OPT2024: 16th Annual Workshop on Optimization for Machine Learning

Aggregating Data for Optimal and Private Learning

Sushant Agarwal$^{\circ *}$ AGARWAL.SUS@NORTHEASTERN.EDU
Yukti Makhija† YUKTIMAKHIJA@GOOGLE.COM
Rishi Saket† RISHISAKET@GOOGLE.COM
Aravindan Raghuveer† ARAGHUVEER@GOOGLE.COM
†*Google DeepMind* $^{\circ}$*Northeastern University*

Abstract

Multiple Instance Regression (MIR) and Learning from Label Proportions (LLP) are learning frameworks arising in many applications, 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 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 collection of labeled *instances* of the form $(\mathbf{x}, y)$, where $\mathbf{x} \in \mathbb{R}^d$ is an instance or feature-vector with label y. In many applications however, due to lack of instrumentation or annotators [10, 13], or privacy constraints [26], instance-wise labels may not be available. Instead, the dataset is partitioned into disjoint sets or *bags* of instances, and for each bag only one *bag-label* is available to the learner. The bag-label is derived from the undisclosed instance-labels present in the bag via some aggregation function depending on the scenario. The goal is to train a model predicting the labels of individual instances. We call this paradigm as learning from aggregate labels, which 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 the bag-label is one of the instance-labels of the bag, and the instance whose label is chosen as the bag-label is not revealed, and (ii) learning from label proportions (LLP) in which the bag-label is the average of the bag's instance-labels. In MIR as well as in LLP, our work considers real-valued instance-labels with regression as the underlying instance-level task.

[*] work done during an internship at Google DeepMind

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 [22]). In many situations, the learner can be an untrusted party, and we wish to protect the privacy of individual instance labels from the learner (and any downstream observer of the learners output), while still allowing the learner to train useful models. We assume the existence of a trusted *aggregator* that has access to all the data, including feature vectors and labels. The aggregator partitions the instances into bags, and along with the bags also releases aggregate labels of each bag (i.e., the bag-label) to the learner. If a bag is of large size, revealing only the aggregate bag-label provides a layer of privacy protection of the labels, while on the other hand, larger bags in the training data lead to a loss in the quality (utility) of the trained model. Apart from the inherent privacy that MIR and LLP offer, the aggregator can further 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.

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 [35]. In such situations, the paradigm of learning from aggregate labels, especially MIR, can be very useful. Given a large amount of unlabeled data, and a limited labeling budget (say m), one could partition the data into m 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. We call this process of partitioning unlabeled data into bags as *label-agnostic* bagging. One might also be interested in the bagging of labeled data, for eg., due to privacy concerns as discussed earlier, which we call *label-dependent* bagging.

For various loss functions in MIR and LLP, we consider the task of optimal bag construction for both the *label-agnostic* and *label-dependent* settings. More specifically, we study the following question; what is the optimal strategy for the aggregator to partition the data into bags, such that the utility of downstream tasks such as linear regression is maximized.

Outline In Section 2, we formally define the problem, and state our main results. We start with the task of linear regression, and define utility to be the closeness of the trained model to the target model (in the realizable setting). In the MIR setting (for the case of instance-level loss, where each instance is assigned their bag label), we show that the optimal bagging strategy corresponds to finding an optimal k-means clustering over the labels. In the LLP setting (for the case of bag-level loss, between the bag-label and average prediction of the bag), we prove that the optimal bagging strategy is label-agnostic, and involves minimizing the condition number of the covariance of the centroids of each bag. For MIR we also consider aggregate-level loss (between the bag-label and prediction of the bag centroid). Here, the utility bound involves both the k-means objective of instance-MIR, and the condition number objective of bag-LLP. In Section 2.4, we also quantify the additional loss in utility incurred due to differential-privacy guarantees, in each of the previous scenarios. In Appendix B, we provide an overview of the analysis for the instance-MIR utility bound, and an upper bound for the condition number objective (which is common to both bag-LLP and aggregate-MIR) based on a random bagging approach. The rest of the proofs are moved to Appendix C . We then study the proposed bagging mechanisms through extensive experimentation in Appendix D, and show that k-means clustering over the instances is an effective label-agnostic bagging heuristic for each of the cases we study. We analyse trends obtained by varying various parameters such as the minimum bag size, number of bags, and privacy budget. In Appendix G, we generalize the previous results to GLM's, which includes popular paradigms such as logistic regression. We discuss the most relevant previous work in Appendix A .

2. Our Results

The training dataset consists of n samples $\in \mathbb{R}^d$ denoted by $n \times d$ matrix X, of rank d, with the corresponding labels denoted by $n \times 1$ matrix Y. X is partitioned into m non-overlapping bags $B = \{B_1, \ldots, B_m\}$, each of size at least k, for some fixed k[*] (Hence, $n \geqslant mk$). We consider the task of linear regression, and adopt a standard way to model it, where label $y_i = x_i^T\theta^* + \gamma_i\,, \gamma_i \sim \mathcal{N}(0, \sigma^2)$, for a fixed underlying model θ^*. We denote the expected value of the label of x_i by $\tilde{y}_i$, i.e., $\tilde{y}_i := x_i^T\theta^*$. An aggregator partitions X into bags, and along with the feature-vectors in each bag also releases aggregate bag-labels of each bag to the learner. The learner's task is to find an estimator $\hat{\theta}$, that is as close as possible to the underlying θ^*. The problem of bag construction is for the aggregator to find an optimal bagging configuration such that a given loss function is minimized, while satisfying the minimum bag size constraint $|B_l| \geqslant k, \forall l \in [m]$. B_l denotes the set of samples in bag l, and $\overline{y}_l$ denotes the aggregate response in bag l. In the case of MIR, we consider the popular case where the aggregate $\overline{y}_l$ is a uniformly random label, and for LLP $\overline{y}_l$ is the mean of the labels. Note that this minimum size constraint for the bags is essential to define a meaningful problem, otherwise the optimal bagging would be the trivial strategy of putting each point in a separate bag.

2.1. MIR, Instance-level loss

Definition 1 (Instance-level loss) *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 ℓ is the squared loss.

In the case of instance-level loss, we basically assign the aggregate label of the bag to each point in the bag. The result below provides an upper bound on the utility. All expectations henceforth are conditioned on a fixed X, unless otherwise stated.

Theorem 2 *For $\hat{\theta}$ as in (1), for a given bagging B,*

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

[*]. We do not use k as the number of clusters or bags, as is common in the use of k-means clustering.

where constants C_1, C_2 are independent of B.

In Appendix C 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 (2) over the set of all baggings amounts to the following optimization problem.

$$\begin{aligned} \min_{B\in\mathcal{B}} \quad & \sum_{l=1}^{m} \sum_{\tilde{y}_i \in B_l} (\tilde{y}_i - \mu_l)^2 \\ \text{subject to} \quad & |B_l| = k \quad \forall l \in [m] \end{aligned} \tag{3}$$

where μ_l is the mean of the labels in B_l, and $\mathcal{B}$ denotes the set of all baggings of the n samples. Note that the optimization involves use of $\tilde{y}$ which is unavailable, but one can instead use y as a proxy, leading to a small additional utility error of $\left(1 - \frac{1}{k}\right)\sigma^2$ (Appendix C). 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 (Appendix C). In Appendix E, we also justify that k-means clustering of the instances X is a good proxy for the k-means clustering of the labels y, leading to a label-agnostic bagging.

2.2. LLP, Bag-level loss

Definition 3 (Bag-level loss) *An estimator $\hat{\theta}$ minimizes bag-level loss, if*

$$\hat{\theta} := \operatorname*{argmin}_{\theta} \frac{1}{m} \sum_{l=1}^{m} \ell\left(\overline{y}_l, \frac{\sum_{i\in B_l} f_\theta(x_i)}{|B_l|}\right). \tag{4}$$

The loss is between the bag-label and mean of the instance level predictions of the bag instances. Below, we provide an upper bound on the utility for equal sized bags (we also show a corresponding result without the equality constraint in Appendix C).

Theorem 4 *For $\hat{\theta}$ as in (4), for a given bagging B such that $|B_l| = k, \forall l \in [m]$,*

$$\mathbb{E}\left[\|\hat{\theta} - \theta^*\|_2^2\right] \leqslant \sigma^2 \frac{m}{k} \left(\frac{\lambda_{max}(f(X))}{\lambda_{min}(f(X))}\right)^2, \tag{5}$$

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-centroid. The optimal bagging strategy involves minimizing the condition number (ratio of the maximum and minimum eigenvalue) of $f(X)$. In Appendix E, we justify that finding an optimal k-means clustering of the instances X is a good proxy for minimizing the condition number. In addition, in Appendix B.2.1, we show that even random bagging gives us a reasonable upper bound. Note that the optimal bagging strategy here does not involve knowledge of the labels, leading to equally good utility for label-agnostic and label-dependent bagging.

2.3. MIR, Aggregate-level loss

Definition 5 (Aggregate-level loss) *An estimator $\hat{\theta}$ minimizes aggregate-level loss, if*

$$\hat{\theta} := \operatorname*{argmin}_{\theta} \frac{1}{m} \sum_{l=1}^{m} \ell \left(\overline{y}_l, f_\theta \left(\frac{\sum_{i \in B_l} x_i}{|B_l|} \right) \right) \tag{6}$$

The loss is between the bag-label and prediction of the centroid of the bag instances. Below, we provide an upper bound on the utility for equal sized bags (we also show a corresponding result without the equality constraint in Appendix C.

Theorem 6 *For $\hat{\theta}$ in (6), given a bagging B such that $|B_l| = k, \forall l \in [m]$,*

$$\mathbb{E}\left[\|\hat{\theta} - \theta^*\|_2^2 \right] \leqslant C_1 \left(\frac{\lambda_{max}(f(X))}{\lambda_{min}(f(X))} \right)^2 \left(C_2 + \sum_{l=1}^{m} \sum_{\tilde{y}_i \in B_l} (\tilde{y}_i - \mu_l)^2 \right) \tag{7}$$

where constants C_1, C_2 are independent of B.

As in the case of bag-LLP, minimizing the first term in (7) corresponds to minimizing the condition number of $f(X)$, and minimizing the second term corresponds to finding the optimal k-means clustering of $\tilde{y}$. In Appendix E, we justify that finding an optimal k-means clustering of the instances X is an effective proxy for minimizing both the terms, providing a label-agnostic bagging. In Appendix F, we give an label-dependent bagging method which combines k-means over the labels, followed by random bagging step, that is also effective.

2.4. Privacy

In each of the previous scenarios, the aggregator can modify the bagging procedure to obtain formal label-differential privacy guarantees [8], defined below.

Definition 7 (Label DP) *A randomized algorithm A taking a dataset as an input is (ϵ, δ)-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,*

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

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. In the results below, we quantify the additional loss in utility that is incurred due to private bagging, for instance-MIR and bag-LLP. We discuss the corresponding result for aggregate-MIR in Appendix C, along with the proofs.

MIR, Instance-level loss

Theorem 8 *There exists a bagging B with $|B_l| = k, \forall l \in [m]$, satisfying (ϵ, δ) label-DP, such that for $\hat{\theta}$ in (1), we have*

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

where $\alpha^2 = \frac{16R^2 \log\left(\frac{1.25}{\delta/2}\right)}{\epsilon^2}$, 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 [4], we add $\mathcal{N}\left(0, \frac{\alpha^2}{k^2}\right)$ noise to the label value to ensure $(\frac{\epsilon}{2}, \frac{\delta}{2})$ label-DP, where $\alpha^2 = \frac{16R^2 \log\left(\frac{1.25}{\delta/2}\right)}{\epsilon^2}$, 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 C, 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 notion of central-DP.

LLP, Bag-level loss

Theorem 9 *There exists a bagging B with $|B_l| = k, \forall l \in [m]$, satisfying (ϵ, δ) label-DP, such that for $\hat{\theta}$ in (4), 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 $\alpha^2 = \frac{4R^2 \log\left(\frac{1.25}{\delta}\right)}{\epsilon^2}$, 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 (ϵ, δ) label-DP, leading to an additional error of $\frac{\alpha^2 m}{k^2}$ over the corresponding non-private bagging mechanism.

3. 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

[1] Apple storekit ad network. https://developer.apple.com/documentation/storekit/skadnetwork/.

[2] Private aggregation api of chrome privacy sandbox. https://developer.chrome.com/docs/privacy-sandbox/aggregation-service/.

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

[4] 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.

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

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

[11] 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.

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

[13] 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.

[14] Cynthia Dwork. Differential privacy. In *Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II 33*, pages 1–12. Springer, 2006.

[15] Hossein Esfandiari, Vahab Mirrokni, Umar Syed, and Sergei Vassilvitskii. Label differential privacy via clustering. In *International Conference on Artificial Intelligence and Statistics*, pages 7055–7075. PMLR, 2022.

[16] Badih Ghazi, Noah Golowich, Ravi Kumar, Pasin Manurangsi, and Chiyuan Zhang. Deep learning with label differential privacy. *Advances in neural information processing systems*, 34:27131–27145, 2021.

[17] Badih Ghazi, Pritish Kamath, Ravi Kumar, Ethan Leeman, Pasin Manurangsi, Avinash Varadarajan, and Chiyuan Zhang. Regression with label differential privacy. *arXiv preprint arXiv:2212.06074*, 2022.

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

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

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

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

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

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

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

[29] 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.

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

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

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

[33] 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.

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

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

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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 and Private Learning: Supplementary Materials

Outline Appendix A includes a detailed discussion of previous work. Appendix B contains the utility analysis for all the settings. In Appendix C, we present supporting proofs along with some additional results that were briefly mentioned in Appendix B. Experiments can be found in Appendix D. Appendix E justifies that k-means of the instances X is an effective label-agnostic bagging heuristic for each setting we consider (instance-MIR, bag-LLP, and aggregate MIR). In Appendix F, we discuss the super-bags algorithm for Agg-MIR which is a combination of label k-means and random bagging. In Appendix G, we generalize previous results to GLM's.

Appendix A. Related Work

Learning from aggregated labels. Learning from label proportions (LLP), in which the bag-labels are the average of the labels within the bag, started with the work of [12] and has been studied in the context of privacy concerns [26], lack of supervision due to cost [10], or coarse instrumentation [13]. While previous works [19, 20, 23, 29, 30, 39] have developed specialized techniques for model training on LLP training data, [40] defined it in the PAC framework, while [27, 28] have shown worst case algorithmic and hardness bounds, and recently [6] gave PAC learning algorithms for Gaussian feature vectors and random bags.
A related formulation is that of multiple instance regression (MIR), introduced in [24], where the bag-label is one of the (real-valued) labels within the bag (in contrast to LLP in which it is their average). For the most part, MIR has been studied in applied settings related to remote sensing and image analysis. Popular baseline techniques apply instance-level regression by assigning the bag-label to the average feature-vector in the bag, called aggregated-MIR, or assigning the bag-label to each feature-vector in the bag, known as instance-MIR [25, 37], whereas several expectation-maximization (EM) based methods have also been proposed [24, 33, 36–38]. Recent work of [9] 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 [1, 2, 22]. With the goal of preserving the utility of models trained on the aggregate labels, model training techniques for either randomly sampled [7] or curated bags [11] have been proposed. More recently, [18] showed that minimizing a natural instance-level loss for LLP yields the best utility when the bags are created by optimizing k-means objective defined over the constituent labels of the bags, for linear regression tasks. We note however, that such a treatment of the equally well used bag-loss method [3] for LLP is lacking, while for MIR this topic of optimal bag creation has not been studied.

Label Differential Privacy. Differential privacy (DP), by now a standard notion of privacy of algorithmic mechanisms, was introduced by [14]. In the context of training datasets, the restricted notion of label-differential privacy (label-DP) was provided by [8]. Recent works have provided label-DP mechanisms for classification [16] and regression [17] while [15] proposed clustering based mechanisms.

Appendix B. Utility analysis

B.1. MIR, Instance-level loss

We denote the uniform distribution by Γ. 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.

$$A_{(i,j)} = \begin{cases} 1 & \text{if } i \in B_l \text{ and } \overline{y}_l = y_j \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

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

$$S_{(i,j)} = \begin{cases} \frac{1}{|B_l|} & \text{if } i, j \in B_l \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

The minimizer of (1) is then given by

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \frac{1}{n} \|Ay - X\theta\|_2^2 = (X^T X)^{-1} X^T A y. \tag{10}$$

We now give a proof sketch for Theorem 2, providing an upper bound for the error of $\hat{\theta}$ (some details are omitted to Appendix B). All the expectations henceforth are over the randomness in A unless otherwise stated.

Proof (of Theorem 2) We begin with the following proposition, and use it to prove the main theorem

Proposition 10

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

Proof (of Proposition 10) By rearranging the terms,

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

γ 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,

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

■

We now upper bound the error in Proposition 10. We simplify the first term.

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

We simplify the RHS above with the following proposition.

Proposition 11

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

Proof

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

Putting the following two lemmas together, we conclude Proposition 11.

Lemma 12 $\mathbb{E}\left[||AX\theta^*||_2^2\right] = ||\tilde{y}||_2^2$.

Proof (of Lemma 12) Let $B(i)$ be the bag containing x_i. Note that $AX\theta^* = \left[\tilde{y}_{\Gamma(B(1))}, \ldots, \tilde{y}_{\Gamma(B(n))}\right]^T$

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

Then we have

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

■

Lemma 13 ${\theta^*}^TX^TSX\theta^* = \sum_{l=1}^{m}\frac{\left(\sum_{i\in B_l}\tilde{y}_i\right)^2}{|B_l|}$.

Proof (of Lemma 13). 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 10, and together with Proposition 11 concludes the proof of Theorem 2.

Proposition 14

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

■

B.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.

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

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

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \frac{1}{m} \|Sy - SX\theta\|_2^2. \tag{12}$$

Theorem 4 provides an upper bound on the error for equal sized bags, showing that

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

We want to develop a bagging algorithm that minimizes the condition number of the covariance of the bag-centroids. Since bounding the condition number as a whole is challenging, we instead find an upper bound for λ_{max} (Lemma 15) and a lower bound λ_{min} of $(SX)^T SX$. Aggregating feature vectors reduces the eigenvalues of the covariance matrix. We propose the following random bagging algorithm (Algorithm 1), which provides a lower bound for the $\lambda_{min}((SX)^T SX)$.

B.2.1. A RANDOM BAGGING APPROACH

Our bagging algorithm considers a fixed random partitioning strategy where the instances are divided into *super*-bags, each containing 2k instances. From each super bag, one k-sized bag is independently sampled, resulting in a collection of $m/2$ bags. We then analyze the minimum eigenvalue of the covariance matrix for this subset of bags. Since covariance matrices are PSD, the minimum eigenvalue of this subset of bags is a lower bound for any collection of m bags formed from the same instances, as adding more covariances will not reduce the minimum eigenvalue.

Lemma 15 $\lambda_{max}\left((SX)^T SX\right) \leqslant \lambda_{max}(X^T X)$.

Let X_l represent the feature matrices of B_l for $l \in [m]$.

$$\lambda_{min}\left((SX)^T SX\right) = \frac{1}{k^2}\lambda_{min}\left(\sum_{l=1}^{m} X_l^T X_l\right)$$

Input: : Instances $\mathcal{X}$, fixed bag size k.
Steps:

1. Randomly partition $\mathcal{X}$ into m' $2k$-sized *super*-bags, where $m' = n/2k$.

$$\mathcal{X} = \bigcup_{l=1}^{m'} \mathcal{X}_l \text{ and } \mathcal{X}_l \bigcap \mathcal{X}_{l'} = \phi \text{ for all } l \neq l'$$

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

Figure 1: Random bagging algorithm for bag-LLP

The feature matrix for bag B_l' sampled using Algorithm 1 can be represented by X_l' for all $l \in [m']$.

$$\frac{1}{k^2}\lambda_{min}\left(\sum_{l=1}^{m} X_l^T X_l\right) \geqslant \frac{1}{k^2}\lambda_{min}\left(\sum_{l=1}^{m} X_l'^T X_l'\right) \tag{13}$$

Let $\mu_{min} = \lambda_{min}\left(\sum_{l=1}^{m'} \mathbb{E}\left[X_l'^T X_l'\right]\right)/k^2$. We expand $X_l'^T X_l'$ and find μ_{min}:

$$\begin{aligned}\mu_{min} &= \frac{1}{k^2}\lambda_{min}\left(\sum_{l=1}^{m'} \mathbb{E}\left[\sum_{x_i,x_j \in B_l'} x_i x_j^T\right]\right)\\ &= \frac{1}{k^2}\lambda_{min}\left(\sum_{l=1}^{m'} \mathbb{E}\left[\sum_{x_i \in B_l'} x_i x_i^T\right] + \mathbb{E}\left[\sum_{i\neq j} x_i x_j^T\right]\right)\end{aligned}$$

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

$$\begin{aligned}
&\mu_{min} \\
&= \frac{\lambda_{min}}{k^2}\left(\sum_{l=1}^{m'} \sum_{x_i \in \mathcal{X}_l} \frac{1}{2} x_i x_i^T + \sum_{(x_i,x_j)\in\mathcal{X}_l} \frac{k-1}{2k-1} x_i x_j^T\right) \\
&= \frac{\lambda_{min}}{k^2}\left(\sum_{l=1}^{m'} \frac{1}{2}\left(1 - \frac{k-1}{2k-1}\right) \sum_{x_i \in \mathcal{X}_l} x_i x_i^T + \frac{k-1}{2(2k-1)} \hat{x}\hat{x}^T\right) \\
&= \frac{\lambda_{min}}{k^2}\left(\sum_{l=1}^{m'} \left(\frac{k}{2(2k-1)}\right) \sum_{x_i \in \mathcal{X}_l} x_i x_i^T + \frac{k-1}{2(2k-1)} \hat{x}\hat{x}^T\right) \\
&= \frac{\lambda_{min}}{2k^2(2k-1)}\left(kX^TX + (k-1)\sum_{l=1}^{m'} \hat{x}\hat{x}^T\right)
\end{aligned}$$

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

Lemma 16 $\lambda_{max}(X_l'^T X_l') \leqslant k\beta$.

Applying Matrix Chernoff (Corollary 5.2 [34]), we get

$$\mathbb{P}\left[\frac{1}{k^2}\lambda_{\min}\left(\sum_{l=1}^{m} X_l'^T X_l'\right) \leqslant (1-\delta)\mu_{\min}\right] \leqslant d \cdot \left[\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right]^{\mu_{\min}/k\beta}$$

Using Equation 13 we get

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

Using Lemma 15 and Equation (14), we get

$$\mathbb{E}\left[\|\hat{\theta} - \theta^*\|_2^2\right] \leqslant \frac{16\sigma^2 n k^2}{(1-\delta)^2}\left(\frac{\lambda_{max}(X^TX)}{\lambda_{min}(X^TX)}\right)^2. \tag{14}$$

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

B.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.

$$A_{(l,i)} = \begin{cases} 1 & \text{if } y_i = \Gamma(B_l), \\ 0 & \text{otherwise.} \end{cases} \tag{15}$$

We denote $\mathbb{E}[A] = S$. This turns out to be the same S as (11), and represents the instances in each bag. The minimizer of the bag-level loss is

$$\hat{\theta} = \operatorname*{argmin}_{\theta} \frac{1}{m} \|Ay - SX\theta\|_2^2. \tag{16}$$

Theorem 6 provides an upper bound on the error for equal sized bags, showing that

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

Appendix C. 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.

C.1. Additional results from Section 2.1

Lemma 17 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 18 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 19 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 17 *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 18 *Given $y_i = \tilde{y}_i + \gamma_i$, where $\gamma_i \sim \mathcal{N}(0, \sigma^2)$. Then, given a clustering B over y,*

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

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

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

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

Proof Follows from Lemma 2.3 in [18]. ■

C.2. Additional Proofs from Section B.1

Proof (of Lemma 14). We use the following inequality:

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

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

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

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

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

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

■

C.3. LLP, Bag-loss

Theorem *[full version of Theorem 4]*
For $\hat{\theta}$ as in (4), for a given bagging B with bagging matrix S, we have

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

For equal sized bags of size k, this simplifies to

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

Proof We start by proving the following lemma

Lemma 20

$$\mathbb{E}\left[\|\hat{\theta} - \theta^*\|_2^2\right] =\sigma^2\|((SX)^TSX)^{-1}(SX)^T(SS^T)^{1/2}\|_F^2 . \tag{17}$$

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

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

By rearranging the terms, we have

$$\begin{aligned}\hat{\theta} - \theta^* &= ((SX)^TSX)^{-1}X^TS^TSy - \theta^* \\ &= ((SX)^TSX)^{-1}X^TS^TSX\theta^* - \theta^* \\ &+ ((SX)^TSX)^{-1}X^TS^TS\gamma \\ &= ((SX)^TSX)^{-1}X^TS^TS\gamma\end{aligned}$$

Since γ is independent of X1, with $\mathbb{E}[\gamma] = 0$, and $\mathbb{E}[\gamma\gamma^T] = \sigma^2\mathcal{I}$, we have

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

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:

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

■

Now we upper bound the RHS.

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

■

C.4. MIR, Aggregate-loss

Theorem *[full version of Theorem 6]For $\hat{\theta}$ in (6), given a bagging B with bagging matrix S,*

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

For equal sized bags, this simplifies to

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

Proof

$$
\begin{aligned}
\hat{\theta} &= \operatorname*{argmin}_{\theta}\frac{1}{m}\|Ay-SX\theta\|_2^2 \\
&= (X^TS^TSX)^{-1}X^TS^TAy.
\end{aligned}
$$

By rearranging the terms, we have

$$
\begin{aligned}
\hat{\theta}-\theta^* &= ((SX)^TSX)^{-1}X^TS^TAy-\theta^* \\
&= ((SX)^TSX)^{-1}X^TS^TAX\theta^*-\theta^*+((SX)^TSX)^{-1}X^TS^TA\gamma
\end{aligned}
$$

γ 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 γ, A are independent. Hence,

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

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

Lemma 21

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

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.

■

C.5. Privacy

In this section, we quantify the additional loss in utility incurred due to label-DP guarantees, for each setting we consider (instance-MIR, bag-LLP, and aggregate MIR). We give full versions of the theorems stated in Section 2.4, along with the proofs.

C.5.1. MIR, INSTANCE-LEVEL

Theorem *[full version of Theorem 8]There exists a bagging B with $|B_l| = k, \forall l \in [m]$, satisfying (ϵ, δ) label-DP, such that for $\hat{\theta}$ in (1), we have*

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

where $\alpha^2 = \frac{16R^2 \log\left(\frac{1.25}{\delta/2}\right)}{\epsilon^2}$, 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 [4, 31], and the fact that $\epsilon << n$ in our setting, we add $\mathcal{N}\left(0, \frac{\alpha^2}{k^2}\right)$ noise to the bag-label value to ensure $\left(\frac{\epsilon}{2}, \frac{\delta}{2}\right)$ label-DP, where $\alpha^2 = \frac{16R^2 \log\left(\frac{1.25}{\delta/2}\right)}{\epsilon^2}$. 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 $\left(\frac{\epsilon}{2}, \frac{\delta}{2}\right)$ label-DP by post-processing. 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 18.

Now, we have two queries, each of which are $\left(\frac{\epsilon}{2}, \frac{\delta}{2}\right)$ label-DP, ensuring (ϵ, δ) 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 [5], 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 [21, 32], for the special case of a single dimension.

C.5.2. LLP, BAG-LEVEL

Theorem *[full version of Theorem 9]There exists a bagging B with $|B_l| = k, \forall l \in [m]$, satisfying (ϵ, δ) label-DP, such that for $\hat{\theta}$ in (4), we have*

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

where $\alpha^2 = \frac{4R^2 \log\left(\frac{1.25}{\delta}\right)}{\epsilon^2}$, 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 (ϵ, δ) label-DP, where $\alpha^2 = \frac{4R^2 \log\left(\frac{1.25}{\delta}\right)}{\epsilon^2}$. 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 $\sigma^2\frac{m}{k}$. ■

C.5.3. MIR, AGGREGATE-LEVEL

Theorem 6 shows that, for $\hat{\theta}$ in (6), given a bagging B, with equal sized bags, we have

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

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 (ϵ, δ) label-DP, where $\alpha^2 = \frac{4R^2 \log\left(\frac{1.25}{\delta}\right)}{\epsilon^2}$, 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 (ϵ, δ) 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.

Appendix D. 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 θ^* from the standard d-dimensional Gaussian distribution, sampling each of the rows of X iid from the standard d-dimensional Gaussian distribution and then setting $y = X\theta^* + \gamma$ where each coordinate of γ is iid drawn from $N(0, \sigma^2)$ where σ 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 also consider the private version of instance-MIR in Table 2. We set $\delta = 10^{-5}$, and vary ϵ. For each mechanism, we see that accuracy drops with a decrease in ϵ. However, the drop is sharper for label k-means, which is expected, 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 $\frac{1}{k}$.

D.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, and provide more details in Appendix E .

Instance-MIR Note that in our setting of linear regression, $\tilde{y} = \theta^*$. In other words, $\tilde{y}$ is just the projection of X along the axis normal to the hyperplane determined by θ^*. 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 θ^* 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 Appendix E , we justify that k-means of the instances X is a good heuristic for the same.

Bag-LLP We want to maximize $\lambda_{\min}((SX)^T SX)$, where $(SX)^T SX$ is the sample covariance matrix of the centroids of each bag. $\lambda_{\min}$ of a covariance matrix measures the variance along the corresponding eigenvector (which is also the direction of least variance). In Appendix E , 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. In order to maximize $\lambda_{\min}((SX)^T SX)$, we maximize variance in every direction. Equivalently, we want to reduce the k-means objective along every direction. In Appendix E , we justify that k-means of the instances X is a good heuristic for the same.

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

Table 1: Non-Private Bagging

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

Table 2: Private Bagging

Aggregate-MIR Note that in order to minimize the error bound, we must simultaneously minimize the condition number of $(SX)^T SX$, 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.

Next, we consider non-isotropic distributions. We generate datasets in the following way:

- *Isotropic*: We independently sample a set $\mathcal{X}$ containing n d-dimensional points from $\mathcal{N}(0, I)$.
- *Non-isotropic (Independent)*: We sample d independent values $\{\lambda_1, \cdots, \lambda_d\}$ from a uniform distribution $U(0.1, 10)$ to be the eigenvalues of the Σ, which is diagonal matrix.
- *Non-isotropic (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.

Once we have sampled feature vectors of the form $X \in \mathbb{R}^{n \times d}$, we sample a random groud truth model θ^* from the standard d-dimensional Gaussian distribution. This true model is then used to generate the true labels $\tilde{y}$. We add noise to $\tilde{y}$ to generate y. We set $y = X\theta^* + \gamma$ where each

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

Table 3: Aggregate-MIR

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

Table 4: Bag-LLP

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

Table 5: Instance-MIR

coordinate of γ is iid drawn from $N(0, \sigma^2)$ where σ is 0.5. We set n to be $50,000$ and d as 32. We also vary k, and use $k = 10, 50$. The result dataset is of the form $(X \in \mathbb{R}^{n \times d}, y \in \mathbb{R}^n)$

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 4) and aggregate-MIR (Table 3), 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 5), all the mechanisms show similar performance, with label k-means showing better performance in many cases.

Appendix E. 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, bag-LLP, and aggregate MIR).

E.1. MIR, Instance-level

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 θ^*. Hence, finding an optimal k-means clustering of $\tilde{y}$ is equivalent to minimizing the k-means objective of the projection of X along this axis. However, if the labels are not given, this axis is unknown, since θ^* 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. We now justify that k-means of the instances X is a good heuristic for the same. First, 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.

Lemma 22 *Consider an orthogonal basis $z_1, \dots z_d$. Fix a clustering S. We can show the following*

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

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 $X = \{X_1, \dots, X_n\}$.

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

■

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.,

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

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 S 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 Σ 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. A random bagging approach would also avoid the problem of directions with large variance dominating for a non-isotropic distribution. However, the k-means objective in every direction will be that of a random clustering, which is sub-optimal.

E.2. LLP, Bag-level

Given a bagging with bagging matrix S, SX is the matrix representing the dataset consisting of the centroids of each bag. We want to maximize $\lambda_{\min}((SX)^T SX)$, where $(SX)^T SX$ is the sample covariance matrix of SX. $\lambda_{\min}$ is the variance along the direction of the corresponding eigen vector (which is also the direction of least variance of dataset SX). We now show that finding a bagging S maximizing the variance of SX along a direction is equivalent to finding an optimal k-means clustering of X projected on that direction.

Lemma 23 *Consider a direction z, and a centred dataset X. Given a bagging S over X with m bags of equal size k,*

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

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 μ_l be the mean of B_l. The variance of the SX along z is

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

■

Earlier, we showed 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. We want to find S such that k-means$(S(X_{z_{\text{min}}}))$ is small along z_{min}, where z_{min} is the direction of least variance of SX. But, since we do not know z_{min}, we find S such that k-means$(S(X_z))$ is small along every direction z. In the previous section, we give instance k-means heuristics for this.

E.3. MIR, Aggregate-level

Note that in order to minimize the error bound, we must simultaneously minimize the condition number of $(SX)^T SX$, 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.

Appendix F. Random Bagging Algorithm for Aggregate-MIR

We propose a random bagging algorithm similar to the one for Bag-LLP (Algorithm 1) for Agg-MIR. The upper bound for Agg-MIR (Theorem 6) is product of the label k-means objective and the condition number. We propose the following algorithm which takes both these objectives into account.

```
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 $m'$ contiguous super-bags of sizes $2k$, where $m' = n/2k$.
        $$\mathcal{X} = \bigcup_{l=1}^{m'} \mathcal{X}_l \text{ and } \mathcal{X}_l \bigcap \mathcal{X}_{l'} = \phi \text{ for all } l \neq l'$$
  3. For $l = 1, \ldots, m'$, a $k$-sized bag $B'_l$ is sampled $u.a.r$ from $\mathcal{X}_l$.
  4. Output $\mathcal{B}'$ where $\mathcal{B}' = \{B'_l\}_{l \in [m']}$
```

Figure 2: Random bagging algorithm for Agg-MIR

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 B.2.1, 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 B.2.1, we get,

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

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

Theorem 24 *For super-bags B'_l as defined in Algorithm 2 with arbitrary non-overlapping partitions $B_l^{(1)}$ and $B_l^{(2)}$, we have*

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

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

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

where, μ denotes the respective cluster means.

Proof

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

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

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

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

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

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.

■

Appendix G. Analysis for GLMs

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.

G.1. MIR

We now generalize our derivation to the class of *generalized linear models* (GLMs). The instance-level labels y_i are conditionally independent given 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 η_i is a location variable and ϕ 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 $\eta_i = x_i^T \theta^*$ for an unknown model θ^*. 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 θ. Our objective is to find a bagging strategy which closes the gap between the true model θ^* and $\hat{\theta}$. For GLMs we achieve this by minimizing the gradient of the loss at θ^*.

G.1.1. ANALYSIS OF INSTANCE-LEVEL LOSS FOR MIR

Lemma 25 *Suppose that the loss $\mathcal{L}$ is strongly convex with parameter μ and $\hat{\theta} = \arg\min_\theta \mathcal{L}(\theta)$. Then, for any model θ^*, we have*

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

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

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

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 25 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 26 *If we consider canonical GLMs with $\eta_i = x_i^T\theta^*$, then we have*

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

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

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

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

■

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 $\mu_i = \mathbb{E}[y_i|x_i] = b'(x_i^T\theta^*)$.

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

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.

G.1.2. ANALYSIS OF AGGREGATE MIR LOSS

Let $\hat\theta$ be the minimizer of the aggregate MIR loss:

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

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 25).

Theorem 27 *If we consider canonical GLMs with $\eta_i = x_i^T\theta^*$, then we get*

$$\mathbb{E}\left[\|\nabla\mathcal{L}(\theta^*)\|_2\right] \leqslant n\lambda_{max}(X^TX)\left(m(\|b'(X\theta^*)\|_2^2 + \|Db''(X\theta^*)\|_1) + \|Sb'(X\theta) - b'(SX\theta)\|_2^2 - \|Sb'(X\theta)\|_2^2\right) \tag{26}$$

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

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

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

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

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

■

We now show how the final objective in Equation 27 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 $b'(t_1) \geqslant b'(t_2) \iff t_1 \geqslant t_2$.

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

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

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

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

We bound the second term:

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

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

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

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

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.