How many labelers do you have? A closer look at gold-standard labels

Chen Cheng* Hilal Asi† John Duchi‡

June 23, 2022

Abstract

The construction of most supervised learning datasets revolves around collecting multiple labels for each instance, then aggregating the labels to form a type of "gold-standard.". We question the wisdom of this pipeline by developing a (stylized) theoretical model of this process and analyzing its statistical consequences, showing how access to non-aggregated label information can make training well-calibrated models easier or—in some cases—even feasible, whereas it is impossible with only gold-standard labels. The entire story, however, is subtle, and the contrasts between aggregated and fuller label information depend on the particulars of the problem, where estimators that use aggregated information exhibit robust but slower rates of convergence, while estimators that can effectively leverage all labels converge more quickly *if* they have fidelity to (or can learn) the true labeling process. The theory we develop in the stylized model makes several predictions for real-world datasets, including when non-aggregate labels should improve learning performance, which we test to corroborate the validity of our predictions.

1 Introduction

The centrality of data collection to the development of machine learning is evident [11], with numerous challenge datasets [22, 20, 1, 18, 10, 31] driving advances. Essential to each of these is the collection of *labeled data*. While in the past, experts could provide reliable labels for reasonably sized datasets, the cost and size of modern datasets often precludes this expert annotation, motivating a growing literature on crowdsourcing and other sophisticated dataset generation strategies that aggregate expert and non-expert feedback [16, 31, 28]. By aggregating multiple labels, one typically hopes to obtain clean, true, "gold-standard" data. Yet most statistical machine learning development—theoretical or methodological—does not investigate this full data generating process, assuming only that data comes in the form of (X, Y) pairs of covariates X and targets (labels) Y [38, 4, 2, 14]. Here, we argue for a more holistic perspective: broadly, that analysis and algorithmic development should focus on the more complete machine learning pipeline, from dataset construction to model output; and more narrowly, questioning such aggregation strategies and the extent to which such cleaned data is essential or even useful.

To that end, we develop a stylized theoretical model to capture uncertainties in the labeling process, allowing us to understand the contrasts, limitations and possible improvements of using aggregated or non-aggregated data in a statistical learning pipeline. We model each example as a pair $(X_i, (Y_{i1}, \ldots, Y_{im}))$ where X_i is a data point and Y_{ij} are noisy labels. In the most basic

*Department of Statistics, Stanford University; email: chencheng@stanford.edu.
†Department of Electrical Engineering, Stanford University; email: asi@stanford.edu.
‡Departments of Statistics and Electrical Engineering, Stanford University; email: jduchi@stanford.edu.

formulation of our results, we compare two methods: empirical risk minimization using all the labels, and empirical risk minimization using cleaned labels $\overline{Y}$ based on majority vote. While this grossly simplifies modern crowdsourcing and other label aggregation strategies [9, 31, 28], the simplicity allows (i) us to understand fundamental limitations of algorithms based on majority-vote aggregation, and (ii) circumventing these limits by using full, non-aggregated information. By carefully analyzing these models, we show (among other results) that training a calibrated model is essentially infeasible using majority-vote aggregation and that classification error of models fit using non-aggregated label information outperforms that of "standard" estimators that use aggregated (cleaned, majority-vote) labels. We develop several extensions to these basic results, including misspecified models, semiparametric scenarios where one must learn link functions, and simple models of learned annotator reliability. The message is consistent throughout: *if* it is possible to model the labeling process with fidelity, using all the labels yields more efficient estimators, while majority-vote estimators provide robust (but slower) convergence.

While our models are stylized, they also make several concrete and testable predictions for real datasets; if our approach provides a useful abstraction, it must suggest improvements in learning even for more complex and challenging to analyze scenarios. Indeed, our theory predicts that methods that fit predictive models on non-aggregated data should both make better-calibrated predictions and, in general, have lower classification error than models that use aggregated clean labels. To that end, we consider two real datasets, and they corroborate the predictions and implications of our theory even beyond logistic models. In particular, majority-vote based algorithms yield uncalibrated models in all experiments, whereas the algorithms that use full-label information train (more) calibrated models. Moreover, the former algorithms exhibit worse classification error in our experiments, with the error gap depending on parameters—such as inherent label noise—that we can also address in our theoretical models.

1.1 Problem formulation

To situate our results, we begin by providing the key ingredients in the paper.

The model with multiple labels. Consider a binary classification problem with data points $X_1, \cdots, X_n \stackrel{\text{iid}}{\sim} \mathbb{P}_X$, $X_i \in \mathbb{R}^d$, with m labelers. We assume each labeler annotates data points independently through a generalized linear model, and the labelers use m possibly different link functions $\sigma_1^\star, \cdots, \sigma_m^\star \in \mathcal{F}_{\mathsf{link}}$, where

$$\mathcal{F}_{\mathsf{link}} := \{\sigma : \mathbb{R} \to [0,1] \mid \sigma(0) = 1/2, \mathsf{sign}(\sigma(t) - 1/2) = \mathsf{sign}(t)\}.$$

Here $\mathsf{sign}(t) = -1$ for $t < 0$, $\mathsf{sign}(t) = 1$ for $t > 0$ and $\mathsf{sign}(0) = 0$. If $\sigma(t) + \sigma(-t) = 1$ for all $t \in \mathbb{R}$, we say the link function is symmetric and denote the class of symmetric functions by $\mathcal{F}_{\mathsf{link}}^0 \subset \mathcal{F}_{\mathsf{link}}$. The link functions generate labels via the distribution

$$\mathbb{P}_{\sigma,\theta}(Y = y \mid X = x) = \sigma(y\langle\theta, x\rangle), \quad \text{for } y \in \{\pm 1\}, \quad x, \theta \in \mathbb{R}^d. \tag{1}$$

Key to our stylized model, and what allows our analysis, is that we assume labelers use the same linear classifier $\theta^\star \in \mathbb{R}^d$—though each labeler j may have a distinct link $\sigma_j^\star$—so we obtain conditionally independent labels $Y_{ij} \sim \mathbb{P}_{\sigma_j^\star,\theta^\star}(\cdot \mid X_i)$. (For example, in the logistic model where labelers have identical link, $\sigma_j^\star(t) = 1/\left(1 + e^{-t}\right)$.) We seek to recover $\theta^\star$ or the direction $u^\star := \theta^\star / \left\|\theta^\star\right\|_2$ from the observations (X_i, Y_{ij}).

Classification and calibration. For an estimator $\widehat{\theta}$ and associated direction $\widehat{u} := \widehat{\theta}/\|\widehat{\theta}\|_2$, we measure performance through

$$\textbf{(i)} \quad \text{The classification error:} \quad \|u^\star - \widehat{u}\|_2.$$
$$\textbf{(ii)} \quad \text{The calibration error:} \quad \|\theta^\star - \widehat{\theta}\|_2.$$

We term these "classification error" and "calibration error" from the rationale that for classification, we only need to control the difference between the directions $\widehat{u}$ and $u^\star$, while calibration—that for a new data point X, the value $\sigma_j^\star(\langle\widehat{\theta}, X\rangle)$ is close to $\mathbb{P}_{\sigma_j^\star,\theta^\star}(Y = 1 \mid X) = \sigma_j^\star(\langle\theta^\star, X\rangle)$—requires controlling the error in $\widehat{\theta}$ as an estimate of $\theta^\star$.

Estimators. We consider two types of estimators: one using aggregated labels and the other using each label from different annotators. At the highest level, the aggregated estimator depends on processed labels $\overline{Y}_i$ for each example X_i, while the non-aggregated estimator uses all labels $Y_{i1}, \ldots, Y_{im}$. To center the discussion, we provide two concrete instantiations via logistic regression (with generalizations in the sequel). For the logistic link $\sigma^{\mathrm{lr}}(t) = \frac{1}{1+e^{-t}}$, define the logistic loss

$$\ell_\theta^{\mathrm{lr}}(y \mid x) = -\log \mathbb{P}_{\sigma^{\mathrm{lr}},\theta}(y \mid x) = \log(1 + e^{-y\langle x,\theta\rangle}).$$

In the non-aggregated model, we let let $\mathbb{P}_{n,m}$ be the empirical measure on $\{(X_i, (Y_{i1}, \ldots, Y_{im}))\}$ and consider the logistic regression estimator

$$\widehat{\theta}_{n,m}^{\mathrm{lr}} = \operatorname*{argmin}_\theta \left\{ \mathbb{P}_{n,m}\ell_\theta^{\mathrm{lr}} = \frac{1}{nm}\sum_{i=1}^n \sum_{j=1}^m \ell_\theta^{\mathrm{lr}}(Y_{ij} \mid X_i) \right\}, \tag{2}$$

which is the maximum likelihood estimator (MLE) assuming the logistic model is true. We focus on the simplest aggregation strategy, where example i has majority vote label

$$\overline{Y}_i = \mathsf{maj}(Y_{i1}, \ldots, Y_{im}).$$

Then letting $\overline{\mathbb{P}}_{n,m}$ be the empirical measure on $\{(X_i, \overline{Y}_i)\}$, the majority-vote estimator solves

$$\widehat{\theta}_{n,m}^{\mathrm{mv}} = \operatorname*{argmin}_\theta \left\{ \overline{\mathbb{P}}_{n,m}\ell_\theta^{\mathrm{lr}} = \frac{1}{n}\sum_{i=1}^n \ell_\theta^{\mathrm{lr}}(\overline{Y}_i \mid X_i) \right\}. \tag{3}$$

Method (3) acts as our proxy for the "standard" data analysis pipeline, with cleaned labels, while method (2) is our proxy for non-aggregated methods using all labels. A more general formulation than the majority vote (3) could allow complex aggregation strategies, e.g., crowdsourcing, but we abstract away details to capture what we view as the essentiae for statistical learning problems (e.g. CIFAR [18] or ImageNet [31]) where only aggregated label information is available.

Our main technical approaches characterize the estimators $\widehat{\theta}_{n,m}^{\mathrm{mv}}$ and $\widehat{\theta}_{n,m}^{\mathrm{lr}}$ via asymptotic calculations. Under appropriate assumptions on the data generating mechanisms (1), which will include misspecification, we both (i) provide consistency results that elucidate the infinite sample limits for $\widehat{\theta}_{n,m}^{\mathrm{mv}}$, $\widehat{\theta}_{n,m}^{\mathrm{lr}}$, and a few more general estimators, and (ii) carefully evaluate their limit distributions via asymptotic normality calculations. The latter allows direct comparisons between the different estimators through their limiting covariances, which exhibit (to us) interestingly varying dependence on m and the scaling of the true parameter $\theta^\star$.

1.2 Summary of theoretical results and implications

We obtain several results clarifying the distinctions between estimators that use aggregate labels from those that treat them individually.

Impossibility of calibration with majority vote We begin in Section 2 with the simple result that any estimator based on majority vote aggregation cannot be calibrated: there exist distributions with distinct numbers of labelers m and link functions σ that induce identical joint distributions on $(X, \overline{Y})$. This contrasts with non-aggregated data, which allows calibrated estimators.

Improved performance using multiple labels for well-specified models As in our discussion above, our main approach to highlighting the import of multiple labels is through asymptotic characterization of the estimators (2)–(3) and similar estimators. We begin this in Section 3 by focusing on the particular case that the labelers follow a logistic model. As specializations of our major results to come, we show that the multi-label MLE (2) is calibrated and enjoys faster rates of convergence (in m) than the majority-vote estimator (3). The improvements depend in subtle ways on the particulars of the underlying distribution, and we connect them to Mammen-Tsybakov-type noise conditions in Propositions 2 and 3. In "standard" cases (say, Gaussian features), the improvement scales as $\sqrt{m}$; for problems with little classification noise, the majority vote estimator becomes brittle (and relatively slower), while the convergence rate gap decreases for noisier problems.

Robustness of majority-vote estimators Nonetheless, our results also evidence support for majority vote estimators of the form (3). Indeed, in Section 4 we provide master results and consequences that hold for both well- and mis-specified losses, which highlight the robustness of the majority vote estimator. While MLE-type estimators (2) enjoy faster rates of convergence when the model is correct, these rates break down, in ways we make precise when the model is mis-specified; in contrast, majority-vote-type estimators (3) maintain their (not quite optimal) convergence guarantees, yielding $\sqrt{m}$-rate improvements.

Semi-parametric approaches The final theoretical component of the paper is to put the pieces together: in Section 5 we show how to achieve the best of both worlds via semi-parametric approaches. We highlight two applications. In the first, we use an initial estimator to fit a link function, then produce a refined estimator minimizing the induced semiparametric loss and recovering efficient estimation. In the second, we highlight how our results provide potential insights into existing crowdsourcing techniques by leveraging a blackbox crowdsourcing algorithm providing measures of labeler reliability to achieve (optimal) estimates.

1.3 Related work

We briefly overview related work, acknowledging that to make our theoretical model tractable, we capture only a few of the complexities inherent in dataset construction. Label aggregation strategies often attempt to evaluate labeler uncertainty, dating to Dawid and Skene [9], who study labeler uncertainty estimation to overcome noise in clinical patient measurements. With the rise of crowdsourcing, such reliability models have attracted substantial recent interest, with approaches for optimal budget allocation [17], addressing untrustworthy or malicious labelers [7], and more broadly an intensive line of work studying crowd labeling and aggregation [41, 40, 35, 28], with substantial applications [10, 31].

The focus in many of these, however, is to obtain a single clean and trustworthy label for each example. Thus, while these aggregation techniques have been successful and important, our work takes a different perspective. First, we target statistical analysis for the full learning pipeline—to understand the theoretical landscape of the learning problem with multiple labels for each example—as opposed to obtaining only clean labels. Moreover, we argue for an increased focus on calibration of the resulting predictive model, which aggregated (clean) labels necessarily impede. We therefore adopt a perspective similar to Peterson et al. and Platanios et al.'s applied work [24, 27], which highlight ways that incorporating human uncertainty into learning pipelines can make classification "more robust" [24]. Instead of splitting the learning procedure into two phases, where the first aggregates labels and the second trains, we simply use non-aggregated labels throughout learning. Platanios et al. [27] propose a richer model than our stylized scenarios, directly modeling labeler reliability, but the simpler approaches we investigate allow us to be theoretically precise about the limiting behavior and performance of the methods.

We mention in passing that our precise consistency results rely on distributional assumptions on the covariates X, for example, that they are Gaussian. That such technical conditions appear may be familiar from recent work, for example, in single-index or nonlinear measurement models [26, 25]. In such settings, one assumes $\mathbb{E}[Y \mid X] = f(\langle \theta^\star, X \rangle)$ for an unknown increasing f, and it is essential that $\mathbb{E}[YX] \propto \theta^\star$ to allow estimation of $\theta^\star$; we leverage similar techniques.

Notation We use $\|x\|_p$ to denote the ℓ_p norm of a vector x. For a matrix M, $\|M\|$ is its spectral norm, and $M^\dagger$ is its Moore-Penrose pseudo inverse. For a unit vector $u \in \mathbb{R}^d$, the projection operator to the orthogonal space of $\mathrm{span}\{u\}$ is $\mathsf{P}_u^\perp = I - uu^\top$. We use the notation $f(n) \asymp g(n)$ for $n \in \mathbb{N}$ and $f(x) \asymp g(x)$ for $x \in \mathbb{R}_+$ if there exist numerical constants c_1, c_2 and $n_0, x_0 \geq 0$ such that $c_1|g(n)| \leq |f(n)| \leq c_2|g(n)|$ and $c_2|g(x)| \leq |f(x)| \leq c_2|g(x)|$ for $n \geq n_0$ and $x \geq x_0$. We also use the empirical process notation $\mathbb{P}Z = \int z d\mathbb{P}(z)$. We let $c = o_m(1)$ denote that $c \to 0$ as $m \to \infty$.

2 Motivation: the impossibility of calibration with majority vote

Our theoretical analysis begins with a few simple observations that at least suggest the importance of using non-aggregated labels: calibration for labeler uncertainty is, in a sense, impossible with only aggregated data, which is in strong contrast to algorithms that use non-aggregated data. Formalizing, we consider a generalized linear model for binary classification, where we have a link function $\sigma^\star \in \mathcal{F}^0_{\mathsf{link}}$. Given $\theta^\star \in \mathbb{R}^d$, we define the model (1) with $\mathbb{P}_{\sigma^\star,\theta^\star}(Y = y \mid X = x) = \sigma^\star(y\langle \theta^\star, x \rangle)$, where $x \in \mathbb{R}^d$ and $y \in \{\pm 1\}$. We generate n data points $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(0, I_d)$, then for each data point X_i we generate m labels $Y_{ij} \overset{\text{iid}}{\sim} \mathbb{P}_{\sigma^\star,\theta^\star}(\cdot \mid X_i)$, $j = 1, \ldots, m$. In this simplified model, no algorithm using majority-vote aggregation can be calibrated without additional information on m and $\sigma^\star$: if $\overline{Y}_i = \mathsf{maj}(Y_{i1}, \ldots, Y_{im})$ denotes the majority-vote (breaking ties randomly), there always exists a calibration function $\overline{\sigma}$ and label size $\overline{m}$ such that $\overline{Y}_i$ has identical distribution under both the model (1) and also with $\overline{\sigma}, \overline{m}$ replacing $\sigma^\star$ and m. Letting $\mathbb{P}^{\sigma^\star,\theta^\star,m}_{(X,\overline{Y})}$ denote the distribution of $(X_i, \overline{Y}_i)$, we have the following result. We defer the proof to Appendix B.

Proposition 1. *Suppose $\sigma^\star \in \mathcal{F}^0_{\mathsf{link}}$ satisfying $\sigma^\star(t) > 0$ for all $t > 0$. For any $\overline{\theta} \in \mathbb{R}^d$ such that $\overline{\theta}/\left\|\overline{\theta}\right\|_2 = \theta^\star/\left\|\theta^\star\right\|_2$ and any positive integer $\overline{m}$, there is another calibration function $\overline{\sigma}$ such that*

$$\mathbb{P}^{\sigma^\star,\theta^\star,m}_{(X,\overline{Y})} \overset{\text{dist}}{=} \mathbb{P}^{\overline{\sigma},\overline{\theta},\overline{m}}_{(X,\overline{Y})}.$$

3 The well-specified logistic model

With the simple impossibility result for majority vote in Proposition 1, we turn our attention to a setting that allows more precise comparisons between a method using aggregated labels and one without by considering the logistic model for the link (1),

$$\sigma^{\mathrm{lr}}(t) = \frac{1}{1+e^{-t}} \in \mathcal{F}^0_{\mathsf{link}}.$$

We present initial results for the estimators (2) and (3) here, as the results highlight many of the conclusions we draw and are relatively clean to present in this setting. In particular, we assume identical links $\sigma_1^\star = \cdots = \sigma_m^\star = \sigma^{\mathrm{lr}}$, have an i.i.d. sample $X_1, \ldots, X_n$, where for each i we draw $(Y_{i1}, \ldots, Y_{im}) \stackrel{\mathrm{iid}}{\sim} \mathbb{P}_{\sigma^{\mathrm{lr}},\theta^\star}(\cdot \mid X_i)$ for a true vector $\theta^\star$.

3.1 The isotropic Gaussian case

To better understand the performance of the full information (2) and majority vote (3) approaches and to deliver the general taste of our results, we start by studying the simplest case when $X \sim \mathsf{N}(0, I_d)$.

Performance with non-aggregated data. We begin with a relatively standard [36] analysis of the non-aggregated MLE estimator $\widehat{\theta}^{\mathrm{lr}}_{n,m}$ in Eq. (2), which we state as a corollary of a result where X has more general distribution in Proposition 2 to come.

Corollary 1. *Let $X \sim \mathsf{N}(0, I_d)$ and $t^\star = \|\theta^\star\|_2$. The maximum likelihood estimator $\widehat{\theta}^{\mathrm{lr}}_{n,m}$ is consistent $\widehat{\theta}^{\mathrm{lr}}_{n,m} \stackrel{p}{\to} \theta^\star$. There exists a function $C(t)$ such that $\lim_{t\to\infty} C(t)t$ exists and is finite and*

$$\sqrt{n}(\widehat{u}^{\mathrm{lr}}_{n,m} - u^\star) \stackrel{d}{\to} \mathsf{N}\left(0, m^{-1}C(t^\star)\mathsf{P}^\perp_{u^\star}\right),$$

where $\mathsf{P}^\perp_{u^\star} = I_d - u^\star u^{\star\top}$.

The first part of Corollary 1 demonstrates that the non-aggregated MLE classifier is calibrated: it recovers both the direction and scale of $\theta^\star$. Moreover, the second part shows that this classifier enjoys convergence rates that roughly scale as $O(1)/\sqrt{nm}$, so that a linear increase in the label size m roughly yields a linear increase in convergence rate.

Performance with majority-vote aggregation. The analysis of the majority vote estimator (3) requires more care, though the assumption that $X \sim \mathsf{N}(0, I_d)$ allows us to calculate the limits explicitly. In brief, we show that when X is Gaussian, the estimator is not calibrated and has slower convergence rates in m for classification error than the non-aggregated classifier. The basic idea is that a classifier fit using majority vote labels $\overline{Y}_i$ should still point in the direction of $\theta^\star$, but it should be (roughly) "calibrated" to the probability of a majority of m labels being correct.

We follow this idea and sketch the derivation here, as it is central to all of our coming theorems, and then state the companion corollary to Corollary 1. Each result depends on the probability

$$\mathbb{P}(\overline{Y} = \mathsf{sign}(\langle x, \theta^\star\rangle) \mid x) = \rho_m(|\langle x, \theta^\star\rangle|)$$

of obtaining a correct label using majority vote, where ρ_m defines the binomial probability function

$$\rho_m(t) = \mathbb{P}\left(\mathsf{Binomial}\Big(m, \frac{1}{1+e^{-|t|}}\Big) \geq \frac{m}{2}\right) = \sum_{i=\lceil m/2\rceil}^{m} \binom{m}{i} \left(\frac{1}{1+e^{-|t|}}\right)^i \left(\frac{e^{-|t|}}{1+e^{-|t|}}\right)^{m-i}, \quad (4)$$

when m is odd. (When m is even the final sum has the additional additive term $\frac{1}{2}\binom{m}{m/2}\frac{e^{-m|t|/2}}{(1+e^{-|t|})^m}$.) Key to the coming result is choosing a parameter to roughly equalize binomial (majority vote) and logistic (Bernoulli) probabilities, and so for $Z \sim \mathsf{N}(0,1)$ we define the function

$$h_m(t) = \mathbb{E}\left[|Z|(1-\rho_m(t^\star|Z|))\right] - \mathbb{E}\left[\frac{|Z|}{1+e^{t|Z|}}\right]. \tag{5}$$

We use h to find the minimizer of population loss $L_m^{\mathrm{mv}}(\theta) = \mathbb{E}[\ell_\theta^{\mathrm{lr}}(\overline{Y} \mid X)]$ by considering the ansatz that $\theta = tu^\star$ for some $t > 0$. Using the definition (4) of ρ_m, we can write

$$L_m^{\mathrm{mv}}(\theta) = \mathbb{E}\left[\log(1+\exp(-S\langle X,\theta\rangle)) \cdot \rho_m(t^\star|\langle X,u^\star\rangle|)\right] + \mathbb{E}\left[\log(1+\exp(S\langle X,\theta\rangle)) \cdot (1-\rho_m(t^\star|\langle X,u^\star\rangle|))\right],$$

where $S = \mathsf{sign}(\langle X,\theta^\star\rangle)$, and compute the gradient

$$\nabla L_m^{\mathrm{mv}}(\theta) = -\mathbb{E}\left[\frac{S}{1+\exp(S\langle X,\theta\rangle)}\rho_m(t^\star|\langle X,u^\star\rangle|)X\right] + \mathbb{E}\left[\frac{S\exp(S\langle X,\theta\rangle)}{1+\exp(S\langle X,\theta\rangle)}(1-\rho_m(t^\star|\langle X,u^\star\rangle|))X\right].$$

We set $Z = \langle X,u^\star\rangle$ and decompose X into the independent sum $X = (X - u^\star Z) + u^\star Z$. Substituting in $\theta = tu^\star$ yields

$$\begin{aligned}
\nabla L_m^{\mathrm{mv}}(tu^\star) &\overset{\text{(i)}}{=} -\mathbb{E}\left[\frac{\mathsf{sign}(Z)}{1+\exp(t|Z|)}\rho_m(t^\star|Z|)X\right] + \mathbb{E}\left[\frac{\mathsf{sign}(Z)\exp(t|Z|)}{1+\exp(t|Z|)}(1-\rho_m(t^\star|Z|))X\right] \\
&= -\mathbb{E}\left[\frac{\mathsf{sign}(Z)Z}{1+\exp(t|Z|)}\rho_m(t^\star|Z|)\right]u^\star + \mathbb{E}\left[\frac{\mathsf{sign}(Z)Z\exp(t|Z|)}{1+\exp(t|Z|)}(1-\rho_m(t^\star|Z|))\right]u^\star \\
&= \left(\mathbb{E}\left[\frac{|Z|\exp(t|Z|)}{1+\exp(t|Z|)}\right] - \mathbb{E}\left[|Z|\rho_m(t^\star|Z|)\right]\right)u^\star \\
&= \left(\mathbb{E}\left[|Z|(1-\rho_m(t^\star|Z|))\right] - \mathbb{E}\left[\frac{|Z|}{1+e^{t|Z|}}\right]\right)u^\star = h_m(t)u^\star,
\end{aligned} \tag{6}$$

where in (i) we substitute $S = \mathsf{sign}(Z)$. As we will present in Corollary 2, $h_m(t) = 0$ has a unique solution $t_m \asymp \sqrt{m}$, and the global minimizer of the population loss L_m^{mv} is thus exactly $t_m u^\star$.

By completing the calculations for the precise value of t_m above and a performing few asymptotic normality calculations, we have the following result, a special case of Proposition 3 to come.

Corollary 2. *Let $X \sim \mathsf{N}(0, I_d)$ and $t^\star = \|\theta^\star\|_2$. There are numerical constants $a, b > 0$ such that the following hold: for the function $h = h_m$ in (5), there is a unique $t_m \geq t_1 = t^\star$ solving $h(t_m) = 0$ and*

$$\widehat{\theta}_{n,m}^{\mathrm{mv}} \xrightarrow{p} t_m u^\star \quad \textit{and} \quad \lim_{m\to\infty} \frac{t_m}{t^\star\sqrt{m}} = a.$$

Moreover, there exists a function $C_m(t) = \frac{b}{t\sqrt{m}}(1+o_m(1))$ as $m \to \infty$ such that $\widehat{u}_{n,m}^{\mathrm{mv}} = \widehat{\theta}_{n,m}^{\mathrm{mv}}/\|\widehat{\theta}_{n,m}^{\mathrm{mv}}\|_2$ satisfies

$$\sqrt{n}\left(\widehat{u}_{n,m}^{\mathrm{mv}} - u^\star\right) \xrightarrow{d} \mathsf{N}\left(0, C_m(t^\star)\mathsf{P}_{u^\star}^{\perp}\right).$$

It is instructive to compare the rates of this estimator to the rates for the non-aggregated MLE in Corollary 1. First, the non-aggregated estimator is calibrated in that $\widehat{\theta}_{n,m}^{\mathrm{lr}} \to \theta^\star$, in contrast to the majority-vote estimator, which roughly "calibrates" to the probability majority vote is correct (cf. (6)) via the convergence $\widehat{\theta}_{n,m}^{\mathrm{mv}} \to c\sqrt{m}\theta^\star$ as $n \to \infty$. The scaling of C_m in Corollary 2 is also important: the majority-vote estimator exhibits worse convergence rates by a factor of $\sqrt{m}$ than the estimator $\widehat{\theta}_{n,m}^{\mathrm{lr}}$: for constants c^{lr} and c^{mv} that depend only on $t^\star = \|\theta^\star\|_2$ and $\Sigma = I - u^\star u^{\star\top}$, we have asymptotic variances differing by $\sqrt{m}$:

$$\sqrt{n}(\widehat{u}_{n,m}^{\mathrm{lr}} - u^\star) \xrightarrow{d} \mathsf{N}\left(0, m^{-1}c^{\mathrm{lr}}\Sigma \cdot (1+o_m(1))\right) \text{ while } \sqrt{n}(\widehat{u}_{n,m}^{\mathrm{mv}} - u^\star) \xrightarrow{d} \mathsf{N}\left(0, m^{-1/2}c^{\mathrm{mv}}\Sigma \cdot (1+o_m(1))\right).$$

3.2 Comparisons for more general feature distributions

The key to the preceding results—and an indication that they are stylized—is that the covariates X decompose into components aligned with $\theta^\star$ and independent noise. Here, we abstract away the Gaussianity assumptions to allow a more general treatment. This generality also allows a more nuanced development carefully tracking label noise, as margin and noise conditions turn out to play a strong role in the relative merits of maximum-likelihood-type (full label information) estimators versus those using cleaned majority-vote labels. The results, as in Sec. 3.1, are consequences of the masters theorem to come in the sequel.

We first make our independence assumption.

Assumption A1. *The covariates X have non-singular covariance Σ and decompose as a sum of independent random vectors in the span of $u^\star$ and its complement*

$$X = W + Zu^\star, \quad \textit{where} \quad W \perp\!\!\!\perp Z, \;\; \langle W, u^\star \rangle = 0, \;\; \mathbb{E}[W] = 0, \;\; \mathbb{E}[Z] = 0.$$

Under these assumptions, we develop a characterization of the limiting behavior of the majority vote and non-aggregated models based on classification difficulty, adopting Mammen and Tsybakov's perspective [21] and measuring difficulty of classification through the proximity of the probability $\mathbb{P}(Y = 1 \mid X = x)$ to $1/2$. Thus, for a *noise exponent* $\beta \in (0, \infty)$, we consider the condition

$$\mathbb{P}\left(\left|\mathbb{P}(Y = 1 \mid X) - \frac{1}{2}\right| \le \epsilon\right) = O(\epsilon^\beta). \tag{M_β}$$

We see that as $\beta \uparrow \infty$ the problem becomes "easier" as it is less likely to have a small margin—in particular, $\beta = \infty$ gives a hard margin that $|\mathbb{P}(Y = 1 \mid X) - \frac{1}{2}| \ge \epsilon$ for all small ϵ. Under the independent decomposition Assumption A1, the noise condition (M_β) solely depends on the covariate's projection onto the signal Z. We therefore consider the following assumption on Z.

Assumption A2. *For a given $\beta > 0$, Z is (β, c_Z)-*regular*, meaning that the absolute value $|Z|$ has density $p(z)$ on $(0, \infty)$, no point mass at 0, and satisfies*

$$\sup_{z \in (0,\infty)} z^{1-\beta} p(z) < \infty, \qquad \lim_{z \to 0} z^{1-\beta} p(z) = c_Z \in (0, \infty).$$

As the logistic function $\sigma^{\mathrm{lr}}(t) = 1/(1 + e^{-t})$ satisfies $\sigma^{\mathrm{lr}\prime}(0) = 1/4$, for $t^\star = \|\theta^\star\|_2$ in our logistic model (1) we have $\mathbb{P}(Y = 1 \mid X = W + u^\star Z, Z = z) = \sigma^{\mathrm{lr}}(t^\star z) = 1/(1 + e^{-t^\star z})$. More generally, for any link function σ differentiable at 0 with $\sigma'(0) > 0$, we have $\mathbb{P}_\sigma(Y = 1 \mid Z = z) = \sigma(t^\star z) = \frac{1}{2} + \sigma'(0) t^\star z + o(t^\star z)$, so that the Mammen-Tsybakov noise condition (M_β) is equivalent to

$$\mathbb{P}\left(|t^\star Z| \le \epsilon\right) = O(\epsilon^\beta).$$

Thus, under Assumption A2, condition (M_β) holds, as by dominated convergence we have

$$\mathbb{P}\left(|t^\star Z| \le \epsilon\right) = \int_0^{\epsilon/t^\star} p(z) dz = \int_0^{\epsilon/t^\star} c_Z (1 + o_\epsilon(1)) z^{\beta - 1} dz = \frac{c_Z}{\beta} \epsilon^\beta \cdot (1 + o_\epsilon(1)),$$

As a concrete case, when the features X are isotropic Gaussian and so $Z \sim \mathsf{N}(0, 1)$, $\beta = 1$. We provide extensions of Corollaries 1 and 2 in the more general cases the noise exponent β allows.

The maximum likelihood estimator retains its convergence guarantees in this setting, and we can be more precise for the analogue of the final claim of Corollary 1 (see Appendix G.1 for a proof):

Proposition 2. *Let Assumptions A1 and A2 hold for some $\beta > 0$ $t^\star = \|\theta^\star\|_2$. Let $L^{\rm lr}(\theta) = \mathbb{E}[\ell_\theta^{\rm lr}(Y \mid X)]$ be the population logistic loss. Then the maximum likelihood estimator (2) satisfies*

$$\sqrt{n}\left(\widehat{\theta}_{n,m}^{\rm lr} - \theta^\star\right) \stackrel{d}{\to} \mathsf{N}(0, m^{-1}\nabla^2 L^{\rm lr}(\theta^\star)^{-1}).$$

Moreover,

$$\sqrt{n}\left(\widehat{u}_{n,m}^{\rm lr} - u^\star\right) \stackrel{d}{\to} \mathsf{N}\left(0, m^{-1}(t^\star)^{-2}\mathsf{P}_{u^\star}^\perp \nabla^2 L^{\rm lr}(\theta^\star)^{-1}\mathsf{P}_{u^\star}^\perp\right),$$

and there exists $C(t)$ such that $\lim_{t\to\infty} C(t)t^{2-\beta}$ exists and is finite such that

$$\sqrt{n}\left(\widehat{u}_{n,m}^{\rm lr} - u^\star\right) \stackrel{d}{\to} \mathsf{N}\left(0, m^{-1}C(t^\star)\left(\mathsf{P}_{u^\star}^\perp \Sigma \mathsf{P}_{u^\star}^\perp\right)^\dagger\right).$$

For majority-vote aggregation, we can in turn generalize Corollary 2. In this case we still have $t_m \asymp \sqrt{m}$. However, the interesting factor here is that the convergence rate now depends on the noise exponent β.

Proposition 3. *Let Assumptions A1 and A2 hold for some $\beta \in (0, \infty)$, and $t^\star = \|\theta^\star\|_2$. Suppose $h = h_m$ is the function (5) with Z defined in Assumption A1. There are constants $a, b > 0$, depending only on β and c_Z, such that the following hold: there is a unique $t_m \geq t_1 = t^\star$ solving $h(t_m) = 0$ and for this t_m we have both*

$$\widehat{\theta}_{n,m}^{\rm mv} \stackrel{p}{\to} t_m u^\star \quad \textit{and} \quad \lim_{m\to\infty} \frac{t_m}{t^\star\sqrt{m}} = a.$$

Moreover, there exists a function $C_m(t) = \frac{b}{(t\sqrt{m})^{2-\beta}}(1 + o_m(1))$ as $m \to \infty$ such that

$$\sqrt{n}\left(\widehat{u}_{n,m}^{\rm mv} - u^\star\right) \stackrel{d}{\to} \mathsf{N}\left(0, C_m(t^\star)\left(\mathsf{P}_{u^\star}\Sigma\mathsf{P}_{u^\star}\right)^\dagger\right).$$

We defer the proof to Appendix G.2.

Paralleling the discussion in Section 3.1, we may compare the performance of the MLE $\widehat{\theta}_{n,m}^{\rm lr}$, which uses all labels, and the majority-vote estimator $\widehat{\theta}_{n,m}^{\rm mv}$ using only the cleaned labels. When the classification problem is hard—meaning that β in Condition (M_β) is near 0 so that that classifying most examples is nearly random chance—we see that the aggregation in the majority vote estimator still allows convergence (nearly) as quickly as the non-aggregated estimator; the problem is so noisy that data "cleaning" by aggregation is helpful. Yet for easier problems, where $\beta \gg 0$, the gap between them grows substantially; this is sensible, as aggregation is likely to force a dataset to be separable, thus making fitting methods unstable (and indeed, a minimizer may fail to exist).

4 Label aggregation and misspecified model

The logistic link provides clean interpretation and results, but it is interesting to move beyond it to more realistic cases where labelers use distinct links, although, to allow precise statements, we still assume the same linear term $x \mapsto \langle\theta^\star, x\rangle$ for each labeler's generalized linear model. We study generalizations of the maximum likelihood and majority vote estimators (2) and (3), highlighting dependence on link fidelity. In this setting, there are m (unknown and possibly distinct) link functions $\sigma_i^\star$, $i = 1, 2, \ldots, m$. We show that the majority-vote estimator $\widehat{\theta}_{n,m}^{\rm mv}$ enjoys better robustness to

model mis-specification than the non-aggregated estimator $\widehat{\theta}^{\rm lr}_{n,m}$, though both use identical losses. In particular, our main result in this section implies

$$\sqrt{n}(\widehat{u}^{\rm lr}_{n,m} - u^\star) \xrightarrow{d} \mathsf{N}\left(0, c\Sigma \cdot (1 + o_m(1))\right) \quad \text{while} \quad \sqrt{n}(\widehat{u}^{\rm mv}_{n,m} - u^\star) \xrightarrow{d} \mathsf{N}\left(0, m^{-1/2} c^{\rm mv}\Sigma \cdot (1 + o_m(1))\right),$$

where c and $c^{\rm mv}$ are constants that depend only on the links σ, $t^\star = \|\theta^\star\|_2$, and $\Sigma = I - u^\star u^{\star\top}$ when $X \sim \mathsf{N}(0, I_d)$. In contrast to the previous section, the majority-vote estimator enjoys roughly $\sqrt{m}$-faster rates than the non-aggregated estimator, maintaining its (slow) improvement with m, which the MLE loses to misspecification.

To set the stage for our results, we define the general link-based loss

$$\ell_{\sigma,\theta}(y \mid x) := -\int_0^{y\langle\theta,x\rangle} \sigma(-v)dv.$$

We then consider the general multi-label estimator and the majority-vote estimator based on the loss $\ell_{\sigma,\theta}$,

$$\widehat{\theta}_{n,m}(\sigma) := \operatorname*{argmin}_\theta \mathbb{P}_{n,m}\ell_{\sigma,\theta}, \qquad \widehat{\theta}^{\rm mv}_{n,m}(\sigma) := \operatorname*{argmin}_\theta \overline{\mathbb{P}}_{n,m}\ell_{\sigma,\theta}. \tag{7}$$

When $\sigma = \sigma^{\rm lr}$ is the logistic link, we recover the logistic loss $\ell_{\sigma^{\rm lr},\theta}(y \mid x) = \ell^{\rm lr}_\theta(y \mid x)$, and thus we recover the results in Section 3. For both the estimators, we suppress the dependence on the link σ to write $\widehat{\theta}_{n,m}, \widehat{\theta}^{\rm mv}_{n,m}$ when the context is clear.

4.1 Master results

To characterize the behavior of multiple label estimators versus majority vote, we provide master results as a foundation for our convergence rate analyses throughout. By a bit of notational chicanery, we consider both the cases that $\overline{Y}$ is a majority vote and that we use multiple (non-aggregated) labels simultaneously. In the case that the estimator uses the majority vote $\overline{Y}$, let

$$\varphi_m(t) = \rho_m(t)\mathbb{1}\{t \geq 0\} + (1 - \rho_m(t))\mathbb{1}\{t < 0\}, \quad \text{where} \quad \rho_m(t) := \mathbb{P}(\overline{Y} = \mathsf{sign}(\langle X, \theta^\star\rangle) \mid \langle X, \theta^\star\rangle = t),$$

and in the case that the estimator uses each label from the m labelers, let

$$\varphi_m(t) = \frac{1}{m}\sum_{j=1}^m \sigma_j^\star(t) = \frac{1}{m}\sum_{j=1}^m \mathbb{P}(Y_j = 1 \mid \langle X, \theta^\star\rangle = t).$$

In either case, we then see that the population loss with the link-based loss $\ell_{\sigma,\theta}$ becomes

$$L(\theta, \sigma) = \mathbb{E}\left[\ell_{\sigma,\theta}(1 \mid X)\varphi_m(\langle X, \theta^\star\rangle) + \ell_{\sigma,\theta}(-1 \mid X)(1 - \varphi_m(\langle X, \theta^\star\rangle))\right], \tag{8}$$

where we have taken a conditional expectation given X. We assume Assumption A1 holds, that X decomposes into the independent sum $X = Zu^\star + W$ with $W \perp u^\star$, and the true link functions $\sigma_j^\star \in \mathcal{F}_{\mathsf{link}}$. We further impose the following assumption for the model link.

Assumption A3. *For each sign $s \in \{-1, 1\}$, the model link function σ satisfies $\lim_{t\to s\cdot\infty} \sigma(t) = 1/2 + sc$ for a constant $0 < c \leq 1/2$ and is a.e. differentiable.*

Minimizer of the population loss. We begin by characterizing—at a somewhat abstract level—the (unique) solutions to the problem of minimizing the population loss (8). To characterize the minimizer $\theta_L^\star := \operatorname{argmin}_\theta L(\theta, \sigma)$, we hypothesize that it aligns with $u^\star = \theta^\star / \|\theta^\star\|_2$, using the familiar ansatz that θ has the form $\theta = tu^\star$. Using the formulation (8), we see that for $t^\star := \|\theta^\star\|_2$,

$$\begin{aligned}\nabla L(\theta, \sigma) &= -\mathbb{E}[\sigma(-\langle\theta, X\rangle)X\varphi_m(\langle X, \theta^\star\rangle)] + \mathbb{E}[\sigma(\langle\theta, X\rangle)X(1 - \varphi_m(\langle X, \theta^\star\rangle))] \\ &= -\mathbb{E}[\sigma(-tZ)X\varphi_m(t^\star Z)] + \mathbb{E}[\sigma(tZ)X(1 - \varphi_m(t^\star Z))] \\ &= \left(-\mathbb{E}[\sigma(-tZ)Z\varphi_m(t^\star Z)] + \mathbb{E}[\sigma(tZ)Z(1 - \varphi_m(t^\star Z))]\right) u^\star = h_m(t)u^\star, \end{aligned} \tag{9}$$

where the final line uses the decomposition $X = Zu^\star + W$ for the random vector $W \perp u^\star$ independent of Z, and we recall expression (5) to define the *calibration function*

$$h_{t^\star,m}(t) := \mathbb{E}[\sigma(tZ)Z(1 - \varphi_m(t^\star Z))] - \mathbb{E}[\sigma(-tZ)Z\varphi_m(t^\star Z)]. \tag{10}$$

The function h measures the gap between the hypothesized link function σ and the label probabilities φ_m, functioning the approximately "calibrate" σ to the observed probabilities. If we presume that a solution to $h_{t^\star,m}(t) = 0$ exists, then evidently $tu^\star$ is *a* minimizer of $L(\theta, \sigma)$. In fact, such a solution exists and is unique (see Appendix C for a proof):

Lemma 4.1. *Let Assumption A1 hold and $h = h_{t^\star,m}$ be the gap function (10). Then there is a unique solution $t_m > 0$ to $h(t) = 0$, and the generic loss (8) has unique minimizer $\theta_L^\star = t_m u^\star$. Define the matrix*

$$\begin{aligned} H_L(t) := \mathbb{E}[(\sigma'(-tZ)\varphi_m(t^\star Z) + \sigma'(tZ)(1 - \varphi_m(t^\star Z)))Z^2]u^\star u^{\star\top} \\ + \mathbb{E}[\sigma'(-tZ)\varphi_m(t^\star Z) + \sigma'(tZ)(1 - \varphi_m(t^\star Z))]\mathsf{P}_{u^\star}^\perp \Sigma \mathsf{P}_{u^\star}^\perp. \end{aligned} \tag{11}$$

Then the Hessian is $\nabla^2 L(\theta_L^\star, \sigma) = H_L(t_m)$.

Asymptotic normality with multiple labels. With the existence of minimizers assured, we turn to their asymptotics. For each of these, we require slightly different calculations, as the resulting covariances are slightly different. To state the result when we have multiple labels, we define the average link function $\overline{\sigma}^\star = \frac{1}{m}\sum_{j=1}^m \sigma_j^\star$ and the three functions

$$\begin{aligned} \mathsf{le}(Z) &:= \sigma(t_m Z)(1 - \overline{\sigma}^\star(t^\star Z)) - \sigma(-t_m Z)\overline{\sigma}^\star(t^\star Z), \\ \mathsf{he}(Z) &:= \sigma'(-t_m Z)\varphi_m(t^\star Z) + \sigma'(t_m Z)(1 - \varphi_m(t^\star Z)), \\ v_j(Z) &:= \sigma_j^\star(t^\star Z)(1 - \sigma_j^\star(t^\star Z))(\sigma(t_m Z) + \sigma(-t_m Z))^2. \end{aligned} \tag{12}$$

The first, the link error le, measures the mis-specification of the link σ relative to the average link $\overline{\sigma}^\star$. The second function, he, is a Hessian term, as $H_L(t_m) = \mathbb{E}[\mathsf{he}(Z)Z^2]u^\star u^{\star\top} + \mathbb{E}[\mathsf{he}(Z)]\mathsf{P}_{u^\star}^\perp \Sigma \mathsf{P}_{u^\star}^\perp$, and the third is a variance term for each labeler j. We have the following theorem, which we prove in Appendix D.

Theorem 1. *Let Assumptions A1 and A3 hold, and let $\widehat{\theta}_{n,m}$ be the multilabel estimator (7). Define the shorthand $\overline{v} = \frac{1}{m}\sum_{j=1}^m v_j$. Then $\widehat{\theta}_{n,m} \xrightarrow{a.s.} \theta_L^\star$, and*

$$\sqrt{n}(\widehat{\theta}_n - \theta_L^\star) \xrightarrow{d} \mathsf{N}\left(0, \frac{\mathbb{E}[\mathsf{le}(Z)^2 Z^2] + m^{-1}\mathbb{E}[\overline{v}(Z)Z^2]}{\mathbb{E}[\mathsf{he}(Z)Z^2]^2} u^\star u^{\star\top} + \frac{\mathbb{E}[\mathsf{le}(Z)^2] + m^{-1}\mathbb{E}[\overline{v}(Z)]}{\mathbb{E}[\mathsf{he}(Z)]^2}\left(\mathsf{P}_{u^\star}^\perp \Sigma \mathsf{P}_{u^\star}^\perp\right)^\dagger\right).$$

Additionally, if $\widehat{u}_n = \widehat{\theta}_{n,m}/\|\widehat{\theta}_{n,m}\|_2$ and t_m is the unique zero of the gap function $h_{t^\star,m}(t) = 0$, then

$$\sqrt{n}(\widehat{u}_n - u^\star) \xrightarrow{d} \mathsf{N}\left(0, \frac{1}{t_m^2}\frac{\mathbb{E}[\mathsf{le}(Z)^2] + m^{-1}\mathbb{E}[\overline{v}(Z)]}{\mathbb{E}[\mathsf{he}(Z)]^2}\left(\mathsf{P}_{u^\star}^\perp \Sigma \mathsf{P}_{u^\star}^\perp\right)^\dagger\right).$$

Theorem 1 exhibits two dependencies: the first on the link error terms $\mathbb{E}[\mathsf{le}(Z)^2]$—essentially, a bias term—and the second on the rescaled average variance $\frac{1}{m}\mathbb{E}[\overline{v}(Z)]$. So the multi-label estimator recovers an optimal $O(1/m)$ covariance if the link errors are negligible, but if they are not, then it necessarily has $O(1)$ asymptotic covariance. The next corollary highlights how things simplify. In the well-specified case that σ is symmetric and $\sigma = \overline{\sigma}^\star$, the zero of the gap function (10) is evidently $t_m = t^\star = \|\theta^\star\|_2$, the error term $\mathsf{le}(Z) = 0$, and $v_j(Z) = \sigma^\star(t^\star Z)(1 - \sigma^\star(t^\star Z))$, and by symmetry, $\sigma'(t) = \sigma'(-t)$ so that $\mathsf{he}(Z) = \sigma'(t^\star Z)$:

Corollary 3 (The well-specified case)**.** *Let the conditions above hold. Then*

$$\sqrt{n}(\widehat{u}_n - u^\star) \xrightarrow{d} \mathsf{N}\left(0, \frac{1}{m} \cdot \frac{1}{\|\theta^\star\|_2^2} \frac{\mathbb{E}[\sigma(t^\star Z)(1 - \sigma(t^\star Z))]}{\mathbb{E}[\sigma'(t^\star Z)]^2} \mathsf{P}_{u^\star}^\perp \Sigma \mathsf{P}_{u^\star}^\perp\right).$$

Asymptotic normality with majority vote. When we use the majority vote estimators, the asymptotics are a bit different: there is no immediate improvement as the number of labelers m increases, because there is no averaging to reduce variance, even in a "well-specified" case. Though, as we shall see, the asymptotic covariance does decrease as m grows, but the dependence is more subtle.

Theorem 2. *Let Assumptions A1 and A3 hold, and let $\widehat{\theta}_n = \widehat{\theta}^{\mathrm{mv}}_{n,m}$ be the general majority vote estimator (7). Let t_m be the zero of the gap function (10), solving $h_{t^\star,m}(t) = 0$. Then $\widehat{\theta}_n \xrightarrow{a.s.} \theta^\star_L = t_m u^\star$, and for $\widehat{u}_n = \widehat{\theta}_n / \|\widehat{\theta}_n\|_2$, we have*

$$\sqrt{n}\left(\widehat{u}_n - u^\star\right) \xrightarrow{d} \mathsf{N}\left(0, \frac{1}{t_m^2} \frac{\mathbb{E}[\sigma(-t_m|Z|)^2 \rho_m(t^\star Z) + \sigma(t_m|Z|)^2(1 - \rho_m(t^\star Z))]}{\mathbb{E}[\mathsf{he}(Z)]^2} \left(\mathsf{P}_{u^\star}^\perp \Sigma \mathsf{P}_{u^\star}^\perp\right)^\dagger\right).$$

We defer the proof to Appendix E. In most cases, we will take the link function σ to be symmetric, so that $\sigma(t) = 1 - \sigma(-t)$, and thus $\sigma'(t) = \sigma'(-t)$, so that $\mathsf{he}(z) = \sigma'(t_m z) \geq 0$. This simplifies the denominator in Theorem 2 to $\mathbb{E}[\sigma'(t_m Z)]^2$. Written differently, we may define a (scalar) variance-characterizing function C_m implicitly as follows: let $t_m = t_m(t)$ be a zero of $h_{t,m}(s) = \mathbb{E}[\sigma(sZ)Z(1 - \varphi_m(tZ))] - \mathbb{E}[\sigma(-sZ)Z\varphi_m(tZ)] = 0$ in s, that is, $h_{t,m}(t_m(t)) = 0$ so that t_m is a function of the size t (recall the gap (10)), and then define

$$C_m(t) := \frac{1}{t_m^2} \frac{\mathbb{E}[\sigma(-t_m|Z|)^2 \rho_m(tZ) + \sigma(t_m|Z|)^2(1 - \rho_m(tZ))]}{\mathbb{E}[\sigma'(t_m Z)]^2} \tag{13}$$

where $t_m = t_m(t)$ above is implicitly defined. Then

$$\sqrt{n}\left(\widehat{u}_n - u^\star\right) \xrightarrow{d} \mathsf{N}\left(0, C_m(t^\star)\left(\mathsf{P}_{u^\star}^\perp \Sigma \mathsf{P}_{u^\star}^\perp\right)^\dagger\right).$$

Each of our main results, including those on well-specified models previously, then follows by characterizing the behavior of $C_m(t)$ in the asymptotics as $m \to \infty$ and the scaling of the solution norm $t_m = \|\theta^\star_L\|_2$, which the calibration gap (10) determines. The key is that the scaling with m varies depending on the fidelity of the model, behavior of the links σ, and the noise exponent (M_β), and our coming consequences of the master theorems 1 and 2 help to reify this scaling.

4.2 Robustness to model mis-specification

Having established the general convergence results for the multi-label estimator $\widehat{\theta}_{n,m}$ and the majority vote estimator $\widehat{\theta}^{\mathrm{mv}}_{n,m}(\sigma)$, we further explicate their performance when we have a mis-specified model—the link σ is incorrect—by leveraging Theorems 1 and 2 to precisely characterize their asymptotics and show the majority-vote estimator can be more robust to model mis-specification.

Multi-label estimator. As our focus here is descriptive, to make interpretable statements about the multi-label estimator $\widehat{\theta}_{n,m}$ in (7), we simplify by assuming that each link $\sigma_j^\star \equiv \sigma^\star \in \mathcal{F}_{\mathsf{link}}$ is identical. Then an immediate corollary of Theorem 1 follows:

Corollary 4. *Let Assumptions A1 and A2 hold for some $\beta \in (0, \infty)$, and $t^\star = \|\theta^\star\|_2$. Then the calibration gap function (10) has unique positive zero $h_{t^\star,m}(t_{\sigma^\star}) = 0$, and the multilabel estimator (7) satisfies*

$$\widehat{\theta}_{n,m} \xrightarrow{p} t_{\sigma^\star} u^\star.$$

Additionally, the normalized estimate $\widehat{u}_{n,m} = \widehat{\theta}_{n,m}/\|\widehat{\theta}_{n,m}\|_2$ satisfies

$$\sqrt{n}\left(\widehat{u}_{n,m} - u^\star\right) \xrightarrow{d} \mathsf{N}\left(0, \frac{\mathbb{E}[\mathsf{le}(Z)^2] + m^{-1}\mathbb{E}[\overline{v}(Z)]}{t_{\sigma^\star}^2 \mathbb{E}[\mathsf{he}(Z)]^2} \left(\mathsf{P}_{u^\star} \Sigma \mathsf{P}_{u^\star}\right)^\dagger\right).$$

So in this simplified case, the asymptotic covariance remains of constant order in m unless $\mathbb{E}[\mathsf{le}(Z)^2] = 0$. In contrast, as we now show, the majority vote estimator exhibits more robustness; this is perhaps expected, as Corollary 2 shows that in the logistic link case, which is *a fortiori* misspecified for majority vote labels, has covariance scaling as $1/\sqrt{m}$, though the generality of the behavior and its distinction from Corollary 4 is interesting.

Majority vote estimator. For the majority-vote estimator, we relax our assumptions and allow $\sigma_j^\star$ to be different, showing how the broad conclusions Corollary 4 suggests continue to hold in some generality: majority vote estimators achieve slower convergence than well-specified (maximum likelihood) estimators using each label, but exhibit more robustness. To characterize the large m behavior, we require the following regularity conditions on the average link $\overline{\sigma}_m^\star = \frac{1}{m}\sum_{j=1}^m \sigma_j^\star$, which we require has a limiting derivative at 0.

Assumption A4. *For a sequence of link functions $\{\sigma_j \mid j \in \mathbb{N}\} \subset \mathcal{F}_{\mathsf{link}}$, let $\overline{\sigma}_m^\star = \frac{1}{m}\sum_{j=1}^m \sigma_j^\star$, there exists $\overline{\sigma}^{\star\prime}(0) > 0$ such that*

$$\textbf{(i)} \qquad \lim_{m\to\infty} \sqrt{m}\left(\overline{\sigma}_m^\star\left(\frac{t}{\sqrt{m}}\right) - \frac{1}{2}\right) = \overline{\sigma}^{\star\prime}(0)t, \quad \textit{for each } t \in \mathbb{R}; \tag{14a}$$

$$\textbf{(ii)} \qquad \liminf_{m\to\infty} \inf_{t\neq 0} \frac{\left|\overline{\sigma}_m^\star(t) - \frac{1}{2}\right|}{\min\{|t|, 1\}} > 0; \tag{14b}$$

$$\textbf{(iii)} \qquad \lim_{t\to 0} \sup_{j\in\mathbb{N}} \left|\sigma_j^\star(t) - \frac{1}{2}\right| = 0. \tag{14c}$$

These assumptions simplify if the links are identical: if $\sigma_j^\star \equiv \sigma^\star$, we only require $\sigma^\star$ is differentiable around 0 with $\sigma^{\star\prime}(0) > 0$ and $|\sigma^\star(t) - \frac{1}{2}| \gtrsim \min\{|t|, 1\}$.

We can apply Theorem 2 to obtain asymptotic normality for the majority vote estimator (7). We recall the probability

$$\rho_m(t) := \mathbb{P}(\overline{Y} = \mathsf{sign}(\langle X, \theta^\star\rangle) \mid \langle X, \theta^\star\rangle = t) \tag{15}$$

of the majority vote being correct given the margin $\langle X, \theta^\star\rangle = t$ and the calibration gap function (10), which by a calculation case resolves to the more convenient form

$$h(t) = h_{t^\star,m}(t) = \mathbb{E}[\sigma(t|Z|)|Z|(1 - \rho_m(t^\star Z))] - \mathbb{E}[\sigma(-t|Z|)|Z|\rho_m(t^\star Z)].$$

The main technical challenge is to characterize the large m behavior for the asymptotic covariance function $C_m(t)$ defined implicitly in the quantity (13). We postpone the details to Appendix G.3 and state the result below, which is a consequence of Theorem 2 and a careful asymptotic expansion of the covariance function (13).

Proposition 4. *Let Assumptions A1 and A2 hold for some $\beta \in (0, \infty)$ with $\int_0^\infty z^{\beta-1}\sigma'(z)dz < \infty$ and $t^\star = \|\theta^\star\|_2$, and in addition that Assumption A4 holds and σ is symmetric. Then there are constants $a, b > 0$, depending only on β, c_Z, σ, and $\overline{\sigma}^{\star\prime}(0)$, such that there is a unique $t_m \geq t_1 = t^\star$ solving $h(t_m) = 0$, and for this t_m we have both*

$$\widehat{\theta}^{\mathrm{mv}}_{n,m} \xrightarrow{p} t_m u^\star \quad \text{and} \quad \lim_{m\to\infty} \frac{t_m}{t^\star\sqrt{m}} = a.$$

Moreover, the covariance (13) has the form $C_m(t) = \frac{b}{(t\sqrt{m})^{2-\beta}}(1 + o_m(1))$, and

$$\sqrt{n}\left(\widehat{u}^{\mathrm{mv}}_{n,m} - u^\star\right) \xrightarrow{d} \mathsf{N}\left(0, C_m(t^\star)\left(\mathsf{P}_{u^\star}\Sigma\mathsf{P}_{u^\star}\right)^\dagger\right).$$

Proposition 4 highlights the robustness of the majority vote estimator: even when the link σ is (more or less) arbitrarily incorrect, the asymptotic covariance still exhibits reasonable scaling. The noise parameter β in Assumption A2, roughly equivalent to the Mammen-Tsybakov noise exponent (M_β), also plays an important role. In typical cases with $\beta = 1$ (e.g., when $X \sim \mathsf{N}(0, I_d)$), we see $C_m(t) \asymp \frac{1}{t\sqrt{m}}$. In noisier cases, corresponding to $\beta \downarrow 0$, majority vote provides substantial benefit approaching a well-specified model; conversely, in "easy" cases where $\beta > 2$, majority vote estimators become *more* unstable, as they make the data (very nearly) separable, which causes logistic-regression and other margin-type estimators to be unstable [5].

5 Semi-parametric approaches

The preceding analysis highlights distinctions between a fuller likelihood-based approach—which uses all the labels, as in (2)—and the robust but somewhat slower rates that majority vote estimators enjoy (as in Proposition 4). That full-label estimators' performance so strongly depends on the fidelity of the link (recall Corollary 4) suggests that we target estimators achieving the best of both worlds: learn both a link function (or collection thereof) and refit the model using *all* the labels. In this section, we develop this more efficient estimation scheme through semiparametric estimation approaches.

We develop a few general convergence results into which we can essentially plug in semiparametric estimators. In distinction from standard results in semiparametric theory (e.g. [36, Ch. 25] or [3]), our results require little more than consistent estimation of the links $\sigma_j^\star$ to recover $1/m$ (optimal) scaling in the asymptotic covariance, as the special structure of our classification problem allows more nuanced calculations; we assume each labeler (link function) generates labels for each of the n datapoints $X_1, \ldots, X_n$, but we could relax the assumption at the expense of extraordinarily cumbersome notation. We give two example applications of the general theory: the first (Sec. 5.2) analyzing a full pipeline for a single index model setting, that robustly estimates direction $u^\star$, the link $\sigma^\star$, and then re-estimates $\theta^\star$; the second assuming a stylized black-box crowdsourcing mechanism that provides estimates of labeler reliability, highlighting how even in some crowdsourcing scenarios, there could be substantial advantages to using full label information.

5.1 Master results

For our specializations, we first provide master results that allow semi-parametric estimation of the link functions. We consider Lipschitz symmetric link functions, where for $\mathsf{L} > 0$ we define

$$\mathcal{F}^{\mathsf{L}}_{\mathsf{link}} := \{\sigma \mid \sigma \not\equiv 1/2 \text{ is non-decreasing, symmetric, and } \mathsf{L}\text{-Lipschitz continuous}\} \subset \mathcal{F}^{0}_{\mathsf{link}}.$$

We consider the general case where there are m distinct labeler link functions $\sigma_1^\star, \ldots, \sigma_m^\star$. To eliminate ambiguity in the links, we assume the model is normalized, $\|\theta^\star\|_2 = 1$ so $\theta^\star = u^\star$. To distinguish from the typical case, we write $\vec{\sigma} = (\sigma_1, \ldots, \sigma_m)$, and for $(x, y) \in \mathbb{R}^d \times \{\pm 1\}^m$ define

$$\ell_{\vec{\sigma},\theta}(y \mid x) := \frac{1}{m} \sum_{j=1}^{m} \ell_{\sigma_j,\theta}(y_j \mid x),$$

which allows us to consider both the standard margin-based loss and the case in which we learn separate measures of quality per labeler. With this notation, we can then naturally define the population loss

$$L(\theta, \vec{\sigma}) := \frac{1}{m} \sum_{j=1}^{m} \mathbb{E}[\ell_{\sigma_j,\theta}(Y_j \mid X)].$$

For any sequence $\{\vec{\sigma}_n\} \subset (\mathcal{F}_{\mathsf{link}}^{\mathsf{L}})^m$ of (estimated) links and data (X_i, Y_i) for $Y_i = (Y_{i1}, \ldots, Y_{im})$, we define the semi-parametric estimator

$$\widehat{\theta}_{n,m}^{\mathrm{sp}} := \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ell_{\vec{\sigma}_n,\theta}(Y_i \mid X_i).$$

We will demonstrate both consistency and asymptotic normality under appropriate convergence and regularity assumptions for the link functions. We assume there is a (semiparametric) collection $\mathcal{F}_{\mathsf{link}}^{\mathsf{sp}} \subset (\mathcal{F}_{\mathsf{link}}^{\mathsf{L}})^m$ of link functions of interest, which may coincide with $(\mathcal{F}_{\mathsf{link}}^{\mathsf{L}})^m$ but may be smaller, making estimation easier. Define the distance $d_{\mathcal{F}_{\mathsf{link}}^{\mathsf{sp}}}$ on $\mathbb{R}^d \times \mathcal{F}_{\mathsf{link}}^{\mathsf{sp}}$ by

$$d_{\mathcal{F}_{\mathsf{link}}^{\mathsf{sp}}}((\theta_1, \vec{\sigma}_1), (\theta_2, \vec{\sigma}_2)) := \|\theta_1 - \theta_2\|_2 + \|\vec{\sigma}_1(-Y\langle X, u^\star\rangle) - \vec{\sigma}_2(-Y\langle X, u^\star\rangle)\|_{L^2(\mathbb{P})}.$$

We make the following assumption.

Assumption A5. *The links $\vec{\sigma}^\star \in \mathcal{F}_{\mathsf{link}}^{\mathsf{sp}}$ are normalized so that $\mathbb{P}(Y_j = y \mid X = x) = \sigma_j^\star(y\langle x, u^\star\rangle)$, and the sequence $\{\vec{\sigma}_n\} \subset \mathcal{F}_{\mathsf{link}}^{\mathsf{sp}}$ is consistent:*

$$\|\vec{\sigma}_n(-Y\langle X, u^\star\rangle) - \vec{\sigma}^\star(-Y\langle X, u^\star\rangle)\|_{L^2(\mathbb{P})} \xrightarrow{p} 0.$$

Additionally, the mapping $(\theta, \vec{\sigma}) \mapsto \nabla_\theta^2 L(\theta, \vec{\sigma})$ is continuous for $d_{\mathcal{F}_{\mathsf{link}}^{\mathsf{sp}}}$ at $(u^\star, \vec{\sigma}^\star)$.

The continuity of $\nabla_\theta^2 L(\theta, \vec{\sigma})$ at $(u^\star, \vec{\sigma}^\star)$ allows us to develop local asymptotic normality. To see that we may expect the assumption to hold, we give reasonably simple conditions sufficient for it, including that the collection of links $\mathcal{F}_{\mathsf{link}}^{\mathsf{sp}}$ is sufficiently smooth or the data distribution is continuous enough. (See Appendix H.1 for a proof.)

Lemma 5.1. *Let $d_{\mathcal{F}_{\mathsf{link}}^{\mathsf{sp}}}$ be the distance in Assumption A5. Let Assumption A1 hold, where $|Z| > 0$ with probability one, has nonzero and continuously differentiable density $p(z)$ on $(0, \infty)$ satisfying $\lim_{z \to s} z^2 p(z) = 0$ for $s \in \{0, \infty\}$. The mapping $(\theta, \vec{\sigma}) \mapsto \nabla_\theta^2 L(\theta, \vec{\sigma})$ is continuous for $d_{\mathcal{F}_{\mathsf{link}}^{\mathsf{sp}}}$ at $(u^\star, \vec{\sigma}^\star)$ whenever $\mathbb{E}[\|X\|_2^4] < \infty$ and either of the following conditions holds:*

1. *For any $\vec{\sigma} = (\sigma_1, \ldots, \sigma_m) \in \mathcal{F}_{\mathsf{link}}^{\mathsf{sp}}$, σ_j' are L'-Lipschitz continuous.*
2. *X has continuous density on $\mathbb{R}^d$.*

We can now present the master result for semi-parametric approaches, which characterizes the asymptotic behavior of the semi-parametric estimator with the variance function

$$C_{m,\vec{\sigma}^\star} := \frac{1}{m} \cdot \frac{\frac{1}{m}\sum_{j=1}^m \mathbb{E}[\sigma_j^\star(Z)(1-\sigma_j^\star(Z))]}{(\frac{1}{m}\sum_{j=1}^m \mathbb{E}[\sigma_j^{\star\prime}(Z)])^2}.$$

Theorem 3. *Let Assumption A1 hold and assume $|Z| > 0$ with probability one and has nonzero and continuous density $p(z)$ on $(0,\infty)$. Let Assumption A5 hold and assume that $\mathbb{E}[\|X\|_2^4] < \infty$. Then $\sqrt{n}(\widehat{\theta}^{\mathrm{sp}}_{n,m} - u^\star)$ is asymptotically normal, and the normalized estimator $\widehat{u}^{\mathrm{sp}}_{n,m} = \widehat{\theta}^{\mathrm{sp}}_{n,m}/\|\widehat{\theta}^{\mathrm{sp}}_{n,m}\|_2$ satisfies*

$$\sqrt{n}\left(\widehat{u}^{\mathrm{sp}}_{n,m} - u^\star\right) \xrightarrow{d} \mathsf{N}\left(0, C_{m,\vec{\sigma}^\star}\left(\mathsf{P}^\perp_{u^\star}\Sigma\mathsf{P}^\perp_{u^\star}\right)^\dagger\right).$$

See Appendix F for proof details. Notably, Theorem 3 exhibits optimal $1/m$ scaling in the covariance whenever $\mathbb{E}[\sigma_j^{\star\prime}(Z)] \gtrsim 1$.

5.2 A single index model

Our first example application of Theorem 3 is to a single index model. We present a multi-phase estimator that first estimates the direction $u^\star = \theta^\star/\|\theta^\star\|_2$, then uses this estimate to find a (consistent) estimate of the link $\sigma^\star$, which we can then substitute directly into Theorem 3. We defer all proofs of this section to Appendix I, which also includes a few auxiliary results that we use to prove the results proper.

We present the the abstract convergence result for link functions first, considering a scenario where we have an initial guess u_n^{init} of the direction $u^\star$, independent of $(X_i, Y_{ij})_{i\le n, j\le m}$, for example constructed via a small held-out subset of the data. We set

$$\vec{\sigma}_n = \operatorname*{argmin}_{\vec{\sigma}\in\mathcal{F}^{\mathsf{sp}}_{\mathsf{link}}} \sum_{i=1}^n\sum_{j=1}^m \left(\sigma_j(\langle u_n^{\mathrm{init}}, X_i\rangle) - Y_{ij}\right)^2,$$

where $\mathcal{F}^{\mathsf{sp}}_{\mathsf{link}} \subset (\mathcal{F}^{\mathsf{L}}_{\mathsf{link}})^m$ and so it consists of nondecreasing L-Lipschitz link functions with $\sigma(0) = \frac{1}{2}$. We assume that for all n, there exists a (potentially random) ϵ_n such that

$$\|u_n^{\mathrm{init}} - u^\star\|_2 \le \epsilon_n.$$

Proposition 5. *Let X_i be vectors with $\mathbb{E}[\|X\|_2^k] < \infty$, where $k \ge 2$. Then with probability 1, there is a finite (random) $C < \infty$ such that for all large enough n,*

$$\|\vec{\sigma}_n(Y\langle u^\star, X\rangle) - \vec{\sigma}^\star(Y\langle u^\star, X\rangle)\|^2_{L^2(\mathbb{P})} \le C\left[n^{\frac{2}{3k}-\frac{2}{3}} + \epsilon_n^2 + Mn^{-\frac{k}{2(k+1)}}\right].$$

The proof is more or less a consequence of standard convergence results for nonparametric function estimation, though we include it for completeness in Appendix I.2 as it includes a few additional technicalities because of the initial estimate of $u^\star$.

Summarizing, we see that a natural procedure is available: if we have models powerful enough to accurately estimate the conditional label probabilities $Y \mid X$, then Proposition 5 coupled with Theorem 3 shows that we can achieve estimation with near-optimal asymptotic covariance. In particular, if u_n^{init} is consistent (so $\epsilon_n \xrightarrow{p} 0$), then $\widehat{\theta}^{\mathrm{sp}}_{n,m}$ induces a normalized estimator $\widehat{u}^{\mathrm{sp}}_n = \widehat{\theta}^{\mathrm{sp}}_{n,m}/\|\widehat{\theta}^{\mathrm{sp}}_{n,m}\|_2$ satisfying $\sqrt{n}(\widehat{u}^{\mathrm{sp}}_n - u^\star) \xrightarrow{d} \mathsf{N}(0, C_{m,\vec{\sigma}}(\mathsf{P}^\perp_{u^\star}\Sigma\mathsf{P}^\perp_{u^\star})^\dagger)$.

5.3 Crowdsourcing model

Crowdsourcing typically targets estimating rater reliability, then using these reliability estimates to recover ground truth labels as accurately as possible, with versions of this approach central since at least Dawid and Skene's Expectation-Maximization-based approaches [9, 41, 29]. We focus here on a simple model of rater reliability, highlighting how—at least in our stylized model of classifier learning—by combining a crowdsourcing reliability model and still using all labels in estimating a classifier, we can achieve asymptotically efficient estimates of $\theta^\star$, rather than the robust but slower estimates $\widehat{\theta}^{\mathrm{mv}}_{n,m}$ arising from "cleaned" labels.

We adopt Whitehill et al.'s roughly "low-rank" model for label generation [41]: for binary classification with m labelers and distinct link functions $\sigma_i^\star$, model the difficulty of X_i by $\beta_i \in (-\infty, \infty)$, where $\mathsf{sign}(\beta_i)$ denotes the true class X_i belongs to. A parameter α_j models the expertise of annotator j, and the probability labeler j correctly classifies X_i is

$$\mathbb{P}(Y_{ij} = 1) = \frac{1}{1 + \exp(-\alpha_i \beta_j)}.$$

(See also Raykar et al. [29].) The focus in these papers was to construct gold-standard labels and datasets (X_i, Y_i); here, we take the alternative perspective we have so far advocated to show how using all labels can yield strong performance.

We thus adopt a semiparametric approach: we model the labelers, assuming a black-box crowdsourcing model that can infer each labeler's ability, then fit the classifier. We represent labeler j's expertise by a scalar $\alpha_j^\star \in (0, \infty)$. Given data $X_i = X$ and the normalized $\theta^\star = u^\star$, we assume a modified logistic link

$$\mathbb{P}(Y_{ij} = 1 \mid X_i = x) = \frac{1}{1 + \exp(-\alpha_j^\star \langle \theta^\star, x \rangle)} = \sigma^{\mathrm{lr}}(\alpha_j^\star \langle u^\star, x \rangle),$$

so $\alpha_j^\star = \infty$ represents an omniscient labeler while $\alpha_j^\star = 0$ means the labeler chooses random labels regardless of the data. Let $\alpha^\star = (\alpha_1^\star, \ldots, \alpha_m^\star) \in \mathbb{R}_+^m$. Then, in keeping with the plug-in approach of the master Theorem 3, we assume the blackbox crowdsourcing model generates an estimate $\alpha_n = (\alpha_{n,1}, \ldots, \alpha_{n,m}) \in \mathbb{R}_+^m$ of $\alpha^\star$ from the data $\{(X_i, (Y_{i1}, \ldots, Y_{im}))\}_{i=1}^n$.

We consider the algorithm using the blackbox crowdsourcing model and empirical risk minimization with the margin-based loss $\ell_{\vec{\sigma}_n, \theta}$ as in Section 5.1, with $\vec{\sigma}_n(t) := (\sigma^{\mathrm{lr}}(\alpha_{n,1} t), \cdots, \sigma^{\mathrm{lr}}(\alpha_{n,m} t))$, or equivalently using the rescaled logistic loss

$$\ell_{\vec{\sigma}_n, \theta}(y \mid x) = \frac{1}{m} \sum_{j=1}^m \ell_\theta^{\mathrm{lr}}(y_j \mid \alpha_{n,j} x).$$

This allows us to apply our general semiparametric Theorem 3 as long as the crowdsourcing model produces a consistent estimate $\alpha_n \xrightarrow{p} \alpha^\star$ (see Appendix H.2 for a proof):

Proposition 6. *Let Assumption A1 hold, $|Z| > 0$ have nonzero and continuous density $p(z)$ on $(0, \infty)$, and $\mathbb{E}[\|X\|_2^4] < \infty$. If $\alpha_n \xrightarrow{p} \alpha^\star \in \mathbb{R}^m$, then $\sqrt{n}(\widehat{\theta}^{\mathrm{sp}}_{n,m} - u^\star)$ is asymptotically normal, and the normalized estimator $\widehat{u}^{\mathrm{sp}}_{n,m} = \widehat{\theta}^{\mathrm{sp}}_{n,m} / \|\widehat{\theta}^{\mathrm{sp}}_{n,m}\|_2$ satisfies*

$$\sqrt{n}\left(\widehat{u}^{\mathrm{sp}}_{n,m} - u^\star\right) \xrightarrow{d} \mathsf{N}\left(0, \frac{1}{\sum_{j=1}^m \mathbb{E}[\sigma^{\mathrm{lr}}(\alpha_j^\star Z)(1 - \sigma^{\mathrm{lr}}(\alpha_j^\star Z))]} \left(\mathsf{P}_{u^\star}^\perp \Sigma \mathsf{P}_{u^\star}^\perp\right)^\dagger\right).$$

By Proposition 6, the semiparametric estimator $\widehat{u}^{\rm sp}_{n,m}$ is efficient when the rater reliability estimates $\alpha_n \in \mathbb{R}^m$ are consistent. It is also immediate that if $\alpha_j^\star \le \alpha_{\max} < \infty$ are bounded, then the asymptotic covariance multiplier $C_{m,\alpha^\star} = (\sum_{j=1}^m \mathbb{E}[\sigma^{\rm lr}(\alpha_j^\star Z)(1-\sigma^{\rm lr}(\alpha_j^\star Z))])^{-1} = O(1/m)$, so we recover the $1/m$ scaling of the MLE, as opposed to the slower rates of the majority vote estimators in Section 4.2. At this point, the refrain is perhaps unsurprising: using all the label information can yield much stronger convergence guarantees.

6 Experiments

We conclude the paper with several experiments to evaluate the methods we propose in this paper and to test whether their (sometimes implicit) methodological suggestions hold merit. Before delving into our experimental results, we detail a few of the expected behaviors our theory suggests; if we fail to see them, then the model we have proposed is too unrealistic to inform practice. First, based on the results in Section 3, we expect the classification error to be better for the non-aggregated algorithm $\widehat{\theta}^{\rm lr}_{n,m}$, and the gap between the two algorithms to become larger for less noisy problems. Moreover, we only expect $\widehat{\theta}^{\rm lr}_{n,m}$ to be calibrated, and our theory predicts that the majority-vote estimator's calibration worsens as the number of labels m increases. More generally, so long as we model uncertainty with enough fidelity, Corollary 4 suggests that multilabel estimators should exhibit better performance than those using majority vote labels $\overline{Y}$.

To that end, we provide two experiments on real datasets: the BlueBirds dataset (Section 6.1) and CIFAR-10H (Section 6.2). Unfortunately, a paucity of large-scale multi-label datasets that we know of precludes more experiments; the ImageNet creators [10, 31] no longer have any intermediate label information from their construction. We consider our two main algorithmic models:

(i) The maximum likelihood estimator $\widehat{\theta}^{\rm lr}_{n,m}$ based on non-aggregated data in Eq. (2).

(ii) The majority-vote based estimator $\widehat{\theta}^{\rm mv}_{n,m}$ of Eq. (3), our proxy for modern data pipelines.

6.1 BlueBirds

We begin with the BlueBirds dataset [40], which is a relatively small dataset consisting of 108 images with ResNet features. The classification problem is challenging, and the task is to classify each image as one of *Indigo Bunting* or *Blue Grosbeak* (two similar-looking blue bird species). For each image, we have 39 labels, obtained through Amazon Mechanical Turk workers. We use a pretrained (on ImageNet) ResNet50 model to generate image features, then apply PCA to reduce the dimensionality from $d_{\rm init} = 2048$ to $d = 25$.

We repeat the following experiment $T = 100$ times. For each number $m = 1, \ldots, 35$ of labelers, we fit the multilabel logistic model (2) and the majority vote estimator (3), finding calibration and classification errors using 10-fold cross validation. We measure calibration error on a held-out example x by $|\operatorname{logit}(\widetilde{p}(x)) - \operatorname{logit}(\widehat{p}(x))|$, where $\widetilde{p}(x)$ is the predicted probability and $\widehat{p}(x)$ is the empirical probability (over the labelers), where $\operatorname{logit}(p) = \log \frac{p}{1-p}$; we measure classification error on example x with labels $(y_1, \ldots, y_m)$ by $\frac{1}{m}\sum_{j=1}^m \mathbb{1}\{y_j \neq \mathsf{sign}(\widetilde{p}(x) - \frac{1}{2})\}$, giving an inherent noise floor because of labeler uncertainty. We report the results in Figure 1. These plots corroborate our theoretical predictions: as the number of labelers m increases, both the majority vote method and the full label method exhibit improved classification error, but considering all labels gives a (significant) improvement in accuracy and in calibration error.

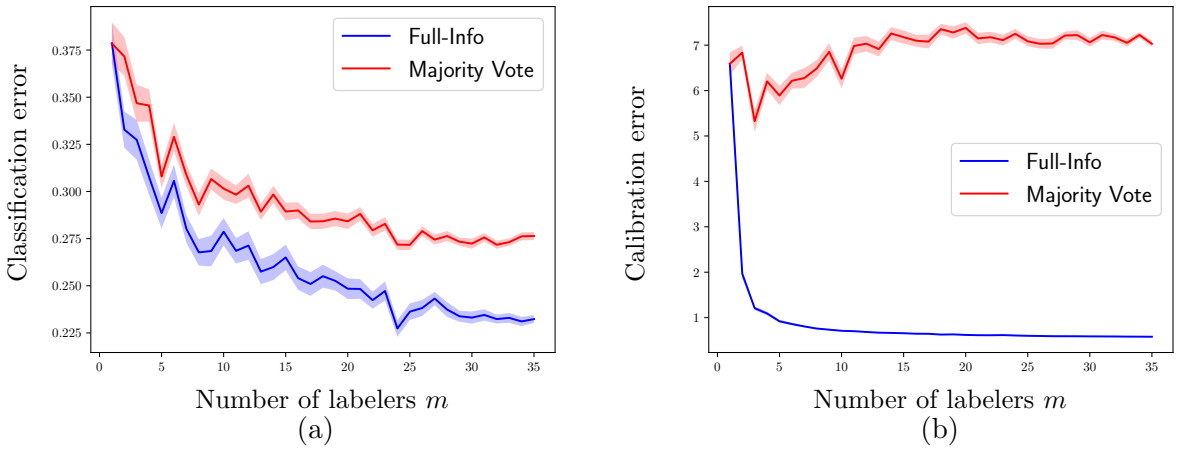

Figure 1. Experiments on BlueBirds dataset. (a) Classification error. (b) Calibration error $|\operatorname{logit}(\widetilde{p}) - \operatorname{logit}(p)|$ with ResNet features reduced via PCA to dimension $d = 25$. Error bars show 2 standard error confidence bands over $T = 100$ trials.

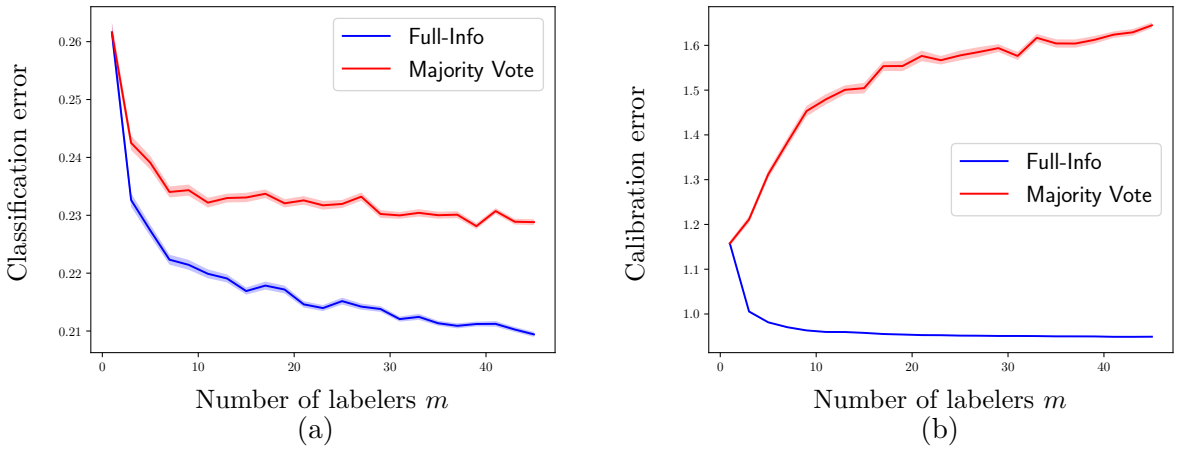

Figure 2. Experiments on CIFAR-10H dataset. (a) Classification error. (b) Calibration error $|\operatorname{logit}(\widetilde{p}) - \operatorname{logit}(p)|$ with ResNet features reduced via PCA to dimension $d = 40$. Error bars show 2 standard error confidence bands over $T = 100$ trials.

6.2 CIFAR-10H

For our second experiment, we consider Peterson et al.'s CIFAR-10H dataset [24], which consists of 10,000 images from CIFAR-10 test set with soft labeling in that for each image, we have approximately 50 labels from different annotators. Each 32×32 image in the dataset belongs to one of the ten classes `airplane`, `automobile`, `bird`, `cat`, `dog`, `frog`, `horse`, `ship`, or `truck`; labelers assign each image to one of the classes. To maintain some fidelity to the binary classification setting we analyze throughout the paper, we transform the problem into a set of 10 binary classification problems. For each class c, we take each initial image/label pair $(x, y) \in \mathbb{R}^{32\times 32} \times \{1, \ldots, 10\}$, assigning binary label 1 if $y = c$ and 0 otherwise (so the annotator labels it as an alternative class $y \neq c$). Most of the images in the dataset are very easy to classify: more than 80% have a unanimous label from each of

the $m = 50$ labelers, meaning that the MLE and majority vote estimators (2) and (3) coincide for these. (In experiments with this full dataset, we saw little difference between the two estimators.)

As our theoretical results highlight the importance of classifier difficulty, we therefore balance the dataset by considering subsets of harder images as follows. For each fixed target c (e.g., `cat`) and for image i, let $\widehat{p}_i$ be the empirical probability of the target among the 50 annotator labels. Then for $p \in [\frac{1}{2}, 1]$, define the subsets

$$\mathcal{S}_p = \{i \in [n] : \max\{\widehat{p}_i, 1 - \widehat{p}_i\} \le p\},$$

so that $p = \frac{1}{2}$ corresponds to images with substantial confusion, and $p = 1$ to all images (most of which are easy). We test on $\mathcal{S}_{0.9}$ (labelers have at most 90% agreement), which consists of with 441 images. For image i, we again generate features $x_i \in \mathbb{R}^d$ by by taking the last layer of a pretrained ResNet50 neural network $\tilde{x}_i \in \mathbb{R}^{d_{\text{init}}}$, using PCA to reduce to a $d = 40$-dimensional feature. We follow the same procedure as in Sec. 6.1, subsampling $m = 1, 2, \ldots, 45$ labelers and using 10-fold cross validation to evaluate classification and calibration error. We report the results in Figure 2. Again we see that—as the number of labelers increases—both aggregated and non-aggregated methods evidence improved classification error, but the majority vote procedure (cleaned data) yields less improvement than one with access to all (uncertain) labels. These results are again consistent with our theoretical predictions.

7 Discussion

In spite of the technical detail we require to prove our results, we view this work as almost preliminary and hope that it inspires further work on the full pipeline of statistical machine learning, from dataset creation to model release. Many questions remain both on the theoretical and applied sides of the work.

On the theoretical side, our main focus has been on a stylized model of label aggregation, with majority vote mostly—with the exception of the crowdsourcing model in Sec. 5.3–functioning as the stand-in for more sophisticated aggregation strategies. It seems challenging to show that *no* aggregation strategy can work as well as multi-label strategies; it would be interesting to more precisely delineate the benefits and drawbacks of more sophisticated denoising and whether it is useful. We focus throughout on low-dimensional asymptotics, using asymptotic normality to compare estimators. While these insights are valuable, and make predictions consistent with the experimental work we provide, investigating how things may change with high-dimensional scaling or via non-asymptotic results might yield new insights both for theory and methodology. As one example, with high-dimensional scaling, classification datasets often become separable [5, 6], which is consistent with modern applied machine learning [42] but makes the asymptotics we derive impossible. One option—for which we have a few preliminary results that we omit, as they are similar to the many results we already include—is to investigate the asymptotics of maximum-margin estimators, which coincide with the limits of ridge-regularized logistic regression [30, 34].

On the methodological and applied side, given the extent to which the test/challenge dataset methodology drives progress in machine learning [11], it seems that developing newer datasets to incorporate labeler uncertainty could yield substantial benefits. A particular refrain is that modern deep learning methods are overconfident in their predictions [13, 42]; perhaps by calibrating them to *labeler* uncertainty we could substantially improve their robustness and performance. We look forward to deeper investigations of the intricacies and intellectual foundations of the full practice of statistical machine learning.

References

[1] A. Asuncion and D. J. Newman. UCI machine learning repository, 2007. URL http://www.ics.uci.edu/~mlearn/MLRepository.html.

[2] P. L. Bartlett, M. I. Jordan, and J. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101:138–156, 2006.

[3] P. Bickel, C. A. J. Klaassen, Y. Ritov, and J. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Springer Verlag, 1998.

[4] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: a survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.

[5] E. Candès and P. Sur. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525, 2019.

[6] E. Candès and P. Sur. The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *Annals of Statistics*, 48(1):27–42, 2020.

[7] M. Charikar, J. Steinhardt, and G. Valiant. Learning from untrusted data. In *Proceedings of the Forty-Ninth Annual ACM Symposium on the Theory of Computing*, 2017.

[8] L. H. Chen, L. Goldstein, and Q.-M. Shao. *Normal Approximation by Stein's method*. Springer, 2010.

[9] A. Dawid and A. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society*, 28:20–28, 1979.

[10] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. ImageNet: a large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[11] D. L. Donoho. 50 years of data science. *Journal of Computational and Graphical Statistics*, 26 (4):745–766, 2017.

[12] E. Giné and R. Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press, 2021.

[13] I. Goodfellow, O. Vinyals, and A. Saxe. Qualitatively characterizing neural network optimization problems. In *Proceedings of the Third International Conference on Learning Representations*, 2015.

[14] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, second edition, 2009.

[15] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

[16] J. Howe. The rise of crowdsourcing. *Wired Magazine*, 14(6):1–4, 2006.

[17] D. Karger, S. Oh, and D. Shah. Budget-optimal task allocation for reliable crowdsourcing systems. *Operations Research*, 62(1):1–24, 2014.

[18] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[19] M. Ledoux and M. Talagrand. *Probability in Banach Spaces*. Springer, 1991.

[20] D. Lewis, Y. Yang, T. Rose, and F. Li. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, 2004.

[21] E. Mammen and A. B. Tsybakov. Smooth discrimination analysis. *Annals of Statistics*, 27: 1808–1829, 1999.

[22] M. Marcus, B. Santorini, and M. Marcinkiewicz. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19:313–330, 1994.

[23] A. Owen. Empirical likelihood ratio confidence regions. *The Annals of Statistics*, 18(1):90–120, 1990.

[24] J. C. Peterson, R. M. Battleday, T. L. Griffiths, and O. Russakovsky. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9617–9626, 2019.

[25] Y. Plan and R. Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Transactions on Information Theory*, 59(1):482–494, 2013.

[26] Y. Plan and R. Vershynin. The generalized lasso with non-linear observations. *IEEE Transactions on Information Theory*, 62(3):1528–1537, 2016.

[27] E. A. Platanios, M. Al-Shedivat, E. Xing, and T. Mitchell. Learning from imperfect annotations. *arXiv:2004.03473 [cs.LG]*, 2020.

[28] A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Ré. Snorkel: rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment*, 11(3):269–282, 2017.

[29] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy. Learning from crowds. *Journal of Machine Learning Research*, 11(4), 2010.

[30] S. Rosset, J. Zhu, and T. Hastie. Boosting as a regularized path to a maximum margin classifier. *Journal of Machine Learning Research*, 5:941–973, 2004.

[31] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[32] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM and Mathematical Programming Society, 2009.

[33] I. Shevtsova. On the absolute constants in the Berry-Esseen-type inequalities. *Doklady Mathematics*, 89(3):378–381, 2014.

[34] D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(18):1–57, 2018.

[35] T. Tian and J. Zhu. Max-margin majority voting for learning from crowds. In *Advances in Neural Information Processing Systems 28*, pages 1621–1629, 2015.

[36] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[37] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York, 1996.

[38] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

[39] M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

[40] P. Welinder, S. Branson, P. Perona, and S. Belongie. The multidimensional wisdom of crowds. In *Advances in Neural Information Processing Systems 23*, pages 2424–2432, 2010.

[41] J. Whitehill, T. fan Wu, J. Bergsma, J. Movellan, and P. Ruvolo. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *Advances in Neural Information Processing Systems 22*, 22:2035–2043, 2009.

[42] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

A Technical lemmas

We collect several technical lemmas and their proofs in this section, which will be helpful in the main proofs. See Appendices A.1, A.2, A.3 and A.4 for their proofs.

Lemma A.1. *Suppose $A, B \in \mathbb{R}^{d\times d}$ are symmetric, $AB = BA = 0$ and the matrix $A + B$ is invertible. Then*

$$(A + B)^{-1} = A^\dagger + B^\dagger.$$

The next two lemmas characterize asymptotic behaviors of expectations involving a fixed function f and some random variable Z that satisfies Assumption A2 for given $\beta > 0$ and $c_Z < \infty$. To facilitate stating the theorems, we recall that such Z are (β, c_Z)-regular.

Lemma A.2. *Let $\beta > 0$ and f be a function on $\mathbb{R}_+$ such that $z^{\beta-1}f(z)$ is integrable. If Z is (β, c_Z)-regular (Assumption A2), then*

$$\lim_{t\to\infty} t^\beta \mathbb{E}[f(t|Z|)] = c_Z \int_0^\infty z^{\beta-1} f(z)dz.$$

Lemma A.3. *Let $\beta > 0$ and $c_Z < \infty$, and Z be (β, c_Z)-regular, let $f : \mathbb{R}_+ \to \mathbb{R}$ satisfy $|f(z)| \le a_0 + a_p z^p$ for some $a_0, a_p < \infty$ and all $z \in \mathbb{R}$, and assume $|Z|$ has finite pth moment. Additionally let $\rho_m(t)$ be the majority vote prediction function (15) and Assumption A4 hold for each $\sigma_j^\star$ with limiting average derivative $\overline{\sigma}^{\star\prime}(0)$ at zero. Then for any $c > 0$*

$$\lim_{m\to\infty} m^{\frac{\beta}{2}} \mathbb{E}\left[f(\sqrt{m}|Z|)(1 - \rho_m(cZ))\right] = c_Z \int_0^\infty z^{\beta-1} f(z)\Phi\left(-2\overline{\sigma}^{\star\prime}(0)cz\right) dz,$$

where $\Phi(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$ is the standard normal cumulative distribution function.

The fourth lemma is a uniform convergence result for the empirical risk in the case that we have potentially distinct link functions (cf. Sec. 5.1).

Lemma A.4. *Assume $\mathbb{E}[\|X\|_2^\gamma] \le M^\gamma$ for some $M \ge 1$ and $\gamma \ge 2$ and let the radius $1 \le r < \infty$. Let $\mathcal{F}_{\mathsf{link}} \subset \{\sigma : \mathbb{R} \to [0, 1], \|\sigma\|_{\mathrm{Lip}} \le \mathsf{L}\}$. Then for a constant $C \lesssim \sqrt{d\mathsf{L}}$, we have*

$$\mathbb{E}\left[\sup_{\|\theta\|_2\le r} \sup_{\vec{\sigma}\in\mathcal{F}_{\mathsf{link}}^m} |P_n\ell_{\vec{\sigma},\theta} - L(\theta, \vec{\sigma})|\right] \le C \cdot (Mr)^{\frac{4\gamma}{3\gamma+1}} \left(\frac{m}{n}\right)^{\frac{\gamma}{3\gamma+1}} \sqrt{\log(rn)}.$$

A.1 Proof of Lemma A.1

As $AB = BA = 0$, the symmetric matrices A and B commute and so are simultaneously orthogonally diagonalizable [15, Thm. 4.5.15]. As $AB = BA = 0$, we can thus write

$$A = U \begin{bmatrix} \Lambda_1 & 0 \\ 0 & 0 \end{bmatrix} U^\top, \qquad B = U \begin{bmatrix} 0 & 0 \\ 0 & \Lambda_2 \end{bmatrix} U^\top,$$

for some orthogonal $U \in \mathbb{R}^{d\times d}$, and as $A + B$ is invertible, Λ_1, Λ_2 are invertible diagonal matrices. We conclude the proof by writing

$$(A + B)^{-1} = U \begin{bmatrix} \Lambda_1^{-1} & 0 \\ 0 & \Lambda_2^{-1} \end{bmatrix} U^\top = U \begin{bmatrix} \Lambda_1^{-1} & 0 \\ 0 & 0 \end{bmatrix} U^\top + U \begin{bmatrix} 0 & 0 \\ 0 & \Lambda_2^{-1} \end{bmatrix} U^\top = A^\dagger + B^\dagger.$$

A.2 Proof of Lemma A.2

By the change of variables $w = tz$, we have

$$t^\beta \mathbb{E}[f(t|Z|)] = \int_0^\infty f(tz) \cdot t^{\beta-1} p(z) \cdot t dz = \int_0^\infty f(w) \cdot t^{\beta-1} p(w/t) dw$$
$$= \int_0^\infty w^{\beta-1} f(w) \cdot (w/t)^{1-\beta} p(w/t) dw.$$

As $|w^{\beta-1} f(w) \cdot (w/t)^{1-\beta} p(w/t)| \le \sup_{z \in (0,\infty)} z^{1-\beta} p(z) \cdot |w^{\beta-1} f(w)|$, where $w^{\beta-1} f(w)$ is integrable by assumption, we can invoke dominated convergence to see that

$$\lim_{t\to\infty} t^\beta \mathbb{E}[f(t|Z|)] = \int_0^\infty w^{\beta-1} f(w) \cdot \lim_{t\to\infty} (w/t)^{1-\beta} p(w/t) dw = c_Z \int_0^\infty w^{\beta-1} f(w) dw.$$

A.3 Proof of Lemma A.3

By rescaling arguments, it suffices to prove the theorem for any function f satisfying $|f(z)| \le 1 + z^p$ on $\mathbb{R}_+$ for any $p \in \mathbb{N}$ such that $|Z|$ has finite pth moment. For such f, we wish to show

$$\lim_{m\to\infty} m^{\frac{\beta}{2}} \mathbb{E}\left[f(\sqrt{m}|Z|)(1 - \rho_m(cZ))\right] = c_Z \int_0^\infty z^{\beta-1} f(z) \Phi\left(-2\overline{\sigma}^{\star\prime}(0) cz\right) dz,$$

where Φ is the standard normal cdf. The key insight is that we can approximate $1 - \rho_m(t)$ by a suitable Gaussian cumulative distribution function (recognizing that $\rho_m(t) > \frac{1}{2}$ for $t \neq 0$ by definition (15) as the probability the majority vote is correct given margin $\langle \theta^\star, X \rangle = t$).

We first assume $Z \ge 0$ with probability 1, as the general result follows by writing $Z = (Z)_+ - (-Z)_+$. We decompose into two expectations, depending on Z being large or small:

$$m^{\frac{\beta}{2}} \mathbb{E}\left[f(\sqrt{m}|Z|)(1 - \rho_m(cZ))\right]$$
$$= \underbrace{m^{\frac{\beta}{2}} \mathbb{E}\left[f(\sqrt{m}Z)(1 - \rho_m(cZ)) \mathbb{1}\left\{0 \le Z \le \frac{M}{\sqrt{m}}\right\}\right]}_{\text{(I)}} + \underbrace{m^{\frac{\beta}{2}} \mathbb{E}\left[f(\sqrt{m}Z)(1 - \rho_m(cZ)) \mathbb{1}\left\{Z > \frac{M}{\sqrt{m}}\right\}\right]}_{\text{(II)}}. \tag{16}$$

The proof consists of three main parts.

1. We approximate $1 - \rho_m(t)$ by a Gaussian cdf.
2. We can approximate term (I) by replacing $1 - \rho_m(t)$ with the Gaussian cdf, showing that
$$\left| \lim_{m\to\infty} \text{(I)} - c_Z \int_0^\infty z^{\beta-1} f(z) \Phi\left(-2\overline{\sigma}^{\star\prime}(0) cz\right) dz \right| = o_M(1). \tag{17}$$
3. For term (II), we show $1 - \rho_m(cz)$ is small when $Z > M/\sqrt{m}$, which allows us to show that
$$\limsup_{m\to\infty} |\text{(II)}| = o_M(1).$$

Thus by adding the two preceding displays and taking $M \to \infty$, we obtain the lemma.

Before we dive into further details, we use the shorthand functions

$$\sigma_m^{\text{std}}(z) :=, \frac{\sum_{j=1}^m (\sigma_j^\star(cz) - \frac{1}{2})}{\sqrt{\sum_{j=1}^m \sigma_j^\star(cz)(1 - \sigma_j^\star(cz))}}, \quad \Delta_m(z) := 1 - \rho_m(cz) - \Phi(-\sigma_m^{\text{std}}(z)), \tag{18}$$

and we also write $p_\infty(\beta) := \sup_{z \in (0,\infty)} z^{1-\beta} p(z) < \infty$.

Part 1. Normal approximation for $1-\rho_m(t)$ when $t = O(1/\sqrt{m})$. Let $p_j = \sigma_j^\star(t)$ for shorthand and $Y_j \sim \text{Bernoulli}(p_j)$ be independent random variables. For $t > 0$, then

$$1-\rho_m(t) = \mathbb{P}\left(Y_1+\cdots+Y_m < \frac{1}{2}\right) = \mathbb{P}\left(\frac{\sum_{j=1}^m (Y_j - p_j)}{\sqrt{\sum_{j=1}^m p_j(1-p_j)}} < -\frac{\sum_{j=1}^m (p_j - \frac{1}{2})}{\sqrt{\sum_{j=1}^m p_j(1-p_j)}}\right).$$

Consider the centered and standardized random variables $\xi_j = (Y_j - p_j)/\sqrt{\sum_{j=1}^m p_j(1-p_j)}$ so that $\xi_1, \cdots, \xi_m$ are zero mean, mutually independent, and satisfy

$$\sum_{j=1}^m \mathsf{Var}\left(\xi_j\right) = 1,$$
$$\sum_{j=1}^m \mathbb{E}\left[|\xi_j|^3\right] = \frac{\sum_{j=1}^m p_j(1-p_j)(p_j^2+(1-p_j)^2)}{(\sum_{j=1}^m p_j(1-p_j))^{3/2}} \le \frac{\max_{1\le j\le m}(p_j^2+(1-p_j)^2)}{\sqrt{\sum_{j=1}^m p_j(1-p_j)}}.$$

By the Berry-Esseen theorem (cf. Chen et al. [8], Shevtsova [33]), for all $t > 0$

$$\left|1-\rho_m(t) - \Phi\left(-\frac{\sum_{j=1}^m \left(p_j - \frac{1}{2}\right)}{\sqrt{\sum_{j=1}^m p_j(1-p_j)}}\right)\right| \le \frac{3}{4}\cdot\frac{\max_{1\le j\le m}(p_j^2+(1-p_j)^2)}{\sqrt{\sum_{j=1}^m p_j(1-p_j)}}.$$

Fix any $M < \infty$. Then for $0 \le t \le cM/\sqrt{m}$ and large enough m, the right hand side of the preceding display has the upper bound $2/\sqrt{m}$, as in numerator we have $p_j^2 + (1-p_j)^2 \le 1$, and in denominator $\min_{1\le j\le m} p_j(1-p_j) \to 1/4$ for all j as $m \to \infty$, which follows from Assumption A4 that

$$\frac{1}{2} \le \limsup_{m\to\infty} \max_{1\le j\le m} p_j \le \limsup_{m\to\infty} \sup_{1\le j<\infty} \sigma_j^\star\left(\frac{cM}{\sqrt{m}}\right) = \frac{1}{2}.$$

By repeating the same argument for $-cM/\sqrt{m} \le t < 0$, we obtain that for large m and $|t| \le cM/\sqrt{m}$,

$$\left|1-\rho_m(t) - \Phi\left(-\left|\frac{\sum_{j=1}^m \left(\sigma_j^\star(t) - \frac{1}{2}\right)}{\sqrt{\sum_{j=1}^m \sigma_j^\star(t)\left(1-\sigma_j^\star(t)\right)}}\right|\right)\right| \le \frac{2}{\sqrt{m}}. \tag{19}$$

Part 2. Approximating (I) by Gaussian cdf. For the first term (I) in (16), we further decompose into a normal approximation term and an error term,

$$\text{(I)} = \underbrace{m^{\frac{\beta}{2}}\mathbb{E}\left[f(\sqrt{m}Z)\Phi(-\sigma_m^{\text{std}}(z))\mathbb{1}\left\{0 \le Z \le \frac{M}{\sqrt{m}}\right\}\right]}_{\text{(III)}} + \underbrace{m^{\frac{\beta}{2}}\mathbb{E}\left[f(\sqrt{m}Z)\Delta_m(z)\mathbb{1}\left\{0 \le Z \le \frac{M}{\sqrt{m}}\right\}\right]}_{\text{(IV)}},$$

where $\Delta_m(z) = 1-\rho_m(cz) - \Phi(-\sigma_m^{\text{std}}(z))$ as in def. (18). We will show (IV) $\to 0$ and so (III) dominates. By the change of variables $w = \sqrt{m}z$, we can further write (III) as

$$\begin{aligned}\text{(III)} &= m^{\frac{\beta}{2}}\int_0^{\frac{M}{\sqrt{m}}} f(\sqrt{m}z)\Phi(-\sigma_m^{\text{std}}(z))\cdot p(z)dz \\ &= \int_0^M w^{\beta-1} f(w)\Phi\left(-\sigma_m^{\text{std}}\left(\frac{w}{\sqrt{m}}\right)\right)\cdot(w/\sqrt{m})^{1-\beta}p(w/\sqrt{m})dw.\end{aligned}$$

We want to take the limit $m \to \infty$ and apply dominated convergence theorem. Because $(w/\sqrt{m})^{1-\beta} p(w/\sqrt{m}) \le p_\infty(\beta) < \infty$ and $\sigma_m^{\mathrm{std}}(w/\sqrt{m}) \ge 0$, we have

$$w^{\beta-1} f(w) \Phi\left(-\sigma_m^{\mathrm{std}}\left(\frac{w}{\sqrt{m}}\right)\right) \cdot \left(-\sigma_m^{\mathrm{std}}\left(\frac{w}{\sqrt{m}}\right)\right) \cdot (w/\sqrt{m})^{1-\beta} p(w/\sqrt{m}) \le w^{\beta-1} f(w) \cdot \Phi(0) p_\infty(\beta).$$

As $\beta > 0$ and $|f(w)| \le 1 + w^p$, $w^{\beta-1} f(w)$ is integrable on $[0, M]$, and by (14a) and (14c) in Assumption A4,

$$\lim_{m\to\infty} \sigma_m^{\mathrm{std}}\left(\frac{w}{\sqrt{m}}\right) = \lim_{m\to\infty} \frac{\sqrt{m}\left(\overline{\sigma}_m^\star\left(\frac{cw}{\sqrt{m}}\right) - \frac{1}{2}\right)}{\sqrt{\frac{1}{m}\sum_{j=1}^m \sigma_j^\star\left(\frac{cw}{\sqrt{m}}\right)\left(1 - \sigma_j^\star\left(\frac{cw}{\sqrt{m}}\right)\right)}} = 2\overline{\sigma}^{\star\prime}(0) cw.$$

Using the above display and that $\lim_{m\to\infty}(w/\sqrt{m})^{1-\beta} p(w/\sqrt{m}) = c_Z$, we can thus apply dominated convergence theorem to conclude that

$$\lim_{m\to\infty} (\mathrm{III}) = c_Z \int_0^M w^{\beta-1} f(w) \cdot \Phi\left(-2\overline{\sigma}^{\star\prime}(0) cw\right) dw = c_Z \int_0^\infty w^{\beta-1} f(w) \Phi\left(-2\overline{\sigma}^{\star\prime}(0) cw\right) dw + o_M(1).$$

Next we turn to the error term (IV). By the bound (19), $|\Delta_m(z)| \le 2/\sqrt{m}$ when $|z| \le M/\sqrt{m}$ for large enough m, and substituting $w = \sqrt{m} z$,

$$\begin{aligned}
|(\mathrm{IV})| &\le m^{\frac{\beta}{2}} \mathbb{E}\left[|f(\sqrt{m}Z)| \cdot \frac{2}{\sqrt{m}} \cdot \mathbb{1}\left\{0 \le Z \le \frac{M}{\sqrt{m}}\right\}\right] \\
&= \frac{2}{\sqrt{m}} \int_0^{\frac{M}{\sqrt{m}}} \sqrt{m} \cdot |f(\sqrt{m}z)| \cdot m^{\frac{\beta-1}{2}} p(z) dz = \frac{2}{\sqrt{m}} \int_0^M w^{\beta-1} |f(w)| \cdot (w/\sqrt{m})^{1-\beta} p(w/\sqrt{m}) dw.
\end{aligned}$$

By using Assumption A2 again that $(w/\sqrt{m})^{1-\beta} p(w/\sqrt{m}) \le p_\infty(\beta) < \infty$, we further have

$$|(\mathrm{IV})| \le \frac{2p_\infty(\beta)}{\sqrt{m}} \cdot \int_0^M w^{\beta-1} |f(w)| dw \le \frac{2p_\infty(\beta)}{\sqrt{m}} \cdot \left(\frac{M^\beta}{\beta} + \frac{M^{p+\beta}}{p+\beta}\right) \to 0,$$

where we use $|f(w)| \le 1 + w^p$. We have thus shown the limit (17).

Part 3. Upper bounding (II). In term (II), when Z is large, the key is that the quantity $1 - \rho_m(t)$ is small when $|t| \ge cM/\sqrt{m}$: Hoeffding's inequality implies the tail bound

$$0 \le 1 - \rho_m(t) \le e^{-2\left(\overline{\sigma}_m^\star(t) - \frac{1}{2}\right)^2 m}.$$

Thus

$$|(\mathrm{II})| \le \int_{\frac{M}{\sqrt{m}}}^1 \sqrt{m} |f(\sqrt{m}z)| e^{-2\left(\overline{\sigma}_m^\star(cz) - \frac{1}{2}\right)^2 m} \cdot m^{\frac{\beta-1}{2}} p(z) dz + \int_1^\infty m^{\frac{\beta}{2}} |f(\sqrt{m}z)| e^{-2\left(\overline{\sigma}_m^\star(cz) - \frac{1}{2}\right)^2 m} \cdot p(z) dz.$$

Using the assumption that $\gamma := \liminf_m \inf_{t \ge c}\left(\overline{\sigma}_m^\star(t) - \frac{1}{2}\right) > 0$ from (14b), that $|f(z)| \le 1 + z^p$ and $|Z|$ has finite pth moment, we observe that

$$\int_1^\infty m^{\frac{\beta}{2}} |f(\sqrt{m}z)| e^{-2\left(\overline{\sigma}_m^\star(cz) - \frac{1}{2}\right)^2 m} \cdot p(z) dz \le m^{\frac{1}{2}(\beta+p)} e^{-2\gamma^2 m} \int_1^\infty (1 + z^p) p(z) dz \to 0,$$

and consequently

$$\limsup_{m\to\infty}|(\mathrm{II})| \le \limsup_{m\to\infty}\int_{\frac{M}{\sqrt{m}}}^{1}\sqrt{m}|f(\sqrt{m}z)|e^{-2\left(\overline{\sigma}_m^\star(cz)-\frac{1}{2}\right)^2 m}\cdot m^{\frac{\beta-1}{2}}p(z)dz$$
$$= \limsup_{m\to\infty}\int_{M}^{\sqrt{m}} w^{\beta-1}|f(w)|e^{-2\left(\overline{\sigma}_m^\star\left(\frac{cw}{\sqrt{m}}\right)-\frac{1}{2}\right)^2 m}\cdot(w/\sqrt{m})^{1-\beta}p(w/\sqrt{m})dw.$$

For $w\in[M,\sqrt{m}]$, we have

$$\left(\overline{\sigma}_m^\star\left(\frac{cw}{\sqrt{m}}\right)-\frac{1}{2}\right)^2 m = \left(\frac{\overline{\sigma}_m^\star\left(\frac{cw}{\sqrt{m}}\right)-\frac{1}{2}}{\frac{cw}{\sqrt{m}}}\right)^2 c^2w^2 \ge \left(\inf_{0<t\le c}\frac{\overline{\sigma}_m^\star(t)-\frac{1}{2}}{t}\right)^2 c^2w^2,$$

while Assumption (14b) gives $\delta := \inf_{0<t\le c}\frac{\overline{\sigma}_m^\star(t)-\frac{1}{2}}{t} > 0$, so

$$\limsup_{m\to\infty}|(\mathrm{II})| \le \int_{M}^{\sqrt{m}} w^{\beta-1}|f(w)|e^{-\delta w^2}\cdot(w/\sqrt{m})^{1-\beta}p(w/\sqrt{m})dw.$$

Using the inequality $(w/\sqrt{m})^{1-\beta}p(w/\sqrt{m})\le p_\infty(\beta)<\infty$, we apply dominated convergence:

$$\lim_{m\to\infty}\int_{M}^{\sqrt{m}} w^{\beta-1}|f(w)|e^{-\delta w^2}\cdot(w/\sqrt{m})^{1-\beta}p(w/\sqrt{m})dw = c_Z\int_{M}^{\infty} w^{\beta-1}|f(w)|e^{-\delta w^2}dw = o_M(1).$$

A.4 Proof of Lemma A.4

We follow a typical symmetrization approach, then construct a covering that we use to prove the lemma. Let $P_n^0 = n^{-1}\sum_{i=1}^n \varepsilon_i 1_{X_i,Y_i}$ be the (random) symmetrized measure with point masses at (X_i,Y_i) for $Y_i=(Y_{i1},\dots,Y_{im})$. Then by a standard symmetrization argument, we have

$$\mathbb{E}\left[\sup_{\|\theta\|_2\le r}\sup_{\vec{\sigma}\in\mathcal{F}_{\mathsf{link}}^m}|P_n\ell_{\vec{\sigma},\theta}-L(\theta,\vec{\sigma})|\right] \le 2\mathbb{E}\left[\sup_{\|\theta\|_2\le r}\sup_{\vec{\sigma}\in\mathcal{F}_{\mathsf{link}}^m}\left|P_n^0\ell_{\vec{\sigma},\theta}\right|\right]. \tag{20}$$

We use a covering argument to bound the symmetrized expectation (20). Let $R<\infty$ to be chosen, and for an (again, to be determined) $\epsilon>0$ let $\mathcal{G}\subset\mathcal{F}_{\mathsf{link}}$ denote an ϵ-cover of $\mathcal{F}_{\mathsf{link}}$ in the supremum norm on $[-R,R]$, that is, $\|g-\sigma\| = \sup_{t\in[-R,R]}|g(t)-\sigma(t)|$, and so for each $\sigma\in\mathcal{F}_{\mathsf{link}}$ there exists $g\in\mathcal{G}$ such that $\|g-\sigma\|\le\epsilon$. Then [37, Ch. 2.7] we have $\log\mathrm{card}(\mathcal{G})\le O(1)\frac{R\mathsf{L}}{\epsilon}$. Let Θ_ϵ be a minimal ϵ-cover of $\{\theta \mid \|\theta\|_2\le r\}$ in $\|\cdot\|_2$, so that $\log\mathrm{card}(\Theta_\epsilon)\le d\log(1+\frac{2r}{\epsilon})$ and $\max_{\theta\in\Theta_\epsilon}\|\theta\|_2\le r$. We claim that for each $\|\theta\|_2\le r$ and $\sigma\in\mathcal{F}_{\mathsf{link}}$, there exists $v\in\Theta_\epsilon$ and $g\in\mathcal{G}$ such that

$$\left|\frac{1}{n}\sum_{i=1}^n\varepsilon_i\left(\ell_{\sigma,\theta}(Y_{ij}\mid X_i)-\ell_{g,v}(Y_{ij}\mid X_i)\right)\right| \le \epsilon+\frac{1}{n}\sum_{i=1}^n\|X_i\|_2\,\epsilon+\frac{1}{n}\sum_{i=1}^n 1\{\|X_i\|_2\ge R/r\}. \tag{21}$$

Indeed, for any $g\in\mathcal{F}_{\mathsf{link}}$ and $\theta,v\in\mathbb{R}^d$, we have

$$|\ell_{\sigma,\theta}(y\mid x)-\ell_{g,v}(y\mid x)| = \left|\int_0^{y\langle x,\theta\rangle}\sigma(-t)dt-\int_0^{y\langle x,v\rangle}g(-t)dt\right|$$
$$\le \sup_{|t|\le r\|x\|_2}|\sigma(t)-g(t)| + \left|\int_{y\langle x,v\rangle}^{y\langle x,\theta\rangle}|\sigma(-t)-g(-t)|dt\right|$$

$$\leq \|\sigma - g\| + 1\{\|x\|_2 \geq R/r\} + |\langle x, \theta - v\rangle|$$
$$\leq \|\sigma - g\| + \|x\|_2 \|\theta - v\|_2 + 1\{\|x\|_2 \geq R/r\},$$

where we have used that $\sigma, g \in [0,1]$. Taking the elements g, v in the respective coverings to minimize the above bound gives the guarantee (21).

We now leverage inequality (21) in the symmetrization step (20). We have

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\|\theta\|_2 \leq r} \sup_{\vec{\sigma} \in \mathcal{F}_{\mathsf{link}}^m} |P_n \ell_{\vec{\sigma},\theta} - L(\theta, \vec{\sigma})|\right] \\
&\lesssim \mathbb{E}\left[\max_{\theta \in \Theta_\epsilon} \max_{\vec{g} \in \mathcal{G}^m} \left|P_n^0 \ell_{\vec{g},\theta}\right|\right] + \epsilon + \mathbb{E}[\|X_1\|_2]\epsilon + \frac{1}{n}\sum_{i=1}^n \mathbb{P}(\|X_i\|_2 \geq R/r) \\
&\overset{(i)}{\lesssim} \sqrt{\frac{dmR\mathsf{L}}{\epsilon} \log\left(1 + \frac{2r}{\epsilon}\right)} \mathbb{E}\left[\frac{r^2}{n}\sum_{i=1}^n \|X_i\|_2^2\right]^{1/2} + \epsilon + \mathbb{E}[\|X_1\|_2]\epsilon + \frac{\mathbb{E}[\|X_i\|_2^\gamma]}{(R/r)^\gamma} \\
&\leq Mr\sqrt{\frac{dmR\mathsf{L}}{n\epsilon} \log\left(1 + \frac{2r}{\epsilon}\right)} + (M+1)\epsilon + \frac{M^\gamma}{(R/r)^\gamma},
\end{aligned}$$

where inequality (i) uses that if Z_i are τ^2-sub-Gaussian, then $\mathbb{E}[\max_{i \leq N} |Z_i|] \leq \sqrt{2\tau^2 \log N}$, and that conditional on $\{X_i, Y_i\}_{i=1}^n$, the symmetrized sum $\sum_{i=1}^n \varepsilon_i \frac{1}{m} \sum_{j=1}^m \ell_{g_j,\theta}(Y_{ij} \mid X_i)$ is $r^2 \sum_{i=1}^n \|X_i\|_2^2$-sub-Gaussian, as $|\ell_{g_j,\theta}(Y_{ij} \mid X_i)| \leq |\langle X_i, \theta\rangle| \leq r\|X_i\|_2$. We optimize this bound to get the final guarantee of the lemma: set $\epsilon = (Rm/n)^{1/3}$ (and note that we will choose $R \geq 1$) to obtain

$$\mathbb{E}\left[\sup_{\|\theta\|_2 \leq r} \sup_{\vec{\sigma} \in \mathcal{F}_{\mathsf{link}}^m} |P_n \ell_{\vec{\sigma},\theta} - L(\theta, \vec{\sigma})|\right] \lesssim \sqrt{d\mathsf{L}\log(rn)} Mr \left(\frac{Rm}{n}\right)^{1/3} + \frac{(Mr)^\gamma}{R^\gamma}.$$

Choose $R = ((Mr)^{3(\gamma-1)} n/m)^{\frac{1}{3\gamma+1}}$.

B Proof of Proposition 1

We assume without loss of generality m is odd. When m is even the proof is identical, except that we randomize $\overline{Y}$ when we have equal votes for both classes. As the marginal distribution of X is $\mathsf{N}(0, I_d)$ for both $\mathcal{P}_{(X,\overline{Y})}^{\sigma^\star,\theta^\star,m}$ and $\mathcal{P}_{(X,\overline{Y})}^{\overline{\sigma},\overline{\theta},\overline{m}}$, we only have to show the existence of a link $\overline{\sigma} \in \mathcal{F}_{\mathsf{link}}^0$ such that the conditional distribution on any $X = x$ is the same, i.e.,

$$\mathbb{P}_{\sigma^\star,\theta^\star,m}(\overline{Y} = 1 \mid X = x) = \mathbb{P}_{\overline{\sigma},\overline{\theta},\overline{m}}(\overline{Y} = 1 \mid X = x). \tag{22}$$

For $m \in \mathbb{N}$, define the one-to-one transformations $T_m : [0,1] \to [0,1]$ by

$$T_m(t) := \sum_{i=\lceil m/2 \rceil}^m \binom{m}{i} t^m (1-t)^{m-i},$$

the probability that a $\mathrm{Binom}(m,t)$ is at least $\lceil m/2 \rceil$. For any m, $T_m(t)$ is monotonically increasing in t, and $T_m(0) = 0, T_m(\frac{1}{2}) = \frac{1}{2}, T_m(1) = 1$, and by symmetry $T_m(t) + T_m(1-t) = 1$. Importantly, by the definition of majority vote, we have

$$\mathbb{P}_{\sigma^\star,\theta^\star,m}(\overline{Y} = 1 \mid X = x) = \sum_{i=\lceil m/2 \rceil}^m \binom{m}{i} \sigma^\star(\langle\theta^\star, x\rangle)^m (1 - \sigma^\star(\langle\theta^\star, x\rangle))^{m-i} = T_m \circ \sigma^\star(\langle\theta^\star, x\rangle).$$

Therefore, the $\overline{\sigma}$ defined as

$$\overline{\sigma}(t) := T_{\overline{m}}^{-1} \circ T_m \circ \sigma^\star \left(\frac{\|\theta\|_2}{\|\overline{\theta}\|_2} t \right)$$

still satisfies $\overline{\sigma}(0) = \frac{1}{2}$ and $\overline{\sigma}(t) - \frac{1}{2} > 0$ for all $t > 0$, since T_m maps $(\frac{1}{2}, 1]$ to $(\frac{1}{2}, 1]$. We also have

$$\begin{aligned}
\overline{\sigma}(t) + \overline{\sigma}(-t) &= T_{\overline{m}}^{-1} \circ T_m \circ \sigma^\star \left(\frac{\|\theta\|_2}{\|\overline{\theta}\|_2} t \right) + T_{\overline{m}}^{-1} \circ T_m \circ \sigma^\star \left(-\frac{\|\theta\|_2}{\|\overline{\theta}\|_2} t \right) \\
&\overset{\text{(i)}}{=} T_{\overline{m}}^{-1} \circ T_m \circ \sigma^\star \left(\frac{\|\theta\|_2}{\|\overline{\theta}\|_2} t \right) + T_{\overline{m}}^{-1} \circ T_m \circ \left(1 - \sigma^\star \left(\frac{\|\theta\|_2}{\|\overline{\theta}\|_2} t \right) \right) \\
&\overset{\text{(ii)}}{=} T_{\overline{m}}^{-1} \circ T_m \circ \sigma^\star \left(\frac{\|\theta\|_2}{\|\overline{\theta}\|_2} t \right) + T_{\overline{m}}^{-1} \circ \left(1 - T_m \circ \sigma^\star \left(\frac{\|\theta\|_2}{\|\overline{\theta}\|_2} t \right) \right) \\
&= 1,
\end{aligned}$$

where (i) and (ii) follow from symmetry of $\sigma^\star$ and T_m, respectively. Thus $\overline{\sigma}$ is a valid link function in $\mathcal{F}^0_{\mathsf{link}}$. Using this $\bar{\sigma}$ yields the desired equality (22)

$$\mathbb{P}_{\overline{\sigma},\overline{\theta},\overline{m}}(\overline{Y} = 1 \mid X = x) = T_{\overline{m}} \circ \overline{\sigma}(\langle \overline{\theta}, x \rangle) = T_{\overline{m}} \circ T_{\overline{m}}^{-1} \circ T_m \circ \sigma^\star \left(\frac{\|\theta\|_2}{\|\overline{\theta}\|_2} \cdot \langle \overline{\theta}, x \rangle \right) = \mathbb{P}_{\sigma^\star,\theta^\star,m}(\overline{Y} = 1 \mid X = x).$$

C Proof of Lemma 4.1

Recall that $h = h_{t^\star,m}$ is the calibration gap function (10). We see that because $\mathbb{E}[Z] = 0$, we have $h(0) = -2\mathbb{E}[Z\varphi_m(t^\star Z)] < 0$, while using Assumption A3 and that $\mathbb{E}[|Z|] < \infty$, we apply dominated convergence and that $\lim_{t\to\infty}(\sigma(tZ) - \frac{1}{2})Z = c|Z|$ with probability 1 to obtain

$$\lim_{t\to\infty} h(t) = \mathbb{E}[c|Z|(1 - \varphi_m(t^\star Z))] + \mathbb{E}[c|Z|\varphi_m(t^\star Z)] = c\mathbb{E}[|Z|] > 0.$$

Because $h'(t) = \mathbb{E}[\sigma'(tZ)Z^2(1 - \varphi_m(t^\star Z))] + \mathbb{E}[\sigma'(-tZ)Z^2\varphi_m(t^\star Z)] > 0$, we see that there is a unique t_m solving $h(t_m) = 0$, and evidently $\theta_L^\star = t_m u^\star$ is a minimizer of L.

We compute the Hessian of L. For this, we again let $\theta = tu^\star$ to write

$$\begin{aligned}
\nabla^2 L(\theta, \sigma) &= \mathbb{E}[\sigma'(-\langle \theta, X\rangle)XX^\top \varphi_m(\langle X, \theta^\star\rangle)] + \mathbb{E}[\sigma'(\langle \theta, X\rangle)XX^\top(1 - \varphi_m(\langle X, \theta^\star\rangle))] \\
&= \mathbb{E}\left[\sigma'(-tZ)\varphi_m(t^\star Z)\left(Z^2 u^\star u^{\star\top} + WW^\top\right)\right] + \mathbb{E}\left[\sigma'(tZ)(1 - \varphi_m(t^\star Z))\left(Z^2 u^\star u^{\star\top} + WW^\top\right)\right] \\
&= \mathbb{E}\left[\left(\sigma'(-tZ)\varphi_m(t^\star Z) + \sigma'(tZ)(1 - \varphi_m(t^\star Z))\right) Z^2\right] u^\star u^{\star\top} \\
&\quad + \mathbb{E}\left[\sigma'(-tZ)\varphi_m(t^\star Z) + \sigma'(tZ)(1 - \varphi_m(t^\star Z))\right] \mathbb{E}[WW^\top],
\end{aligned}$$

which gives $\nabla^2 L(\theta, \sigma) \succ 0$ and so $\theta_L^\star$ is unique. Finally, the desired form of the Hessian follows as $\mathbb{E}[WW^\top] = \mathsf{P}_{u^\star}^\perp \Sigma \mathsf{P}_{u^\star}^\perp$.

D Proof of Theorem 1

Let t_m be the solution to $h_{t^\star,m}(t) = 0$ as in Lemma 4.1. The consistency argument is immediate: the losses ℓ are convex, continuous, and locally Lipschitz, so Shapiro et al. [32, Thm. 5.4] gives

$\widehat{\theta}_n \stackrel{a.s.}{\to} \theta_L^\star$. By an appeal to standard M-estimator theory [e.g. 36, Thm. 5.23], we thus obtain

$$\sqrt{n}\big(\widehat{\theta}_n - \theta_L^\star\big) \stackrel{d}{\to} \mathsf{N}\left(0, \nabla^2 L(\theta_L^\star, \sigma)^{-1} \mathsf{Cov}\left(\frac{1}{m}\sum_{j=1}^m \nabla \ell_{\sigma,\theta_L^\star}(Y_j \mid X)\right) \nabla^2 L(\theta_L^\star, \sigma)^{-1}\right). \tag{23}$$

We expand the covariance term to obtain the first main result of the theorem. For shorthand, let $G_j = \nabla \ell_{\sigma,\theta_L^\star}(Y_j \mid X)$, so that $\sum_{j=1}^m \mathbb{E}[G_j] = 0$ and the G_j are conditionally independent given X. Applying the law of total covariance, we have

$$\begin{aligned}
\mathsf{Cov}\left(\frac{1}{m}\sum_{j=1}^m G_j\right) &= \mathsf{Cov}\left(\frac{1}{m}\sum_{j=1}^m \mathbb{E}[G_j \mid X]\right) + \mathbb{E}\left[\mathsf{Cov}\left(\frac{1}{m}\sum_{j=1}^m G_j \mid X\right)\right] \\
&= \underbrace{\mathsf{Cov}\left(\frac{1}{m}\sum_{j=1}^m \mathbb{E}[G_j \mid X]\right)}_{\text{(I)}} + \frac{1}{m^2}\sum_{j=1}^m \underbrace{\mathbb{E}\left[\mathsf{Cov}(G_j \mid X)\right]}_{\text{(II)}},
\end{aligned}$$

where we have used the conditional independence of the Y_j conditional on X. We control each of terms (I) and (II) in turn.

For the first, we have by the independent decomposition $X = Zu^\star + \mathsf{P}_{u^\star}^\perp W$ that

$$\mathbb{E}[G_j \mid X] = \big(\sigma(t_m Z)(1 - \sigma_j^\star(t^\star Z)) - \sigma(-t_m Z)\sigma_j^\star(t^\star Z)\big) X,$$

and so

$$\begin{aligned}
\text{(I)} &= \mathbb{E}\left[\left(\sigma(t_m Z)(1 - \overline{\sigma}^\star(t^\star Z)) - \sigma(-t_m Z)\overline{\sigma}^\star(t^\star Z)\right)^2 XX^\top\right] \\
&= \mathbb{E}\left[\left(\sigma(t_m Z)(1 - \overline{\sigma}^\star(t^\star Z)) - \sigma(-t_m Z)\overline{\sigma}^\star(t^\star Z)\right)^2 Z^2\right] u^\star u^{\star\top} \\
&\quad + \mathbb{E}\left[\left(\sigma(t_m Z)(1 - \overline{\sigma}^\star(t^\star Z)) - \sigma(-t_m Z)\overline{\sigma}^\star(t^\star Z)\right)^2\right] \mathsf{P}_{u^\star}^\perp \Sigma \mathsf{P}_{u^\star}^\perp \\
&= \mathbb{E}[\mathsf{le}(Z)^2 Z^2] u^\star u^{\star\top} + \mathbb{E}[\mathsf{le}(Z)^2] \mathsf{P}_{u^\star}^\perp \Sigma \mathsf{P}_{u^\star}^\perp,
\end{aligned}$$

where we used the independence of W and Z. For term (II) above, we see that conditional on X, $\nabla \ell_{\sigma,\theta}(Y_j \mid X)$ is a binary random variable taking values in $\{-\sigma(-t_m Z)X, \sigma(t_m Z)X\}$ with probabilities $\sigma_j^\star(t^\star Z)$ and $1 - \sigma_j^\star(t^\star Z)$, respectively, so that a calculation leveraging the variance of Bernoulli random variables yields

$$\mathsf{Cov}(G_j \mid X) = \sigma_j^\star(t^\star Z)(1 - \sigma_j^\star(t^\star Z))\left(\sigma(t_m Z) + \sigma(-t_m Z)\right)^2 XX^\top = v_j(Z) XX^\top,$$

where we used the definition (12) of the variance terms. A similar calculation to that we used for term (I) then gives that

$$\text{(II)} = \frac{1}{m^2}\sum_{j=1}^m \mathbb{E}[v_j(Z) Z^2] u^\star u^{\star\top} + \frac{1}{m^2}\sum_{j=1}^m \mathbb{E}[v_j(Z)] \mathsf{P}_{u^\star}^\perp \Sigma \mathsf{P}_{u^\star}^\perp.$$

Applying Lemma A.1 then allows us to decompose the the covariance in expression (23) into terms in the span of $u^\star u^{\star\top}$ and those perpendicular to it, so that the asymptotic covariance is

$$\frac{\mathbb{E}[\mathsf{le}(Z)^2 Z^2] + \frac{1}{m^2}\sum_{j=1}^m \mathbb{E}[v_j(Z) Z^2]}{\mathbb{E}[\mathsf{he}(Z) Z^2]^2} u^\star u^{\star\top} + \frac{\mathbb{E}[\mathsf{le}(Z)^2] + \frac{1}{m^2}\sum_{j=1}^m \mathbb{E}[v_j(Z)]}{\mathbb{E}[\mathsf{he}(Z)]^2} \left(\mathsf{P}_{u^\star}^\perp \Sigma \mathsf{P}_{u^\star}^\perp\right)^\dagger,$$

by applying Lemma 4.1 for the form of the Hessian $\nabla^2 L(\theta_L^\star, \sigma)$ and Lemma A.1 for the inverse, giving the first result of the theorem.

To obtain the second result, we apply the delta method with the mapping $\phi(x) = x/\left\|x\right\|_2$, which satisfies $\nabla\phi(x) = (I - \phi(x)\phi(x)^\top)/\left\|x\right\|_2$, so that for $\widehat{u}_n = \widehat{\theta}_n/\|\widehat{\theta}_n\|_2$ we have

$$\sqrt{n}(\widehat{u}_n - u^\star) \xrightarrow{d} \mathsf{N}\left(0, \frac{1}{t_m^2}\frac{\mathbb{E}[\mathsf{le}(Z)^2] + \frac{1}{m^2}\sum_{j=1}^m \mathbb{E}[v_j(Z)]}{\mathbb{E}[\mathsf{he}(Z)]^2}\left(\mathsf{P}_{u^\star}^\perp \Sigma \mathsf{P}_{u^\star}^\perp\right)^\dagger\right)$$

as desired.

E Proof of Theorem 2

As in the proof of Theorem 1, we begin with a consistency result. Let t_m be the solution to $h_{t^\star,m}(t) = 0$ as in Lemma 4.1. The once again, Shapiro et al. [32, Thm. 5.4] shows that $\widehat{\theta}_n \xrightarrow{a.s.} \theta_L^\star$. As previously, appealing to standard M-estimator theory [e.g. 36, Thm. 5.23], we obtain

$$\sqrt{n}\big(\widehat{\theta}_n - \theta_L^\star\big) \xrightarrow{d} \mathsf{N}\left(0, \nabla^2 L(\theta_L^\star, \sigma)^{-1}\mathsf{Cov}\left(\nabla \ell_{\sigma,\theta_L^\star}(\overline{Y} \mid X)\right)\nabla^2 L(\theta_L^\star, \sigma)^{-1}\right), \tag{24}$$

the difference from the asymptotic (23) appearing in the covariance term. Here, we recognize that conditional on $X = Zu^\star + W$, the vector $\nabla \ell_{\sigma,\theta_L^\star}(\overline{Y} \mid X)$ takes on the values $\{-\sigma(-t_m Z)X, \sigma(t_m Z)X\}$ each with probabilities $\varphi_m(t^\star Z)$ and $1 - \varphi_m(t^\star Z)$, respectively, while $\nabla \ell_{\sigma,\theta_L^\star}(\overline{Y} \mid X)$ is (unconditionally) mean zero. Thus we have

$$\begin{aligned}
\mathsf{Cov}\left(\nabla \ell_{\sigma,\theta_L^\star}(\overline{Y} \mid X)\right) &= \mathbb{E}\left[\sigma(-t_m Z)^2 \varphi_m(t^\star Z) XX^\top + \sigma(t_m Z)^2(1 - \varphi_m(t^\star Z))XX^\top\right] \\
&= \mathbb{E}\left[\left(\sigma(-t_m Z)^2\varphi_m(t^\star Z) + \sigma(t_m Z)^2(1 - \varphi_m(t^\star Z))\right) Z^2\right] u^\star u^{\star\top} \\
&\quad + \mathbb{E}\left[\sigma(-t_m Z)^2\varphi_m(t^\star Z) + \sigma(t_m Z)^2(1 - \varphi_m(t^\star Z))\right] \mathsf{P}_{u^\star}^\perp \Sigma \mathsf{P}_{u^\star}^\perp.
\end{aligned}$$

Applying Lemma A.1 as in the proof of Theorem 1 to decompose the covariance terms in the asymptotic (24), and substituting in $\rho_m(t) = \sigma(t)\mathbb{1}\{t \geq 0\} + (1 - \sigma(t))\mathbb{1}\{t < 0\}$, the limiting covariance in expression (24) becomes

$$\begin{aligned}
&\frac{\mathbb{E}[(\sigma(-t_m|Z|)^2\rho_m(t^\star Z) + \sigma(t_m|Z|)^2(1 - \rho_m(t^\star Z)))Z^2]}{\mathbb{E}[\mathsf{he}(Z)Z^2]^2} u^\star u^{\star\top} \\
&\quad + \frac{\mathbb{E}[\sigma(-t_m|Z|)^2\rho_m(t^\star Z) + \sigma(t_m|Z|)^2(1 - \rho_m(t^\star Z))]}{\mathbb{E}[\mathsf{he}(Z)]^2}\left(\mathsf{P}_{u^\star}^\perp \Sigma \mathsf{P}_{u^\star}^\perp\right)^\dagger.
\end{aligned}$$

Lastly, we apply the delta method to $\phi(x) = x/\left\|x\right\|_2$, exactly as in the proof of Theorem 1, which gives the theorem.

F Proof of Theorem 3

We divide the theorem into two main parts, consistent with the typical division of asymptotic normality results into a consistency result and a distributional result. Recall the notation $L(\theta, \vec{\sigma}) = \mathbb{E}[\ell_{\vec{\sigma},\theta}(Y \mid X)]$ and $\|\vec{\sigma} - \vec{g}\|_{L^2(\mathbb{P})}^2 = \mathbb{E}[\|\vec{\sigma}(Z) - \vec{g}(Z)\|_2^2]$, where Z has the distribution that Assumption A1 specifies.

F.1 Proof of consistency

We demonstrate the consistency $\widehat{\theta}^{\mathrm{sp}}_{n,m} \xrightarrow{p} u^\star$ in three parts, which we present as Lemmas F.1, F.2, and F.4. The first presents an analogue of Lemma 4.1 generalized to the case in which there are m distinct link functions, allowing us to characterize the link-dependent minimizers

$$\theta^\star_{\vec{\sigma}} := \operatorname*{argmin}_\theta L(\theta, \vec{\sigma})$$

via a one-dimensional scalar on the line $\{tu^\star \mid t \in \mathbb{R}_+\}$. The second, Lemma F.2, then shows that $\theta^\star_{\vec{\sigma}} \to u^\star$ as $\vec{\sigma}$ approaches $\vec{\sigma}^\star$, where the scaling that $\|\theta^\star_{\vec{\sigma}}\|_2 \to 1$ follows by the normalization Assumption A5. Finally, the third lemma demonstrates the probabilistic convergence $\widehat{\theta}^{\mathrm{sp}}_{n,m} - \theta^\star_{\vec{\sigma}_n} \xrightarrow{p} 0$ whenever $\vec{\sigma}_n \xrightarrow{p} \vec{\sigma}^\star$ in $L^2(P)$.

We begin with the promised analogue of Lemma 4.1.

Lemma F.1. *Define the calibration gap function*

$$h_{\vec{\sigma}}(t) := \frac{1}{m} \sum_{j=1}^m \mathbb{E}\left[\left(\sigma_j(tZ) - \sigma_j^\star(Z)\right) Z\right]. \tag{25}$$

Then the loss $L(\theta, \vec{\sigma})$ has unique minimizer $\theta^\star_{\vec{\sigma}} = t_{\vec{\sigma}} u^\star$ for the unique $t_{\vec{\sigma}} \in (0, \infty)$ solving $h_{\vec{\sigma}}(t) = 0$. Additionally, taking $\mathsf{he}_j(t, z) := \sigma_j'(-tz)\sigma_j^\star(z) + \sigma_j'(tz)\sigma_j^\star(-z)$, we have

$$\nabla^2 L(tu^\star, \vec{\sigma}) = \frac{1}{m} \sum_{j=1}^m \left(\mathbb{E}[\mathsf{he}_j(t, Z) Z^2] u^\star u^{\star\top} + \mathbb{E}[\mathsf{he}_j(t, Z)] \mathsf{P}^\perp_{u^\star} \Sigma \mathsf{P}^\perp_{u^\star}\right) \succ 0.$$

Proof. We perform a derivation similar to that we used to derive the gap function (10), with a few modifications to allow collections of m link functions. Note that

$$L(\theta, \vec{\sigma}) = \frac{1}{m} \sum_{j=1}^m \mathbb{E}[\ell_{\sigma_j, \theta}(1 \mid X) \sigma_j^\star(\langle X, u^\star \rangle) + \ell_{\sigma_j, \theta}(-1 \mid X) \sigma_j^\star(-\langle X, u^\star \rangle)],$$

so that leveraging the usual ansatz that $\theta = tu^\star$, we have

$$\begin{aligned}
\nabla L(\theta, \vec{\sigma}) &= \mathbb{E}\left[\frac{1}{m} \sum_{j=1}^m \left(\sigma_j^\star(-\langle X, u^\star \rangle) \sigma_j(\langle \theta, X \rangle) - \sigma_j^\star(\langle X, u^\star \rangle) \sigma_j(-\langle \theta, X \rangle)\right) X\right] \\
&= \frac{1}{m} \sum_{j=1}^m \mathbb{E}\left[\left(\sigma_j^\star(-Z)\sigma_j(tZ) - \sigma_j^\star(Z)\sigma_j(-tZ)Z\right)\right] u^\star \\
&= h_{\vec{\sigma}}(t) u^\star,
\end{aligned}$$

where $h_{\vec{\sigma}}(t) = \frac{1}{m} \sum_{j=1}^m \mathbb{E}[(\sigma_j^\star(-Z)\sigma_j(tZ) - \sigma_j^\star(Z)\sigma_j(-tZ))Z]$ is the immediate generalization of the gap (10). We now simplify it to the form (25). Using the symmetries $\sigma^\star(z) = 1 - \sigma^\star(-z)$ and $\sigma(z) = 1 - \sigma(-z)$ for any $\sigma, \sigma^\star \in \mathcal{F}_{\mathsf{link}}$, we observe that

$$\sigma(tz)(1 - \sigma^\star(z)) - \sigma(-tz)\sigma^\star(z) = \sigma(tz) - (\sigma(tz) + \sigma(-tz))\sigma^\star(z) = \sigma(tz) - \sigma^\star(z),$$

and so $h_{\vec{\sigma}}(t) = \frac{1}{m} \sum_{j=1}^m \mathbb{E}[(\sigma_j(tZ) - \sigma_j^\star(Z))Z]$. Of course, $h_{\vec{\sigma}}(0) = -\frac{1}{m} \sum_{j=1}^m \mathbb{E}[\sigma_j^\star(Z)Z] < 0$, as $\mathbb{E}[Z] = 0$, while $\lim_{t \to \infty} h_{\vec{\sigma}}(t) = \frac{1}{m} \sum_{j=1}^m \mathbb{E}[|Z| - \sigma_j^\star(Z)Z] > 0$. Then as $h'_{\vec{\sigma}}(t) = \frac{1}{m} \sum_{j=1}^m \mathbb{E}[\sigma_j'(tZ)Z^2] > 0$, there exists a unique $t_{\vec{\sigma}} \in (0, \infty)$ satisfying $h_{\vec{\sigma}}(t_{\vec{\sigma}}) = 0$.

We turn to the Hessian derivation, where as in the proof of Lemma 4.1, we write for $\theta = tu^\star$ that

$$\nabla^2 L(\theta, \vec{\sigma}) = \frac{1}{m}\sum_{j=1}^m \mathbb{E}\left[(\sigma_j'(-tZ)\sigma_j^\star(Z) + \sigma_j'(tZ)\sigma_j^\star(-Z))Z^2\right] u^\star u^{\star\top} + \frac{1}{m}\sum_{j=1}^m \mathbb{E}\left[(\sigma_j'(-tZ)\sigma_j^\star(Z) + \sigma_j'(tZ)\sigma_j^\star(-Z))\right] \mathsf{P}_{u^\star}^\perp \Sigma \mathsf{P}_{u^\star}^\perp,$$

and so $\nabla^2 L(\theta, \vec{\sigma}) \succ 0$ and $L(\theta, \vec{\sigma})$ has unique minimizer $\theta_{\vec{\sigma}}^\star = t_{\vec{\sigma}} u^\star$. □

With Lemma F.1 serving as the analogue of Lemma 4.1, we can now show the continuity of the optimizing parameter $\theta_{\vec{\sigma}}^\star$ in $\vec{\sigma}$:

Lemma F.2. *As $\|\vec{\sigma} - \vec{\sigma}^\star\|_{L^2(\mathbb{P})} \to 0$, we have $\theta_{\vec{\sigma}}^\star - u^\star \to 0$.*

Proof. Via Lemma F.1, it is evidently sufficient to show that the solution $t_{\vec{\sigma}}$ to $h_{\vec{\sigma}}(t) = 0$ converges to 1. To show this, note the expansion

$$m h_{\vec{\sigma}}(t) = \sum_{j=1}^m \left(\mathbb{E}[(\sigma_j(tZ) - \sigma_j^\star(tZ))Z] + \mathbb{E}[(\sigma_j^\star(tZ) - \sigma_j^\star(Z))Z]\right). \tag{26}$$

We use the following claim, which shows that the first term tends to 0 uniformly in t near 1:

Claim F.3. *For any $\sigma, \sigma^\star \in \mathcal{F}_{\mathsf{link}}$, we have $\sup_{t\in[\frac12,2]} \|\sigma(tZ) - \sigma^\star(tZ)\|_{L^2(\mathbb{P})} \to 0$ whenever $\|\sigma^\star - \sigma\|_{L^2(\mathbb{P})} \to 0$.*

Proof. We use that the density $p(z)$ of $|Z|$ is continuous and nonzero on $(0,\infty)$. Take any $0 < M_0 < M_1 < \infty$. Then

$$\begin{aligned}
&\sup_{t\in[\frac12,2]} \|\sigma^\star(tZ) - \sigma(tZ)\|_{L^2(\mathbb{P})} = \sup_{t\in[\frac12,2]} \sqrt{\int_0^\infty (\sigma^\star(tz) - \sigma(tz))^2 p(z) dz} \\
&\overset{(i)}{\le} \mathbb{P}(|Z| \le M_0) + \mathbb{P}(|Z| \ge M_1) + \sup_{t\in[\frac12,2]} \sqrt{\int_{M_0}^{M_1} (\sigma^\star(tz) - \sigma(tz))^2 p(tz) \cdot \frac{p(z)}{p(tz)} dz} \\
&\le \mathbb{P}(|Z| \le M_0) + \mathbb{P}(|Z| \ge M_1) + \sup_{\frac{M_0}{2} \le z, z' \le 2M_1} \sqrt{\frac{p(z)}{p(z')}} \sup_{t\in[\frac12,2]} \sqrt{\int_{M_0}^{M_1} (\sigma^\star(tz) - \sigma(tz))^2 p(tz) dz},
\end{aligned}$$

where in (i) we use that the link functions $\sigma^\star$ and σ are bounded within $[0,1]$. For any fixed $0 < M_0 \le M_1 < \infty$, the ratio $\frac{p(z)}{p(z')}$ is bounded for $z, z' \in [\frac12 M_0, 2M_1]$, and using the substitution $tz \mapsto z$ we have

$$\sup_{t\in[\frac12,2]} \sqrt{\int_{M_0}^{M_1} (\sigma^\star(tz) - \sigma(tz))^2 p(tz) dz} \le \sqrt{2} \cdot \|\sigma^\star - \sigma\|_{L^2(\mathbb{P})}.$$

We thus have that $\|\sigma(tZ) - \sigma^\star(tZ)\|_{L^2(\mathbb{P})} \le K \|\sigma - \sigma^\star\|_{L^2(\mathbb{P})} + \mathbb{P}(|Z| \notin [M_0, M_1])$, where K depends only on M_0, M_1 and the distribution of Z. Take $M_0 \downarrow 0$ and $M_1 \uparrow \infty$. □

Leveraging the expansion (26) preceding Claim F.3 and the claim itself, we see that

$$h_{\vec{\sigma}}(t) = \mathbb{E}[Z^2] \cdot o(1) + \frac{1}{m} \sum_{j=1}^{m} \mathbb{E}[(\sigma_j^\star(tZ) - \sigma_j^\star(Z))Z]$$

uniformly in $t \in [\frac{1}{2}, 2]$ as $\|\vec{\sigma}^\star - \vec{\sigma}\|_{L^2(\mathbb{P})} \to 0$. The monotonicity of each $\sigma_j^\star$ guarantees that if $f_j(t) = \mathbb{E}[(\sigma_j^\star(tZ) - \sigma_j^\star(Z))Z]$, then $f_j'(t) = \mathbb{E}[\sigma_j^{\star\prime}(tZ)Z^2] > 0$, and so $t = 1$ uniquely solves $f_j(t) = 0$ and we must have $t_{\vec{\sigma}} \to 1$ as $\|\vec{\sigma} - \vec{\sigma}^\star\|_{L^2(\mathbb{P})} \to 0$. □

Finally, we proceed to the third part of the consistency argument: the convergence in probability.

Lemma F.4. *If* $\|\vec{\sigma}_n - \vec{\sigma}\|_{L^2(\mathbb{P})} \xrightarrow{p} 0$, *then* $\widehat{\theta}^{\mathrm{sp}}_{n,m} - \theta^\star_{\vec{\sigma}_n} \xrightarrow{p} 0$.

Proof. By Lemma F.1 and the assumed continuity of the population Hessian, there exists $\delta > 0$ such that

$$L(\theta, \vec{\sigma}) \geq L(\theta^\star_{\vec{\sigma}}, \vec{\sigma}) + \frac{\lambda}{2} \|\theta - \theta^\star_{\vec{\sigma}}\|_2^2 \tag{27}$$

whenever both $\|\vec{\sigma} - \vec{\sigma}^\star\|_{L^2(\mathbb{P})} \leq \delta$ and $\left\|\theta^\star_{\vec{\sigma}} - \theta\right\|_2 \leq \delta$. Applying the uniform convergence Lemma A.4, we see that for any $r < \infty$, we have

$$\sup_{\|\theta\|_2 \leq r, \vec{\sigma} \in \mathcal{F}^m_{\mathsf{link}}} |P_n \ell_{\vec{\sigma},\theta} - L(\theta, \vec{\sigma})| \xrightarrow{p} 0.$$

For $\delta > 0$, define the events

$$\mathcal{E}_n(\delta) := \left\{ \|\theta^\star_{\vec{\sigma}} - u^\star\|_2 \leq \delta, \|\vec{\sigma}_n - \vec{\sigma}^\star\|_{L^2(\mathbb{P})} \leq \delta \right\},$$

where Lemma F.2 and the assumption that $\|\vec{\sigma}_n - \vec{\sigma}^\star\|_{L^2(\mathbb{P})} \xrightarrow{p} 0$ imply that $\mathbb{P}(\mathcal{E}_n(\delta)) \to 1$ for all $\delta > 0$. By the growth condition (27) and uniform convergence $P_n \ell_{\vec{\sigma},\theta} - L(\theta, \vec{\sigma}) \xrightarrow{p} 0$ over $\|\theta\|_2 \leq r$, we therefore have that with probability tending to 1,

$$\inf_{\|\theta - \theta^\star_{\vec{\sigma}_n}\|_2 = \delta} \left\{ P_n \ell_{\theta,\vec{\sigma}_n} - P_n \ell_{\theta^\star_{\vec{\sigma}_n},\vec{\sigma}_n} \right\} \geq \frac{\lambda}{4} \delta^2.$$

The convexity of the losses $\ell_{\theta,\vec{\sigma}}$ in θ and that $\widehat{\theta}^{\mathrm{sp}}_{n,m}$ minimizes $P_n \ell_{\theta,\vec{\sigma}_n}$ then guarantee the desired convergence $\widehat{\theta}^{\mathrm{sp}}_{n,m} - \theta^\star_{\vec{\sigma}_n} \xrightarrow{p} 0$. □

F.2 Asymptotic normality via Donsker classes

While we do not have $\|\vec{\sigma}_n - \vec{\sigma}^\star\|_{L^2(\mathbb{P})} = o_P(n^{-1/2})$, which allows the cleanest and simplest asymptotic normality results with nuisance parameters (e.g. [36, Thm. 25.54]), we still expect $\sqrt{n}(\widehat{\theta}^{\mathrm{sp}}_{n,m} - \theta^\star_{\vec{\sigma}_n})$ to be asymptotically normal, and therefore, as $\theta^\star_{\vec{\sigma}} = tu^\star$ for some $t > 0$, the normalized estimators $\widehat{\theta}^{\mathrm{sp}}_{n,m}/\|\widehat{\theta}^{\mathrm{sp}}_{n,m}\|_2$ should be asymptotically normal to $u^\star$. To develop the asymptotic normality results, we perform an analysis of the empirical process centered at the *estimators* $\widehat{\theta}^{\mathrm{sp}}_{n,m}$, rather than the "true" parameter $u^\star$ as would be typical.

We begin with an expansion. We let $\theta^\star_n = \theta^\star_{\vec{\sigma}_n}$ for shorthand. Then as $\mathbb{P}_{n,m} \nabla_\theta \ell_{\widehat{\theta}^{\mathrm{sp}}_{n,m},\vec{\sigma}_n} = 0$ and $\mathbb{P} \nabla_\theta \ell_{\theta^\star_n,\vec{\sigma}_n} = 0$, we can derive

$$\mathbb{G}_{n,m} \nabla_\theta \ell_{\widehat{\theta}^{\mathrm{sp}}_{n,m},\vec{\sigma}_n} = \sqrt{n} \left(\mathbb{P}_{n,m} \nabla_\theta \ell_{\widehat{\theta}^{\mathrm{sp}}_{n,m},\vec{\sigma}_n} - \mathbb{P} \nabla_\theta \ell_{\widehat{\theta}^{\mathrm{sp}}_{n,m},\vec{\sigma}_n} \right)$$

$$
\begin{aligned}
&= \sqrt{n}\left(\mathbb{P}\nabla_\theta \ell_{\theta_n^\star, \vec{\sigma}_n} - \mathbb{P}\nabla_\theta \ell_{\widehat{\theta}^{\mathrm{sp}}_{n,m}, \vec{\sigma}_n}\right) \\
&= \sqrt{n}\left(\nabla_\theta L(\theta_n^\star, \vec{\sigma}_n) - \nabla_\theta L(\widehat{\theta}^{\mathrm{sp}}_{n,m}, \vec{\sigma}_n)\right) \\
&= \left(\int_0^1 \nabla_\theta^2 L((1-t)\widehat{\theta}^{\mathrm{sp}}_{n,m} + t\theta_n^\star, \vec{\sigma}_n) dt\right) \cdot \sqrt{n}(\theta_n^\star - \widehat{\theta}^{\mathrm{sp}}_{n,m}).
\end{aligned}
$$

The assumed continuity of $\nabla_\theta^2 L(\theta, \vec{\sigma})$ at $(u^\star, \vec{\sigma}^\star)$ (recall Assumption A5) then implies that

$$
\begin{aligned}
\sqrt{n}(\theta_n^\star - \widehat{\theta}^{\mathrm{sp}}_{n,m}) &= \left(\int_0^1 \nabla_\theta^2 L((1-t)\theta_n^\star + t\widehat{\theta}^{\mathrm{sp}}_{n,m}, \vec{\sigma}_n) dt\right)^{-1} \cdot \mathbb{G}_{n,m} \nabla_\theta \ell_{\widehat{\theta}^{\mathrm{sp}}_{n,m}, \vec{\sigma}_n} \\
&= \left(\nabla_\theta^2 L(u^\star, \vec{\sigma}^\star) + o_P(1)\right)^{-1} \cdot \mathbb{G}_{n,m} \nabla_\theta \ell_{\widehat{\theta}^{\mathrm{sp}}_{n,m}, \vec{\sigma}_n}, \qquad (28)
\end{aligned}
$$

where we have used the consistency guarantees $\theta_n^\star \xrightarrow{p} u^\star, \widehat{\theta}^{\mathrm{sp}}_{n,m} \xrightarrow{p} u^\star$ by Lemmas F.2 and F.4 and that $\|\vec{\sigma}_n - \vec{\sigma}\|_{L^2(\mathbb{P})} \xrightarrow{p} 0$ by assumption.

The expansion (28) forms the basis of our asymptotic normality result; while $\widehat{\theta}^{\mathrm{sp}}_{n,m}$ and $\vec{\sigma}_n$ may be data dependent, by leveraging uniform central limit theorems and the theory of Donsker function classes, we can show that $\mathbb{G}_{n,m}\nabla\ell$ has an appropriate normal limit. To that end, define the function classes

$$
\mathcal{F}_\delta := \left\{\nabla_\theta \ell_{\theta,\sigma} \mid \|\theta - u^\star\|_2 \le \delta, \sigma \in \mathcal{F}_{\mathsf{link}}\right\},
$$

where we leave the Lipschitz constant L in $\mathcal{F}_{\mathsf{link}}$ tacit. By Assumption A5 and that $\widehat{\theta}^{\mathrm{sp}}_{n,m} \xrightarrow{p} u^\star$, we know with probability tending to 1 we have the membership $\nabla_\theta \ell_{\widehat{\theta}^{\mathrm{sp}}_{n,m}, \vec{\sigma}_n} \in \mathcal{F}_\delta$. The key result is then that $\mathcal{F}_\delta$ is a Donsker class:

Lemma F.5. *Assume that $\mathbb{E}[\|X\|_2^4] < \infty$. Then $\mathcal{F}_\delta$ is a Donsker class, and moreover, if $d_{\mathcal{F}_{\mathsf{link}}}((\widehat{\theta}^{\mathrm{sp}}_{n,m}, \vec{\sigma}_n), (u^\star, \vec{\sigma}^\star)) \xrightarrow{p} 0$, then*

$$
\mathbb{G}_{n,m} \nabla_\theta \ell_{\widehat{\theta}^{\mathrm{sp}}_{n,m}, \vec{\sigma}_n} - \mathbb{G}_{n,m} \nabla_\theta \ell_{u^\star, \vec{\sigma}^\star} \xrightarrow{p} 0.
$$

Temporarily deferring the proof of Lemma F.5, let us see how it leads to the proof of Theorem 3. Using Lemma F.5 and Slutsky's lemmas in the equality (28), we obtain

$$
\begin{aligned}
\sqrt{n}(\widehat{\theta}^{\mathrm{sp}}_{n,m} - \theta_n^\star) &= -\left(\nabla_\theta^2 L(u^\star, \vec{\sigma}^\star) + o_P(1)\right)^{-1} \cdot \mathbb{G}_{n,m} \nabla_\theta \ell_{u^\star, \vec{\sigma}^\star} + o_P(1) \\
&\xrightarrow{d} \mathsf{N}\left(0, \nabla^2 L(u^\star, \vec{\sigma}^\star) \mathsf{Cov}\left(\frac{1}{m}\sum_{j=1}^m \nabla \ell_{u^\star, \sigma_j^\star}(Y_j \mid X)\right) \nabla^2 L(u^\star, \vec{\sigma}^\star)\right).
\end{aligned}
$$

Calculations completely similar to those we use in the proof of Theorem 1 then give Theorem 3: because $\mathbb{P}(Y_j = y \mid X = x) = \sigma_j^\star(y\langle u^\star, x\rangle)$ by assumption,

$$
\mathsf{Cov}\left(\frac{1}{m}\sum_{j=1}^m \nabla \ell_{u^\star, \sigma_j^\star}(Y_j \mid X)\right) = \frac{1}{m^2}\sum_{j=1}^m \mathsf{Cov}(\nabla \ell_{u^\star, \sigma_j^\star}(Y_j \mid X))
$$

because $Y_j \mid X$ are conditionally independent, while

$$
\begin{aligned}
\mathsf{Cov}(\nabla \ell_{u^\star, \sigma_j^\star}(Y_j \mid X)) &= \mathbb{E}[\sigma_j^\star(Z)(1-\sigma_j^\star(Z)) X X^\top] \\
&= \mathbb{E}[\sigma_j^\star(Z)(1-\sigma_j^\star(Z)) Z^2] u^\star u^{\star\top} + \mathbb{E}[\sigma_j^\star(Z)(1-\sigma_j^\star(Z))] \mathsf{P}_{u^\star}^\perp \Sigma \mathsf{P}_{u^\star}^\perp.
\end{aligned}
$$

When $\sigma_j = \sigma_j^\star$, the Hessian function he_j in Lemma F.1 simplifies to $\mathsf{he}_j(1, z) = \sigma_j^{\star\prime}(z)$ as $\sigma_j^\star$ is symmetric about 0. We then apply the delta method as in the proof of Theorem 1.

Finally, we return to the proof of Lemma F.5.

Proof of Lemma F.5. To prove $\mathcal{F}_\delta$ is Donsker, we show that each coordinate of $\nabla_\theta \ell_{\theta,\sigma} \in \mathcal{F}_\delta$ is, and as $\nabla_\theta \ell_{\theta,\sigma}(y \mid x) = -y\sigma(-y\langle x, \theta\rangle)x$, this amounts to showing the coordinate functions $f^{(i)}_{\theta,\sigma}(v) = v_i \sigma(-\langle v, \theta\rangle)$ form a Donsker class when v has distribution $V = YX$. Let

$$\mathcal{F}^{(i)}_\delta := \left\{ f^{(i)}_{\theta,\sigma}(\cdot) \mid \|\theta - u^\star\|_2 \le \delta, \ \sigma \in \mathcal{F}_{\mathsf{link}} \right\},$$

so it is evidently sufficient to prove that $\mathcal{F}^{(1)}_\delta$ forms a Donsker class.

We use bracketing and entropy numbers [37] to control the $\mathcal{F}^{(i)}_\delta$. Recall that for a function class $\mathcal{F}$, an ϵ-*bracket* of $\mathcal{F}$ in $L^q(\mathbb{P})$ is a collection of functions $\{(l_i, u_i)\}$ such that for each $f \in \mathcal{F}$, there exists i such that $\ell_i \le f \le u_i$ and $\|u_i - l_i\|_{L^q(\mathbb{P})} \le \epsilon$. The *bracketing number* $N_{[]}(\epsilon, \mathcal{F}, L^q(\mathbb{P}))$ is the cardinality of the smallest such ϵ-bracket, and the *bracketing entropy* is

$$J_{[]}(\mathcal{F}, L^q(\mathbb{P})) := \int_0^\infty \sqrt{\log N_{[]}(\epsilon, \mathcal{F}, L^q(\mathbb{P}))} d\epsilon.$$

To show that $\mathcal{F}^{(i)}_\delta$ is Donsker, it is sufficient [37, Ch. 2.5.2] to show that $J_{[]}(\mathcal{F}^{(i)}_\delta, L^2(\mathbb{P})) < \infty$. Our approach to demonstrate that $\mathcal{F}^{(1)}_\delta$ is Donsker is thus to construct an appropriate ϵ-bracket of $\mathcal{F}^{(1)}_\delta$, which we do by first covering ℓ_2-balls in $\mathbb{R}^d$, then for vectors $\theta \in \mathbb{R}^d$, constructing a bracketing of the induced function class $\{f^{(1)}_{\theta,\sigma}\}_{\sigma \in \mathcal{F}_{\mathsf{link}}}$, which we combine to give the final bracketing of $\mathcal{F}^{(1)}_\delta$.

We proceed with this two-stage covering and bracketing. Let $\epsilon, \gamma > 0$ be small numbers whose values we determine later. Define the ℓ_2 ball $\mathbb{B}^d_2 = \{x \in \mathbb{R}^d \mid \|x\|_2 \le 1\}$, and for any $0 < \epsilon < \delta$, let $\mathcal{N}_\epsilon$ be a minimal ϵ-cover of $\delta\mathbb{B}^d_2$ in the Euclidean norm of size $N = N(\epsilon, \delta\mathbb{B}^d_2, \|\cdot\|_2)$, $\mathcal{N}_\epsilon = \{\theta_1, \ldots, \theta_N\}$, so that for any θ with $\|\theta\|_2 \le \delta$ there exists $\theta_i \in \mathcal{N}_\epsilon$ with $\|\theta - \theta_i\|_2 \le \epsilon$. Standard bounds [39, Lemma 5.7]) give

$$\log N(\epsilon, \delta\mathbb{B}^d_2, \|\cdot\|_2) \le d \log\left(1 + \frac{2\delta}{\epsilon}\right). \tag{29}$$

For simplicity of notation and to avoid certain tedious negations, we define the "flipped" monotone function family

$$\mathcal{F}_{\mathsf{flip}} := \{g : \mathbb{R} \to \mathbb{R} \mid g(t) = \sigma(-t)\}_{\sigma \in \mathcal{F}_{\mathsf{link}}}.$$

Now, for any $\theta \in \mathbb{R}^d$, let μ_θ denote the pushforward measure of $\langle V, \theta\rangle = Y\langle X, \theta\rangle$. For $\theta \in \mathbb{R}^d$, we then let $\mathcal{N}_{[],\gamma,\theta}$ be a minimal γ-bracketing of $\mathcal{F}_{\mathsf{flip}}$ in the $L^4(\mu_\theta)$ norm. That is, for $N = N_{[]}(\gamma, \mathcal{F}_{\mathsf{flip}}, L^2(\mu_\theta))$, we have $\mathcal{N}_{[],\gamma,\theta} = \{(l_{\theta,i}, u_{\theta,i})\}_{i=1}^N$, and for each $\sigma \in \mathcal{F}_{\mathsf{link}}$, there exists $i = i(\sigma)$ such that

$$l_{\theta,i}(t) \le \sigma(-t) \le u_{\theta,i}(t) \quad \text{and} \quad \|u_{\theta,i} - l_{\theta,i}\|_{L^4(\mu_\theta)} \le \gamma.$$

Because elements of $\mathcal{F}_{\mathsf{flip}} \subset \mathbb{R} \to [0,1]$ are monotone, van der Vaart and Wellner [37, Thm. 2.7.5] guarantee there exists a universal constant $K < \infty$ such that

$$\sup_Q \log N_{[]}(\gamma, \mathcal{F}_{\mathsf{flip}}, L^4(Q)) \le \frac{K}{\gamma}, \tag{30}$$

and in particular, $\log N_{[]}(\gamma, \mathcal{F}_{\mathsf{flip}}, L^4(\mu_\theta)) \le \frac{K}{\gamma}$ for each $\theta \in \mathbb{R}^d$.

With the covering $\mathcal{N}_\epsilon$ and induced bracketing collections $\mathcal{N}_{[],\gamma,\theta}$, we now turn to a construction of the actual bracketing of the class $\mathcal{F}^{(1)}_\delta$. For any $\theta \in \mathbb{R}^d$ and bracket $(l_{\theta,i}, u_{\theta,i}) \in \mathcal{N}_{[],\gamma,\theta}$, define the functionals $\widehat{l}_{\theta,j}, \widehat{u}_{\theta,j} : \mathbb{R}^d \to \mathbb{R}$ by

$$\widehat{l}_{\theta,j}(v) := (v_1)_+ \max\left\{l_{\theta,j}(\langle v, \theta\rangle) - \mathsf{L}\|v\|_2 \epsilon, 0\right\} - (-v_1)_+ \min\left\{u_{\theta,j}(\langle v, \theta\rangle) + \mathsf{L}\|v\|_2 \epsilon, 1\right\},$$

$$\widehat{u}_{\theta,j}(v) \coloneqq (v_1)_+ \min\left\{u_{\theta,j}(\langle v,\theta\rangle) + \mathsf{L}\left\|v\right\|_2 \epsilon, 1\right\} - (-v_1)_+ \max\left\{l_{\theta,j}(\langle v,\theta\rangle) - \mathsf{L}\left\|v\right\|_2 \epsilon, 0\right\}.$$

The key is that these functions form a bracketing of $\mathcal{F}_\delta^{(1)}$:

Lemma F.6. *Define the set*

$$\mathcal{B}_{\epsilon,\gamma} \coloneqq \left\{\left(\widehat{l}_{\theta_i,j}, \widehat{u}_{\theta_i,j}\right) \mid \theta_i \in \mathcal{N}_\epsilon,\ 1 \le j \le N_{[]}(\gamma, \mathcal{F}_{\mathsf{flip}}, L^4(\mu_{\theta_i}))\right\}.$$

Then $\mathcal{B}_{\epsilon,\gamma}$ is a

$$2\mathsf{L}\mathbb{E}[\|X\|_2^4]^{1/2} \cdot \epsilon + \mathbb{E}[\|X\|_2^4]^{1/4} \cdot \gamma$$

bracketing of $\mathcal{F}_\delta^{(1)}$ with cardinality at most $\log \operatorname{card}(\mathcal{B}_{\epsilon,\gamma}) \le \frac{K}{\gamma} + d\log(1+\frac{\delta}{\epsilon})$.

Proof. Let $f_{\theta,\sigma}^{(1)}(v) \in \mathcal{F}_\delta^{(1)}$. Take $\theta_i \in \mathcal{N}_\epsilon$ satisfying $\|\theta - \theta_i\|_2 \le \epsilon$ and $(l_{\theta_i,j}, u_{\theta,j}) \in \mathcal{N}_{[],\gamma,\theta_i}$ such that $l_{\theta_i,j}(t) \le \sigma(-t) \le u_{\theta_i,j}(t)$ for all t, where $\|u_{\theta_i,j} - l_{\theta_i,j}\|_{L^4(\mu_{\theta_i})} \le \gamma$. We first demonstrate the bracketing guarantee

$$\widehat{l}_{\theta_i,j}(v) \le f_{\theta,\sigma}^{(1)}(v) = v_1\sigma(-\langle v,\theta\rangle) \le \widehat{u}_{\theta_i,j}(v) \ \text{ for all } v \in \mathbb{R}^d.$$

For the upper bound, we have

$$\begin{aligned}
&f_{\theta,\sigma}^{(1)}(v) = v_1\sigma(-\langle v,\theta\rangle)\\
&\overset{\text{(i)}}{\le} (v_1)_+ \min\left\{\sigma(-\langle v,\theta_i\rangle) + \mathsf{L}|\langle v,\theta_i-\theta\rangle|, 1\right\} - (-v_1)_+ \max\left\{\sigma(-\langle v,\theta_i\rangle) - \mathsf{L}|\langle v,\theta_i-\theta\rangle|, 0\right\}\\
&\overset{\text{(ii)}}{\le} (v_1)_+ \min\left\{\sigma(-\langle v,\theta_i\rangle) + \mathsf{L}\left\|v\right\|_2 \epsilon, 1\right\} - (-v_1)_+ \max\left\{\sigma(-\langle v,\theta_i\rangle) - \mathsf{L}\left\|v\right\|_2 \epsilon, 0\right\}\\
&\overset{\text{(iii)}}{\le} (v_1)_+ \min\left\{u_{\theta_i,j}(\langle v,\theta_i\rangle) + \mathsf{L}\left\|v\right\|_2 \epsilon, 1\right\} - (-v_1)_+ \max\left\{l_{\theta_i,j}(\langle v,\theta_i\rangle) - \mathsf{L}\left\|v\right\|_2 \epsilon, 0\right\}\\
&= \widehat{u}_{\theta_i,j}(v),
\end{aligned}$$

where step (i) follows from the L-Lipschitz continuity of σ, (ii) from the Cauchy-Schwarz inequality and that $\|\theta - \theta_i\|_2 \le \epsilon$, while step (iii) follows by the construction that $l_{\theta_i,j}(t) \le \sigma(-t) \le u_{\theta_i,j}(t)$ for all $t \in \mathbb{R}$. Similarly, we obtain the lower bound

$$f_{\theta,\sigma}^{(1)}(v) = v_1\sigma(-\langle v,\theta\rangle) \ge \widehat{l}_{\theta_i,j}(v),$$

again valid for all $v \in \mathbb{R}^d$.

The second part of the proof is to bound the distance between the upper and lower elements in the bracketing. By definition, $\widehat{u}_{\theta_i,j} - \widehat{l}_{\theta_i,j}$ has the pointwise upper bound

$$\left(\widehat{u}_{\theta_i,j}(v) - \widehat{l}_{\theta_i,j}(v)\right)^2 \le \left(|v_1|\left(u_{\theta_i,j}(\langle v,\theta_i\rangle) - l_{\theta_i,j}(\langle v,\theta_i\rangle) + 2\mathsf{L}\left\|v\right\|_2 \epsilon\right)\right)^2.$$

Recalling that $V = YX$, by the Minkowski and Cauchy-Schwarz inequalities, we thus obtain

$$\begin{aligned}
\left\|\widehat{u}_{\theta_i,j}(V) - \widehat{l}_{\theta_i,j}(V)\right\|_{L^2(\mathbb{P})} &\le \left\||V_1|\left(u_{\theta_i,j}(\langle V,\theta_i\rangle) - l_{\theta_i,j}(\langle V,\theta_i\rangle)\right)\right\|_{L^2(\mathbb{P})} + \left\||V_1| \cdot 2\mathsf{L}\left\|V\right\|_2 \epsilon\right\|_{L^2(\mathbb{P})}\\
&\le \left\||V_1|\right\|_{L^4(\mathbb{P})} \cdot \left(\left\|u_{\theta_i,j}(\langle V,\theta_i\rangle) - l_{\theta_i,j}(\langle V,\theta_i\rangle)\right\|_{L^4(\mathbb{P})} + 2\mathsf{L}\epsilon \cdot \left\|\left\|V\right\|_2\right\|_{L^4(\mathbb{P})}\right).
\end{aligned}$$

Noting the trivial bounds $\left\||V_1|\right\|_{L^4(\mathbb{P})} \le \|X\|_{L^4(\mathbb{P})} < \infty$ and the assumed bracketing distance

$$\left\|u_{\theta_i,j}(\langle V,\theta_i\rangle) - l_{\theta_i,j}(\langle V,\theta_i\rangle)\right\|_{L^4(\mathbb{P})} = \left\|u_{\theta_i,j} - l_{\theta_i,j}\right\|_{L^4(\mu_{\theta_i})} \le \gamma,$$

we have the desired bracketing distance $\|\widehat{u}_{\theta_i,j} - \widehat{l}_{\theta_i,j}\|_{L^2(\mathbb{P})} \le 2\mathsf{L}\mathbb{E}[\|X\|_2^4]^{1/2}\epsilon + \mathbb{E}[\|X\|_2^4]^{1/4}\gamma$.

The final cardinality bound is immediate via inequalities (29) and (30). □

Lemma F.6 will yield the desired entropy integral bound. Fix any $t > 0$, and note that if we take $\epsilon = \epsilon(t) \coloneqq t/(4\mathsf{L}\mathbb{E}[\|X\|_2^4]^{1/2})$ and $\gamma = \gamma(t) \coloneqq t/(2\mathbb{E}[\|X\|_2^4]^{1/4})$, then the set $\mathcal{B}_{\epsilon,\gamma}$ is a t-bracketing of $\mathcal{F}_\delta^{(1)}$ in L^2, and moreover, we have the cardinality bound

$$\log N_{[]}(t, \mathcal{F}_\delta^{(1)}, L^2(\mathbb{P})) \le d \log\left(1 + \frac{2\delta}{\epsilon(t)}\right) + \frac{K}{\gamma(t)} \le \frac{8\mathsf{L}d\delta \cdot \mathbb{E}[\|X\|_2^4]^{1/2} + 2K\mathbb{E}[\|X\|_2^4]^{1/4}}{t}.$$

Additionally, as covering numbers are necessarily integer, we have $\log N_{[]}(t) = 0$ whenever $t > (8\mathsf{L}d\delta \cdot \mathbb{E}[\|X\|_2^4]^{1/2} + 2K\mathbb{E}[\|X\|_2^4]^{1/4})/\log 2$. This gives the entropy integral bound

$$J_{[]}(\mathcal{F}_\delta^{(1)}, L^2(\mathbb{P})) = \int_0^\infty \sqrt{\log N_{[]}(t, \mathcal{F}_\delta^{(1)}, L^2(\mathbb{P}))}dt < \infty,$$

and consequently (cf. [36, Thm. 19.5] or [37, Ch. 2.5.2]), $\mathcal{F}_\delta^{(1)}$ is a Donsker class. A completely identical argument shows that $\mathcal{F}_\delta^{(i)}, i = 2, 3, \cdots, d$ are Donsker, and so $\mathcal{F}_\delta$ is a Donsker class, completing the proof of the first claim in Lemma F.5.

To complete the proof of Lemma F.5, we need to show that

$$\mathbb{G}_{n,m}(\nabla_\theta \ell_{\widehat{\theta}^{\mathrm{sp}}_{n,m},\vec{\sigma}_n} - \nabla_\theta \ell_{u^\star,\vec{\sigma}^\star}) = \frac{1}{m}\sum_{j=1}^m \mathbb{G}_n^{(j)}(\nabla_\theta \ell_{\widehat{\theta}^{\mathrm{sp}}_{n,m},\sigma_{n,j}} - \nabla_\theta \ell_{u^\star,\sigma_j^\star}) \xrightarrow{p} 0,$$

where $\mathbb{G}_n^{(j)}$ denotes the empirical process on $(X_i, Y_{ij})_{i=1}^n$. Notably, because m is finite, it is sufficient to show that

$$\mathbb{G}_n^{(j)}(\nabla_\theta \ell_{\widehat{\theta}^{\mathrm{sp}}_{n,m},\sigma_{n,j}} - \nabla_\theta \ell_{u^\star,\sigma_j^\star}) \xrightarrow{p} 0, \quad j = 1, \ldots, m.$$

To that end, we suppress dependence on j for notational simplicity and simply write $\mathbb{G}_n$ and $\ell_{\widehat{\theta}^{\mathrm{sp}}_{n,m},\sigma_n}$, where $\|\sigma_n - \sigma^\star\|_{L^2(\mathbb{P})} \xrightarrow{p} 0$. For any Donsker class $\mathcal{F} \subset \mathcal{X} \to \mathbb{R}^d$ and $\epsilon > 0$, we have

$$\limsup_{\delta \downarrow 0} \limsup_{n\to\infty} \mathbb{P}\left(\sup_{\|f-g\|_{L^2(\mathbb{P})} \le \delta} \mathbb{G}_n(f - g) \ge \epsilon\right) = 0,$$

(see [12, Thm. 3.7.31]), and so in turn it is sufficient to prove that

$$\left\|\nabla \ell_{\widehat{\theta}^{\mathrm{sp}}_{n,m},\sigma_n} - \nabla \ell_{u^\star,\sigma^\star}\right\|_{L^2(\mathbb{P})} \xrightarrow{p} 0. \tag{31}$$

To demonstrate the convergence (31), let M be finite, and note that for any fixed $\theta, \sigma \in \mathcal{F}_{\mathsf{link}}$ that for $V = YX$ we have

$$\begin{aligned}
\|\nabla \ell_{\theta,\sigma} - \nabla \ell_{u^\star,\sigma^\star}\|_{L^2(\mathbb{P})}^2 &= \mathbb{E}\left[\|V\sigma(-\langle V, \theta\rangle) - V\sigma^\star(-\langle V, u^\star\rangle)\|_2^2\right] \\
&\le \mathbb{E}\left[\|V\|_2^2 1\{\|V\|_2 \ge M\}\right] + M^2 \|\sigma(-\langle V, \theta\rangle) - \sigma^\star(-\langle V, u^\star\rangle)\|_{L^2(\mathbb{P})}^2 \\
&\le \mathbb{E}\left[\|X\|_2^2 1\{\|X\|_2 \ge M\}\right] + 2M^2 \|\sigma(-\langle V, u^\star\rangle) - \sigma^\star(-\langle V, u^\star\rangle)\|_{L^2(\mathbb{P})}^2 \\
&\qquad + 2M^2 \|\sigma(-\langle V, u^\star\rangle) - \sigma(-\langle V, \theta\rangle)\|_{L^2(\mathbb{P})}^2
\end{aligned}$$

by the triangle inequality. As

$$\|\sigma(-\langle V, u^\star\rangle) - \sigma(-\langle V, \theta\rangle)\|_{L^2(\mathbb{P})} \le \mathsf{L}\|\theta - u^\star\|_2 \cdot \|V\|_{L^2(\mathbb{P})} = \mathsf{L}\|\theta - u^\star\|_2 \cdot \|X\|_{L^2(\mathbb{P})}$$

and $\widehat{\theta}^{\mathrm{sp}}_{n,m} - u^\star \xrightarrow{p} 0$ and

$$\|\sigma_n(-\langle V, u^\star\rangle) - \sigma^\star(-\langle V, u^\star\rangle)\|_{L^2(\mathbb{P})} = \|\sigma_n - \sigma^\star\|_{L^2(\mathbb{P})} \xrightarrow{p} 0$$

by assumption, it follows that for any $\epsilon > 0$ that

$$\mathbb{P}\left(\left\|\nabla_\theta \ell_{\widehat{\theta}^{\mathrm{sp}}_{n,m},\sigma_n} - \nabla_\theta \ell_{u^\star,\sigma^\star}\right\|_{L^2(\mathbb{P})} \geq \mathbb{E}[\|X\|_2^2\, 1\{\|X\|_2 \geq M\}] + \epsilon\right) \to 0.$$

Taking $M \uparrow \infty$ gives the convergence (31), completing the proof.

G Proofs of asymptotic normality

In this appendix, we include proofs of the convergence results in Propositions 2, 3, and 4. In each, we divide the proof into three steps: we characterize the loss minimizer, apply one of the master Theorems 1 or 2 to obtain asymptotic normality, and then characterize the behavior of the asymptotic covariance as $m \to \infty$.

G.1 Proof of Proposition 2

Asymptotic normality of the MLE. The asymptotic normality result is an immediate consequence of the classical asymptotics for maximum likelihood estimators [36, Thm. 5.29].

Normalized estimator. For the normalized estimator, we appeal to the master results developed in Section 4.1. In particular, since we are in the well-specified logistic model, we can invoke Corollary 3 and write directly that

$$\sqrt{n}(\widehat{u}^{\mathrm{lr}}_{n,m} - u^\star) \xrightarrow{d} \mathsf{N}\left(0, \frac{1}{m} \cdot \frac{1}{t^{\star 2}} \frac{\mathbb{E}[\sigma^{\mathrm{lr}}(t^\star Z)(1-\sigma^{\mathrm{lr}}(t^\star Z))]}{\mathbb{E}[\sigma^{\mathrm{lr}\prime}(t^\star Z)]^2} \mathsf{P}^\perp_{u^\star} \Sigma \mathsf{P}^\perp_{u^\star}\right),$$

which immediately implies

$$C(t) = \frac{1}{t^2} \frac{\mathbb{E}[\sigma^{\mathrm{lr}}(tZ)(1-\sigma^{\mathrm{lr}}(tZ))]}{\mathbb{E}[\sigma^{\mathrm{lr}\prime}(tZ)]^2} = \frac{1}{t^2 \mathbb{E}\left[\frac{e^{tZ}}{(1+e^{tZ})^2}\right]},$$

and that further $C(t)t^{2-\beta} = t^{-\beta}\mathbb{E}[\frac{e^{t|Z|}}{(1+e^{t|Z|})^2}]^{-1}$. To compute the limit when $t \to \infty$, we invoke Lemma A.2 and we conclude that

$$\lim_{t\to\infty} C(t)t^{2-\beta} = \lim_{t\to\infty} \frac{1}{t^\beta \mathbb{E}[\frac{e^{t|Z|}}{(1+e^{t|Z|})^2}]} = \frac{1}{c_Z \int_0^\infty \frac{z^{\beta-1}e^z}{(1+e^z)^2} dz}.$$

G.2 Proof of Proposition 3

Minimizer of the population loss. We can see identity (6) still holds with the calibration gap $h(t) = h_m(t) = \mathbb{E}[|Z|(1-\rho_m(t^\star|Z|))] - \mathbb{E}[\frac{|Z|}{1+e^{t|Z|}}]$ in Eq. (5) as $X - u^\star Z$ and $u^\star Z$ are independent. The function $h(t)$ is monotonically increasing in t with $h(\infty) = \mathbb{E}\left[|Z|(1-\rho_m(t^\star|Z|))\right] > 0$, while $1 - \rho_m(t|Z|) \leq 1 - \rho_1(t|Z|) = \frac{1}{1+e^{t|Z|}}$, we must have $h(t^\star) \leq 0$. Therefore there must be a unique zero point $t_m \geq t^\star$ of $h(t)$, and so $t_m u^\star$ is the unique minimizer of the population loss $L^{\mathrm{mv}}_m(\theta)$.

Asymptotic variance. As t_m solves $h_m(t_m) = 0$, Eq. (6) guarantees that $t_m u^\star$ is the global minimizer of the population loss L_m^{mv}. Appealing to Theorem 2, it follows that $\widehat{\theta}_{n,m}^{\text{mv}} \xrightarrow{p} t_m u^\star$, and

$$\sqrt{n}(\widehat{u}_{n,m}^{\text{mv}} - u^\star) \xrightarrow{d} \mathsf{N}\left(0, C_m(t^\star)\left(\mathsf{P}_{u^\star}\Sigma\mathsf{P}_{u^\star}\right)^\dagger\right)$$

for the variance function (13), which in this case simplifies to

$$C_m(t^\star) = \frac{\mathbb{E}[\frac{1}{(1+e^{t_m|Z|})^2}\rho_m(t^\star|Z|) + \frac{1}{(1+e^{-t_m|Z|})^2}(1-\rho_m(t^\star|Z|))]}{t_m^2\mathbb{E}[\frac{e^{t_m Z}}{(1+e^{t_m Z})^2}]^2}$$

via the symmetry $\rho_m(t) = \rho_m(-t)$.

Large m behavior. The remainder of the proof is to characterize the behavior of $C_m(t^\star)$ as $m \to \infty$. We first derive asymptotics for t_m. To simplify notation, we let $\|\theta\|_2 = t = t^\star$. Because t_m solves $h_m(t_m) = 0$ we have

$$\mathbb{E}\left[\frac{|Z|}{1+e^{t_m|Z|}}\right] = \mathbb{E}\left[|Z|(1-\rho_m(t|Z|))\right], \tag{32}$$

we must have $t_m \to \infty$ as $m \to \infty$ because $\rho_m(t) \to 1$ for any $t > 0$ as $m \to \infty$, so the right side of equality (32) converges to 0 by the dominated convergence theorem, and hence so must the left hand side. Invoking Lemma A.2 for the left hand side, it follows that

$$\lim_{m\to\infty} t_m^{\beta+1}\mathbb{E}\left[\frac{|Z|}{1+e^{t_m|Z|}}\right] = \lim_{m\to\infty} t_m^{\beta}\mathbb{E}\left[\frac{t_m|Z|}{1+e^{t_m|Z|}}\right] = c_Z\int_0^\infty \frac{z^\beta}{1+e^z}dz,$$

while invoking Lemma A.3 for the right hand side of (32), it follows that

$$\begin{aligned}\lim_{m\to\infty} m^{\frac{\beta+1}{2}}\mathbb{E}\left[|Z|(1-\rho_m(t|Z|))\right] &= \lim_{m\to\infty} m^{\frac{\beta}{2}}\mathbb{E}\left[\sqrt{m}|Z|(1-\rho_m(t|Z|))\right] \\ &= c_Z\int_0^\infty z^\beta\Phi\left(-\frac{tz}{2}\right)dz = c_Z t^{-\beta-1}\int_0^\infty z^\beta\Phi\left(-\frac{z}{2}\right)dz,\end{aligned}$$

where the last line follows from change of variables $tz \mapsto z$. The identity (32) implies that the ratio $\mathbb{E}[|Z|/(1+e^{t_m|Z|})]/\mathbb{E}[|Z|(1-\rho_m(t|Z|))] = 1$ and so we have that as $m \to \infty$,

$$\frac{t_m}{\sqrt{m}} = \left(\frac{t_m^{\beta+1}}{m^{\frac{\beta+1}{2}}}\right)^{\frac{1}{\beta+1}} = \left(\frac{t_m^{\beta+1}\mathbb{E}\left[\frac{|Z|}{1+e^{t_m|Z|}}\right]}{m^{\frac{\beta+1}{2}}\mathbb{E}\left[|Z|(1-\rho_m(t|Z|))\right]}\right)^{\frac{1}{\beta+1}} \to \left(\frac{\int_0^\infty \frac{z^\beta}{1+e^z}dz}{\int_0^\infty z^\beta\Phi\left(-\frac{z}{2}\right)dz}\right)^{\frac{1}{\beta+1}}\cdot t =: at.$$

In particular, $t_m/t^\star\sqrt{m} = a(1+o_m(1))$.

We finally proceed to compute asymptotic behavior of $C_m(t^\star)$, the variance (13). By Lemma A.2 the limit of its denominator as $t_m \to \infty$ satisfies

$$\lim_{m\to\infty} \underbrace{t_m^{2\beta}\mathbb{E}\left[\frac{e^{t_m Z}}{(1+e^{t_m Z})^2}\right]^2}_{:=\mathsf{den}(C_m(t))} = \lim_{t\to\infty}\left(t^\beta\mathbb{E}\left[\frac{e^{tZ}}{(1+e^{tZ})^2}\right]\right)^2 = \left(c_Z\int_0^\infty \frac{z^{\beta-1}e^z}{(1+e^z)^2}dz\right)^2.$$

We decompose the numerator into the two parts

$$m^{\frac{\beta}{2}}\mathbb{E}\left[\frac{1}{\left(1+e^{t_m|Z|}\right)^2}\rho_m(t|Z|) + \frac{1}{\left(1+e^{-t_m|Z|}\right)^2}(1-\rho_m(t|Z|))\right]$$

$$= \underbrace{m^{\frac{\beta}{2}}\mathbb{E}\left[\left(\frac{1}{\left(1+e^{-t_m|Z|}\right)^2}-\frac{1}{\left(1+e^{t_m|Z|}\right)^2}\right)(1-\rho_m(t|Z|))\right]}_{\text{(I)}} + \underbrace{m^{\frac{\beta}{2}}\mathbb{E}\left[\frac{1}{\left(1+e^{t_m|Z|}\right)^2}\right]}_{\text{(II)}}.$$

As we have already shown that $m^{-\frac{1}{2}}t_m \to at$, we know for any $\epsilon > 0$ that for large enough m, $(1-\epsilon)at\sqrt{m} \le t_m \le (1+\epsilon)at\sqrt{m}$. We can thus invoke Lemma A.2 to get

$$\lim_{m\to\infty}\text{(II)} = \lim_{m\to\infty} m^{\frac{\beta}{2}}\mathbb{E}\left[\frac{1}{\left(1+e^{\sqrt{m}at|Z|}\right)^2}\right] = c_Z\int_0^\infty \frac{z^{\beta-1}}{(1+e^{atz})^2}dz = c_Z t^{-\beta}\int_0^\infty \frac{z^{\beta-1}}{(1+e^{az})^2}dz.$$

With the same argument, we apply Lemma A.3 to establish the convergence

$$\begin{aligned}\lim_{m\to\infty}\text{(I)} &= c_Z\int_0^\infty z^{\beta-1}\left(\frac{1}{(1+e^{-atz})^2}-\frac{1}{(1+e^{atz})^2}\right)\Phi\left(-\frac{tz}{2}\right)dz \\ &= c_Z\int_0^\infty z^{\beta-1}\frac{e^{atz}-1}{e^{atz}+1}\Phi\left(-\frac{tz}{2}\right)dz = c_Z t^{-\beta}\int_0^\infty z^{\beta-1}\frac{e^{az}-1}{e^{az}+1}\Phi\left(-\frac{z}{2}\right)dz,\end{aligned}$$

where we use the change of variables $tz \mapsto z$. Taking limits, we have

$$\begin{aligned}\lim_{m\to\infty} m^{1-\frac{1}{2}\beta}C_m(t) &= \lim_{m\to\infty}\frac{m^{\frac{\beta}{2}}\mathbb{E}[\frac{1}{(1+e^{t_m|Z|})^2}\rho_m(t|Z|)+\frac{1}{(1+e^{-t_m|Z|})^2}(1-\rho_m(t|Z|))]}{m^{\beta-1}t_m^{2-2\beta}\cdot t_m^{2\beta}\mathbb{E}[\frac{e^{t_mZ}}{(1+e^{t_mZ})^2}]^2} \\ &= \lim_{m\to\infty}\left(\frac{t_m}{\sqrt{m}}\right)^{2\beta-2}\cdot\frac{\lim_{m\to\infty}\text{(I)}+\lim_{m\to\infty}\text{(II)}}{\mathsf{den}(C_m(t))} \\ &= (at)^{2\beta-2}\cdot\frac{c_Z t^{-\beta}\int_0^\infty z^{\beta-1}\left(\frac{1}{(1+e^{az})^2}+\frac{e^{az}-1}{e^{az}+1}\Phi\left(-\frac{z}{2}\right)\right)dz}{(c_Z\int_0^\infty\frac{z^{\beta-1}e^z}{(1+e^z)^2}dz)^2},\end{aligned}$$

where we used that $t_m/\sqrt{m} \to at$ as above.

G.3 Proof of Proposition 4

Minimizer of the population loss. By Lemma 4.1, we know the gap $h(t) = 0$ has a unique solution t_m, with $t_m u^\star$ minimizing the population loss $L(\theta, \sigma)$.

Asymptotic variance. Directly invoking Theorem 2 yields asymptotic normality:

$$\sqrt{n}\left(\widehat{u}_{n,m}^{\text{mv}} - u^\star\right) \xrightarrow{d} \mathsf{N}\left(0, C_m(t^\star)\left(\mathsf{P}_{u^\star}^{\perp}\Sigma\mathsf{P}_{u^\star}^{\perp}\right)^\dagger\right),$$

where the covariance function (13) has the form

$$C_m(t^\star) = \frac{1}{t_m^2}\frac{\mathbb{E}[\sigma(-t_m|Z|)^2\rho_m(t^\star Z)+\sigma(t_m|Z|)^2(1-\rho_m(t^\star Z))]}{\mathbb{E}[\sigma'(t_m Z)]^2} \tag{33}$$

and again t_m is the implicitly defined zero of $h(t) = 0$.

Large m behavior. We derive the large m asymptotics of t_m and C_m under Assumption A4 and using the shorthand $\|\theta^\star\|_2 = t$. The proof is essentially identical to that of Proposition 3 in Appendix G.2. First, recalling the probability (15), $\rho_m(t) = \mathbb{P}(\overline{Y} = \mathsf{sign}(\langle X, \theta^\star\rangle) \mid \langle X, \theta^\star\rangle = t)$, we see that $1 - \rho_m(tz) \to 0$ for any $z \neq 0$ as $m \to \infty$, and thus by dominated convergence, $\mathbb{E}\left[|Z|\left(1 - \rho_m(tZ)\right)\right] \to 0$. The analogue of the identity (32) in the proof of Proposition 3, that t_m is the zero of $h(t) = \mathbb{E}[\sigma(t|Z|)|Z|(1-\rho_m(t^\star Z))] - \mathbb{E}[\sigma(-t|Z|)|Z|\rho_m(t^\star Z)]$, implies

$$\mathbb{E}[\sigma(t_m|Z|)|Z|(1-\rho_m(t^\star Z))] = \mathbb{E}\left[\sigma(-t_m|Z|)|Z|\rho_m(t^\star Z)\right]. \tag{34}$$

As σ is bounded and $\mathbb{E}[\sigma(-t|Z|)|Z|\rho_m(t^\star|Z|)] \to \mathbb{E}[\sigma(-t|Z|)|Z|]$ for any t, the convergence of the left hand side of equality (34) to 0 as $m \to \infty$ means we must have $t_m \to \infty$. Invoking Lemma A.2 yields

$$\lim_{m\to\infty} t_m^{\beta+1}\mathbb{E}\left[|Z|\sigma(-t_m|Z|)\right] = \lim_{m\to\infty} t_m^{\beta}\mathbb{E}\left[t_m|Z|\sigma(-t_m|Z|)\right] = c_Z\int_0^\infty z^\beta\sigma(-z)dz.$$

Applying Lemma A.3 gives

$$m^{\frac{\beta+1}{2}}\mathbb{E}\left[|Z|(1-\rho_m(tZ))\right] = m^{\frac{\beta}{2}}\mathbb{E}\left[\sqrt{m}|Z|(1-\rho_m(tZ))\right] \to c_Z t^{-\beta-1}\int_0^\infty z^\beta\Phi\left(-2\overline{\sigma}^{\star\prime}(0)z\right)dz.$$

Rewriting the identity (34) using the symmetry of σ, so that $\sigma(t)+\sigma(-t) = 1$, we have the equivalent statement that $\mathbb{E}[|Z|(1-\rho_m(t^\star Z))] = \mathbb{E}[\sigma(-t_m|Z|)|Z|]$, or $\mathbb{E}[\sigma(-t_m|Z|)|Z|]/\mathbb{E}[|Z|(1-\rho_m(t^\star Z))] = 1$. Using this identity ratio, we find that

$$\frac{t_m}{\sqrt{m}} = \left(\frac{t_m^{\beta+1}\mathbb{E}\left[|Z|\sigma(-t_m|Z|)\right]}{m^{\frac{\beta+1}{2}}\mathbb{E}\left[|Z|(1-\rho_m(tZ))\right]}\right)^{\frac{1}{\beta+1}} \to \left(\frac{\int_0^\infty z^\beta\sigma(-z)dz}{\int_0^\infty z^\beta\Phi(-2\overline{\sigma}^{\star\prime}(0)z)dz}\right)^{\frac{1}{\beta+1}}\cdot t^\star =: at^\star.$$

This concludes the asymptotic characterization that $t_m = \sqrt{m}at^\star\cdot(1+o_m(1))$.

Finally, we turn to the asymptotics for $C_m(t^\star)$ in (33). By Lemma A.2 its denominator has limit

$$\lim_{m\to\infty}\underbrace{t_m^{2\beta}\mathbb{E}\left[\sigma'(t_mZ)\right]^2}_{:=\mathsf{den}C_m(t^\star)} = \lim_{t\to\infty}\left(t^\beta\mathbb{E}\left[\sigma'(tZ)\right]\right)^2 = \left(c_Z\int_0^\infty z^{\beta-1}\sigma'(z)dz\right)^2.$$

We decompose the (rescaled) numerator of the variance (33) into the two parts

$$\underbrace{m^{\frac{\beta}{2}}\mathbb{E}\left[\left(\sigma(t_m|Z|)^2-\sigma(-t_m|Z|)^2\right)\left(1-\rho_m(t|Z|)\right)\right]}_{\text{(I)}} + \underbrace{m^{\frac{\beta}{2}}\mathbb{E}\left[\sigma(-t_m|Z|)^2\right]}_{\text{(II)}}.$$

Lemmas A.2 and that $t_m = a\sqrt{m}t^\star(1+o_m(1)) \to \infty$, coupled with the dominated convergence theorem, establishes the convergence

$$\lim_{m\to\infty}\text{(II)} = \lim_{m\to\infty} m^{\frac{\beta}{2}}\mathbb{E}\left[\sigma(-\sqrt{m}at^\star|Z|)^2\right] = c_Z t^{-\beta}\int_0^\infty\frac{z^{\beta-1}}{\left(1+e^{az}\right)^2}dz.$$

Similarly, Lemma A.3 and that $t_m = a\sqrt{m}t^\star(1+o_m(1))$ gives that

$$\begin{aligned}\lim_{m\to\infty}\text{(I)} &= \lim_{m\to\infty} m^{\frac{\beta}{2}}\mathbb{E}\left[\left(\sigma(at^\star\sqrt{m}|Z|)^2-\sigma(-at^\star\sqrt{m}|Z|)^2\right)\left(1-\rho_m(tZ)\right)\right]\\ &= c_Z\int_0^\infty z^{\beta-1}\left(\sigma(at^\star z)^2-\sigma(-at^\star z)^2\right)\Phi\left(-2\overline{\sigma}^{\star\prime}(0)tz\right)dz\end{aligned}$$

$$= c_Z t^{\star-\beta} \int_0^\infty z^{\beta-1} \left(\sigma(az)^2 - \sigma(-az)^2\right) \Phi\left(-2\overline{\sigma}^{\star\prime}(0) z\right) dz,$$

where in the last line we use change of variables $tz \mapsto z$. Hence we have

$$\begin{aligned}\lim_{m\to\infty} m^{1-\frac{1}{2}\beta} C_m(t^\star) &= \lim_{m\to\infty} \frac{1}{m^{\beta-1} t_m^{2-2\beta}} \cdot \frac{\lim_{m\to\infty}(\mathrm{I}) + \lim_{m\to\infty}(\mathrm{II})}{\mathsf{den}(C_m(t^\star))} \\ &= \lim_{m\to\infty} \left(\frac{t_m}{\sqrt{m}}\right)^{2\beta-2} \cdot \frac{c_Z t^{\star-\beta} \int_0^\infty z^{\beta-1}(\sigma(az)^2 + (\sigma(az)^2 - \sigma(-az)^2)\Phi(-2\overline{\sigma}^{\star\prime}(0)z))dz}{(c_Z \int_0^\infty \frac{z^{\beta-1}e^z}{(1+e^z)^2} dz)^2}.\end{aligned}$$

Finally, as $t_m/\sqrt{m} = at^\star(1+o_m(1))$ we obtain that $m^{1-\beta/2} C_m(t^\star) = t^{\star\beta-2} b$ for some constant b depending only on $\beta, c_Z, \overline{\sigma}^{\star\prime}(0)$, and σ.

H Proofs for semiparametric approaches

H.1 Proof of Lemma 5.1

As σ and $\sigma^\star$ are L-Lipschitz, we may without loss of generality assume that $\mathsf{L} = 1$ and so $\left\|\sigma_j'\right\|_\infty \le 1$ and $\left\|\sigma_j^{\star\prime}\right\|_\infty \le 1$. We can compute the Hessian at any $\theta \in \mathbb{R}^d$ and $\vec{\sigma} = (\sigma_1, \cdots, \sigma_m)$,

$$\begin{aligned}\nabla^2 L(\theta, \vec{\sigma}) &= \mathbb{E}\left[\frac{1}{m}\sum_{j=1}^m \left(\sigma_j^\star(-\langle X, u^\star\rangle)\sigma_j'(\langle\theta, X\rangle) + \sigma_j^\star(\langle X, u^\star\rangle)\sigma_j'(-\langle\theta, X\rangle)\right) XX^\top\right] \\ &= \frac{1}{m}\sum_{j=1}^m \mathbb{E}\left[\sigma_j'(-Y\langle X, \theta\rangle) XX^\top\right],\end{aligned}$$

where in the last line we use that $\sigma_j^\star$ and σ_j are symmetric for $j = 1, \cdots, m$. Therefore we can upper bound the distance between Hessians by

$$\left\|\nabla^2 L(\theta, \vec{\sigma}) - \nabla^2 L(u^\star, \vec{\sigma}^\star)\right\| \le \frac{1}{m}\sum_{j=1}^m \underbrace{\left\|\mathbb{E}\left[\sigma_j'(-Y\langle X, \theta\rangle) XX^\top\right] - \mathbb{E}\left[\sigma_j^{\star\prime}(-Y\langle X, u^\star\rangle) XX^\top\right]\right\|}_{:=\delta_j},$$

and thus we only need to prove each quantity $\delta_j \to 0$ if $d_{\mathcal{F}_{\mathsf{link}}^{\mathsf{sp}}}((\theta, \vec{\sigma}), (u^\star, \vec{\sigma}^\star)) \to 0$, that is, if $\|\theta - u^\star\|_2 \to 0$ and $\left\|\sigma_j(-Y\langle X, u^\star\rangle) - \sigma_j^\star(-Y\langle X, u^\star\rangle)\right\|_{L^2(\mathbb{P})} \to 0$. In the following, we will show $\delta_j \to 0$ under the two different conditions. To further simplify the quantity, we claim it is sufficient to show $\xi_j \coloneqq \left\|\mathbb{E}[\sigma_j'(-Y\langle X, \theta\rangle) XX^\top] - \mathbb{E}[\sigma_j'(-Y\langle X, u^\star\rangle) XX^\top]\right\| \to 0$. Indeed, we have

Lemma H.1. *If $\xi_j \to 0$ and $\left\|\sigma_j(-Y\langle X, u^\star\rangle) - \sigma_j^\star(-Y\langle X, u^\star\rangle)\right\|_{L^2(\mathbb{P})} \to 0$, then $\delta_j \to 0$.*

Proof. By the triangle inequality and the independent decomposition $X = Zu^\star + W$, we have

$$\begin{aligned}\delta_j &\le \xi_j + \left\|\mathbb{E}\left[\sigma_j'(-Y\langle X, u^\star\rangle) XX^\top\right] - \mathbb{E}\left[\sigma_j^{\star\prime}(-Y\langle X, u^\star\rangle) XX^\top\right]\right\| \\ &= \xi_j + \left\|\mathbb{E}\left[(\sigma_j'(Z) - \sigma_j^{\star\prime}(Z)) Z^2\right] \cdot u^\star u^{\star\top}\right\| = \xi_j + \left|\mathbb{E}\left[(\sigma_j'(Z) - \sigma_j^{\star\prime}(Z)) Z^2\right]\right|.\end{aligned}$$

It remains to show $\mathbb{E}\left[(\sigma_j'(Z) - \sigma_j^{\star\prime}(Z))Z^2\right] \to 0$. Using the symmetry of σ_j and $\sigma_j^\star$, so $\sigma_j'(t) = \sigma_j'(-t)$, we can replace Z by $|Z|$. Then integrating by parts, for any $0 < \epsilon < M < \infty$ we have

$$\begin{aligned}\mathbb{E}\left[\sigma_j'(Z)Z^2\mathbb{1}\{\epsilon \le |Z| \le M\}\right] &= \int_\epsilon^M \sigma_j'(z)z^2p(z)dz \\ &= \sigma_j(M)M^2p(M) - \sigma_j(\epsilon)\epsilon^2p(\epsilon) - \int_\epsilon^M \sigma_j(z)(2zp(z) + z^2p'(z))dz.\end{aligned}$$

By our w.l.o.g. assumption that $\|\sigma'\|_\infty \le 1$, we have $|\mathbb{E}[\sigma_j'(Z)Z^2] - \mathbb{E}[\sigma_j'(Z)Z^2\mathbb{1}\{\epsilon \le |Z| \le M\}]| \le \mathbb{E}[Z^2\mathbb{1}\{|Z| < \epsilon \text{ or } |Z| > M\}]$. Thus, recognizing the trivial bound $\|\sigma_j\|_\infty \le 1$, we have

$$\begin{aligned}\left|\mathbb{E}\left[(\sigma_j'(Z) - \sigma_j^{\star\prime}(Z))Z^2\right]\right| \le 2\mathbb{E}[Z^2\mathbb{1}\{|Z| < \epsilon \text{ or } |Z| > M\}] + 2\left(\epsilon^2p(\epsilon) + M^2p(M)\right) + \\ + \underbrace{\left|\int_\epsilon^M \sigma_j(z)(2zp(z) + z^2p'(z)) - \int_\epsilon^M \sigma_j^\star(z)(2zp(z) + z^2p'(z))\right|}_{(\star)}.\end{aligned}$$

We show for any fixed $0 < \epsilon < M < \infty$, $(\star) \to 0$. Applying the Cauchy-Schwarz inequality twice, we have the bounds

$$\begin{aligned}\left|\int_\epsilon^M (\sigma_j(z) - \sigma_j^\star(z))zp(z)dz\right| &\le \left\|\sigma_j(Z) - \sigma_j^\star(Z)\right\|_{L^2(\mathbb{P})} \cdot \sqrt{\mathbb{E}[Z^2]} \to 0, \\ \left|\int_\epsilon^M (\sigma_j(z) - \sigma_j^\star(z))z^2p'(z)dz\right| &\le \left\|\sigma_j(Z) - \sigma_j^\star(Z)\right\|_{L^2(\mathbb{P})} \cdot \sqrt{\int_\epsilon^M z^4\left(\frac{p'(z)}{p(z)}\right)^2 p(z)dz} \\ &\le \left\|\sigma_j(Z) - \sigma_j^\star(Z)\right\|_{L^2(\mathbb{P})} \cdot \sup_{z\in[\epsilon,M]}\left|\frac{p'(z)}{p(z)}\right| \sqrt{\mathbb{E}[Z^4]} \to 0,\end{aligned}$$

where for the final inequality we use that $p(z)$ is nonzero and continuously differentiable.

As $(\star) \to 0$, we evidently have $\limsup |\mathbb{E}[(\sigma_j'(Z) - \sigma_j^{\star\prime}(Z))Z^2]| \le 2\mathbb{E}[Z^2(\mathbb{1}\{|Z| < \epsilon\} + \mathbb{1}\{|Z| > M\})] + 2(\epsilon^2p(\epsilon) + M^2p(M))$ for arbitrary $0 < \epsilon < M < \infty$. Using the assumptions that $\mathbb{E}[Z^2] \le \mathbb{E}[\|X\|_2^2] < \infty$ and $\lim_{z\to s} z^2p(z) = 0$ for $s \in \{0, \infty\}$, we conclude the proof by taking $\epsilon \to 0$ and $M \to \infty$. □

Finally we prove $\xi_j := \left\|\mathbb{E}[\sigma_j'(-Y\langle X, \theta\rangle)XX^\top] - \mathbb{E}[\sigma_j'(-Y\langle X, u^\star\rangle)XX^\top]\right\| \to 0$ under the two conditions in the statement of Lemma 5.1: that σ_j' are Lipschitz or X has a continuous density.

Condition 1. The links have Lipschitz derivatives We apply Jensen's inequality to write

$$\begin{aligned}\xi_j \le \mathbb{E}\left[|\sigma_j'(-Y\langle X, \theta\rangle) - \sigma_j'(-Y\langle X, u^\star\rangle)| \cdot \left\|XX^\top\right\|\right] &= \mathbb{E}\left[|\sigma_j'(-Y\langle X, \theta\rangle) - \sigma_j'(-Y\langle X, u^\star\rangle)| \cdot \|X\|_2^2\right] \\ &\le \mathsf{L}' \|\theta - u^\star\|_2 \cdot \mathbb{E}\left[\|X\|_2^3\right]\end{aligned}$$

by the L'-Lipschitz continuity of σ'. Taking $\theta \to u^\star$ completes the proof for this case.

Condition 2. The covariates have continuous density Let X have density $q(x)$. We rewrite the convergence $\theta \to u^\star$ instead as $\theta = Vu^\star$ where $V \to I_d$ is invertible. We again divide the expectation into large $\|X\|_2$ part and small $\|X\|_2$ part. Let $M < \infty$ be large enough that $\mathbb{E}[\|X\|_2^2\,\mathbb{1}\{\|X\|_2 > M\}] \le \epsilon$. Then using $\|\sigma_j'\|_\infty \le \mathsf{L}'$, we obtain

$$\xi_j \le \left\|\mathbb{E}\left[(\sigma_j'(-Y\langle X, \theta\rangle) - \sigma_j{}'(-Y\langle X, u^\star\rangle))XX^\top\right]\mathbb{1}\{\|X\|_2 \le M\}\right\| + 2\mathsf{L}'\mathbb{E}\left[\|X\|_2^2\,\mathbb{1}\{\|X\|_2 > M\}\right]$$

$$\leq \left\| \left(V^{-\top} \mathbb{E}\left[\sigma_j'(-\langle V^\top X, u^\star \rangle) V^\top X X V \right] V - \mathbb{E}\left[\sigma_j{}'(-\langle X, u^\star \rangle) X X^\top \right] \right) \mathbb{1}\{\|X\|_2 \leq M\} \right\| + 2\mathsf{L}'\epsilon.$$

By the linear transformation of variables $X' := V^\top X$, the first term in the above display is

$$\begin{aligned}
&\left\| V^{-\top} \mathbb{E}\left[\sigma_j'(-Y\langle V^\top X, u^\star \rangle) V^\top X X V \right] V - \mathbb{E}\left[\sigma_j{}'(-Y\langle X, u^\star \rangle) X X^\top \right] \right\| \\
&= \left\| \det(V^{-1}) \cdot V^{-\top} \left(\int_{\|V^\top x\|_2 \leq M} \sigma_j'(-x^\top u^\star) x x^\top p(V^{-\top} x) dx \right) V - \int_{\|x\|_2 \leq M} \sigma_j'(-x^\top u^\star) x x^\top p(x) dx \right\|.
\end{aligned}$$

Since $p(x)$ is absolutely continuous on any compact set, the above term converges to 0 as $V \to I_d$. Finally, as $\epsilon > 0$ was arbitrary, we take $M \to \infty$ to conclude that $\xi_j \to 0$.

H.2 Proof of Proposition 6

We apply Theorem 3. We first give the specialization of $C_{m,\vec{\sigma}^\star}$ that the assumptions of the proposition imply, recognizing that as ${\sigma^{\mathrm{lr}}}' = \sigma^{\mathrm{lr}}(1 - \sigma^{\mathrm{lr}})$, we have

$$C_{m,\vec{\sigma}^\star} = \frac{\sum_{j=1}^m \mathbb{E}[\sigma^{\mathrm{lr}}(\alpha_j^\star Z)(1 - \sigma^{\mathrm{lr}}(\alpha_j^\star Z))]}{(\sum_{j=1}^m \mathbb{E}[{\sigma^{\mathrm{lr}}}'(\alpha_j^\star Z)])^2} = \left(\sum_{j=1}^m \mathbb{E}[\sigma^{\mathrm{lr}}(\alpha_j^\star Z)(1 - \sigma^{\mathrm{lr}}(\alpha_j^\star Z))] \right)^{-1}.$$

We now argue that we can actually invoke Theorem 3, which requires verification of Assumption A5. Because for any $M < \infty$, the link functions $t \mapsto \sigma^{\mathrm{lr}}(\alpha t)$ have Lipschitz continuous derivatives for $|\alpha| \leq M$, so when $\vec{\sigma}_n$ has form $\vec{\sigma}_n = [\sigma^{\mathrm{lr}}(\alpha_{n,j} \cdot)]_{j=1}^m$, Lemma 5.1 implies the continuity of the mapping $(\theta, \vec{\sigma}) \mapsto \nabla_\theta^2 L(\theta, \vec{\sigma})$ for $d_{\mathcal{F}_{\mathsf{link}}^{\mathsf{sp}}}$ at $(u^\star, \vec{\sigma}^\star)$. Then recognizing that by Lipschitz continuity of σ^{lr} we have

$$\|\vec{\sigma}_n(Z) - \vec{\sigma}^\star(Z)\|_{L^2(\mathbb{P})} \leq \|\alpha_n - \alpha^\star\|_2^2 \, \|Z\|_{L^2(\mathbb{P})}^2 \xrightarrow{p} 0$$

whenever $\alpha_n \in \mathbb{R}^m$ satisfies $\alpha_n \xrightarrow{p} \alpha^\star$, we obtain the proposition.

I Proofs of nonparametric convergence results

In this technical appendix, we include proofs of the results from Section 5.2 as well as a few additional results, which are essentially corollaries of results on localized complexities and nonparametric regression models, though we require a few modifications because our setting is slightly non-standard.

I.1 Preliminary results

To set notation and to make reading it self-contained, we provide some definitions. The $L^r(P)$ norm of a function or random vector is $\|f\|_{L^r(P)} = (\int |f|^r dP)^{1/r}$, so that its $L^2(P_n)$-norm is $\|f\|_{L^2(\mathbb{P}_n)}^2 = \frac{1}{n}\sum_{i=1}^n f(X_i)^2$. We consider the following abstract nonparametric regression setting: we have a function class $\mathcal{F} \subset \{\mathbb{R} \to \mathbb{R}\}$ with $f^\star \in \mathcal{F}$, and our observations follow the model

$$Y_i = f^\star(X_i) + \xi_i, \tag{35}$$

but instead of observing (X_i, Y_i) pairs we observe $(\widetilde{X}_i, Y_i)$ pairs, where $\widetilde{X}_i$ may not be identical to X_i (these play the roll of $\langle u^\star, X_i \rangle$ versus $\langle u_n^{\mathrm{init}}, X_i \rangle$ in the results to come). We assume that ξ_i are bounded so that $\sup \xi - \inf \xi \leq 1$, independent, and satisfy the conditional mean-zero property that

$\mathbb{E}[\xi_i \mid X_i] = 0$ (though ξ_i may not be independent of X_i). For a (thus far unspecified) function class $\mathcal{F}$ we set

$$\widehat{f} = \operatorname*{argmin}_{f \in \mathcal{F}} P_n(Y - f(\widetilde{X}))^2.$$

We now demonstrate that the error $\|\widehat{f} - f^\star\|_{L^2(\mathbb{P}_n)}^2 = \frac{1}{n}\sum_{i=1}^n (\widehat{f}(X_i) - f^\star(X_i))^2$ can be bounded by a combination of the errors $\widetilde{X}_i - X_i$ and local complexities of the function class $\mathcal{F}$. Our starting point is a local complexity bound analagous to the localization results available for in-sample prediction error in nonparametric regression [cf. 39, Thm. 13.5]. To present the results, for a function class $\mathcal{H}$ we define the localized ξ-complexity

$$\mathcal{R}_n(u; \mathcal{H}) := \mathbb{E}\left[\sup_{h \in \mathcal{H}, \|h\|_{L^2(\mathbb{P}_n)} \le u} \left| n^{-1} \sum_{i=1}^n \xi_i h(x_i) \right| \right],$$

where we treat the expectation conditionally on X_i and ξ_i are random (i.e. $\xi_i = Y_i - f^\star(X_i)$). For the model (35), we define the centered class $\mathcal{F}^\star = \{f - f^\star \mid f \in \mathcal{F}\}$, which is star-shaped as $\mathcal{F}$ is a convex set.[1] We say that δ satisfies the *critical radius inequality* if

$$\frac{1}{\delta}\mathcal{R}_n(\delta; \mathcal{F}^\star) \le \delta. \tag{36}$$

With this, we can provide a proposition giving a high-probability bound on the in-sample prediction error of the empirical estimator $\widehat{f}$, which is essentially identical to [39, Thm. 13.5], though we require a few modifications to address that we observe $\widetilde{X}_i$ and not X_i and that the noise ξ_i are bounded but not Gaussian.

Proposition 7. *Let $\mathcal{F}$ be a convex function class, $\delta_n > 0$ satisfy the critical inequality (36), and let $\gamma^2 = \sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^n (f(X_i) - f(\widetilde{X}_i))^2$. Then for $t \ge \delta_n$,*

$$\mathbb{P}\left(\|\widehat{f} - f^\star\|_{L^2(\mathbb{P}_n)}^2 \ge 30 t\delta_n + 25\gamma \max\{\gamma, \|\xi\|_{L^2(\mathbb{P}_n)}\} \mid X_1^n, \widetilde{X}_1^n\right) \le \exp\left(-\frac{nt\delta_n}{4}\right).$$

See Section I.3 for a proof of the proposition.

Revisiting the critical radius inequality (36), we can also apply [39, Corollary 13.7], which allows us to use an entropy integral to guarantee bounds on the critical radius. Here, we again fix $X_1^n = x_1^n$, and for a function class $\mathcal{H}$ we let $\mathcal{B}_n(\delta; \mathcal{H}) = \{h \in \operatorname{Star}(\mathcal{H}) \mid \|h\|_{L^2(\mathbb{P}_n)} \le \delta\}$. Let $N_n(t; \mathcal{B})$ denote the t-covering number of $\mathcal{B}$ in $\|n\cdot\|_{L^2(\mathbb{P}_n)}$-norm. Then modifying a few numerical constants, we have

Corollary 5 (Wainwright [39], Corollary 13.7)**.** *Let the conditions of Proposition 7 hold. Then for a numerical constant $C \le 16$, any $\delta \in [0, 1]$ satisfying*

$$\frac{C}{\sqrt{n}} \int_{\delta^2/4}^{\delta} \sqrt{\log N_n(t; \mathcal{B}_n(\delta, \mathcal{F}^\star))}\, dt \le \frac{\delta^2}{2}$$

satisfies the critical inequality (36).

As an immediate consequence of this inequality, we have the following:

[1] A set $\mathcal{H}$ is *star-shaped* if for all $h \in \mathcal{H}$, if $\alpha \in [0, 1]$ then $\alpha h \in \mathcal{H}$.

Corollary 6. *Assume $|x_i| \le b$ for all $i \in [n]$ and that $\mathcal{F}$ is contained in the class of M-Lipschitz functions with $f(0) = 0$ (or any other fixed constant). Then for a numerical constant $c < \infty$, the choice*

$$\delta_n = c\left(\frac{Mb}{n}\right)^{1/3}$$

satisfies the critical inequality (36).

Proof. The covering numbers N_∞ for the class $\mathcal{F}$ of M-Lipschitz functions on $[-b, b]$ satisfying $f(0) = 0$ in supremum norm $\|f\|_\infty = \sup_{x\in[-b,b]} |f(x)|$ satisfy $\log N_\infty(t; \mathcal{F}) \lesssim \frac{Mb}{t}$ [cf. 39, Example 5.10]. Using that $N_n \le N_\infty$, we thus have

$$\int_{\delta^2/4}^{\delta} \sqrt{\log N_n(t; \mathcal{B}_n(\delta, \mathcal{F}^\star))} \lesssim \int_{\delta^2/4}^{\delta} \sqrt{\frac{Mb}{t}} dt = 2\sqrt{Mb}\left(\sqrt{\delta} - \delta/4\right) \le 2\sqrt{Mb\delta}$$

whenever $\delta \le 1$. For a numerical constant $c > 0$ it suffices in Corollary 5 to choose δ satisfying $c\frac{1}{\sqrt{n}}\sqrt{Mb\delta} \le \delta^2$, or $\delta = c(\frac{Mb}{n})^{1/3}$. □

I.2 Proof of Proposition 5

We assume without loss of generality that $m = 1$, as nothing in the proof changes except notationally (as we assume m is fixed). We apply Proposition 7 and Corollary 6. For notational simplicity, let Q denote the measure on $\mathbb{R}$ that $Y\langle u^\star, X\rangle$ induces for $X \sim P$, and Q_n similarly for P_n. We first show that $\|\sigma_n - \sigma_\star\|_{L^2(Q_n)}$ converges quickly. First, we recall [23, Lemma 3] that $\max_{i\le n} n^{-1/k} \|X_i\|_2 \stackrel{a.s.}{\to} 0$ as $n \to \infty$. Thus there is a (random) $B < \infty$ such that $\max_{i\le n} \|X_i\|_2 \le Bn^{1/k}$ for all n. Therefore, Corollary 6 implies that the choice $\delta_n = c(\frac{MB}{n^{1-1/k}})^{1/3}$ satisfies the critical inequality (36), and taking $\gamma_n^2 = \frac{M^2}{n}\sum_{i=1}^n \langle u_n^{\text{init}} - u^\star, X_i\rangle^2 \le M^2\epsilon_n^2 \|X\|^2_{L^2(\mathbb{P}_n)}$, we have that for $t \ge \delta_n$,

$$\|\sigma_n - \sigma_\star\|^2_{L^2(\mathbb{P}_n)} \lesssim t\delta_n + M^2\epsilon_n^2 \|X\|^2_{L^2(\mathbb{P}_n)} \quad \text{with probability at least } 1 - \exp\left(-\frac{nt\delta_n}{4}\right)$$

on the event that $\max_{i\le n} \|X_i\|_2 \le Bn^{1/k}$ for all n, where we have conditioned (in the probability) on X_i. As $\|X\|^2_{L^2(\mathbb{P}_n)} \stackrel{a.s.}{\to} \mathbb{E}[\|X\|_2^2]$, we may choose $t = \delta_n \gg 1/n^{1/3}$ and find that the Borell-Cantelli lemma then implies that with probability 1, there is a random $C < \infty$ such that

$$\|\sigma_n - \sigma^\star\|^2_{L^2(Q_n)} \le C\left(n^{\frac{2}{3k}-\frac{2}{3}} + \epsilon_n^2\right) \tag{37}$$

for all n with probability 1.

Finally we argue that $\|\sigma_n - \sigma^\star\|_{L^2(Q)} \stackrel{a.s.}{\to} 0$. Let $b < \infty$ be otherwise arbitrary, and let $\mathcal{G}_b$ be the collection of $2M$-Lipschitz functions on $[-b, b]$ with $\|g\|_\infty \le 1$ and $g(0)$ for $g \in \mathcal{G}_b$, noting that $\sigma_n - \sigma^\star$ restricted to $[-b, b]$ evidently belongs to $\mathcal{G}_b$. Then we have

$$\begin{aligned}
&\|\sigma_n - \sigma_\star\|^2_{L^2(Q)} \\
&= \|\sigma_n - \sigma_\star\|^2_{L^2(Q_n)} + \int_{|t|\le b} (\sigma_n(t) - \sigma_\star(t))^2(dQ(t) - dQ_n(t)) + \int_{|t|>b} (\sigma_n(t) - \sigma_\star(t))^2(dQ(t) - dQ_n(t)) \\
&\le \|\sigma_n - \sigma_\star\|^2_{L^2(Q_n)} + \sup_{g\in\mathcal{G}_b} |Qg^2 - Q_n g^2| + Q([-b, b]^c) + Q_n([-b, b]^c),
\end{aligned} \tag{38}$$

where the inequality used that $\|\sigma_n - \sigma_\star\|_\infty \le 1$ by construction. The first term in inequality (38) we have already controlled. We may control the second supremum term almost immediately using Dudley's entropy integral and a Rademacher contraction inequality. Indeed, we have

$$\mathbb{E}[\sup_{g\in\mathcal{G}_b} |Q_n g^2 - Qg^2|] \overset{(i)}{\le} 2\mathbb{E}[\sup_{g\in\mathcal{G}_b} |Q_n^0 g^2|] \overset{(ii)}{\le} 4\mathbb{E}[\sup_{g\in\mathcal{G}_b} |Q_n^0|] \overset{(iii)}{\lesssim} \frac{1}{\sqrt{n}} \int_0^\infty \sqrt{\log N_\infty(t; \mathcal{G}_b)} dt,$$

where inequality (i) is a standard symmetrization inequality, (ii) is the Rademacher contraction inequality [19, Ch. 4] applied to the function $t \mapsto t^2$, which is 2 Lipschitz for $t \in [-1, 1]$, and (iii) is Dudley's entropy integral bound. As the sup-norm log-covering numbers of M-Lipschitz functions on $[-b, b]$ scale as $\frac{Mb}{t}$ for $t \le Mb$ and are 0 otherwise, we obtain $\mathbb{E}[\sup_{g\in\mathcal{G}_b} |Q_n g^2 - Qg^2|] \lesssim \frac{Mb}{\sqrt{n}}$. The bounded-differences inequality then implies that for any $t > 0$,

$$\mathbb{P}\left(\sup_{g\in\mathcal{G}_b} |(Q_n - Q)g^2| \ge c\frac{Mb}{\sqrt{n}} + t\right) \le \mathbb{P}\left(\sup_{g\in\mathcal{G}_b} |(Q_n - Q)g^2| \ge \mathbb{E}[\sup_{g\in\mathcal{G}_b} |(Q_n - Q)g^2|] + t\right) \le \exp(-cnt^2).$$

Finally, the final term in the bound (38) evidently satisfies $Q([-b, b]^c) \le \frac{\mathbb{E}[\|X\|_2^k]}{b^k}$ and $\sup_b |Q_n([-b, b]^c) - Q([-b, b])| \le 2\sqrt{t/n}$ with probability at least $1 - e^{-2t^2}$ by the DKW inequality. We thus find by the Borel-Cantelli lemma that simultaneously for all $b < \infty$, with probability at least $1 - e^{-nt^2}$ we have

$$\|\sigma_n - \sigma_\star\|_{L^2(Q)}^2 \le \|\sigma_n - \sigma_\star\|_{L^2(Q_n)}^2 + \frac{CMb}{\sqrt{n}} + Ct + \frac{\mathbb{E}[\|X\|_2^k]}{b^k},$$

where C is a numerical constant.

Substituting inequality (37) into the preceding display and taking $b = n^{-\frac{1}{2(k+1)}}$, we get the result.

I.3 Proof of Proposition 7

We begin with an extension of the familiar basic inequality [e.g. 39, Eq. (13.18)], where we see by convexity that for any $\eta > 0$ we have

$$\begin{aligned}
\sum_{i=1}^n (Y_i - \widehat{f}(X_i))^2 &\le (1+\eta)\sum_{i=1}^n (Y_i - \widehat{f}(\widetilde{X}_i))^2 + (1 + 1/\eta)\sum_{i=1}^n (\widehat{f}(\widetilde{X}_i) - \widehat{f}(X_i))^2 \\
&\le (1+\eta)\sum_{i=1}^n (Y_i - f^\star(\widetilde{X}_i))^2 + (1 + 1/\eta)\sum_{i=1}^n (\widehat{f}(\widetilde{X}_i) - \widehat{f}(X_i))^2 \\
&\le (1+\eta)^2\sum_{i=1}^n (Y_i - f^\star(X_i))^2 + (2 + \eta + 1/\eta)\sum_{i=1}^n \left[(\widehat{f}(\widetilde{X}_i) - \widehat{f}(X_i))^2 + (f^\star(\widetilde{X}_i) - f^\star(X_i))^2\right].
\end{aligned}$$

Noting that $Y_i = f^\star(X_i) + \xi_i$, algebraic manipulations yield that if $\Delta = [f^\star(X_i) - \widehat{f}(X_i)]_{i=1}^n$ is the error vector, then

$$\|\Delta\|_{L^2(\mathbb{P}_n)}^2 - \frac{2}{n}\xi^T\Delta + \|\xi\|_{L^2(\mathbb{P}_n)}^2 \le (1+\eta)^2 \|\xi\|_{L^2(\mathbb{P}_n)}^2 + \frac{2 + \eta + 1/\eta}{n}\sum_{i=1}^n \left[(\widehat{f}(\widetilde{X}_i) - \widehat{f}(X_i))^2 + (f^\star(\widetilde{X}_i) - f^\star(X_i))^2\right].$$

Simplifying, we obtain the following:

Lemma I.1. *Let* $\gamma^2 = \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (f(X_i) - f(\widetilde{X}_i))^2$ *and* $\Delta = [\widehat{f}(X_i) - f^\star(X_i)]_{i=1}^n$. *Then*

$$\begin{aligned} \|\Delta\|_{L^2(\mathbb{P}_n)}^2 &\le \inf_\eta \left\{ (2\eta + \eta^2) \|\xi\|_{L^2(\mathbb{P}_n)}^2 + \frac{2}{n} \xi^T \Delta + (4 + 2\eta + 2/\eta)\gamma^2 \right\} \\ &\le \frac{2}{n} \xi^T \Delta + 11\gamma \max\left\{\gamma, \|\xi\|_{L^2(\mathbb{P}_n)}\right\}. \end{aligned}$$

Proof. The first inequality is algebraic manipulations and uses that our choice of η was arbitrary. For the second, we consider two cases: that $\|\xi\|_{L^2(\mathbb{P}_n)} \ge \gamma$ and that $\|\xi\|_{L^2(\mathbb{P}_n)} \le \gamma$. In the first case, we consider $\eta \le 1$, yielding the simplified bound that $\|\Delta\|_{L^2(\mathbb{P}_n)}^2 \le \frac{2}{n} \xi^T \Delta + 3\eta \|\xi\|_{L^2(\mathbb{P}_n)}^2 + \frac{8\gamma^2}{\eta}$ for $\eta \le 1$. Taking $\eta = \gamma / \|\xi\|_{L^2(\mathbb{P}_n)}$ gives that

$$\|\Delta\|_{L^2(\mathbb{P}_n)}^2 \le \frac{2}{n} \xi^T \Delta + 11 \|\xi\|_{L^2(\mathbb{P}_n)} \gamma.$$

In the case that $\gamma \ge \|\xi\|_{L^2(\mathbb{P}_n)}$, we choose $\eta = 1$, and the bound simplifies to $\|\Delta\|_{L^2(\mathbb{P}_n)}^2 \le \frac{2}{n} \xi^T \Delta + 3 \|\xi\|_{L^2(\mathbb{P}_n)}^2 + 8\gamma^2 \le \frac{2}{n} \xi^T \Delta + 11\gamma^2$. □

We now return to the proof of the proposition proper. We begin with an essentially immediate extension of the result [39, Lemma 13.12]. We let $\mathcal{H}$ be an arbitrary star-shaped function class. Define the event

$$A(u) := \left\{ \text{there exists } g \in \mathcal{H} \text{ s.t. } \|g\|_{L^2(\mathbb{P}_n)} \ge u \text{ and } |P_n \xi g| \ge 2 \|g\|_{L^2(\mathbb{P}_n)} u \right\}, \tag{39}$$

which we treat conditionally on $X_1^n = x_1^n$ as in the definition of the local complexity $\mathcal{R}_n$. (Here the noise ξ_i are still random, taken conditionally on $X_1^n = x_1^n$.)

Lemma I.2 (Modification of Lemma 13.12 of Wainwright [39])**.** *Let* $\mathcal{H}$ *be a start-shaped function class and let* $\delta_n > 0$ *satisfy the critical radius inequality*

$$\frac{1}{\delta} \mathcal{R}_n(\delta; \mathcal{H}) \le \delta.$$

Then for all $u \ge \delta_n$*, we have*

$$\mathbb{P}(A(u)) \le \exp\left(-\frac{nu^2}{4}\right).$$

Deferring the proof of Lemma I.2 to Section I.3.1, we can parallel the argument for [39, Thm. 13.5] to obtain our proposition.

Let $\mathcal{H} = \mathcal{F}^\star$ in Lemma I.2. Whenever $t \ge \delta_n$, we have $\mathbb{P}(A(\sqrt{t\delta_n})) \le e^{-nt\delta_n/4}$. We consider the two cases that $\|\Delta\|_{L^2(\mathbb{P}_n)} \lesseqgtr \sqrt{t\delta_n}$. In the former that $\|\Delta\|_{L^2(\mathbb{P}_n)} \le \sqrt{t\delta_n}$, we have nothing to do. In the latter, we have $\widehat{f} - f^\star \in \mathcal{F}^\star$ while $\|\Delta\|_{L^2(\mathbb{P}_n)} > \sqrt{t\delta_n}$, so that if $A(\sqrt{t\delta_n})$ fails then we must have

$$\left| \frac{1}{n} \sum_{i=1}^n \xi_i \Delta(x_i) \right| \le 2 \|\Delta\|_{L^2(\mathbb{P}_n)} \sqrt{t\delta_n}.$$

From the extension of the basic inequality in Lemma I.1 we see that

$$\|\Delta\|_{L^2(\mathbb{P}_n)}^2 \le 4 \|\Delta\|_{L^2(\mathbb{P}_n)} \sqrt{t\delta_n} + 11 \max\left\{\gamma^2, \gamma \|\xi\|_{L^2(\mathbb{P}_n)}\right\}.$$

Solving for $\|\Delta\|_{L^2(\mathbb{P}_n)}$ then yields

$$\|\Delta\|_{L^2(\mathbb{P}_n)} \le \frac{4\sqrt{t\delta_n} + \sqrt{16t\delta_n + 44\gamma \max\{\gamma, \|\xi\|_{L^2(\mathbb{P}_n)}\}}}{2} \le 4\sqrt{t\delta_n} + \sqrt{11\gamma \max\{\gamma, \|\xi\|_{L^2(\mathbb{P}_n)}\}}.$$

Simplifying gives Proposition 7.

I.3.1 Proof of Lemma I.2

Mimicking the proof of [39, Lemma 13.12], we begin with [39, Eq. (13.40)]:

$$\mathbb{P}(A(u)) \le \mathbb{P}(Z_n(u) \ge 2u^2) \quad \text{for} \quad Z_n(u) \coloneqq \sup_{g \in \mathcal{H}, \|g\|_{L^2(\mathbb{P}_n)} \le u} \left| n^{-1} \sum_{i=1}^n \xi_i g(x_i) \right|.$$

Now note that if $\|g\|_{L^2(\mathbb{P}_n)} \le u$, then the function $\xi \mapsto |n^{-1} \sum_{i=1}^n \xi_i g(x_i)|$ is $u/\sqrt{n}$-Lipschitz with respect to the ℓ_2-norm, so that convex concentration inequalities [e.g. 39, Theorem 3.4] imply that $\mathbb{P}(Z_n(u) \ge \mathbb{E}[Z_n(u)] + t) \le \exp(-\frac{t^2 n}{4b^2u^2})$ whenever $\sup \xi - \inf \xi \le b$, and so for $b = 1$ we have

$$\mathbb{P}(Z_n(u) \ge \mathbb{E}[Z_n(u)] + u^2) \le \exp\left(-\frac{nu^2}{4}\right).$$

As $\mathbb{E}[Z_n(u)] = \mathcal{R}_n(u)$, we finally use that the normalized complexity $t \mapsto \frac{\mathcal{R}_n(t)}{t}$ is non-decreasing [39, Lemma 13.6] to obtain that for $u \ge \delta_n$, $\frac{1}{u}\mathcal{R}_n(u) \le \frac{1}{\delta_n}\mathcal{R}_n(\delta_n) \le \delta_n$, the last inequality by assumption. In particular, we find that for $u \ge \delta_n$ we have $\mathbb{E}[Z_n(u)] = \mathcal{R}_n(u) \le u\delta_n \le u^2$, and so

$$\mathbb{P}(Z_n(u) \ge 2u^2) \le \mathbb{P}(Z_n(u) \ge \mathbb{E}[Z_n(u)] + u^2) \le \exp\left(-\frac{nu^2}{4}\right),$$

as desired.