACY

In this section, we quantify the additional loss in utility incurred due to label-DP guarantees, for each setting we consider. We give full versions of the theorems stated in Section 4, along with the proofs.

## D.1 MIR, INSTANCE-LEVEL

**Theorem** (full version of Theorem 5)**.** *There exists a bagging $B$ with $|B_l| = k, \forall l \in [m]$, satisfying $(\varepsilon, \delta)$ label-DP, such that for $\hat{\theta}$ in* (1)*, we have*

$$\mathbb{E}\left[||\hat{\theta} - \theta^*||_2^2\right] \leq ||(X^T X)^{-1} X^T||_{op}^2 \left(2\left(OPT + n\left(1 - \frac{1}{k}\right)\alpha^2\right) + d\left(\sigma^2 + \frac{\alpha^2}{k^2}\right)\right),$$

*where $\alpha^2 = \frac{16R^2 \log\left(\frac{1.25}{\delta/2}\right)}{\varepsilon^2}$, and $OPT$ is the objective value of the optimal $k$-means clustering over $\tilde{y}$.*

*Proof.* The error due to privacy can be decomposed into two parts.

We need to add noise to the bag-labels before releasing them. MIR outputs one label at random, hence the sensitivity of the output is $2R$. Due to privacy amplification via subsampling [Balle et al., 2018, Steinke, 2022], and the fact that $\varepsilon << n$ in our setting, we add $\mathcal{N}\left(0, \frac{\alpha^2}{k^2}\right)$ noise to the bag-label value to ensure $\left(\frac{\varepsilon}{2}, \frac{\delta}{2}\right)$ label-DP, where $\alpha^2 = \frac{16R^2 \log\left(\frac{1.25}{\delta/2}\right)}{\varepsilon^2}$. Note that we assume addition of $\mathcal{N}\left(0, \sigma^2\right)$ noise to each $\tilde{y}_i$. Adding $\mathcal{N}\left(0, \frac{\alpha^2}{k^2}\right)$ to each bag-label is equivalent to adding $\mathcal{N}\left(0, \frac{\alpha^2}{k^2}\right)$ to each label $y_i$, hence leading to a total noise of $\mathcal{N}\left(0, \sigma^2 + \frac{\alpha^2}{k^2}\right)$ to each $\tilde{y}_i$, leading to an additional error of $d\frac{\alpha^2}{k^2}$ over the intital $d\sigma^2$.

In addition, since the objective here is a label-dependent clustering, we must use a differentially private $k$-means algorithm, leading to additional loss in utility. Adding $\mathcal{N}\left(0, \alpha^2\right)$ noise to each label, and then find an optimal clustering over the noise labels, satisfies $\left(\frac{\varepsilon}{2}, \frac{\delta}{2}\right)$ label-DP by postprocessing. If $OPT$ is the objective value of the optimal $k$-means clustering over $\tilde{y}$, this private clustering method will lead to an additional error of $\left(1 - \frac{1}{k}\right) \alpha^2$, due to Lemma 10.

Now, we have two queries, each of which are $\left(\frac{\varepsilon}{2}, \frac{\delta}{2}\right)$ label-DP, ensuring $(\varepsilon, \delta)$ label-DP in total due to composition. □

**Private clustering** Note that it is possible to further reduce the error $n\left(1 - \frac{1}{k}\right) \alpha^2$ due to private clustering. Note that the above method for private clustering satisfies the more stringent notion of local-DP [Bebensee, 2019], while we only need to satisfy the standard notion of central-DP. Hence, while it is easy to analyse, we can potentially find a much more accurate private clustering mechanism, suitably modifying existing algorithms in the rich literature on differentially-private $k$-means clustering [Su et al., 2015, Lu and Shen, 2020], for the special case of a single dimension.

### D.2 LLP, BAG-LEVEL

**Theorem** (full version of Theorem 6)**.** *There exists a bagging $B$ with $|B_l| = k, \forall l \in [m]$, satisfying $(\varepsilon, \delta)$ label-DP, such that for $\hat{\theta}$ in* (2)*, we have*

$$\mathbb{E}\left[\|\hat{\theta} - \theta^*\|_2^2\right] = OPT\left(\sigma^2 + \frac{\alpha^2}{k}\right)\frac{m}{k},$$

*where $\alpha^2 = \frac{4R^2 \log\left(\frac{1.25}{\delta}\right)}{\varepsilon^2}$, and $OPT$ is the optimal value of $\left(\frac{\lambda_{max}(f(X))}{\lambda_{min}(f(X))}\right)^2$.*

*Proof.* In this case, the optimal bagging strategy in independent of the labels. Hence, we just need to add noise to the bag-labels before releasing them, and not add noise for a private clustering of the labels. Each bag-label here is the mean of $k$ labels, hence the sensitivity of the output is $\frac{2R}{k}$. We add $\mathcal{N}\left(0, \frac{\alpha^2}{k^2}\right)$ noise to the label value to ensure $(\varepsilon, \delta)$ label-DP, where $\alpha^2 = \frac{4R^2 \log\left(\frac{1.25}{\delta}\right)}{\varepsilon^2}$. This is equivalent to adding $\mathcal{N}\left(0, \frac{\alpha^2}{k}\right)$ noise to each of the $k$ labels, and then averaging them. Note that we assume addition of $\mathcal{N}\left(0, \sigma^2\right)$ noise to each $\tilde{y}_i$. Adding $\mathcal{N}\left(0, \frac{\alpha^2}{k}\right)$ to each label $y_i$, leads to a total noise of $\mathcal{N}\left(0, \sigma^2 + \frac{\alpha^2}{k}\right)$ to each $\tilde{y}_i$, leading to an additional error of $\frac{\alpha^2}{k}\frac{m}{k}$ over the intital $\sigma^2\frac{m}{k}$.

□

### D.3 MIR, AGGREGATE-LEVEL

Theorem 3 shows that, for $\hat{\theta}$ in (3), given a bagging $B$, with equal sized bags, we have

$$\mathbb{E}\left[\|\hat{\theta} - \theta^*\|_2^2\right] \leq \frac{1}{k}\|((SX)^T SX)^{-1}(SX)^T\|_{op}^2 \left(\sum_{l=1}^{m} \sum_{\tilde{y}_i \in B_l} (\tilde{y}_i - \mu_l)^2 + \sigma^2 nk\right),$$

If we want a private bagging $B$, the error due to privacy can be decomposed into two parts. We need to add noise to the bag-labels before releasing them. As in the case of Instance-MIR, we add $\mathcal{N}\left(0, \frac{\alpha^2}{k^2}\right)$ noise to the bag-labels value to ensure $(\varepsilon, \delta)$ label-DP, where $\alpha^2 = \frac{4R^2 \log\left(\frac{1.25}{\delta}\right)}{\varepsilon^2}$, leading to an additional error of $nk\frac{\alpha^2}{k^2}$ over the intital $nk\sigma^2$.

Now, there are two terms that contribute to the clustering error, term 1 $\left(\|((SX)^T SX)^{-1}(SX)^T\|_{op}^2\right)$, and term 2 $\left(\sum_{l=1}^{m} \sum_{\tilde{y}_i \in B_l} (\tilde{y}_i - \mu_l)^2\right)$. Term 1 is involved in Bag-LLP, and minimizes the condition number of the bag-centroids. Term 2 is also involved in Instance-MIR, and minimizes a label-dependent $k$-means clustering objective. If we minimize Term 1, the optimal bagging strategy in independent of the labels. Hence, we just need to add noise to the bag-labels before releasing them, and not add noise for a private clustering of the labels. However, in this case, the value of Term 2 could be suboptimal.

If we minimize Term 2, we must use a differentially private $k$-means algorithm, leading to additional loss in utility. Adding $\mathcal{N}\left(0, \alpha^2\right)$ noise to each label, and then find an optimal clustering over the noise labels, satisfies $(\varepsilon, \delta)$ label-DP. As in the case of Instance-MIR, this private clustering method will lead to an additional error of $n\left(1 - \frac{1}{k}\right)\alpha^2$. Note that since we now have two private queries, we would have to split the privacy budget amongst them. However, minimizing term 2 might lead to a suboptimal value of Term 1.

# E ADDITIONAL EXPERIMENTAL RESULTS

## E.1 STATISTICAL SIGNIFICANCE

Table 7 has statistical significance scores for the results in Table 1 of the paper. There is one row for each bag-size $\{10, 50\}$ and three settings of LLP with Bag Loss, MIR with Instance Loss and MIR with Aggregate Loss. The columns **I**$k$**M**, **L**$k$**M** and **Rand** contain the mean errors of the bagging methods: Instance $k$-means, Label $k$-means and Random. All methods are evaluated on 15 independent trial datasets for each row. In column **(L**$k$**M vs I**$k$**M)-T** we present the paired $T$-value for Label $k$-means and Instance $k$-means. In column **S1 (90%)** we check whether the magnitude of this $T$-value is greater than the critical-$T = 1.760$ indicating whether there is a significant difference in the means with $90\%$ confidence. In column **(Rand vs best)-T** we present the paired $T$-value for Random Bagging vs. the better (i.e., lower error) of Label $k$-means and Instance $k$-means. Column **S2 (90%)** indicates whether there is a significant difference in the means of Random vs the better of Label $k$-means and Instance $k$-means, with $90\%$ confidence.

Table 8 similarly has the confidence scores for the results in Table 2 (private bagging) of the paper.

**Takeaways**: We see from Table 7 that Instance-$k$-means has statistically significant better performance over Label $k$-means as well as Random for MIR with Aggregate Loss and for LLP with Bag Loss (bag-size 10). For MIR with Instance loss on the other hand there is no statistically significant difference between the methods. In Table 8 we see that Label $k$-means is statistically significantly better than Instance $k$-means for all settings. However, Label $k$-means has statistically significant better performance over Random for 2 settings with bag-size 10. Overall we see that in many settings our results provide statistically significant separation between the techniques.

# F ANALYSIS FOR GLRS

We generalize the previous results for linear regression to the setting of Generalized Linear Model's (GLMs), which includes popular paradigms such as logistic regression. We study both instance-level and aggregate-level losses for MIR under the GLM framework. For Instance-MIR, we derive an upper bound that leads to label $k$-means clustering as the optimal bagging strategy. This result holds across all distributions within the exponential family. For Aggregate-MIR, our objective suggests minimizing the range between the maximum and minimum expected instance-labels within a bag, implying that features with similar expected labels should be grouped together, yielding a clustering-based outcome. This result holds for exponential distributions which have a monotonic first derivative. The detailed analysis is provided below.

The instance-level labels $y_i$ are conditionally independent given $\mathbf{x}_i$ in GLMs, and are drawn from a specific distribution within the exponential family. The corresponding log-likelihood function can be expressed as:

$$\log p(y_i \mid \eta_i, \phi) = \frac{y_i \eta_i - b(\eta_i)}{a_i(\phi)} + c(y_i, \phi)\,, \tag{20}$$

where $\eta_i$ is a location variable and $\phi$ is the scaling variable. The functions $a_i$, $b$, and $c$ are provided. We can take $a_i(\phi) = \phi/w_i$, where $w_i$ is a constant prior information. We analyse canonical GLMs, in which $\eta_i = \mathbf{x}_i^T \theta^*$ for an unknown model $\theta^*$. Some properties of GLMs are $\mu = \mathbb{E}[y|x] = b'(x^T\theta^*)$ and $Var(y|x) = a(\phi)b''(x^T\theta^*)$. We consider $\mathcal{L}$ to the negative log likelihood and we can ignore the term $c(y_i, \phi)$ as it does not depend on $\theta$. Our objective is to find a bagging strategy which closes the gap between the true model $\theta^*$ and $\hat{\theta}$. For GLMs we achieve this by minimizing the gradient of the loss at $\theta^*$.

## F.1 INSTANCE-MIR

**Lemma 14.** *Suppose that the loss $\mathcal{L}$ is strongly convex with parameter $\mu$ and $\hat{\theta} = \arg\min_\theta \mathcal{L}(\theta)$. Then, for any model $\theta^*$, we have*

$$\|\hat{\theta} - \theta^*\|_2 \leq \frac{1}{\mu}\|\mathcal{L}(\theta^*)\|_2.$$

*In addition, if $\mathcal{L}$ has a Lipschitz continuous gradient with parameter $L$, we have*

$$\frac{1}{L}\|\mathcal{L}(\theta^*)\|_2 \leq \|\hat{\theta} - \theta^*\|_2.$$

| Data | k | $\sigma$ | Bagging Method | $\|\hat{\theta} - \theta^*\|_2^2$ |
|---|---|---|---|---|
| Isotropic | 10 | 0.5 | Instance $k$-means | $0.010693 \pm 0.00167$ |
| | | | Label $k$-means | $0.044320 \pm 0.00720$ |
| | | | Label $k$-means super-bags | $0.040845 \pm 0.01104$ |
| | | | Random | $0.022352 \pm 0.00447$ |
| | | 2 | Instance $k$-means | $0.037875 \pm 0.00494$ |
| | | | Label $k$-means | $0.056199 \pm 0.01042$ |
| | | | Label $k$-means super-bags | $0.059399 \pm 0.01304$ |
| | | | Random | $0.053995 \pm 0.01119$ |
| | 50 | 0.5 | Instance $k$-means | $0.046242 \pm 0.00773$ |
| | | | Label $k$-means | $0.064936 \pm 0.01016$ |
| | | | Label $k$-means super-bags | $0.058051 \pm 0.00631$ |
| | | | Random | $0.057210 \pm 0.00981$ |
| | | 2 | Instance $k$-means | $0.056337 \pm 0.01002$ |
| | | | Label $k$-means | $0.065491 \pm 0.00853$ |
| | | | Label $k$-means super-bags | $0.061981 \pm 0.00991$ |
| | | | Random | $0.065836 \pm 0.01079$ |
| Non-isotropic (Independent) | 10 | 0.5 | Instance $k$-means | $0.014946 \pm 0.00421$ |
| | | | Label $k$-means | $0.040369 \pm 0.00990$ |
| | | | Label $k$-means super-bags | $0.042778 \pm 0.00804$ |
| | | | Random | $0.020230 \pm 0.00506$ |
| | | | Scaled Instance $k$-means | $0.012608 \pm 0.00354$ |
| | | 2 | Instance $k$-means | $0.039141 \pm 0.00884$ |
| | | | Label $k$-means | $0.048532 \pm 0.01083$ |
| | | | Label $k$-means super-bags | $0.052560 \pm 0.01105$ |
| | | | Random | $0.058208 \pm 0.00860$ |
| | | | Scaled Instance $k$-means | $0.042403 \pm 0.00573$ |
| | 50 | 0.5 | Instance $k$-means | $0.041916 \pm 0.00736$ |
| | | | Label $k$-means | $0.062490 \pm 0.00929$ |
| | | | Label $k$-means super-bags | $0.060436 \pm 0.01054$ |
| | | | Random | $0.055356 \pm 0.01085$ |
| | | | Scaled Instance $k$-means | $0.047906 \pm 0.00964$ |
| | | 2 | Instance $k$-means | $0.059583 \pm 0.00788$ |
| | | | Label $k$-means | $0.062350 \pm 0.01028$ |
| | | | Label $k$-means super-bags | $0.062662 \pm 0.01306$ |
| | | | Random | $0.065602 \pm 0.00934$ |
| | | | Scaled Instance $k$-means | $0.059133 \pm 0.01235$ |
| Non-isotropic (Non-independent) | 10 | 0.5 | Instance $k$-means | $0.031268 \pm 0.00649$ |
| | | | Label $k$-means | $0.052303 \pm 0.01065$ |
| | | | Label $k$-means super-bags | $0.049302 \pm 0.00531$ |
| | | | Random | $0.034642 \pm 0.01052$ |
| | | | Scaled Instance $k$-means | $0.022451 \pm 0.00636$ |
| | | 2 | Instance $k$-means | $0.043493 \pm 0.00732$ |
| | | | Label $k$-means | $0.054761 \pm 0.01151$ |
| | | | Label $k$-means super-bags | $0.056316 \pm 0.01127$ |
| | | | Random | $0.055723 \pm 0.01053$ |
| | | | Scaled Instance $k$-means | $0.039650 \pm 0.00781$ |
| | 50 | 0.5 | Instance $k$-means | $0.052643 \pm 0.01071$ |
| | | | Label $k$-means | $0.060606 \pm 0.00677$ |
| | | | Label $k$-means super-bags | $0.059758 \pm 0.00977$ |
| | | | Random | $0.057136 \pm 0.00876$ |
| | | | Scaled Instance $k$-means | $0.046376 \pm 0.00642$ |
| | | 2 | Instance $k$-means | $0.058460 \pm 0.01074$ |
| | | | Label $k$-means | $0.060828 \pm 0.00811$ |
| | | | Label $k$-means super-bags | $0.065220 \pm 0.00745$ |
| | | | Random | $0.067064 \pm 0.01064$ |
| | | | Scaled Instance $k$-means | $0.059597 \pm 0.00908$ |

Table 4: Aggregate-MIR

| Data | k | $\sigma$ | Bagging Method | $\|\hat{\theta} - \theta^*\|_2^2$ |
|---|---|---|---|---|
| Isotropic | 10 | 0.5 | Instance $k$-means | $0.007562 \pm 0.00137$ |
| | | | Label $k$-means | $0.043625 \pm 0.00722$ |
| | | | Label $k$-means super-bags | $0.044586 \pm 0.00906$ |
| | | | Random | $0.009745 \pm 0.00206$ |
| | | 2 | Instance $k$-means | $0.014722 \pm 0.00329$ |
| | | | Label $k$-means | $0.056195 \pm 0.01101$ |
| | | | Label $k$-means super-bags | $0.056651 \pm 0.01085$ |
| | | | Random | $0.026405 \pm 0.00502$ |
| | 50 | 0.5 | Instance $k$-means | $0.037432 \pm 0.00721$ |
| | | | Label $k$-means | $0.063826 \pm 0.00800$ |
| | | | Label $k$-means super-bags | $0.058686 \pm 0.01111$ |
| | | | Random | $0.046269 \pm 0.00830$ |
| | | 2 | Instance $k$-means | $0.040709 \pm 0.00964$ |
| | | | Label $k$-means | $0.063859 \pm 0.00486$ |
| | | | Label $k$-means super-bags | $0.058983 \pm 0.00880$ |
| | | | Random | $0.049042 \pm 0.00872$ |
| Non-isotropic (Independent) | 10 | 0.5 | Instance $k$-means | $0.009739 \pm 0.00201$ |
| | | | Label $k$-means | $0.042496 \pm 0.00626$ |
| | | | Label $k$-means super-bags | $0.044571 \pm 0.00929$ |
| | | | Random | $0.010518 \pm 0.00339$ |
| | | | Scaled Instance $k$-means | $0.008552 \pm 0.00191$ |
| | | 2 | Instance $k$-means | $0.018930 \pm 0.00425$ |
| | | | Label $k$-means | $0.049482 \pm 0.01074$ |
| | | | Label $k$-means super-bags | $0.055759 \pm 0.01066$ |
| | | | Random | $0.030314 \pm 0.00652$ |
| | | | Scaled Instance $k$-means | $0.014849 \pm 0.00286$ |
| | 50 | 0.5 | Instance $k$-means | $0.036923 \pm 0.00536$ |
| | | | Label $k$-means | $0.059834 \pm 0.00598$ |
| | | | Label $k$-means super-bags | $0.062452 \pm 0.01025$ |
| | | | Random | $0.039461 \pm 0.00760$ |
| | | | Scaled Instance $k$-means | $0.038586 \pm 0.00784$ |
| | | 2 | Instance $k$-means | $0.043048 \pm 0.01045$ |
| | | | Label $k$-means | $0.058143 \pm 0.01113$ |
| | | | Label $k$-means super-bags | $0.059907 \pm 0.00812$ |
| | | | Random | $0.054860 \pm 0.00659$ |
| | | | Scaled Instance $k$-means | $0.045390 \pm 0.00617$ |
| Non-isotropic (Non-independent) | 10 | 0.5 | Instance $k$-means | $0.032367 \pm 0.00835$ |
| | | | Label $k$-means | $0.052438 \pm 0.00936$ |
| | | | Label $k$-means super-bags | $0.050445 \pm 0.01255$ |
| | | | Random | $0.024585 \pm 0.00755$ |
| | | | Scaled Instance $k$-means | $0.024811 \pm 0.00498$ |
| | | 2 | Instance $k$-means | $0.033099 \pm 0.01050$ |
| | | | Label $k$-means | $0.057081 \pm 0.00955$ |
| | | | Label $k$-means super-bags | $0.057327 \pm 0.01297$ |
| | | | Random | $0.032676 \pm 0.00675$ |
| | | | Scaled Instance $k$-means | $0.029420 \pm 0.00755$ |
| | 50 | 0.5 | Instance $k$-means | $0.051425 \pm 0.00895$ |
| | | | Label $k$-means | $0.061918 \pm 0.00820$ |
| | | | Label $k$-means super-bags | $0.058320 \pm 0.01040$ |
| | | | Random | $0.048222 \pm 0.01074$ |
| | | | Scaled Instance $k$-means | $0.049910 \pm 0.00773$ |
| | | 2 | Instance $k$-means | $0.051430 \pm 0.00661$ |
| | | | Label $k$-means | $0.065289 \pm 0.01090$ |
| | | | Label $k$-means super-bags | $0.069147 \pm 0.01071$ |
| | | | Random | $0.059075 \pm 0.00885$ |
| | | | Scaled Instance $k$-means | $0.047859 \pm 0.00678$ |

Table 5: Bag-LLP

| Data | k | $\sigma$ | Bagging Method | $\|\hat{\theta} - \theta^*\|_2^2$ |
|---|---|---|---|---|
| Isotropic | 10 | 0.5 | Instance $k$-means | $0.008894 \pm 0.00168$ |
| | | | Label $k$-means | $0.007597 \pm 0.00197$ |
| | | | Random | $0.007997 \pm 0.00174$ |
| | | 2 | Instance $k$-means | $0.019629 \pm 0.00410$ |
| | | | Label $k$-means | $0.010983 \pm 0.00239$ |
| | | | Random | $0.010078 \pm 0.00190$ |
| | 50 | 0.5 | Instance $k$-means | $0.039916 \pm 0.00828$ |
| | | | Label $k$-means | $0.040155 \pm 0.00986$ |
| | | | Random | $0.044420 \pm 0.00472$ |
| | | 2 | Instance $k$-means | $0.049003 \pm 0.01167$ |
| | | | Label $k$-means | $0.040044 \pm 0.00608$ |
| | | | Random | $0.040281 \pm 0.00600$ |
| Non-isotropic (Independent) | 10 | 0.5 | Instance $k$-means | $0.008672 \pm 0.00215$ |
| | | | Label $k$-means | $0.007790 \pm 0.00158$ |
| | | | Random | $0.008808 \pm 0.00174$ |
| | | | Scaled Instance $k$-means | $0.009683 \pm 0.00102$ |
| | | 2 | Instance $k$-means | $0.018395 \pm 0.00421$ |
| | | | Label $k$-means | $0.012217 \pm 0.00205$ |
| | | | Random | $0.011335 \pm 0.00198$ |
| | | | Scaled Instance $k$-means | $0.022363 \pm 0.00499$ |
| | 50 | 0.5 | Instance $k$-means | $0.042065 \pm 0.00686$ |
| | | | Label $k$-means | $0.041108 \pm 0.00867$ |
| | | | Random | $0.038124 \pm 0.00552$ |
| | | | Scaled Instance $k$-means | $0.037391 \pm 0.00674$ |
| | | 2 | Instance $k$-means | $0.043934 \pm 0.00901$ |
| | | | Label $k$-means | $0.041059 \pm 0.00527$ |
| | | | Random | $0.044340 \pm 0.00826$ |
| | | | Scaled Instance $k$-means | $0.047298 \pm 0.00768$ |
| Non-isotropic (Non-independent) | 10 | 0.5 | Instance $k$-means | $0.023122 \pm 0.00747$ |
| | | | Label $k$-means | $0.023248 \pm 0.00916$ |
| | | | Random | $0.022115 \pm 0.00565$ |
| | | | Scaled Instance $k$-means | $0.019744 \pm 0.00628$ |
| | | 2 | Instance $k$-means | $0.035530 \pm 0.01027$ |
| | | | Label $k$-means | $0.027272 \pm 0.00708$ |
| | | | Random | $0.026394 \pm 0.00626$ |
| | | | Scaled Instance $k$-means | $0.034814 \pm 0.00768$ |
| | 50 | 0.5 | Instance $k$-means | $0.049454 \pm 0.00978$ |
| | | | Label $k$-means | $0.048404 \pm 0.00920$ |
| | | | Random | $0.048654 \pm 0.01101$ |
| | | | Scaled Instance $k$-means | $0.051057 \pm 0.00644$ |
| | | 2 | Instance $k$-means | $0.049799 \pm 0.00843$ |
| | | | Label $k$-means | $0.045538 \pm 0.00981$ |
| | | | Random | $0.047661 \pm 0.00710$ |
| | | | Scaled Instance $k$-means | $0.048617 \pm 0.00801$ |

Table 6: Instance-MIR

| Loss | $k$ | I$k$M | L$k$M | Rand | (L$k$M vs I$k$M)-T | S1 (90%) | (Rand vs best)-T | S2 (90%) |
|---|---|---|---|---|---|---|---|---|
| Bag | 10 | $0.0082 \pm 0.002$ | $0.0458 \pm 0.012$ | $0.0099 \pm 0.002$ | 12.301 | Yes | 2.004 | Yes |
| LLP | 50 | $0.0392 \pm 0.008$ | $0.0629 \pm 0.008$ | $0.0423 \pm 0.009$ | 7.062 | Yes | 1.261 | No |
| Instance | 10 | $0.0088 \pm 0.002$ | $0.0072 \pm 0.002$ | $0.0085 \pm 0.002$ | -1.688 | No | 1.332 | No |
| MIR | 50 | $0.0388 \pm 0.006$ | $0.0404 \pm 0.007$ | $0.0419 \pm 0.006$ | 0.643 | No | 1.172 | No |
| Aggregate | 10 | $0.0102 \pm 0.002$ | $0.0453 \pm 0.008$ | $0.0221 \pm 0.004$ | 15.85 | Yes | 8.284 | Yes |
| MIR | 50 | $0.0437 \pm 0.008$ | $0.0601 \pm 0.008$ | $0.0619 \pm 0.012$ | 5.339 | Yes | 4.505 | Yes |

Table 7: Statistical Significant for Non-Private Bagging

| $\varepsilon$ | $k$ | I$k$M | L$k$M | Rand | (L$k$M vs I$k$M)-T | S1 (90%) | (Rand vs best)-T | S2 (90%) |
|---|---|---|---|---|---|---|---|---|
| 0.5 | 10 | $0.0619 \pm 0.012$ | $0.0505 \pm 0.005$ | $0.0553 \pm 0.008$ | -4.105 | Yes | 1.761 | Yes |
| | 50 | $0.0656 \pm 0.012$ | $0.0559 \pm 0.008$ | $0.0564 \pm 0.007$ | -2.297 | Yes | 0.208 | No |
| 1 | 10 | $0.0537 \pm 0.009$ | $0.0362 \pm 0.006$ | $0.0397 \pm 0.010$ | -5.513 | Yes | 1.189 | No |
| | 50 | $0.0595 \pm 0.012$ | $0.0480 \pm 0.005$ | $0.0447 \pm 0.005$ | -3.029 | Yes | -1.689 | No |
| 2 | 10 | $0.0390 \pm 0.008$ | $0.0189 \pm 0.004$ | $0.0216 \pm 0.005$ | -9.182 | Yes | 2.148 | Yes |
| | 50 | $0.0521 \pm 0.009$ | $0.0431 \pm 0.006$ | $0.0434 \pm 0.008$ | -3.513 | Yes | 0.1569 | No |

Table 8: Statistical Significant for Private Bagging

Let $\hat{\theta}$ be the minimizer of the instance-level loss:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \frac{1}{n} \sum_{l=1}^{m} \sum_{i \in B_l} \frac{\overline{y_l}\eta_i - b(\eta_i)}{a_i(\phi)} \tag{21}$$

We find the optimal $\hat{\theta}$ by solving $\nabla\mathcal{L}(\hat{\theta}) = \mathbf{0}$. We use Lemma 14 which states that $\|\hat{\theta} - \theta^*\|_2$ is lower bounded by $\|\nabla\mathcal{L}(\theta^*)\|_2$ for strongly convex functions.

We define a instance-level attribution matrix for MIR, $A \in \{0,1\}^{n\times n}$, which assigns the bag-label to each feature vector in the bag. The prime feature vector is chosen uniformly at random. Let $\overline{y} = [\overline{y_1}, \ldots, \overline{y_m}]$, where $\overline{y_l} = y(\Gamma(B_l))$ as previously defined.

$$A_{(i,j)} = \begin{cases} 1 & \text{if } i \in B_l \text{ and } \overline{y_l} = y(x_j) \\ 0 & \text{otherwise.} \end{cases} \tag{22}$$

We define $S \in [0,1]^{n\times n}$ as the expectation of $A$:

$$S_{(i,j)} = \begin{cases} \frac{1}{|B_l|} & \text{if } i, j \in B_l \\ 0 & \text{otherwise.} \end{cases} \tag{23}$$

**Theorem 10.** *If we consider canonical GLMs with $\eta_i = \mathbf{x}_i^T\theta^*$, then we have*

$$\mathbb{E}\left[\|\nabla\mathcal{L}(\theta^*)\|_2\right] \leq m(\|b'(X\theta^*)\|_2^2 + \|Db''(X\theta^*)\|_1) + \|(S-I)b'(X\theta^*)\|_2^2 - \|Sb'(X\theta^*)\|_2^2 \tag{24}$$

*where, $D = \mathit{Diag}(\{a_i(\phi)\})$.*

*Proof.* We begin by computing $\nabla\mathcal{L}(\theta)$ and expressing it in the matrix format:

$$\begin{aligned} \nabla\mathcal{L}(\theta) &= \frac{1}{n} \sum_{l=1}^{m} \sum_{i \in B_l} \frac{(\overline{y_l} - b'(x_i^T\theta))x_i}{a_i(\phi)} \\ &= X^T D^{-1}(Ay - b'(X\theta)) \end{aligned}$$

where, $D := \text{Diag}(\{a_i(\phi)\})$.

$$\begin{aligned}
\mathbb{E}\left[\|\nabla\mathcal{L}(\theta)\|_2^2|X\right] &= \mathbb{E}\left[\|X^T D^{-1}(Ay - b'(X\theta))\|_2^2|X\right] \\
&\leq \|X^T D^{-1}\|_{op}^2 \mathbb{E}\left[\|Ay - b'(X\theta)\|_2^2|X\right] \\
&= \|X^T D^{-1}\|_{op}^2 \mathbb{E}\left[(Ay - b'(X\theta))^T(Ay - b'(X\theta))|X\right] \\
&= \|X^T D^{-1}\|_{op}^2 \mathbb{E}\left[(Ay)^T(Ay) - b'(X\theta)^T Ay - (Ay)^T b'(X\theta) + b'(X\theta)^T b'(X\theta)|X\right] \\
&= \|X^T D^{-1}\|_{op}^2 \left(\mathbb{E}\left[(Ay)^T(Ay)|X\right] - b'(X\theta)^T Sy - (Sy)^T b'(X\theta) + b'(X\theta)^T b'(X\theta)\right) \\
&= \|X^T D^{-1}\|_{op}^2 \big(\mathbb{E}\left[(Ay)^T(Ay)|X\right] - b'(X\theta)^T Sy - (Sy)^T b'(X\theta) + b'(X\theta)^T b'(X\theta) \\
&\quad + (Sb'(X\theta))^T(Sb'(X\theta)) - (Sb'(X\theta))^T(Sb'(X\theta))\big) \\
&= \|X^T D^{-1}\|_{op}^2 \left(\mathbb{E}\left[\|Ay\|_2^2|X\right] + \|(S-I)b'(X\theta)\|_2^2 - \|Sb'(X\theta)\|_2^2\right) \\
&\leq \|X^T D^{-1}\|_{op}^2 \left(\mathbb{E}\left[\|A\|_{op}^2\|y\|_2^2|X\right] + \|(S-I)b'(X\theta)\|_2^2 - \|Sb'(X\theta)\|_2^2\right) \\
&\leq \|X^T D^{-1}\|_{op}^2 \left(m(\|b'(X\theta^*)\|_2^2 + \|Db''(X\theta^*)\|_1) + \|(S-I)b'(X\theta)\|_2^2 - \|Sb'(X\theta)\|_2^2\right)
\end{aligned}$$

□

Note that the term $\|X^T D^{-1}\|_{op}^2$ is constant and the first term $m(\|b'(X\theta^*)\|_2^2 + \|Db''(X\theta^*)\|_1)$ is independent of the bagging strategy, it can be disregarded. Thus, we focus on the remaining terms to derive a clustering objective for event-level MIR. To proceed, we expand the matrix notation and express these terms as a summation over instances. We define $\mu_l := \frac{\mu_i}{|B_l|}$, where $\mu_i = \mathbb{E}[y_i|x_i] = b'(x_i^T\theta^*)$.

$$\min_{(B_1,\ldots,B_m)\in\mathcal{B}} \quad \|(S-I)b'(X\theta)\|_2^2 - \|Sb'(X\theta)\|_2^2 = \min_{(B_1,\ldots,B_m)\in\mathcal{B}} \quad \sum_{l=1}^{m}\sum_{i\in B_l}(\mu_i - \mu_l)^2 - \sum_{l=1}^{m}|B_l|\mu_l$$

Minimizing the first term in the objective is similar to performing $1d$ $k$-means clustering and maximizing the second term forces the bags to be of larger sizes.

### F.2 AGGREGATE-MIR

Let $\hat{\theta}$ be the minimizer of the Aggregate-MIR loss:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \frac{1}{m}\sum_{l=1}^{m} \frac{\overline{y_l}\sum_{i\in B_l}\frac{\eta_i}{|B_l|} - b(\sum_{i\in B_l}\frac{\eta_i}{|B_l|})}{a_l(\phi)} \tag{25}$$

The steps involved in analysing this function are similar to the instance-loss function. We find the optimal $\hat{\theta}$ by solving $\nabla\mathcal{L}(\hat{\theta}) = \mathbf{0}$ and then minimize $\|\nabla\mathcal{L}(\theta^*)\|_2$ to approximate $\|\hat{\theta} - \theta^*\|_2$ (Lemma 14).

**Theorem 11.** *If we consider canonical GLMs with $\eta_i = \mathbf{x}_i^T\theta^*$, then we get*

$$\mathbb{E}\left[\|\nabla\mathcal{L}(\theta^*)\|_2\right] \leq n\lambda_{max}(X^T X)\left(m(\|b'(X\theta^*)\|_2^2 + \|Db''(X\theta^*)\|_1) + \|Sb'(X\theta) - b'(SX\theta)\|_2^2 - \|Sb'(X\theta)\|_2^2)\right) \tag{26}$$

*where, $D = \text{Diag}(\{a_i(\phi)\})$.*

*Proof.* We begin by computing $\nabla\mathcal{L}(\theta)$ and expressing it in the matrix format:

$$\begin{aligned}
\nabla\mathcal{L}(\theta) &= \frac{1}{n}\sum_{l=1}^{m}\frac{(\overline{y_l} - b'(\sum_{i\in B_l}\frac{x_i^T\theta}{|B_l|}))\sum_{i\in B_l}\frac{x_i^T\theta}{|B_l|}}{a_l(\phi)} \\
&= (SX)^T D^{-1}(Ay - b'(SX\theta))
\end{aligned}$$

where, $D := \text{Diag}(\{a_l(\phi)\})$.

$$
\begin{aligned}
\mathbb{E}\left[\|\nabla\mathcal{L}(\theta)\|_2^2|X\right] &= \mathbb{E}\left[\|(SX)^T D^{-1}(Ay - b'(SX\theta))\|_2^2|X\right] \\
&\leq \|(SX)^T D^{-1}\|_{op}^2 \mathbb{E}\left[\|Ay - b'(SX\theta)\|_2^2|X\right] \\
&= \|(SX)^T D^{-1}\|_{op}^2 \mathbb{E}\left[(Ay - b'(SX\theta))^T(Ay - b'(SX\theta))|X\right] \\
&= \|(SX)^T D^{-1}\|_{op}^2 \mathbb{E}\left[(Ay)^T(Ay) - b'(SX\theta)^T Ay - (Ay)^T b'(SX\theta) + b'(SX\theta)^T b'(SX\theta)|X\right] \\
&= \|(SX)^T D^{-1}\|_{op}^2 \left(\mathbb{E}\left[(Ay)^T(Ay)|X\right] - b'(SX\theta)^T Sy - (Sy)^T b'(SX\theta) + b'(SX\theta)^T b'(SX\theta)\right) \\
&= \|(SX)^T D^{-1}\|_{op}^2 \big(\mathbb{E}\left[(Ay)^T(Ay)|X\right] - b'(SX\theta)^T Sb'(X\theta) - (Sb'(X\theta))^T b'(SX\theta) + b'(SX\theta)^T b'(SX\theta) + \\
&\quad (Sb'(X\theta))^T(Sb'(X\theta)) - (Sb'(X\theta))^T(Sb'(X\theta))\big) \\
&= \|(SX)^T D^{-1}\|_{op}^2 \left(\mathbb{E}\left[\|Ay\|_2^2|X\right] + \|Sb'(X\theta) - b'(SX\theta)\|_2^2 - \|Sb'(X\theta)\|_2^2\right) \\
&\leq \|(SX)^T D^{-1}\|_{op}^2 \left(\mathbb{E}\left[\|A\|_{op}^2\|y\|_2^2|X\right] \|Sb'(X\theta) - b'(SX\theta)\|_2^2 - \|Sb'(X\theta)\|_2^2\right) \\
&\leq \|(SX)^T D^{-1}\|_{op}^2 \left(m(\|b'(X\theta^*)\|_2^2 + \|Db''(X\theta^*)\|_1) + \|Sb'(X\theta) - b'(SX\theta)\|_2^2 - \|Sb'(X\theta)\|_2^2\right) \\
&\leq \|D^{-1}\|_{op}^2 \lambda_{max}(X^T X) \left(m(\|b'(X\theta^*)\|_2^2 + \|Db''(X\theta^*)\|_1) + \|Sb'(X\theta) - b'(SX\theta)\|_2^2 - \|Sb'(X\theta)\|_2^2\right)
\end{aligned}
$$

□

We now show how the final objective in Equation 11 leads to a clustering objective. The key term in this objective which depends on $S$ is $\|Sb'(X\theta) - b'(SX\theta)\|_2^2$. Our task is to determine the optimal bagging matrix $S$ that would minimize this term. To simplify this expression and develop an interpretable algorithm, we assume that the function $b'(.)$ is monotonic. Focusing on the case where $b'(.)$ is an increasing function, we know that $b'(t_1) \geq b'(t_2) \iff t_1 \geq t_2$.

$$
\left\|(Sb'(X\theta) - b'(SX\theta)\right\|_2^2 = \sum_{l=1}^{m}\left(\sum_{x\in B_l}\frac{b'(x^T\theta^*)}{|B_l|} - b'(\sum_{x\in B_l}\frac{x^T\theta^*}{|B_l|})\right)^2
$$

Since $b'$ is an increasing function, the inequality $b'(\max_{x'\in B_l} x'^T\theta^*) \geq b'(x^T\theta^*)$ holds true for all $x \in B_l$ (and $\max_{x'\in B_l} x'^T\theta^* \geq x^T\theta^*$). Similarly, $b'(x^T\theta^*) \geq b'(\min_{x'\in B_l} x'^T\theta^*)$ would hold true for all $x \in B_l$ $x^T\theta^* \geq \min_{x'\in B_l})$ We now look at the first term:

$$
\frac{b'(\min_{x'\in B_l} x'^T\theta^*)}{|B_l|} \leq \sum_{x\in B_l}\frac{b'(x^T\theta^*)}{|B_l|} \leq \frac{b'(\sum_{x\in B_l}\max_{x'\in B_l} x'^T\theta^*)}{|B_l|}
$$

$$
b'(\min_{x'\in B_l} x'^T\theta^*) \leq \sum_{x\in B_l}\frac{b'(x^T\theta^*)}{|B_l|} \leq b'(\max_{x'\in B_l} x'^T\theta^*)
$$

We bound the second term:

$$
b'(\sum_{x\in B_l}\frac{\min_{x'\in B_l} x'^T\theta^*}{|B_l|}) \leq b'(\sum_{x\in B_l}\frac{x^T\theta^*}{|B_l|}) \leq b'(\frac{\sum_{x\in B_l}\max_{x'} x'T\theta^*}{|B_l|})
$$

$$
b'(\min_{x'\in B_l} x'^T\theta^*) \leq b'(\sum_{x\in B_l}\frac{x^T\theta^*}{|B_l|}) \leq b'(\max_{x'\in B_l} x'^T\theta^*)
$$

It is easy to see that the difference $\|Sb'(X\theta) - b'(SX\theta)\|_2^2$ has an upper bound:

$$
\sum_{l=1}^{m}\left(\sum_{x\in B_l}\frac{b'(x^T\theta^*)}{|B_l|} - b'(\sum_{x\in B_l}\frac{x^T\theta^*}{|B_l|})\right)^2 \leq \sum_{l=1}^{m}\left(b'(\max_{x'\in B_l} x'^T\theta^*) - b'(\min_{x'\in B_l} x'^T\theta^*)\right)^2 \tag{27}
$$

If $n = mk$ and we need to construct-equal sized bags having k instances each, then the minimization of Equation 27 can be achieved by sorting $b'(x^T\theta^*)$ for all $x \in X$, and dividing the points into contiguous chunks of size $k$. This process resembles the $1d$ clustering objective with an equal-size constraint.

The monotonicity condition holds true for majority of the distributions belonging to the exponential family including normal, poisson, logistic and inverse gaussian.