PriorBoost: An Adaptive Algorithm for Learning from Aggregate Responses

Adel Javanmard [*1,2] **Matthew Fahrbach** [*2] **Vahab Mirrokni** [2]

Abstract

This work studies algorithms for learning from aggregate responses. We focus on the construction of aggregation sets (called *bags* in the literature) for event-level loss functions. We prove for linear regression and generalized linear models (GLMs) that the optimal bagging problem reduces to one-dimensional size-constrained k-means clustering. Further, we theoretically quantify the advantage of using curated bags over random bags. We then propose the `PriorBoost` algorithm, which adaptively forms bags of samples that are increasingly homogeneous with respect to (unobserved) individual responses to improve model quality. We study label differential privacy for aggregate learning, and we also provide extensive experiments showing that `PriorBoost` regularly achieves optimal model quality for event-level predictions, in stark contrast to non-adaptive algorithms.

1. Introduction

In supervised learning, the learner is given a training dataset of n i.i.d pairs $(\boldsymbol{x}_i, y_i)$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ is a feature vector and y_i is the corresponding response. Responses are real-valued for regression problems, and belong to a finite discrete set for multi-class classification. The fundamental problem in supervised learning is to (1) train a model with this data, and (2) use this model to infer the response/label of unseen test instances. However, in many practical applications (e.g., medical tests and elections), the responses contain sensitive information, but the features are far less sensitive (e.g., demographic information or zip codes/regions). In such applications, there are valid concerns about revealing *individual responses* to the learning algorithm, even if it is a trusted party.

A popular approach to mitigate this privacy concern in practice is to let the learner access responses in an *aggregate manner*. In the framework of *learning from aggregate responses* (LAR), the learner is given access to a collection of unlabeled feature vectors called *bags* and an aggregate summary of the responses in each bag. A widely used choice is the mean response or label proportions of each bag (Yu et al., 2014). The learner then fits a model using the aggregate responses with the goal of accurately *predicting individual responses* on future data.

The problem of learning from aggregate responses (a.k.a. learning from label proportions in the context of classification) dates back to at least Wein & Zenios (1996) in the context of group testing, a technique used in many different fields including medical diagnostics, population screening, and quality control. The idea is to combine multiple samples into a group and test them together rather than individually. This approach has been widely adopted in cases where testing resources are limited or the prevalence of the condition being tested for is low. LAR has also been studied in other earlier work (de Freitas & Kück, 2005; Musicant et al., 2007; Quadrianto et al., 2008; Rueping, 2010; Patrini et al., 2014) for settings where direct access to the individual responses is not possible (e.g., in political party elections where aggregate votes are only available at discrete district levels).

Recently, there has been a resurgence in the LAR framework primarily due to the rise of privacy concerns; see (Scott & Zhang, 2020; Saket, 2022; Zhang et al., 2022; Busa-Fekete et al., 2024; Chen et al., 2023; Brahmbhatt et al., 2023; Javanmard et al., 2024; Li et al., 2024) for a non-exhaustive list. Specifically, if the aggregation bags are large enough and have no (or little) overlap, revealing only the aggregate responses provides a layer of privacy protection, often formalized in terms of k-*anonymity* (Sweeney, 2002). Large tech companies have recently deployed aggregate learning frameworks, including Apple's SKAdNetwork library (Kollnig et al., 2022) and the Private Aggregation API in the Google Privacy Sandbox (Geradin et al., 2020). Aggregate responses can further be perturbed to provide label differential privacy (Chaudhuri & Hsu, 2011), a popular notion of privacy that measures the leakage of personal label/response information, which we discuss in detail in Section 6.

In some applications, the bagging configurations are naturally determined by the problem at hand (e.g., in the voting

[*]Equal contribution [1]University of Southern California [2]Google Research. Correspondence to: Adel Javanmard <ajavanma@usc.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

example above the bags are defined based on districts). In other applications, however, the learner has the flexibility of curating bags of query samples to maximize model utility while complying with privacy or legal constraints imposed by the data regulators. Our work focuses on the problem of *bag curation* in the framework of learning from aggregate responses.

1.1. Problem statement

We first describe the process of learning from aggregate responses, for a given collection of bags. Consider a partition of n samples into m non-overlapping bags, each of size at least k, for a prespecified k (and hence $n \geq mk$). We focus on training a model by minimizing the following *event-level loss*:

$$\widehat{\boldsymbol{\theta}} := \arg\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{\ell=1}^{m} \sum_{i \in B_\ell} \mathcal{L}(\overline{y}_\ell, f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)) \,, \tag{1}$$

where B_ℓ is the set of samples in bag ℓ and $\overline{y}_\ell$ is the mean response in bag ℓ. In words, with this approach the model is learned by fitting individual predictions to the average response of its bag.

The problem of bag curation is to find an optimal bagging configuration that maximizes model utility (in terms of minimizing estimation error), while satisfying the minimum bag size constraint $|B_\ell| \geq k$. Note that this min-size constraint implies k-anonymity in the sense of that any response in the (aggregate) dataset is shared by at least k individuals. Larger values of k offer higher protection of individual responses.

1.2. Overview of our approach and contributions

This work focuses on event-level loss and the problem of bag curation. To control privacy leakage, we require the bags to be non-overlapping and of size at least k. An important property of our mechanism is the following: The learner *never sees an individual response*. Conceptually, the learner always constructs a query of fresh samples to send to an oracle, who then returns the aggregate response.

Our key insight is to leverage available prior information about $\mathbb{E}[y \mid \boldsymbol{x}]$ to construct better bags for the learner. Such prior information can be based on domain knowledge, models trained on public data, or even *previous iterations of an aggregate learning algorithm*.

We summarize our contributions as follows.

- **Reduction to size-constrained k-means clustering.** We first present our method assuming access to a prior. We start with linear regression and characterize the dependence of the model estimation error on the bag construction. We then show that finding optimal bags reduces to a one-dimensional size-constrained k-means clustering problem that involves prior information on the *expected* response of samples. In Section 3, we then extend our derivations to the family of generalized linear models.

- **Advantage over random bagging.** In Section 4, we theoretically demonstrate the improvement of our bagging approach over schemes that construct bags independently of data (including random bagging).

- **Iterative prior-boosting algorithm.** In Section 5, we propose an adaptive algorithm called `PriorBoost`, which constructs a good prior from the aggregate data itself. It can be used even in settings where no public prior distribution is available. `PriorBoost` partitions the training data across multiple stages: it start with random bagging, and then *iteratively refines the prior* by constructing more consistent bags on the remaining data.

- **Differentially private LAR.** In Section 6, we propose a mechanism that adds Laplace noise to aggregate responses to ensure label differential privacy. We observe an intriguing tradeoff on the choice of minimum bag size k. On the one hand, larger k implies less sensitivity of aggregate responses to individual substitution and hence less noise is needed to ensure privacy. On the other hand, smaller k results in smaller bias of the trained model. The optimal choice of k (for a fixed privacy budget ε) depends on how these two effects contribute to the model test loss. We showcase this tradeoff empirically and discuss how the optimal k varies with the sample size n, features dimension d, and bag construction algorithm.

- **Experiments.** We study `PriorBoost` through extensive experiments in Section 7. This includes a comparison with random bagging for linear and logistic regression tasks, as well as a careful exploration into label differential privacy with Laplace noise for different privacy budgets.

1.3. Other related work

An active line of work in LAR is centered around the design of new loss functions. In addition to the the event-level loss in (1), another popular choice is *bag-level loss* (or aggregate likelihood), which measures the mismatch between the aggregate responses $\overline{y}_\ell$ and the *average* model predictions $1/|B_\ell| \sum_{i \in B_\ell} f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$ across bags $\ell \in [m]$ (Rueping, 2010; Yu et al., 2014). Javanmard et al. (2024) study the statistical properties of both losses and show that for quadratic loss functions $\ell(x, y) = (x - y)^2$, the event-level loss can be seen as a regularized form of the bag-level loss. They propose a novel interpolating loss that optimally adjusts the strength of the regularization.

It is worth noting that in many large-scale production ML systems, models are often trained online (Anil et al., 2022; Fahrbach et al., 2023; Coleman et al., 2024), and event-level loss is more amenable to online optimization. A separate system can be in charge of bagging and generating aggregate responses without the learner needing to know the bagging structure. In contrast, bag-level loss minimization requires computing average predictions for each bag, making it more challenging to implement, especially with mini-batch SGD where all samples in a bag must be in the same batch.

(Li et al., 2024) studies the problem of learning from label proportions and various learning rules that acheive PAC learning guarantees for classification loss. It also proposes novel debiasing techniques to achieve optimistic rates in both the realizable and agnostic setting.

We note that the works discussed above mainly consider random bagging. Closer to our goal, Chen et al. (2023) study the problem of bag curation, but they take a different approach than ours by grouping samples by common features instead of predicted response values.

2. Warm-up: Linear regression

The high-level intuition behind our approach is that useful bagging configurations are ones where aggregate responses are close to their individual responses. This allows for the estimator to be close to the *empirical risk minimizer* (ERM), similar to teacher-student knowledge distillation (Hinton et al., 2015). Our goal is therefore to use available predictions $\widetilde{y} \approx \mathbb{E}[y \mid \boldsymbol{x}]$ based on prior information to construct better bags for the aggregate learner.

To illustrate this idea, we start with a linear regression setup where response y_i is generated as

$$y_i = \boldsymbol{x}_i^\intercal \boldsymbol{\theta}^* + \varepsilon_i\,, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2). \tag{2}$$

The design matrix is $\boldsymbol{X} = \begin{bmatrix}\boldsymbol{x}_1 & \dots & \boldsymbol{x}_n\end{bmatrix}^\intercal \in \mathbb{R}^{n\times d}$, the response vector is $\boldsymbol{y} = (y_1, \dots, y_n)^\intercal$, and the noise vector is $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^\intercal$. We assume $\boldsymbol{\varepsilon}$ is independent of $\boldsymbol{X}$, and that $\mathbb{E}[\boldsymbol{\varepsilon}] = \boldsymbol{0}$ and $\mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\intercal] = \sigma^2 \boldsymbol{I}$. Letting m denote the number of bags, we encode the assignment of samples to bags with a matrix $\boldsymbol{S} \in \mathbb{R}^{m\times n}$, where

$$\boldsymbol{S}_{\ell,i} = \begin{cases} \frac{1}{\sqrt{|B_\ell|}} & \text{if } i \in B_\ell, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Consider the event-level loss minimizer of (1) with $\mathcal{L}$ being least squares loss, which we can write as

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \frac{1}{n} \|\boldsymbol{S}^\intercal \boldsymbol{S}\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2\,. \tag{4}$$

2.1. Bounding the estimator error

Our next result characterizes the error of this estimator. All proofs in this section are deferred to Appendix A.

Theorem 2.1. *If the design matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ has rank d, then for the estimator $\widehat{\boldsymbol{\theta}}$ given by Eq. (4), we have*

$$\mathbb{E}\Big[\, \Big\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\Big\|_2^2 \;\Big|\; \boldsymbol{X}\Big] = \big\|(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal(\boldsymbol{S}^\intercal\boldsymbol{S} - \boldsymbol{I})\boldsymbol{X}\boldsymbol{\theta}^*\big\|_2^2 + \sigma^2 \big\|(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\boldsymbol{S}^\intercal\big\|_{\mathrm{F}}^2\,. \tag{5}$$

An optimal bagging configuration (in the sense of minimizing the estimation error) is one whose matrix $\boldsymbol{S}$ minimizes (5) among all feasible partitions. The first term of the right-hand side is the (conditional) bias of $\widehat{\boldsymbol{\theta}}$ and the second term is its variance. As we can see, the choice of $\boldsymbol{S}$ affects both terms.

Instead of solving for an optimal $\boldsymbol{S}$, which can be challenging due to its partition structure, we first develop an upper bound on the error, and then we minimize this bound over $\boldsymbol{S}$ to give guidance on how to design aggregation bags.

Corollary 2.2. *The estimation error $\mathbb{E}[\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2^2 \mid \boldsymbol{X}]$ in Eq. (5) is upper bounded by*

$$\big\|(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\big\|_{\mathrm{op}}^2 (\|(\boldsymbol{S}^\intercal\boldsymbol{S} - \boldsymbol{I})\boldsymbol{X}\boldsymbol{\theta}^*\|_2^2 + \sigma^2 \min(m, d)).$$

2.2. Reducing to size-constrained k-means clustering

Next observe that $\boldsymbol{I} - \boldsymbol{S}^\intercal\boldsymbol{S}$ is a projection matrix given by

$$(\boldsymbol{I} - \boldsymbol{S}^\intercal\boldsymbol{S})_{i,j} = \begin{cases} 1 - \frac{1}{|B_\ell|} & \text{if } i,j \in B_\ell \text{ and } i = j, \\ -\frac{1}{|B_\ell|} & \text{if } i,j \in B_\ell \text{ and } i \neq j, \\ 0 & \text{otherwise.} \end{cases}$$

Specifically, $\boldsymbol{I} - \boldsymbol{S}^\intercal\boldsymbol{S}$ is the projection onto the space of vectors that have zero mean within each bag.

Let $\widetilde{y}_i := \mathbb{E}[y_i \mid \boldsymbol{x}_i] = \boldsymbol{x}_i^\intercal\boldsymbol{\theta}$ be the conditional expected response of sample $\boldsymbol{x}_i$ according to the prior model $\boldsymbol{\theta} \in \mathbb{R}^d$. Letting $\widetilde{\boldsymbol{y}} = (\widetilde{y}_1, \dots, \widetilde{y}_n)$, we then have

$$\|(\boldsymbol{I} - \boldsymbol{S}^\intercal\boldsymbol{S})\widetilde{\boldsymbol{y}}\|_2^2 = \sum_{\ell=1}^{m} \sum_{i\in B_\ell} (\widetilde{y}_i - \mu_\ell)^2\,, \tag{6}$$

where $\mu_\ell = \frac{1}{|B_\ell|}\sum_{i\in B_\ell} \widetilde{y}_i$ is the mean of the entries of $\widetilde{\boldsymbol{y}}$ in bag ℓ. Observe that (6) is the one-dimensional k-means objective.

To summarize, let $\mathcal{B}$ denote the set of all partitions of the n samples. Minimizing the upper bound in Corollary 2.2 over the set of non-overlapping bags of size at least k amounts to the following optimization problem:

$$\begin{aligned} \min_{(B_1,\dots,B_m)\in\mathcal{B}} \quad & \sum_{\ell=1}^{m} \sum_{i\in B_\ell} (\widetilde{y}_i - \mu_\ell)^2 + \sigma^2 \min(m, d) \\ \text{subject to} \quad & |B_\ell| \geq k \quad \forall \ell \in [m] \end{aligned}$$

This problem exhibits an interesting tradeoff with the number of bags m. The first term in the objective is the *bias of the estimator* $\widehat{\boldsymbol{\theta}}$, which measures the within-bag deviation of $\widetilde{\boldsymbol{y}}$. If we require larger bags (and hence a smaller m), this term increases since there will be more heterogeneity within bags. Decreasing m, however, reduces the second term in the objective, which is the *variance of the estimator* $\widehat{\boldsymbol{\theta}}$. The reason is that the aggregate responses $\overline{y}_\ell$ are averaged across larger bags and thus have lower variance. This reduction in the variance of the aggregated responses corresponds to a reduction in the estimator variance.

Focusing on the case where $m \geq d$, we can drop the second term in the objective to get the following one-dimensional k-means clustering problem with minimum size constraints:[1]

$$\begin{aligned} \min_{(B_1,\ldots,B_m)\in\mathcal{B}} \quad & \sum_{\ell=1}^{m} \sum_{i\in B_\ell} (\widetilde{y}_i - \mu_\ell)^2 \\ \text{subject to} \quad & |B_\ell| \geq k \quad \forall \ell \in [m] \end{aligned} \tag{7}$$

The next result establishes a structural property about optimal solutions to this problem.

Lemma 2.3 (Sorting structure)**.** *Consider the optimization problem* (7) *and sort the values* $\widetilde{y}_i$ *in non-increasing order as* $\widetilde{y}_{(1)} \geq \cdots \geq \widetilde{y}_{(n)}$*. There exists an optimal solution* $\{B_\ell^* : \ell \in [m]\}$ *with the following property: if* $\widetilde{y}_{(i)}$ *and* $\widetilde{y}_{(j)}$ *are in a bag* B_ℓ^**, then* $\widetilde{y}_{(k)} \in B_\ell^*$ *for all* $k \in \{i, i+1, \ldots, j\}$*.*

We discuss the algorithmic consequences of Lemma 2.3 in more detail in Section 5.

3. Extension to GLMs

We next extend our derivation to the family of *generalized linear models* (GLMs). In a GLM, the response variables y_i are conditionally independent given $\boldsymbol{x}_i$, and generated from a particular distribution in the exponential family where the log-likelihood function is written as:

$$\log p(y_i \mid \eta_i, \phi) = \frac{y_i \eta_i - b(\eta_i)}{a_i(\phi)} + c(y_i, \phi)\,, \tag{8}$$

where η_i is the location parameter and ϕ is the scale parameter. The functions $a_i(\cdot)$, $b(\cdot)$, and $c(\cdot,\cdot)$ are known. It is sometimes assumed that $a_i(\phi)$ has the form $a_i(\phi) = \phi/w_i$, where w_i is a known prior weight. We consider canonical GLMs, in which the location parameter has the form $\eta_i = \boldsymbol{x}_i^\intercal \boldsymbol{\theta}^*$ for an unknown model parameter $\boldsymbol{\theta}^*$. GLMs include several well-known statistical models, including linear regression, logistic regression, and Poisson regression.

[1] More accurately, this is a one-dimensional m-means clustering problem with size constraints. We use k to denote the minimum bag size to agree with the notion of k-anonymity.

Let $\widehat{\boldsymbol{\theta}}$ be the minimizer of the event-level loss in (1) with $\mathcal{L}$ the negative log-likelihood. Concretely,

$$\begin{aligned} \widehat{\boldsymbol{\theta}} &= \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) \\ &:= \arg\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{\ell=1}^{m} \sum_{i\in B(\ell)} \frac{\overline{y}_\ell \boldsymbol{x}_i^\intercal \boldsymbol{\theta} - b(\boldsymbol{x}_i^\intercal \boldsymbol{\theta})}{a_i(\phi)}\,, \end{aligned} \tag{9}$$

where we drop the term $c(y_i, \phi)$ as it does not depend on $\boldsymbol{\theta}$.

By the optimality of $\widehat{\boldsymbol{\theta}}$, we have $\nabla \mathcal{L}(\widehat{\boldsymbol{\theta}}) = \mathbf{0}$. Our goal is to find a bagging configuration that makes $\widehat{\boldsymbol{\theta}}$ close to the ground truth model $\boldsymbol{\theta}^*$. A natural approach towards this goal is to make the gradient of the loss at $\boldsymbol{\theta}^*$ small. As we show in Lemma B.2, for strongly convex losses, the estimation error $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2$ can be controlled by $\|\nabla \mathcal{L}(\boldsymbol{\theta}^*)\|_2$.

Our next result characterizes the norm of the loss gradient at $\boldsymbol{\theta}^*$, connecting it to the bagging matrix $\boldsymbol{S}$. Throughout, we use the following convention: For a function $f : \mathbb{R} \to \mathbb{R}$, when f is applied to a vector, it is applied to each entry of that vector, i.e., $f(\boldsymbol{v}) = (f(v_1), \ldots, f(v_n))$.

Theorem 3.1. *Consider the GLM family in* (8) *with canonical link functions* ($\eta_i = \boldsymbol{x}_i^\intercal \boldsymbol{\theta}^*$)*. For negative log-likelihood loss in* (9)*, we have*

$$\begin{aligned} \mathbb{E}\Big[\|\nabla \mathcal{L}(\boldsymbol{\theta}^*)\|_2^2 \mid \boldsymbol{X} \Big] &= \big\| \boldsymbol{X}^\intercal \boldsymbol{D}^{-1} (\boldsymbol{S}^\intercal \boldsymbol{S} - \boldsymbol{I}) b'(\boldsymbol{X}\boldsymbol{\theta}^*) \big\|_2^2 \\ &\quad + \Big\| \boldsymbol{X}^\intercal \boldsymbol{D}^{-1} \boldsymbol{S}^\intercal \boldsymbol{S} \boldsymbol{D}^{1/2} \mathrm{diag}(b''(\boldsymbol{X}\boldsymbol{\theta}^*))^{1/2} \Big\|_{\mathrm{F}}^2\,, \end{aligned} \tag{10}$$

where $\boldsymbol{D} = \mathrm{diag}(\{a_i(\phi)\})$.

We defer all proofs in this section to Appendix B. Note that $\mathbb{E}[y \mid \boldsymbol{x}] = b'(\boldsymbol{x}^\intercal \boldsymbol{\theta}^*)$ and $\mathrm{Var}(y \mid \boldsymbol{x}) = a(\phi) b''(\boldsymbol{x}^\intercal \boldsymbol{\theta}^*)$ are available from the given prior and therefore, in principle, the right-hand side of (10) can be minimized over the choice of bagging matrix $\boldsymbol{S}$.

However, similar to the case of linear regression, we start by upper bounding (10), and then we minimize this upper bound over the choice of $\boldsymbol{S}$. This provides guidance for how to construct the bags, and is easier to compute while being more interpretable.

Corollary 3.2. *Define* $\mu_i := \mathbb{E}[y_i \mid \boldsymbol{x}_i] = b'(\boldsymbol{x}_i^\intercal \boldsymbol{\theta}^*)$ *and* $v_i := \mathrm{Var}(y_i \mid \boldsymbol{x}_i) = a_i(\phi) b''(\boldsymbol{x}_i^\intercal \boldsymbol{\theta}^*)$*, and let their vector forms be* $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ *and* $\boldsymbol{v} = (v_1, \ldots, v_n)$*. Then,*

$$\begin{aligned} \mathbb{E}\Big[\|\nabla \mathcal{L}(\boldsymbol{\theta}^*)\|_2^2 \mid \boldsymbol{X} \Big] &\leq \big\| \boldsymbol{X}^\intercal \boldsymbol{D}^{-1} \big\|_{\mathrm{op}}^2 \\ &\quad \cdot \Big\{ \|(\boldsymbol{S}^\intercal \boldsymbol{S} - \boldsymbol{I})\boldsymbol{\mu}\|_2^2 + \min\Big(\sum_{\ell=1}^{m} \sum_{i\in B_\ell} \frac{v_i}{|B_\ell|}, d \|\boldsymbol{v}\|_\infty \Big) \Big\}\,. \end{aligned} \tag{11}$$

In the case of linear regression, we have $v_i = \sigma^2$, so the term involving v_i becomes $\sigma^2 \min(m, d)$ like in Corollary 2.2, which only depends on the number of bags.

Further, if $m/d \geq \max(v_i)/\min(v_i)$, the min term in (11) is achieved by $d\|\boldsymbol{v}\|_\infty$, so this term can be dropped from the objective, bringing us to the familiar size-constrained clustering problem:

$$\begin{aligned} \min_{(B_1,\ldots,B_m)\in\mathcal{B}} \quad & \sum_{\ell=1}^{m}\sum_{i\in B_\ell}(\mu_i-\overline{\mu}_\ell)^2 \\ \text{subject to} \quad & |B_\ell|\geq k \quad \forall \ell\in[m] \end{aligned} \tag{12}$$

We conclude by showing that we can drop the variance term from the bound in (11) for logistic and Poisson regression, i.e., that (12) is the correct objective function.

Logistic regression. In this case we have $y \in \{0,1\}$, so the log-likelihood becomes:

$$\log p(y \mid \eta) = y\eta - \log(1+e^\eta)\,,$$

which corresponds to $b(\eta)=\log(1+e^\eta)$, $a(\phi)=1$, and $c(y,\phi)=0$. Therefore, $\mu = b'(\eta) = 1/(1+e^{-\eta})$ and $v = e^\eta/(1+e^\eta)^2$. Then, for any $i,j\in[n]$, we have

$$\begin{aligned} \frac{v_i}{v_j} &= e^{\eta_i-\eta_j}\left(\frac{1+e^{\eta_j}}{1+e^{\eta_i}}\right)^2 \leq e^{\eta_i-\eta_j}e^{2(\eta_j-\eta_i)_+} \\ &\leq e^{|\eta_i-\eta_j|} \leq e^{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2\|\boldsymbol{\theta}^*\|_2}, \end{aligned}$$

where we used $\eta_i = \boldsymbol{x}_i^\intercal\boldsymbol{\theta}^*$. Therefore, if $\|\boldsymbol{x}_i\|_2 \leq B$, we have $\max(v_i)/\min(v_i) \leq \exp(2B\|\boldsymbol{\theta}^*\|_2)$, so for $m/d \geq \exp(2B\|\boldsymbol{\theta}^*\|_2)$, we can drop the variance term from the objective function.

Poisson regression. In this case we have $y\in\mathbb{Z}_{\geq 0}$, so the log-likelihood reads as:

$$\log p(y\mid\eta) = y\eta - e^\eta - \log(y!)\,,$$

which corresponds to $b(\eta)=e^\eta$, $c(y,\phi)=-\log(y!)$, and $a(\phi)=1$. Thus, $\mu=b'(\eta)=e^\eta$ and $v=a(\phi)b''(\eta)=e^\eta$. Then, similar to the previous example, $\max(v_i)/\min(v_i)\leq \exp(2B\|\boldsymbol{\theta}^*\|_2)$ and so for $m/d\geq\exp(2B\|\boldsymbol{\theta}^*\|_2)$, we can drop the variance term from the objective function.

4. Comparison with random bagging

We now theoretically justify the benefit of our prior-based bagging approach for aggregate learning compared to random bagging by proving a separation in the estimator error for linear models. An analogous but more involved analysis can also be carried out for GLMs. Before we present our results, we neet to establish some definitions and state our assumptions.

Definition 4.1. A random variable X is η-subgaussian if $\mathbb{E}[\exp(X^2/\eta^2)]\leq 2$. A random vector $\boldsymbol{x}$ is η-subgaussian if all of the one-dimensional marginals are η-subgaussian, i.e., $\boldsymbol{x}^\intercal\boldsymbol{v}$ is η-subgaussian for all $\boldsymbol{v}$ with $\|\boldsymbol{v}\|_2=1$.

Some examples of subgaussian random variables include Gaussian, Bernoulli, and all bounded random variables.

Assumption 4.2. The features vectors $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\in\mathbb{R}^d$ are drawn i.i.d from a centered κ-subgaussian distribution with covariance matrix $\boldsymbol{\Sigma}:=\mathbb{E}[\boldsymbol{x}_i\boldsymbol{x}_i^\intercal]\in\mathbb{R}^{d\times d}$.

Assumption 4.3. We consider an asymptotic regime where the sample size n and the features dimension d both grow to infinity. We assume that the eignevalues of $\boldsymbol{\Sigma}$ remain bounded and also away from zero in this asymptotic regime, i.e., $\sigma_{\min}(\boldsymbol{\Sigma})\geq C_{\min}>0$ and $\|\boldsymbol{\Sigma}\|_{\text{op}}\leq C_{\max}<\infty$ for some constants $C_{\min}$ and $C_{\max}$.

Our first theorem upper bounds the estimator error when the bags are formed using the ground truth model $\boldsymbol{\theta}^*$.

Theorem 4.4. *Consider the linear model (2) under Assumptions 4.2 and 4.3. Suppose that the dimension d and the sample size n grow to infinity and $n=\Omega(d)$. For the bagging matrix $\boldsymbol{S}$ constructed by solving problem (7), the following holds true with probability at least $1-1/n-2e^{-cd}$,*

$$\mathbb{E}\Big[\Big\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\Big\|_2^2\;\Big|\;\boldsymbol{X}\Big]\leq C\left(\frac{k\log(n)\|\boldsymbol{\theta}^*\|_2^2+\sigma^2 d}{n\sigma_{\min}(\boldsymbol{\Sigma})}\right),$$

for some constants $c,C>0$ that depend only on the subgaussian norm κ.

Out next result lower bounds the estimator error when the bags are chosen *independently of the data*. This applies to random bags as a special case.

Theorem 4.5. *Consider the linear model (2) under Assumptions 4.2 and 4.3. Suppose the dimension d and the sample size n grow to infinity and $n=\Omega(d^2\log d)$. If the bags are constructed independent of data and each of size k, the following holds true with probability at least $1-2e^{-c_1 d}-2d^{-c}$,*

$$\begin{aligned} \mathbb{E}\Big[\Big\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\Big\|_2^2\;\Big|\;\boldsymbol{X}\Big] \geq \Bigg[&\left(1-\frac{1}{k}-\frac{Cd\sqrt{\log d}}{\sigma_{\min}(\boldsymbol{\Sigma})\sqrt{n}}\right)^2\|\boldsymbol{\theta}^*\|_2^2 \\ &+\frac{\sigma^2}{kn}\frac{\text{trace}(\boldsymbol{\Sigma})-\sqrt{d\log d}}{(\|\boldsymbol{\Sigma}\|_{\text{op}}+c_0\sqrt{\frac{d}{n}})^2}\Bigg], \end{aligned}$$

where $c,c_0,c_1,C>0$ are constants that only depend on κ, the subgaussian norm of the features vectors.

Remark 4.6. Theorems 4.4 and 4.5 quantify the improvement we get in model risk when using the bag construction from constrained k-means instead of random bags. Note that in the asymptotic regime where $n,d\to\infty$ with $n=\Omega(d^2\log d)$, the model risk under Theorem 4.4 converges to zero, while the risk under Theorem 4.5 is lower bounded by $(1-\frac{1}{k})^2\|\boldsymbol{\theta}^*\|_2^2$. We note that $C_{\min}d\leq\text{trace}(\boldsymbol{\Sigma})\leq c_{\max}d$. In other words, the bias of the estimated model remains non-vanishing under random bags, whereas it vanishes asymptotically when the bags are constructed via size-constrained k-means.

Remark 4.7. Theorem 4.4 considers bagging configurations based on k-means with a minimum group size constraint in (7). It assumes access to an oracle model that gives the correct ordering of (unobserved) responses y_i. However, as stated in our methodology, we use a prior model to compute the conditional expected responses $\widetilde{y}_i$, and because of this there may be a mismatch between the ordering of y_i's and $\widetilde{y}_i$'s. We denote by $\boldsymbol{S}$ and $\widetilde{\boldsymbol{S}}$ the corresponding bagging matrices. Our next theorem shows how the estimator error inflates with respect to the mismatch quantity $\|\boldsymbol{S}\boldsymbol{S}^\intercal - \widetilde{\boldsymbol{S}}\widetilde{\boldsymbol{S}}^\intercal\|_{\text{op}}$.

Theorem 4.8. *Consider the linear model (2) under Assumptions 4.2 and 4.3. Suppose that the dimension d and the sample size n grow to infinity, and $n = \Omega(d)$. Let $\widetilde{\boldsymbol{S}}$ be the bagging configuration based on problem (7) using the predicted responses $\widetilde{y}_i$ by a prior model. Similarly, let $\boldsymbol{S}$ be the corresponding bagging configuration by an oracle model who has access to individual responses y_i. If we have a mismatch $\|\boldsymbol{S}\boldsymbol{S}^\intercal - \widetilde{\boldsymbol{S}}\widetilde{\boldsymbol{S}}^\intercal\|_{\text{op}} \le \varepsilon$, then the following holds true with probability at least $1 - 1/n - 2e^{-cd}$,*

$$\mathbb{E}\Big[\, \Big\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\Big\|_2^2 \;\Big|\; \boldsymbol{X}\Big] \le C\left(\frac{k\log(n)\,\|\boldsymbol{\theta}^*\|_2^2 + \sigma^2 d}{n\sigma_{\min}(\boldsymbol{\Sigma})}\right) + \frac{\sigma_{\max}(\boldsymbol{\Sigma}) - C'\sqrt{d/n}}{\sigma_{\min}(\boldsymbol{\Sigma}) + C'\sqrt{d/n}}\,\|\boldsymbol{\theta}^*\|_2^2\,\varepsilon^2,$$

for some constants $c, C, C' > 0$ that depend only on the subgaussian norm κ.

5. Algorithm

We now present the `PriorBoost` algorithm. The high-level idea is to partition the data $\boldsymbol{X}$ into T parts, and use each slice $\boldsymbol{X}^{(t)}$ together with last round's model $\widehat{\boldsymbol{\theta}}^{(t-1)}$ to form better bags $\boldsymbol{S}^{(t)}$, and hence learn a stronger event-level model $\widehat{\boldsymbol{\theta}}^{(t)}$ at each step. This is an iterative and adaptive procedure. However, since we get one aggregate response per sample (non-overlapping bags), taking more steps means using less data per step. We compare `PriorBoost` to the random bagging algorithm in Section 7 that uses all available data in a one non-adaptive round.

Concretely, the first step of `PriorBoost` uses random bagging to learn $\widehat{\boldsymbol{\theta}}^{(1)}$ from the aggregate responses of the first slice $\boldsymbol{X}^{(1)}$. In each subsequent step, we use $\widehat{\boldsymbol{\theta}}^{(t-1)}$ to predict the individual responses $\widetilde{\boldsymbol{y}}^{(t)}$ for this round of data $\boldsymbol{X}^{(t)}$. Based on these predictions, we form aggregation bags by solving the one-dimensional size-constrained k-means clustering problem in (7). Recall that our goal is for bags to be homogoneous with respect to the true responses, which the learner never sees. The learner then gets the aggregate response of each bag, learns a better model $\boldsymbol{\theta}^{(t)}$, and repeats the process. We give pseudocode for `PriorBoost` in Algorithm 1 and summarize its core clustering subroutine below.

Algorithm 1 `PriorBoost`

Input: data $\boldsymbol{X}$, model $\mathcal{L}(\cdot, f_{\boldsymbol{\theta}}(\cdot))$, number of steps T

Split $\boldsymbol{X}$ into T equal-sized parts $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(T)}$
Get aggregate responses $\overline{\boldsymbol{y}}^{(1)}$ for $(\boldsymbol{X}^{(1)}, \boldsymbol{S}^{(\text{random})})$
Update $\widehat{\boldsymbol{\theta}}^{(1)} \leftarrow \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\overline{\boldsymbol{y}}^{(1)}, f_{\boldsymbol{\theta}}(\boldsymbol{X}^{(1)}))$
for $t = 2$ to T **do**
  Predict $\widetilde{\boldsymbol{y}}^{(t)} \leftarrow f_{\widehat{\boldsymbol{\theta}}^{(t-1)}}(\boldsymbol{X}^{(t)})$
  Sort samples by $\widetilde{y}_i^{(t)}$ and solve (7) to get bags $\boldsymbol{S}^{(t)}$ using Lemma 5.1
  Get aggregate responses $\overline{\boldsymbol{y}}^{(t)}$ for $(\boldsymbol{X}^{(t)}, \boldsymbol{S}^{(t)})$
  Update $\widehat{\boldsymbol{\theta}}^{(t)} \leftarrow \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\overline{\boldsymbol{y}}^{(t)}, f_{\boldsymbol{\theta}}(\boldsymbol{X}^{(t)}))$
end for
return $\widehat{\boldsymbol{\theta}}^{(T)}$

Lemma 5.1. *The clustering problem in (7) with bags of minimum size k can be solved in time $O(nk + n\log n)$.*

This subroutine exploits the sorted structure of an optimal partition (Lemma 2.3) and uses dynamic programming with a constant-time update for the sum of squared distance term for the last cluster in the recurrence (Wang & Song, 2011). We describe this algorithm in more detail and give a proof of the lemma in Appendix D.

Remark 5.2. If we have a weak model for predicting event-level responses (e.g., using prior $\widehat{\boldsymbol{\theta}}^{(0)}$ or transfer learning), we can use its predictions for $\widetilde{y}_i$ to sort $\boldsymbol{X}^{(1)}$ and apply Lemma 5.1 in step $t = 1$. This warm starts `PriorBoost` compared to random bagging $\boldsymbol{S}^{(\text{random})}$ and allows the algorithm to use fewer adaptive rounds.

6. Differential privacy for aggregate responses

As previously explained, aggregate learning offers a degree of privacy protection by obscuring individual responses and only disclosing aggregated responses for each bag. If the bags do not overlap and each bag has a minimum size k, substituting individual responses with the aggregated ones ensures k-anonymity, a privacy concept asserting that any given response is indistinguishable from at least $k-1$ other responses.

Another widely used notion of privacy that formalizes the privacy protection of responses/labels is *label differential privacy* (label DP), introduced by Chaudhuri & Hsu (2011). In simple terms, a mechanism, or data processing algorithm, is deemed label DP if its output distribution remains largely unchanged if a single response/label is altered in the input dataset. The concept of label differential privacy is derived from (full) differential privacy (Dwork et al., 2006a;b), focusing specifically on preserving the privacy of responses rather than all features. It is important to note that dif-

ferential privacy provides a guarantee for data processing algorithms, whereas k-anonymity is a property of datasets. We recall the formal definition of label DP from (Chaudhuri & Hsu, 2011).

Definition 6.1 (Label differential privacy)**.** Consider a randomized mechanism $\mathcal{M} : \mathcal{D} \to \mathcal{O}$ that takes as input dataset D and outputs into space $\mathcal{O}$. A mechanism $\mathcal{M}$ is called ε-*label DP* if for any two datasets (D, D') that differ in the label of a single example and any subset $O \subseteq \mathcal{O}$, we have

$$\Pr(\mathcal{M}(D) \in O) \leq e^{\varepsilon} \Pr(\mathcal{M}(D') \in O)\,,$$

where ε is the privacy budget.

It is easy to see that learning from aggregate responses, in the form described so far, is not label DP. However, we can use the Laplace mechanism on top of aggregation to ensure label DP. We empirically study the optimal size of the bags, in terms of minimizing model estimation error, for a given privacy budget in Section 7.3.

In the Laplace mechanism, the magnitude of the noise being added depends on the privacy guarantee ε and the sensitivity of the output to each single change in the dataset. Suppose that the responses/labels are bounded $|y_i| \leq B$ by some value B that is independent of data (and hence can be used without sacrificing any data privacy). The sensitivity of an aggregate response, for a bag of size k, is then given by B/k. Therefore, to ensure ε-label DP, we add independent draws $Z_\ell \sim \text{Laplace}(0, B/\varepsilon k)$ to the aggregate responses $\overline{y}_\ell$, for each $\ell \in [m]$. By Dwork et al. (2006b, Proposition 1) these noisy aggregated responses are ε-DP, and by closure of DP under post-processing (Dwork et al., 2014, Proposition 2.1) any learning algorithm that only uses the noisy aggregate responses is ε-DP.

7. Experiments

We empirically study linear regression, logistic regression, and label DP in the aggregate learning framework. For these tasks, we compare three algorithms:

- `PriorBoost`: Pseudocode presented in Algorithm 1.
- `OneShot`: Random bagging on all of the training data. This is equivalent to Algorithm 1 with $T = 1$ (i.e., a non-adaptive version).
- `PBPrefix`: Variant of Algorithm 1 where at each step t, the model trains on all data seen so far. Specifically, the data used to learn $\widehat{\boldsymbol{\theta}}^{(t)}$ in Line 8 is $\bigcup_{i=1}^{t}\{(\boldsymbol{X}^{(i)}, \overline{\boldsymbol{y}}^{(i)})\}$.

Our experiments use NumPy (Harris et al., 2020) and scikit-learn's `LogisticRegression` (Pedregosa et al., 2011).

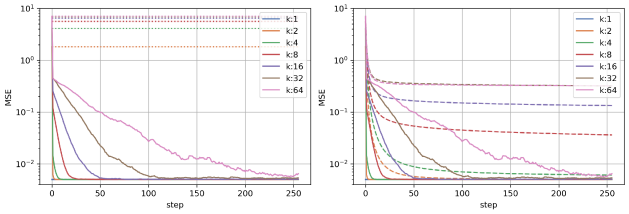

Figure 1: Linear regression. Compares `PriorBoost` (solid) with `OneShot` (left, dotted) and `PBPrefix` (right, dashed) by plotting test MSE at each step t for different bag sizes k.

7.1. Linear regression

We start by generating a dataset $(\boldsymbol{X}, \boldsymbol{y})$ with $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ as follows. First, sample a ground truth model $\boldsymbol{\theta}^* \sim \mathcal{N}_d(\mathbf{0}, \boldsymbol{I})$. Next, generate a design matrix $\boldsymbol{X}$ of n i.i.d. feature vectors $\boldsymbol{x}_i \sim \mathcal{N}_d(\mathbf{0}, \boldsymbol{I})$ and get their responses $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$, where each $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ is i.i.d. Gaussian noise with $\sigma = 0.1$.

To study the convergence of `PriorBoost` and `PBPrefix`, we set $T = 256$. Then we set $n = T \cdot 4096 = 2^{20}$ and $d = 8$ so that both algorithms get 4096 new samples per step. We generate an independent test set of n samples from the same model and plot the test mean squared error (MSE) at each step of the algorithm (using the entire test set) in Figure 1. `OneShot`, as described above, creates random bags of size k across all of the training data, gets the mean response of each bag, and fits a linear regression model with least squares loss. We run each algorithm for bags of size $k \in \{1, 2, 4, 8, 16, 32, 64\}$.

In Figure 1, `PriorBoost` converges to optimal model quality (i.e., the loss when $k = 1$) for all bag sizes k. This is in stark contrast to non-adaptive `OneShot` (i.e., random bagging), whose test loss gets worse as k increases. In the right subplot, `PBPrefix` converges much slower than `PriorBoost`—and to suboptimal solutions. This is because the aggregate responses obtained in early steps of the algorithms are noisy, as the prior $\widehat{\boldsymbol{\theta}}^{(t)}$ is weaker. Noisy responses are helpful for constructing better bags in the next iteration (and hence allowing us to learn a stronger prior), but they can be actively unhelpful if they remain in the training set for too long (e.g., the data that `PBPrefix` trains on in later steps). `PriorBoost`, however, only trains on the last slice of aggregate data $(\boldsymbol{X}^{(t)}, \overline{\boldsymbol{y}}^{(t)})$, and therefore "forgets" early/noisy mean responses, leading to better final model quality while also using fewer samples per step.

7.2. Logistic regression

For our first logistic regression experiment, we use the same ground truth model weights, design matrix, and Gaussian noise $(\boldsymbol{\theta}^*, \boldsymbol{X}, \boldsymbol{\varepsilon})$ as in linear regression, but now we create binary labels by sending them through a sigmoid function and rounding: $y_i = \texttt{round}(\sigma(\boldsymbol{x}_i^\intercal \boldsymbol{\theta}^* + \varepsilon_i)) \in \{0, 1\}$. After

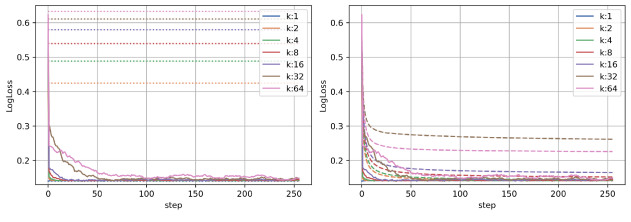

Figure 2: Logistic regression. Compare `PriorBoost` (solid) with `OneShot` (left, dotted) and `PBPrefix` (right, dashed) by plotting test log loss at each step t for different bag sizes k.

each aggregation step t, the oracle rounds the mean response of each bag $\texttt{round}(\overline{y}_\ell) \in \{0, 1\}$ to get back to a binary label, which is an additional source of noise.[2] All three algorithms fit logistic regression models with binary cross-entropy loss and L2 regularization penalty $\frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2$ for $\lambda = 10$.

Similar to the linear regression experiment, Figure 2 shows that `PriorBoost` converges to optimality for all bag sizes. In contrast, `OneShot` steadily degrades as k increases. We also see that by training on all aggregate responses available at step t to learn $\widehat{\boldsymbol{\theta}}^{(t)}$, `PBPrefix` converges slower and to suboptimal solutions for $k \geq 16$.

7.3. Differential privacy

We now modify `PriorBoost` by adding Laplace noise to the aggregate responses to make the algorithm ε-label DP as described in Section 6. The key observation is that for binary labels and bags of size at least k, we can reduce the scale of the Laplace noise by a factor of k. We use the same experimental setup as in Section 7.2, geometrically sweep over privacy budgets $0.01 \leq \varepsilon \leq 100$, and compare the test loss of `PriorBoost` to an ε-label DP version of random bagging. Error bars are computed over 10 realizations.

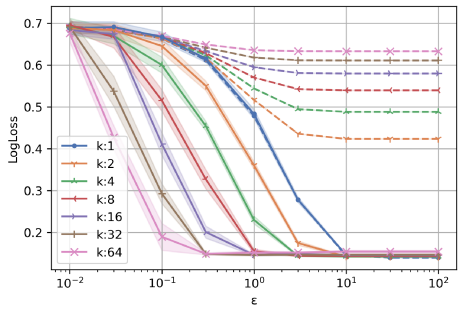

Figure 3: ε-label differentially private logistic regression. Compares final `PriorBoost` (solid) test log loss to `OneShot` (dashed) for different bag sizes k.

As we increase the privacy loss ε, the test loss decreases for each value of k. This is expected because the privacy constraint becomes more relaxed, allowing for better model quality. Note that $\varepsilon = \infty$ corresponds to not using any differential privacy. For each value of ε, the utility of `PriorBoost` *improves* as the bag size k increases, whereas the utility of `OneShot` degrades as we increase k. This may initially seem surprising, but recall that `PriorBoost` actively reduces the bias of the estimated model by forming homogeneous bags with respect to labels. As we increase k, `PriorBoost` (1) effectively maintains a slow growth rate for the bias, (2) can afford to reduce the scale of its Laplace noise by a factor of k, and (3) gets low-variance mean labels since they are averaged over larger bags. Therefore, all in all, `PriorBoost` favors larger k in this setup. In particular, it approaches its non-private loss at a faster rate in ε, for larger k. For example with $k = 64$, it already nearly achieves the non-private loss for $\varepsilon \geq 0.3$. For `OneShot` though, larger k significantly increases the bias of the estimated model, which outweighs the variance reduction and results in a larger loss. Also note that for $k = 1$ (singleton bags), `PriorBoost` and `OneShot` match. In summary, this experiment shows we can achieve more utility for a fixed ε by using `PriorBoost` to learn large curated bags whose mean labels require less random noise.

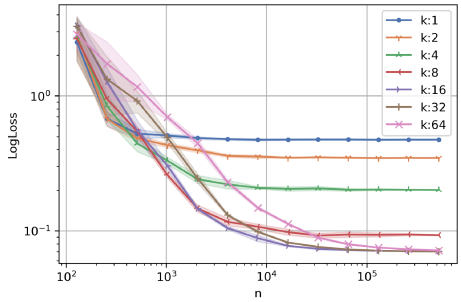

Figure 4: Optimal bag sizes k for ε-label DP `PriorBoost` for logistic regression. Test loss for $\varepsilon = 1$ as the number of samples n increases.

Optimal bag sizes. The plots in Figure 3 are for fixed n and d with $n \gg d$. To better understand the effect of bag size on the bias-variance tradeoff, we next run the same logistic regression experiment for $d = 64$, $T = 128$, $\varepsilon = 1$, while varying the total number of samples n. An intriguing observation from Figure 4 is that the optimal bag size (i.e., the one minimizing the loss) grows with n. The crossover points for optimal k also become farther apart as n grows. For example, $k = 4$ is optimal for a smaller range of n compared to $k = 16$. As discussed earlier, a larger k yields larger bias while reducing the variance. However, a virtue of `PriorBoost` is that both the bias and variance decrease in the sample size n. The loss plots in Figure 4 suggest that the decay rate (in n) of the bias is faster than the decay rate of the variance, and so as n grows, the optimal bag size for `PriorBoost` becomes larger.

[2] We randomly round $\overline{y}_\ell = 1/2$ to 0 or 1 in a consistent way to avoid biasing the distribution of binary aggregate labels.

Conclusion

This work proposes a novel method for using available prior information for expected responses of samples to construct bags for aggregate learning. We devise the multi-stage algorithm `PriorBoost` to obtain good priors from the aggregate data itself if no public prior is available. We also propose a differentially private version, as well as intriguing observations about optimal bag sizes. Our analysis provably shows the advantage of our approach over random bagging, which we back up with strong numerical experiments.

Acknowledgement

We would like to thank Lorne Applebaum, Ashwinkumar Badanidiyuru, Lin Chen, Alessandro Epasto, Thomas Fu, Xiaoen Ju, Nick Ondo, Fariborz Salehi, and Dinah Shender for helpful discussions related to this work. Adel Javanmard is supported in part by the NSF CAREER Award DMS-1844481, the NSF Award DMS-2311024, and the Sloan fellowship in Mathematics.

Impact statement

This work proposes a method to construct aggregation sets (bags) for training models from aggregate responses. We develop tools to improve model quality while preserving the privacy of users, as individual responses are not revealed in this setting. We also design mechanisms to be used on top of aggregation that provide differential privacy guarantees. Our work contributes to the field of machine learning and privacy-prese rving systems. We do not anticipate this work to have any adverse societal or ethical repercussions.

References

Anil, R., Gadanho, S., Huang, D., Jacob, N., Li, Z., Lin, D., Phillips, T., Pop, C., Regan, K., Shamir, G. I., et al. On the factory floor: ML engineering for industrial-scale ads recommendation models. In *Proceedings of the 5th Workshop on Online Recommender Systems and User Modeling co-located with the 16th ACM Conference on Recommender Systems*. CEUR-WS, 2022.

Brahmbhatt, A., Pokala, M., Saket, R., and Raghuveer, A. LLP-Bench: A large scale tabular benchmark for learning from label proportions. *arXiv preprint arXiv:2310.10096*, 2023.

Busa-Fekete, R., Choi, H., Dick, T., Gentile, C., and Munoz Medina, A. Easy learning from label proportions. *Advances in Neural Information Processing Systems*, 36, 2024.

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

Chen, L., Fu, G., Karbasi, A., and Mirrokni, V. Learning from aggregated data: Curated bags versus random bags. *arXiv preprint arXiv:2305.09557*, 2023.

Coleman, B., Kang, W.-C., Fahrbach, M., Wang, R., Hong, L., Chi, E., and Cheng, D. Unified Embedding: Battle-tested feature representations for web-scale ML systems. *Advances in Neural Information Processing Systems*, 36, 2024.

de Freitas, N. and Kück, H. Learning about individuals from group statistics. In *Proceedings of the 21st Conference in Uncertainty in Artificial Intelligence*, pp. 332–339. AUAI Press, 2005.

Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., and Naor, M. Our data, ourselves: Privacy via distributed noise generation. In *Proceedings of the 25th Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 486–503. Springer, 2006a.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Theory of Cryptography Conference*, pp. 265–284. Springer, 2006b.

Dwork, C., Roth, A., et al. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Fahrbach, M., Javanmard, A., Mirrokni, V., and Worah, P. Learning rate schedules in the presence of distribution shift. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 9523–9546. PMLR, 2023.

Geradin, D., Katsifis, D., and Karanikioti, T. Google as a de facto privacy regulator: Analyzing chrome's removal of third-party cookies from an antitrust perspective. 2020.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. Array programming with NumPy. *Nature*, 585(7825):357–362, 2020.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Javanmard, A., Chen, L., Mirrokni, V., Badanidiyuru, A., and Fu, G. Learning from aggregate responses: Instance level versus bag level loss functions. In *Twelfth International Conference on Learning Representations*, 2024.

Kollnig, K., Shuba, A., Van Kleek, M., Binns, R., and Shadbolt, N. Goodbye tracking? Impact of iOS app tracking transparency and privacy labels. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 508–520, 2022.

Li, G., Chen, L., Javanmard, A., and Mirrokni, V. Optimistic rates for learning from label proportions. In *The 37th Annual Conference on Learning Theory*, 2024.

Musicant, D. R., Christensen, J. M., and Olson, J. F. Supervised learning by training on aggregate outputs. In *Seventh IEEE International Conference on Data Mining*, pp. 252–261. IEEE, 2007.

Patrini, G., Nock, R., Rivera, P., and Caetano, T. (almost) no label no cry. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Quadrianto, N., Smola, A. J., Caetano, T. S., and Le, Q. V. Estimating labels from label proportions. In *Proceedings of the 25th International Conference on Machine learning*, pp. 776–783, 2008.

Rueping, S. SVM classifier estimation from group probabilities. In *Proceedings of the 27th International Conference on Machine Learning*, pp. 911–918, 2010.

Saket, R. Algorithms and hardness for learning linear thresholds from label proportions. *Advances in Neural Information Processing Systems*, 35:1267–1279, 2022.

Scott, C. and Zhang, J. Learning from label proportions: A mutual contamination framework. *Advances in Neural Information Processing Systems*, 33:22256–22267, 2020.

Sweeney, L. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing: Theory and Applications*, pp. 210–268. Cambridge University Press, 2012.

Vershynin, R. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Wang, H. and Song, M. Ckmeans.1d.dp: Optimal k-means clustering in one dimension by dynamic programming. *The R Journal*, 3(2):29, 2011.

Wein, L. M. and Zenios, S. A. Pooled testing for hiv screening: capturing the dilution effect. *Operations Research*, 44(4):543–569, 1996.

Yu, F. X., Choromanski, K., Kumar, S., Jebara, T., and Chang, S.-F. On learning from label proportions. *arXiv preprint arXiv:1402.5902*, 2014.

Zhang, J., Wang, Y., and Scott, C. Learning from label proportions by learning with label noise. *Advances in Neural Information Processing Systems*, 35:26933–26942, 2022.

A. Missing analysis from Section 2

A.1. Proof of Theorem 2.1

The derivative of the loss at the minimizer is zero, which gives us:

$$\boldsymbol{X}^\intercal(\boldsymbol{S}^\intercal\boldsymbol{S}\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}}) = \boldsymbol{0}.$$

By rearranging the terms, we have

$$\begin{aligned}\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* &= (\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\boldsymbol{S}^\intercal\boldsymbol{S}\boldsymbol{y} - \boldsymbol{\theta}^* \\ &= (\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\boldsymbol{S}^\intercal\boldsymbol{S}\boldsymbol{X}\boldsymbol{\theta}^* - \boldsymbol{\theta}^* + (\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\boldsymbol{S}^\intercal\boldsymbol{S}\boldsymbol{\varepsilon} \\ &= (\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal(\boldsymbol{S}^\intercal\boldsymbol{S} - \boldsymbol{I}_n)\boldsymbol{X}\boldsymbol{\theta}^* + (\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\boldsymbol{S}^\intercal\boldsymbol{S}\boldsymbol{\varepsilon}\,.\end{aligned}$$

Since the noise vector $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is independent of the design matrix $\boldsymbol{X}$ with $\mathbb{E}[\boldsymbol{\varepsilon}] = \boldsymbol{0}$ and $\mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\intercal] = \sigma^2\boldsymbol{I}$, we have

$$\mathbb{E}\Big[\,\Big\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\Big\|_2^2 \;\Big|\; \boldsymbol{X}\Big] = \big\|(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal(\boldsymbol{S}^\intercal\boldsymbol{S} - \boldsymbol{I}_n)\boldsymbol{X}\boldsymbol{\theta}^*\big\|_2^2 + \sigma^2\mathrm{trace}((\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\boldsymbol{S}^\intercal\boldsymbol{S}\boldsymbol{S}^\intercal\boldsymbol{S}\boldsymbol{X}(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1})\,. \quad (13)$$

Further, since the bags are non-overlapping, we have $\boldsymbol{S}\boldsymbol{S}^\intercal = \boldsymbol{I}_m$, by which we get

$$\begin{aligned}\mathrm{trace}((\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\boldsymbol{S}^\intercal\boldsymbol{S}\boldsymbol{S}^\intercal\boldsymbol{S}\boldsymbol{X}(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}) &= \mathrm{trace}((\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\boldsymbol{S}^\intercal\boldsymbol{S}\boldsymbol{X}(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}) \\ &= \big\|(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\boldsymbol{S}^\intercal\big\|_{\mathrm{F}}^2\,. \quad (14)\end{aligned}$$

Substituting into (13), we prove the claim.

A.2. Proof of Corollary 2.2

By definition of the operator norm, we have

$$\big\|(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal(\boldsymbol{S}^\intercal\boldsymbol{S} - \boldsymbol{I}_n)\boldsymbol{X}\boldsymbol{\theta}^*\big\|_2 \le \big\|(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\big\|_{\mathrm{op}}\,\big\|(\boldsymbol{S}^\intercal\boldsymbol{S} - \boldsymbol{I}_n)\boldsymbol{X}\boldsymbol{\theta}^*\big\|_2\,.$$

Next, we upper bound the variance term in Eq. (14) using the inequality

$$\|\boldsymbol{A}\boldsymbol{B}\|_{\mathrm{F}}^2 \le \min\Big(\|\boldsymbol{A}\|_{\mathrm{op}}^2\,\|\boldsymbol{B}\|_{\mathrm{F}}^2\,,\,\|\boldsymbol{B}\|_{\mathrm{op}}^2\,\|\boldsymbol{A}\|_{\mathrm{F}}^2\Big)\,.$$

We have $\|\boldsymbol{S}\|_{\mathrm{F}}^2 = m$ and $\|\boldsymbol{S}\|_{\mathrm{op}} = 1$ by the Cauchy–Schwarz inequality. We also assumed $\mathrm{rank}(\boldsymbol{X}) \le d$, which implies

$$\big\|(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\big\|_{\mathrm{F}} \le \sqrt{d}\,\big\|(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\big\|_{\mathrm{op}}\,.$$

Therefore, we obtain

$$\big\|(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\boldsymbol{S}^\intercal\big\|_{\mathrm{F}}^2 \le \big\|(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\big\|_{\mathrm{op}}^2\,\min(m,d)\,.$$

Combining these two bounds with Theorem 2.1 gives the result.

A.3. Proof of Lemma 2.3

The key idea is to rewrite the optimization problem by "lifting" the space of optimization variables as follows:

$$\begin{aligned}\min_{(B_1,\ldots,B_m)\in\mathcal{B}} \quad & \sum_{\ell=1}^{m}\sum_{i\in B_\ell}(\widetilde{y}_i - c_\ell)^2 \quad (15)\\ \text{subject to} \quad & |B_\ell| \ge k \qquad \forall \ell \in [m] \\ & c_\ell \in \mathbb{R} \qquad \forall \ell \in [m]\end{aligned}$$

In words, we introduce the additional variables $c_\ell \in \mathbb{R}$, for $\ell \in [m]$. It is easy to see that problems (7) and (15) have the same optimal bagging configurations. Now, suppose that $\{(B_\ell^*, c_\ell^*) : \ell \in [m]\}$ is an optimal solution to (15). If the claim is not true, then there exists $\widetilde{y}_i > \widetilde{y}_j$ and $c_\ell > c_{\ell'}$ such that $\widetilde{y}_i \in B_{\ell'}^*$ and $\widetilde{y}_j \in B_\ell^*$. We then argue that by assigning $\widetilde{y}_i$ to B_ℓ^* and $\widetilde{y}_j$ to $B_{\ell'}^*$, we can reduce the objective value, which is a contradiction. To show this, we must prove that

$$\begin{aligned}(\widetilde{y}_i - c_{\ell'})^2 + (\widetilde{y}_j - c_\ell)^2 > (\widetilde{y}_i - c_\ell)^2 + (\widetilde{y}_j - c_{\ell'})^2 &\iff -\widetilde{y}_i c_{\ell'} - \widetilde{y}_j c_\ell > -\widetilde{y}_i c_\ell - \widetilde{y}_j c_{\ell'} \\ &\iff (\widetilde{y}_i - \widetilde{y}_j)(c_\ell - c_{\ell'}) > 0\,,\end{aligned}$$

which is true by our assumption.

B. Missing analysis from Section 3

We recall the notion of strong convexity below.

Definition B.1. A function $f : \mathbb{R}^n \to \mathbb{R}$ is *strongly convex* with parameter μ if the following holds for all $x, y \in \mathbb{R}$:

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + s_{\boldsymbol{x}}^{\intercal}(\boldsymbol{y} - \boldsymbol{x}) + \frac{\mu}{2} \|\boldsymbol{y} - \boldsymbol{x}\|_2^2 \,,$$

for any $s_{\boldsymbol{x}} \in \partial f(\boldsymbol{x})$, where $\partial f(\boldsymbol{x})$ denotes the set of subgradients of f at $\boldsymbol{x}$.

The next lemma states that controlling the estimation error for a strongly convex loss function reduces to controlling the norm of the gradient of the loss at the true model.

Lemma B.2. *Suppose that the loss $\mathcal{L}$ is strongly convex with parameter μ and $\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta})$. Then, for any model $\boldsymbol{\theta}^*$, we have*

$$\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right\|_2 \leq \frac{1}{\mu} \|\mathcal{L}(\boldsymbol{\theta}^*)\|_2 \,.$$

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

$$\frac{1}{L} \|\mathcal{L}(\boldsymbol{\theta}^*)\|_2 \leq \left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right\|_2 .$$

Proof. By writing the definition of strong convexity for $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^*$, and noting that $\nabla\mathcal{L}(\widehat{\boldsymbol{\theta}}) = \mathbf{0}$, we get

$$\mathcal{L}(\boldsymbol{\theta}^*) \geq \mathcal{L}(\widehat{\boldsymbol{\theta}}) + \frac{\mu}{2} \left\|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}\right\|_2^2 .$$

Likewise, by changing the role of $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^*$, we have

$$\mathcal{L}(\widehat{\boldsymbol{\theta}}) \geq \mathcal{L}(\boldsymbol{\theta}^*) + \nabla\mathcal{L}(\boldsymbol{\theta}^*)^{\intercal}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) + \frac{\mu}{2} \left\|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}\right\|_2^2 .$$

By adding the above two inequalities and rearranging the terms, we arrive at

$$\nabla\mathcal{L}(\boldsymbol{\theta}^*)^{\intercal}(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}) \geq \mu \left\|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}\right\|_2^2 .$$

Next, by Cauchy–Schwarz inequality, $\nabla\mathcal{L}(\boldsymbol{\theta}^*)^{\intercal}(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}) \leq \|\mathcal{L}(\boldsymbol{\theta}^*)\|_2 \left\|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}\right\|_2$, which along with the previous inequality proves the first claim.

The second claim follows easily from Lipschitz condition. We write

$$\|\nabla\mathcal{L}(\boldsymbol{\theta}^*)\|_2 = \left\|\nabla\mathcal{L}(\boldsymbol{\theta}^*) - \nabla\mathcal{L}(\widehat{\boldsymbol{\theta}})\right\|_2 \leq L \left\|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}\right\|_2 ,$$

which completes the proof. □

B.1. Proof of Theorem 3.1

The gradient of the loss in (9) reads as

$$\begin{aligned}\nabla\mathcal{L}(\boldsymbol{\theta}) &= \frac{1}{n} \sum_{\ell=1}^{m} \sum_{i \in B(\ell)} \frac{1}{a_i(\phi)} (\overline{y}_\ell - b'(\boldsymbol{\theta}^{\intercal}\boldsymbol{x}_i))\boldsymbol{x}_i \\ &= \boldsymbol{X}^{\intercal}\boldsymbol{D}^{-1}(\boldsymbol{S}^{\intercal}\boldsymbol{S}\boldsymbol{y} - b'(\boldsymbol{X}\boldsymbol{\theta})) \,.\end{aligned}$$

We next consider the following bias-variance decomposition:

$$\mathbb{E}\Big[\, \|\nabla\mathcal{L}(\boldsymbol{\theta}^*)\|_2^2 \;\Big|\; \boldsymbol{X}\Big] = \Big\|\mathbb{E}\Big[\nabla\mathcal{L}(\boldsymbol{\theta}^*) \;\Big|\; \boldsymbol{X}\Big]\Big\|_2^2 + \operatorname{trace}(\operatorname{Cov}(\nabla\mathcal{L}(\boldsymbol{\theta}^*) \mid \boldsymbol{X})) \,. \tag{16}$$

Under the GLM, the responses y_i are independent conditioned on $\boldsymbol{x}_i$. In addition,

$$\mathbb{E}[y_i \mid \boldsymbol{x}_i] = b'(\boldsymbol{\theta}^\intercal \boldsymbol{x}_i), \quad \mathrm{Var}(y_i \mid \boldsymbol{x}_i) = a_i(\phi) b''(\boldsymbol{\theta}^\intercal \boldsymbol{x}_i)\,.$$

We therefore get

$$\begin{aligned}\mathbb{E}\Big[\nabla \mathcal{L}(\boldsymbol{\theta}^*) \;\Big|\; \boldsymbol{X}\Big] &= \boldsymbol{X}^\intercal \boldsymbol{D}^{-1}(\boldsymbol{S}^\intercal \boldsymbol{S} b'(\boldsymbol{X}\boldsymbol{\theta}^*) - b'(\boldsymbol{X}\boldsymbol{\theta}^*)) \\ &= \boldsymbol{X}^\intercal \boldsymbol{D}^{-1}(\boldsymbol{S}^\intercal \boldsymbol{S} - \boldsymbol{I}) b'(\boldsymbol{X}\boldsymbol{\theta}^*)\,. \end{aligned} \tag{17}$$

In addition,

$$\begin{aligned}\mathrm{Cov}(\nabla \mathcal{L}(\boldsymbol{\theta}^*) \mid \boldsymbol{X}) &= \mathbb{E}\Big[\boldsymbol{X}^\intercal \boldsymbol{D}^{-1} \boldsymbol{S}^\intercal \boldsymbol{S}(\boldsymbol{y} - b'(\boldsymbol{X}\boldsymbol{\theta}^*))(\boldsymbol{y} - b'(\boldsymbol{X}\boldsymbol{\theta}^*))^\intercal \boldsymbol{S}^\intercal \boldsymbol{S} \boldsymbol{D}^{-1} \boldsymbol{X} \;\Big|\; \boldsymbol{X}\Big] \\ &= \boldsymbol{X}^\intercal \boldsymbol{D}^{-1} \boldsymbol{S}^\intercal \boldsymbol{S} \boldsymbol{D} \mathrm{diag}(b''(\boldsymbol{X}\boldsymbol{\theta}^*)) \boldsymbol{S}^\intercal \boldsymbol{S} \boldsymbol{D}^{-1} \boldsymbol{X}\,.\end{aligned}$$

Therefore,

$$\mathrm{trace}(\mathrm{Cov}(\nabla \mathcal{L}(\boldsymbol{\theta}^*) \mid \boldsymbol{X})) = \left\| \boldsymbol{X}^\intercal \boldsymbol{D}^{-1} \boldsymbol{S}^\intercal \boldsymbol{S} \boldsymbol{D}^{1/2} \mathrm{diag}(b''(\boldsymbol{X}\boldsymbol{\theta}^*))^{1/2} \right\|_{\mathrm{F}}^2 . \tag{18}$$

Combining (17) and (18) into (16) completes the proof.

B.2. Proof of Corollary 3.2

We upper bound each term of (10) separately. For the first term, we have

$$\left\| \boldsymbol{X}^\intercal \boldsymbol{D}^{-1}(\boldsymbol{S}^\intercal \boldsymbol{S} - \boldsymbol{I})\boldsymbol{\mu} \right\|_2 \le \left\| \boldsymbol{X}^\intercal \boldsymbol{D}^{-1} \right\|_{\mathrm{op}} \left\| (\boldsymbol{S}^\intercal \boldsymbol{S} - \boldsymbol{I})\boldsymbol{\mu} \right\|_2 .$$

For the second term, we develop two upper bounds and take the minimum of the two.

For the first upper bound we have

$$\begin{aligned}\left\| \boldsymbol{X}^\intercal \boldsymbol{D}^{-1} \boldsymbol{S}^\intercal \boldsymbol{S} \boldsymbol{D}^{1/2} \mathrm{diag}(b''(\boldsymbol{X}\boldsymbol{\theta}^*))^{1/2} \right\|_{\mathrm{F}}^2 &\le d \left\| \boldsymbol{X}^\intercal \boldsymbol{D}^{-1} \boldsymbol{S}^\intercal \boldsymbol{S} \boldsymbol{D}^{1/2} \mathrm{diag}(b''(\boldsymbol{X}\boldsymbol{\theta}^*))^{1/2} \right\|_{\mathrm{op}}^2 \\ &\le d \left\| \boldsymbol{X}^\intercal \boldsymbol{D}^{-1} \right\|_{\mathrm{op}}^2 \left\| \boldsymbol{S}^\intercal \boldsymbol{S} \right\|_{\mathrm{op}}^2 \left\| \boldsymbol{D}^{1/2} \mathrm{diag}(b''(\boldsymbol{X}\boldsymbol{\theta}^*))^{1/2} \right\|_{\mathrm{op}} \\ &\le d \left\| \boldsymbol{X}^\intercal \boldsymbol{D}^{-1} \right\|_{\mathrm{op}}^2 \left\| \boldsymbol{v} \right\|_\infty .\end{aligned}$$

For the second upper bound we have

$$\left\| \boldsymbol{X}^\intercal \boldsymbol{D}^{-1} \boldsymbol{S}^\intercal \boldsymbol{S} \boldsymbol{D}^{1/2} \mathrm{diag}(b''(\boldsymbol{X}\boldsymbol{\theta}^*))^{1/2} \right\|_{\mathrm{F}}^2 \le \left\| \boldsymbol{X}^\intercal \boldsymbol{D}^{-1} \right\|_{\mathrm{op}}^2 \left\| \boldsymbol{S}^\intercal \boldsymbol{S} \boldsymbol{D}^{1/2} \mathrm{diag}(b''(\boldsymbol{X}\boldsymbol{\theta}^*))^{1/2} \right\|_{\mathrm{F}}^2 , \tag{19}$$

using the inequality $\|\boldsymbol{A}\boldsymbol{B}\|_{\mathrm{F}}^2 \le \|\boldsymbol{A}\|_{\mathrm{op}}^2 \|\boldsymbol{B}\|_{\mathrm{F}}^2$.

We also note that

$$\begin{aligned}\left\| \boldsymbol{S}^\intercal \boldsymbol{S} \boldsymbol{D}^{1/2} \mathrm{diag}(b''(\boldsymbol{X}\boldsymbol{\theta}^*))^{1/2} \right\|_{\mathrm{F}}^2 &= \mathrm{trace}\left(\boldsymbol{S}^\intercal \boldsymbol{S} \boldsymbol{D} \mathrm{diag}(b''(\boldsymbol{X}\boldsymbol{\theta}^*)) \boldsymbol{S}^\intercal \boldsymbol{S} \right) \\ &= \mathrm{trace}\left(\boldsymbol{D} \mathrm{diag}(b''(\boldsymbol{X}\boldsymbol{\theta}^*)) \boldsymbol{S}^\intercal \boldsymbol{S} \boldsymbol{S}^\intercal \boldsymbol{S} \right) \\ &= \mathrm{trace}\left(\boldsymbol{D} \mathrm{diag}(b''(\boldsymbol{X}\boldsymbol{\theta}^*)) \boldsymbol{S}^\intercal \boldsymbol{S} \right) \\ &= \mathrm{trace}\left(\boldsymbol{D} \mathrm{diag}(b''(\boldsymbol{X}\boldsymbol{\theta}^*)) \boldsymbol{S}^\intercal \boldsymbol{S} \right) \\ &= \sum_{\ell=1}^m \sum_{i \in B_\ell} \frac{a_i(\phi) b''(\boldsymbol{x}_i^\intercal \boldsymbol{\theta}^*)}{|B_\ell|} = \sum_{\ell=1}^m \sum_{i \in B_\ell} \frac{v_i}{|B_\ell|}\,.\end{aligned}$$

Combining the above bounds, we obtain

$$\left\| \boldsymbol{X}^\intercal \boldsymbol{D}^{-1} \boldsymbol{S}^\intercal \boldsymbol{S} \boldsymbol{D}^{1/2} \mathrm{diag}(b''(\boldsymbol{X}\boldsymbol{\theta}^*))^{1/2} \right\|_{\mathrm{F}}^2 \le \left\| \boldsymbol{X}^\intercal \boldsymbol{D}^{-1} \right\|_{\mathrm{op}}^2 \min\Big(\sum_{\ell=1}^m \sum_{i \in B_\ell} \frac{v_i}{|B_\ell|}, d \left\| \boldsymbol{v} \right\|_\infty \Big)\,. \tag{20}$$

This completes the proof of the corollary.

C. Missing analysis from Section 4

C.1. Proof of Theorem 4.4

We prove the claim using the result of Corollary 2.2. Recall the notation $\boldsymbol{\mu} := \boldsymbol{X}\boldsymbol{\theta}^*$. By Lemma 2.3, we know that the solution $\boldsymbol{S}$ given by (7) has a simple sorting structure. We use that structure to construct a bagging scheme to upper bound the term $\|(\boldsymbol{S}^\intercal\boldsymbol{S}-\boldsymbol{I})\boldsymbol{\mu}\|_2^2$. Without loss of generality, assume that n is divisible by k. Sort the entries of $\boldsymbol{\mu}$ and construct the bags as $B(\ell) = \{(\ell-1)k+1,\dots,\ell k\}$ for $\ell = 1,\dots,m := n/k$. In addition, let $\bar{\mu}_\ell$ indicate the average of μ_i's over bag ℓ. This construction of bags satisfy the constraint of (7) and so we have

$$\begin{aligned}
\|(\boldsymbol{S}^\intercal\boldsymbol{S}-\boldsymbol{I})\boldsymbol{\mu}\|_2^2 &\le \sum_{\ell=1}^{m}\sum_{j=(\ell-1)k+1}^{\ell k}(\mu_j-\bar{\mu}_\ell)^2\\
&\le k\sum_{\ell=1}^{m}(\mu_{\ell k}-\mu_{(\ell-1)k+1})^2\\
&\le k\Big(\sum_{\ell=1}^{m}\mu_{\ell k}-\mu_{(\ell-1)k+1}\Big)^2\\
&= k(\mu_{(1)}-\mu_{(n)})^2\\
&\le 4k\,\|\boldsymbol{\mu}\|_\infty^2\,.
\end{aligned}$$

where $\bar{\mu}_i$ in the first inequality denotes the average of μ_i's over bag i.

We next bound $\|\boldsymbol{\mu}\|_\infty^2$. Since $\boldsymbol{x}_i$'s are κ-subgaussian, we have that μ_i is $\kappa\,\|\boldsymbol{\theta}^*\|_2$-subgaussian, for $i\in[n]$.

Lemma C.1. *Suppose that $\xi_1,\dots,\xi_n$ are centered η-subgaussian random variables. Then, with probability at least $1-\frac{1}{n}$, we have*

$$\max_{i\in[n]}\xi_i^2 \le 2\eta^2\log n.$$

By using Lemma C.1 we obtain

$$\|(\boldsymbol{S}^\intercal\boldsymbol{S}-\boldsymbol{I})\boldsymbol{\mu}\|_2^2 \le 4k\,\|\boldsymbol{\mu}\|_\infty^2 \le 8k\kappa^2\,\|\boldsymbol{\theta}^*\|_2^2\log n\,, \tag{21}$$

with probability at least $1-1/n$. We next use the concentration bounds on the singular values of matrices with i.i.d. subgaussian rows to bound $\big\|(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\big\|_{\rm op}$. Specifically, we use Vershynin (2012, Equation (5.25)), which states that with probability at least $1-2e^{-c_1t^2}$, the following holds true:

$$\left\|\frac{1}{n}\boldsymbol{X}^\intercal\boldsymbol{X}-\boldsymbol{\Sigma}\right\|_{\rm op} \le \max(\delta_1,\delta_1^2),\quad \delta_1 = C_1\sqrt{\frac{d}{n}}+\frac{t}{\sqrt{n}}\,, \tag{22}$$

for constants $c_1, C_1 > 0$ that depend only on κ. We define the probabilistic event $\mathcal{E}$ as follows:

$$\mathcal{E}_1 := \left\{\left\|\frac{1}{n}\boldsymbol{X}^\intercal\boldsymbol{X}-\boldsymbol{\Sigma}\right\|_{\rm op} \le C\sqrt{\frac{d}{n}}\right\}\,, \tag{23}$$

for some fixed constant $C>0$. Then using (22) we have $\Pr(\mathcal{E}_1)\ge 1-2e^{-c_1d}$, for some constant depending on C and κ. Under the event $\mathcal{E}$, and by using Weyl's inequality for singular values, we have

$$\sigma_{\min}(\boldsymbol{X}) \ge \sqrt{n\sigma_{\min}(\boldsymbol{\Sigma})-C\sqrt{dn}}\,. \tag{24}$$

Combining (21) and (37) in Corollary 2.2, we obtain the result.

Proof of Lemma C.1. Since ξ_i is η-subgaussian, by definition $\mathbb{E}[\exp(X^2/\eta^2)]\le 2$. Exponentiating and using Markov's inequality, we obtain

$$\Pr(|\xi_i|\ge t) = \Pr(e^{\xi_i^2/\eta^2}\ge e^{t^2/\eta^2}) \le e^{-t^2/\eta^2}\mathbb{E}[e^{\xi_i^2/\eta^2}] \le 2e^{-t^2/\eta^2}\,.$$

Choosing $t = \eta\sqrt{2\log n}$ and union bounding over $i \in [n]$, we get

$$\Pr\Big(\max_{i\in[n]} |\xi_i| \geq \eta\sqrt{2\log n}\Big) \leq \frac{2}{n}\,,$$

which completes the proof of lemma. □

C.2. Proof of Theorem 4.5

We recall the characterization of the risk given by Theorem 2.1:

$$\mathbb{E}\Big[\,\Big\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\Big\|_2^2 \;\Big|\; \boldsymbol{X}\Big] = \big\|(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal(\boldsymbol{S}^\intercal\boldsymbol{S} - \boldsymbol{I}_n)\boldsymbol{X}\boldsymbol{\theta}^*\big\|_2^2 + \sigma^2\,\big\|(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\boldsymbol{S}^\intercal\big\|_{\mathrm{F}}^2\,. \tag{25}$$

We introduce the shorthand $\boldsymbol{\Lambda} := \boldsymbol{S}^\intercal\boldsymbol{S} - \boldsymbol{I}_n$. Our next lemma lower bounds the expected bias.

Lemma C.2. *Under the assumptions of Theorem 4.5, the following holds with probability at least* $1 - 2e^{-c_1 d} - 2d^{-c}$*,*

$$\mathbb{E}\Big[\,\big\|(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\boldsymbol{\Lambda}\boldsymbol{X}\boldsymbol{\theta}^*\big\|_2^2 \;\Big|\; \boldsymbol{X}\Big] \geq \left(1 - \frac{1}{k} - C\frac{d\sqrt{\log d}}{\sigma_{\min}(\boldsymbol{\Sigma})\sqrt{n}}\right)^2 \|\boldsymbol{\theta}^*\|_2^2\,,$$

where constants $C, c, c_1 > 0$ *only depend on the subgaussian norm* κ*.*

Our next lemma lower bound the variance term in (25).

Lemma C.3. *Under the assumptions of Theorem 4.5, the following holds with probability at least* $1 - 2e^{-c_1 d} - 2d^{-c}$*,*

$$\mathbb{E}\Big[\,\big\|(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\boldsymbol{S}\big\|_{\mathrm{F}}^2 \;\Big|\; \boldsymbol{X}\Big] \geq \frac{1}{kn}\cdot\frac{\mathrm{trace}(\boldsymbol{\Sigma}) - \sqrt{d\log d}}{(\|\boldsymbol{\Sigma}\|_{\mathrm{op}} + c_0\sqrt{\frac{d}{n}})^2}\,,$$

where constants $c, c_0, c_1 > 0$ *only depend on the subgaussian norm* κ*.*

Proof of Theorem 4.5 follows by using Lemma C.2 and Lemma C.3 in the decomposition (25).

C.3. Proof of Lemma C.2

Consider the following optimization problem

$$\widehat{\boldsymbol{\alpha}} = \frac{1}{2n}\arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^d} \|\boldsymbol{X}\boldsymbol{\alpha} - \boldsymbol{\Lambda}\boldsymbol{X}\boldsymbol{\theta}^*\|_2^2\,. \tag{26}$$

It is easy to see that by the KKT condition $\widehat{\boldsymbol{\alpha}} = (\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\boldsymbol{\Lambda}\boldsymbol{X}\boldsymbol{\theta}^*$, and so we are interested in the norm of the solution to the above optimization problem. In order to do this, we define $\boldsymbol{\alpha}_* := \frac{\mathrm{trace}(\boldsymbol{\Lambda})}{n}\boldsymbol{\theta}^*$. As we will see later this is indeed the solution of the population version of the above loss (when $n \to \infty$). The strategy is to upper bound $\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_*\|_2$ from which we obtain a lower bound on $\|\widehat{\boldsymbol{\alpha}}\|_2$.

By the optimality of $\widehat{\boldsymbol{\alpha}}$ we have

$$\begin{aligned}
0 &\leq \frac{1}{2n}\|\boldsymbol{X}\boldsymbol{\alpha}_* - \boldsymbol{\Lambda}\boldsymbol{X}\boldsymbol{\theta}^*\|_2^2 - \frac{1}{2n}\|\boldsymbol{X}\widehat{\boldsymbol{\alpha}} - \boldsymbol{\Lambda}\boldsymbol{X}\boldsymbol{\theta}^*\|_2^2 \\
&= \frac{1}{n}(\boldsymbol{\alpha}_* - \widehat{\boldsymbol{\alpha}})^\intercal\boldsymbol{X}^\intercal(\boldsymbol{X}\boldsymbol{\alpha}_* - \boldsymbol{\Lambda}\boldsymbol{X}\boldsymbol{\theta}^*) - \frac{1}{2n}\|\boldsymbol{X}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_*)\|_2^2\,.
\end{aligned}$$

Rearranging the terms we get

$$\frac{1}{2n}\|\boldsymbol{X}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_*)\|_2^2 \leq \|\boldsymbol{\alpha}_* - \widehat{\boldsymbol{\alpha}}\|_2\,\frac{1}{n}\|\boldsymbol{X}^\intercal(\boldsymbol{X}\boldsymbol{\alpha}_* - \boldsymbol{\Lambda}\boldsymbol{X}\boldsymbol{\theta}^*)\|_2\,.$$

The left-hand side can be also lower bounded by

$$\frac{1}{2}\sigma_{\min}\Big(\frac{1}{n}\boldsymbol{X}^\intercal\boldsymbol{X}\Big)\,\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_*\|_2^2 \leq \frac{1}{2n}\|\boldsymbol{X}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_*)\|_2^2\,.$$

Combining the last two inequalities we arrive at

$$\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_*\|_2 \le \frac{2\,\|\boldsymbol{X}^\intercal(\boldsymbol{X}\boldsymbol{\alpha}_* - \boldsymbol{\Lambda}\boldsymbol{X}\boldsymbol{\theta}^*)\|_2}{n\sigma_{\min}(\boldsymbol{X}^\intercal\boldsymbol{X}/n)}\,. \tag{27}$$

By using concentration bound on the singular values of matrices with i.i.d subgaussian rows, see Vershynin (2012, Equation (5.25)), we have that with probability at least $1 - 2e^{-c_1 t^2}$,

$$\left\|\frac{1}{n}\boldsymbol{X}^\intercal\boldsymbol{X} - \boldsymbol{\Sigma}\right\|_{\rm op} \le \max(\delta_1, \delta_1^2), \quad \delta_1 = C_1\sqrt{\frac{d}{n}} + \frac{t}{\sqrt{n}}\,, \tag{28}$$

for constants $c_1, C_1 > 0$ which depend only on κ. We define the probabilistic event $\mathcal{E}_1$ as follows:

$$\mathcal{E}_1 := \left\{\left\|\frac{1}{n}\boldsymbol{X}^\intercal\boldsymbol{X} - \boldsymbol{\Sigma}\right\|_{\rm op} \le C\sqrt{\frac{d}{n}}\right\}\,,$$

for some fixed constant $C > C_1$. Then using (28) we have $\Pr(\mathcal{E}_1) \ge 1 - 2e^{-c_1 d}$, for some constant depending on C and κ.

We next bound the numerator of the right-hand side of (27). We write

$$\frac{1}{n}\,\|\boldsymbol{X}^\intercal(\boldsymbol{X}\boldsymbol{\alpha}_* - \boldsymbol{\Lambda}\boldsymbol{X}\boldsymbol{\theta}^*)\|_2 \le \left\|\Big(\frac{1}{n}\boldsymbol{X}^\intercal\boldsymbol{X} - \boldsymbol{\Sigma}\Big)\boldsymbol{\alpha}_*\right\|_2 + \left\|\boldsymbol{\Sigma}\boldsymbol{\alpha}_* - \frac{1}{n}\boldsymbol{X}^\intercal\boldsymbol{\Lambda}\boldsymbol{X}\boldsymbol{\theta}^*\right\|_2\,. \tag{29}$$

Under event $\mathcal{E}_1$ the first term is bounded by $C\sqrt{d/n}\,\|\boldsymbol{\alpha}_*\|_2$.

In addition, by its definition it is easy to see that $\boldsymbol{\Lambda}$ is a projection matrix of rank $n - m$. More specifically, it projects onto the space of vectors which are zero mean on each of the m bags. Therefore $\mathrm{trace}(\boldsymbol{\Lambda}) = n - m$ and so

$$\|\boldsymbol{\alpha}_*\|_2 = \frac{n-m}{n}\,\|\boldsymbol{\theta}^*\|_2 = (1 - \frac{1}{k})\,\|\boldsymbol{\theta}^*\|_2\,. \tag{30}$$

Hence, under the event $\mathcal{E}_1$ the first term in (29) is bounded by

$$\left\|\Big(\frac{1}{n}\boldsymbol{X}^\intercal\boldsymbol{X} - \boldsymbol{\Sigma}\Big)\boldsymbol{\alpha}_*\right\|_2 \le \left\|\Big(\frac{1}{n}\boldsymbol{X}^\intercal\boldsymbol{X} - \boldsymbol{\Sigma}\Big)\right\|_{\rm op}\|\boldsymbol{\alpha}_*\|_2 < C\,\|\boldsymbol{\theta}^*\|_2\sqrt{\frac{d}{n}}\,. \tag{31}$$

To bound the second term in the (29), we note that by definition of $\boldsymbol{\alpha}_*$,

$$\begin{aligned}\left\|\boldsymbol{\Sigma}\boldsymbol{\alpha}_* - \frac{1}{n}\boldsymbol{X}^\intercal\boldsymbol{\Lambda}\boldsymbol{X}\boldsymbol{\theta}^*\right\|_2 &= \frac{1}{n}\left\|\Big(\mathrm{trace}(\boldsymbol{\Lambda})\boldsymbol{\Sigma} - \boldsymbol{X}^\intercal\boldsymbol{\Lambda}\boldsymbol{X}\Big)\boldsymbol{\theta}^*\right\|_2 \\ &\le \frac{\|\boldsymbol{\theta}^*\|_2}{n}\,\|\mathrm{trace}(\boldsymbol{\Lambda})\boldsymbol{\Sigma} - \boldsymbol{X}^\intercal\boldsymbol{\Lambda}\boldsymbol{X}\|_{\rm op} \\ &\le \|\boldsymbol{\theta}^*\|_2\,\frac{d}{n}\Big|\mathrm{trace}(\boldsymbol{\Lambda})\boldsymbol{\Sigma} - \boldsymbol{X}^\intercal\boldsymbol{\Lambda}\boldsymbol{X}\Big|_\infty\,,\end{aligned} \tag{32}$$

where for a matrix A, the notation $|\boldsymbol{A}|_\infty$ refers to the maximum absolute values of its entries. In the last step we used the inequality $\|\boldsymbol{A}\|_{\rm op} \le d|\boldsymbol{A}|_\infty$, for symmetric $\boldsymbol{A} \in \mathbb{R}^{d\times d}$.

We next proceed by upper bounding the right-hand side of (32). We first show that the matrix of interest side has zero mean. To see this note that for any $i, j \in [d]$ we have

$$\mathbb{E}[(\boldsymbol{X}^\intercal\boldsymbol{\Lambda}\boldsymbol{X})_{ij}] = \mathbb{E}[\tilde{\boldsymbol{x}}_i^\intercal\boldsymbol{\Lambda}\tilde{\boldsymbol{x}}_j] = \mathrm{trace}(\boldsymbol{\Lambda}\mathbb{E}(\tilde{\boldsymbol{x}}_j\tilde{\boldsymbol{x}}_i^\intercal)) = \mathrm{trace}(\boldsymbol{\Lambda})\Sigma_{ij}\,,$$

where $\tilde{\boldsymbol{x}}_i$ denotes the i-th column of $\boldsymbol{X}$. Therefore, $\mathbb{E}[\boldsymbol{X}^\intercal\boldsymbol{\Lambda}\boldsymbol{X}] = \mathrm{trace}(\boldsymbol{\Lambda})\boldsymbol{\Sigma}$. We next use the (asymmetric version of) Hanson–Wright inequality (see, e.g, Vershynin (2018, Theorem 6.2.1)), by which we get that for any fixed $i, j \in [d]$,

$$\Pr\left\{|(\boldsymbol{X}^\intercal\boldsymbol{\Lambda}\boldsymbol{X})_{ij} - \mathrm{trace}(\boldsymbol{\Lambda})\Sigma_{ij}| \ge t\right\} \le 2\exp\left\{-c\min\Big(\frac{t^2}{\kappa^2 n(1-1/k)}, \frac{t}{\kappa^2}\Big)\right\}\,, \tag{33}$$

where we used the fact that $\|\mathbf{\Lambda}\|_{\rm F} = n - m = n - n/k$, since it is a projection matrix of rank $n - m$. By union bounding over the d^2 coordinates $i, j \in [d]$, we get

$$\Pr\left\{|\boldsymbol{X}^\intercal \mathbf{\Lambda} \boldsymbol{X} - \operatorname{trace}(\mathbf{\Lambda})\mathbf{\Sigma}|_\infty \geq t\right\} \leq 2d^2 \exp\left\{-c_0 \min\left(\frac{t^2}{\kappa^2 n(1-1/k)}, \frac{t}{\kappa^2}\right)\right\}. \tag{34}$$

Fix a constant $C > \sqrt{\frac{2}{c_0}}\kappa$ and define the event $\mathcal{E}_2$ as follows

$$\mathcal{E}_2 := \left\{|\boldsymbol{X}^\intercal \mathbf{\Lambda} \boldsymbol{X} - \operatorname{trace}(\mathbf{\Lambda})\mathbf{\Sigma}|_\infty \leq C\sqrt{n \log d}\right\}.$$

Using the deviation bound (34) we have $\Pr(\mathcal{E}_2) \geq 1 - 2d^{-c}$ with $c = \frac{C^2 c_0}{\kappa^2(1-1/k)} - 2 > 0$. Recalling the bound (32), on the event $\mathcal{E}_2$ we have

$$\left\|\mathbf{\Sigma}\boldsymbol{\alpha}_* - \frac{1}{n}\boldsymbol{X}^\intercal \mathbf{\Lambda} \boldsymbol{X}\boldsymbol{\theta}^*\right\|_2 \leq C \|\boldsymbol{\theta}^*\|_2 \, d\sqrt{\frac{\log d}{n}}. \tag{35}$$

Putting together equations (29), (31), (35), we obtain that on the event $\mathcal{E} := \mathcal{E}_1 \cap \mathcal{E}_2$,

$$\frac{1}{n}\|\boldsymbol{X}^\intercal(\boldsymbol{X}\boldsymbol{\alpha}_* - \Lambda \boldsymbol{X}\boldsymbol{\theta}^*)\|_2 \leq C \|\boldsymbol{\theta}^*\|_2 \, d\sqrt{\frac{\log d}{n}}, \tag{36}$$

for a constant C depending on the subgaussian norm κ. In addition on the event $\mathcal{E}_1$, we have

$$\sigma_{\min}\left(\frac{1}{n}\boldsymbol{X}^\intercal \boldsymbol{X}\right) \geq \sigma_{\min}(\mathbf{\Sigma}) - \left\|\frac{1}{n}\boldsymbol{X}^\intercal \boldsymbol{X} - \mathbf{\Sigma}\right\|_{\rm op} \geq \sigma_{\min}(\mathbf{\Sigma}) - C\sqrt{\frac{d}{n}}. \tag{37}$$

Next by combining (37) and (36) into (27), we get that

$$\|\boldsymbol{\alpha}_* - \widehat{\boldsymbol{\alpha}}\|_2 \leq \frac{Cd}{\sigma_{\min}(\mathbf{\Sigma})}\sqrt{\frac{\log d}{n}} \|\boldsymbol{\theta}^*\|_2, \tag{38}$$

for some constant $C > 0$. Note that here we used the fact that $d = o(n)$. Therefore, by using triangle inequality, on the event $\mathcal{E}$

$$\|\widehat{\boldsymbol{\alpha}}\|_2 \geq \|\boldsymbol{\alpha}_*\|_2 - \|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_*\|_2 \geq \left(1 - \frac{1}{k} - C\frac{d\sqrt{\log d}}{\sigma_{\min}(\mathbf{\Sigma})\sqrt{n}}\right)\|\boldsymbol{\theta}^*\|_2,$$

for a constant $C > 0$ that depends on the subgaussian norm κ.

We also have

$$\Pr(\mathcal{E}) = 1 - \Pr(\mathcal{E}_1^c \cup \mathcal{E}_2^c) \geq 1 - \Pr(\mathcal{E}_1^c) - \Pr(\mathcal{E}_2^c) \geq 1 - 2e^{-c_1 d} - 2d^{-c},$$

which along with the previous equation gives the desired result.

C.4. Proof of Lemma C.3

Write $\boldsymbol{S}^\intercal = [\boldsymbol{s}_1|\ldots|\boldsymbol{s}_m]$ with $\boldsymbol{s}_i \in \mathbb{R}^n$ and $\|\boldsymbol{s}_i\|_2 = 1$. We then have

$$\left\|(\boldsymbol{X}^\intercal \boldsymbol{X})^{-1}\boldsymbol{X}^\intercal \boldsymbol{S}\right\|_{\rm F}^2 = \sum_{i=1}^n \left\|(\boldsymbol{X}^\intercal \boldsymbol{X})^{-1}\boldsymbol{X}^\intercal \boldsymbol{s}_i\right\|_2^2. \tag{39}$$

We next show that for any unit vector $\boldsymbol{s}$ which is independent of data $(\boldsymbol{y}, \boldsymbol{X})$ we have

$$\left\|(\boldsymbol{X}^\intercal \boldsymbol{X})^{-1}\boldsymbol{X}^\intercal \boldsymbol{s}\right\|_2^2 \geq \frac{\operatorname{trace}(\mathbf{\Sigma}) - \sqrt{d\log d}}{n^2(\|\mathbf{\Sigma}\|_{\rm op} + c_0\sqrt{\frac{d}{n}})^2}\left(1 - 2e^{-c_1 d} - 2d^{-c}\right), \tag{40}$$

which together with (39) and the fact that $m = n/k$, implies the claim of Lemma C.3.

Define $\boldsymbol{v} := (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1}\boldsymbol{X}^\intercal \boldsymbol{s}$. Therefore, $\frac{1}{n}\boldsymbol{X}^\intercal \boldsymbol{X}\boldsymbol{v} = \frac{1}{n}\boldsymbol{X}^\intercal \boldsymbol{s}$, which implies that

$$\left\|\frac{1}{n}\boldsymbol{X}^\intercal \boldsymbol{X}\right\|_{\rm op}^2 \|\boldsymbol{v}\|_2^2 \geq \left\|\frac{1}{n}\boldsymbol{X}^\intercal \boldsymbol{s}\right\|_2^2 . \tag{41}$$

Our strategy to lower bound $\mathbb{E}[\|\boldsymbol{v}\|_2^2]$ is to upper bound the left-hand side of (41) and lower bound it right-hand side.

For the first task, recall the concentration bound (28). By taking $t = c'\sqrt{d}$ in that bound, we obtain

$$\Pr\left(\left\|\frac{1}{n}\boldsymbol{X}^\intercal \boldsymbol{X} - \boldsymbol{\Sigma}\right\|_{\rm op} \leq c_0\sqrt{\frac{d}{n}}\right) \geq 1 - 2e^{-c_1 d}, \tag{42}$$

for some constants c_0, c_1 depending on κ, the subgaussian norm of rows of $\boldsymbol{X}$. We refer to the probabilistic event in (42) by $\mathcal{E}_1$.

We then proceed to the second task, i.e., lower bounding $\left\|\frac{1}{n}\boldsymbol{X}^\intercal \boldsymbol{s}\right\|_2$. To do this, denote the columns of $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ by $\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_d \in \mathbb{R}^n$. In this notation,

$$\|\boldsymbol{X}^\intercal \boldsymbol{s}\|_2^2 = \sum_{\ell=1}^d (\tilde{\boldsymbol{x}}_\ell^\intercal \boldsymbol{s})^2 := \sum_{\ell=1}^d Z_\ell^2 . \tag{43}$$

By assumption, $Z_\ell = \tilde{\boldsymbol{x}}_\ell^\intercal \boldsymbol{s}$ are independent subgaussian random variables with $\mathbb{E}[Z_\ell^2] = \Sigma_{\ell,\ell}$ and the subgaussian norm $\|Z_\ell\|_{\psi_2} \leq C\kappa$ for a universal constant $C > 0$. Therefore, by Vershynin (2012, Remark 5.18 and Lemma 5.14), $Z_\ell^2 - \Sigma_{\ell,\ell}$ are independent centered sub-exponential random variables with $\|Z_\ell^2 - \Sigma_{\ell,\ell}\|_{\psi_1} \leq 2\|Z_\ell^2\|_{\psi_1} \leq 4\|Z_\ell\|_{\psi_2}^2 \leq 4C^2\kappa^2 := C_0$. Here, $\|\cdot\|_{\psi_1}$ refers to the subexponential norm of a random variable. We can therefore use an exponential deviation inequality, Vershynin (2012, Corollary 5.17) to control sum (43). This gives us for every $\varepsilon \geq 0$,

$$\Pr\left(\Big|\, \|\boldsymbol{X}^\intercal \boldsymbol{s}\|_2^2 - \mathrm{trace}(\boldsymbol{\Sigma})\,\Big| \geq \varepsilon d\right) = \Pr\left(\Big|\sum_{\ell=1}^d Z_\ell^2 - \mathrm{trace}(\boldsymbol{\Sigma})\,\Big| \geq \varepsilon d\right) \leq 2\exp\Big[-c\min\Big(\frac{\varepsilon^2}{C_0^2}, \frac{\varepsilon}{C_0}\Big)d\Big],$$

where $c > 0$ is an absolute constant. We take $\varepsilon = \sqrt{(\log d)/d}$ and define the probabilistic event

$$\mathcal{E}_2 := \left\{\Big|\, \|\boldsymbol{X}^\intercal \boldsymbol{s}\|_2^2 - \mathrm{trace}(\boldsymbol{\Sigma})\Big| \leq \sqrt{d\log d}\right\} .$$

By the above deviation bound we have $\Pr(\mathcal{E}_2) \geq 1 - 2d^{-c}$ for some constant $c > 0$.

We next consider the event $\mathcal{E} := \mathcal{E}_1 \cap \mathcal{E}_2$. Using (42) and the above bound on $\Pr(\mathcal{E}_2)$ we get

$$\Pr(\mathcal{E}) = 1 - \Pr(\mathcal{E}_1^c \cup \mathcal{E}_2^c) \geq 1 - \Pr(\mathcal{E}_1^c) - \Pr(\mathcal{E}_2^c) \geq 1 - 2e^{-c_1 d} - 2d^{-c} .$$

Further, on the event $\mathcal{E}$ we have

$$\left\|\frac{1}{n}\boldsymbol{X}^\intercal \boldsymbol{X}\right\|_{\rm op} \leq \|\boldsymbol{\Sigma}\|_{\rm op} + \left\|\frac{1}{n}\boldsymbol{X}^\intercal \boldsymbol{X} - \boldsymbol{\Sigma}\right\|_{\rm op} \leq \|\boldsymbol{\Sigma}\|_{\rm op} + c_0\sqrt{\frac{d}{n}} . \tag{44}$$

$$\|\boldsymbol{X}^\intercal \boldsymbol{s}\|_2^2 \geq \mathrm{trace}(\boldsymbol{\Sigma}) - \Big|\, \|\boldsymbol{X}^\intercal \boldsymbol{s}\|_2^2 - \mathrm{trace}(\boldsymbol{\Sigma})\Big| \geq \mathrm{trace}(\boldsymbol{\Sigma}) - \sqrt{d\log d} . \tag{45}$$

Therefore, by invoking (41), on the event $\mathcal{E}$ we have

$$\|\boldsymbol{v}\|_2^2 \geq \frac{\left\|\frac{1}{n}\boldsymbol{X}^\intercal \boldsymbol{s}\right\|_2^2}{\left\|\frac{1}{n}\boldsymbol{X}^\intercal \boldsymbol{X}\right\|_{\rm op}^2} \geq \frac{\mathrm{trace}(\boldsymbol{\Sigma}) - \sqrt{d\log d}}{n^2(\|\boldsymbol{\Sigma}\|_{\rm op} + c_0\sqrt{\frac{d}{n}})^2} . \tag{46}$$

Since $\|\boldsymbol{v}\|_2^2$ is non-negative by an application of Markov's inequality we get

$$\mathbb{E}[\|\boldsymbol{v}\|_2^2] \geq \frac{\mathrm{trace}(\boldsymbol{\Sigma}) - \sqrt{d\log d}}{(\|\boldsymbol{\Sigma}\|_{\rm op} + c_0\sqrt{\frac{d}{n}})^2}\Pr(\mathcal{E}) \geq \frac{\mathrm{trace}(\boldsymbol{\Sigma}) - \sqrt{d\log d}}{n^2(\|\boldsymbol{\Sigma}\|_{\rm op} + c_0\sqrt{\frac{d}{n}})^2}\left(1 - 2e^{-c_1 d} - 2d^{-c}\right) .$$

This completes the proof of (40) and concludes the proof of Lemma C.3.

C.5. Proof of Theorem 4.8

The proof is similar to the proof of Theorem 4.4. We consider the bias-variance decomposition of the upper bound given in Corollary (2.2).

We have

$$\left\|(\tilde{\boldsymbol{S}}^\intercal\tilde{\boldsymbol{S}}-\boldsymbol{I}_n)\boldsymbol{X}\boldsymbol{\theta}^*\right\|_2^2 \le 2\left\|(\boldsymbol{S}^\intercal\boldsymbol{S}-\boldsymbol{I}_n)\boldsymbol{X}\boldsymbol{\theta}^*\right\|_2^2 + 2\left\|(\boldsymbol{S}^\intercal\boldsymbol{S}-\tilde{\boldsymbol{S}}^\intercal\tilde{\boldsymbol{S}})\boldsymbol{X}\boldsymbol{\theta}^*\right\|_2^2 . \tag{47}$$

The first term is bounded in Theorem 4.4. For the second term, we bound it as

$$\begin{aligned}\left\|(\boldsymbol{S}^\intercal\boldsymbol{S}-\tilde{\boldsymbol{S}}^\intercal\tilde{\boldsymbol{S}})\boldsymbol{X}\boldsymbol{\theta}^*\right\|_2^2 &\le \left\|\boldsymbol{S}^\intercal\boldsymbol{S}-\tilde{\boldsymbol{S}}^\intercal\tilde{\boldsymbol{S}}\right\|_{\mathrm{op}}^2 \|\boldsymbol{X}\|_{\mathrm{op}}^2 \|\boldsymbol{\theta}^*\|_2^2 \\ &\le \varepsilon^2 \|\boldsymbol{\theta}^*\|_2^2 \,\sigma_{\max}(\boldsymbol{X}^\intercal\boldsymbol{X}) . \end{aligned} \tag{48}$$

Combining (47) and (48) into Corollary 2.2, we see that the mismatch between bagging configuration $\boldsymbol{S}$ and $\tilde{\boldsymbol{S}}$ contributes an inflation term to the model risk which is upper bounded by

$$\begin{aligned}&\left\|(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\right\|_{\mathrm{op}}^2 \left\|(\boldsymbol{S}\boldsymbol{S}^\intercal-\tilde{\boldsymbol{S}}\tilde{\boldsymbol{S}}^\intercal)\boldsymbol{X}\boldsymbol{\theta}^*\right\|_2^2 \\ &\le \sigma_{\max}((\boldsymbol{X}^\intercal\boldsymbol{X})^{-1})\,\varepsilon^2\,\|\boldsymbol{\theta}^*\|_2^2\,\sigma_{\max}(\boldsymbol{X}^\intercal\boldsymbol{X}) \\ &= \frac{\sigma_{\max}\left(\frac{1}{n}\boldsymbol{X}^\intercal\boldsymbol{X}\right)}{\sigma_{\min}\left(\frac{1}{n}\boldsymbol{X}^\intercal\boldsymbol{X}\right)}\varepsilon^2\,\|\boldsymbol{\theta}^*\|_2^2 .\end{aligned}$$

Note that the result of Theorem 4.4 is under an event with probability at least $1-1/n-2e^{-cd}$. Under this same event, we have $\left\|\frac{1}{n}\boldsymbol{X}^\intercal\boldsymbol{X}-\boldsymbol{\Sigma}\right\|_{\mathrm{op}} \le C\sqrt{\frac{d}{n}}$ (see Equation (23)). Therefore, by Weyl's inequality for singular values we have

$$\sigma_{\max}\left(\frac{1}{n}\boldsymbol{X}^\intercal\boldsymbol{X}\right) \le \sigma_{\max}(\boldsymbol{\Sigma}) + C\sqrt{\frac{d}{n}}, \quad \sigma_{\min}\left(\frac{1}{n}\boldsymbol{X}^\intercal\boldsymbol{X}\right) \ge \sigma_{\min}(\boldsymbol{\Sigma}) - C\sqrt{\frac{d}{n}},$$

which completes the proof of theorem.

D. Missing analysis from Section 5

D.1. Proof of Lemma 5.1

We build on the observation in Lemma 2.3 about the sorted structure of an optimal solution. First, sort the points by their $\widetilde{y}_i$ value in $O(n\log n)$ time. Next, we present a dynamic programming algorithm that optimally slices the sorted list, i.e., a stars-and-bars partition where each part has size at least k.

Define the function $f_k(i)$ to be the objective of an optimal solution for the subproblem defined by the first i points. It follows that

$$f_k(i) = \begin{cases} \infty & \text{if } i<0 \\ 0 & \text{if } i=0 \\ \min_{k\le s\le i}\left\{f_k(i-s)+\sum_{j=i-s+1}^{i}(\widetilde{y}_j-\mu_{i,s})^2\right\} & \text{if } i\ge 1 \end{cases}$$

where

$$\mu_{i,s} = \frac{1}{s}\sum_{j=i-s+1}^{i}\widetilde{y}_j.$$

This recurrence considers all suffixes of size $s\ge k$ as the last cluster, computes their sum of squares error, and recursively solves the subproblem on the remaining points via $f_k(i-s)$. This naively leads to an $O(n^3)$-time dynamic programming algorithm. However, there are two observations that allow us to reduce the running time to $O(nk)$:

1. We can assume each cluster in an optimal solution has size $k\le s<2k$. If not, we can split a cluster of size $s\ge 2k$ into two parts without increasing the objective. It follows that we can compute each $f_k(i)$ by considering $O(k)$ recursive states.

2. We can iteratively compute the sum of squared errors $d(i,s) := \sum_{j=i-s+1}^{i} (\widetilde{y}_j - \mu_{i,s})^2$ in constant time, as shown in Wang & Song (2011):

$$d(i,s) = d(i,s-1) + \frac{s-1}{s}(\widetilde{y}_{i-s+1} - \mu_{i,s-1})^2$$
$$\mu_{i,s} = \frac{\widetilde{y}_{i-s+1} + (s-1)\mu_{i,s-1}}{s}$$

   This means each value of $f_k(i)$ can be computed in $O(k)$ time.

Putting everything together, we can compute $f_k(n)$ and reconstruct an optimal clustering in $O(nk)$ time after sorting.