Proceedings of Machine Learning Research vol 291:1–100, 2025 38th Annual Conference on Learning Theory

Learning sparse generalized linear models with binary outcomes via iterative hard thresholding

Namiko Matsumoto NMATSUMO@UCSD.EDU
Arya Mazumdar ARYA@UCSD.EDU
University of California, San Diego

Editors: Nika Haghtalab and Ankur Moitra

Abstract

In statistics, generalized linear models (GLMs) are widely used for modeling data and can expressively capture potential nonlinear dependence of the model's outcomes on its covariates. Within the broad family of GLMs, those with binary outcomes, which include logistic and probit regressions, are motivated by common tasks such as binary classification with (possibly) non-separable data. In addition, in modern machine learning and statistics, data is often high-dimensional yet has a low intrinsic dimension, making sparsity constraints in models another reasonable consideration. In this work, we propose to use and analyze an iterative hard thresholding (projected gradient descent on the ReLU loss) algorithm, called *binary iterative hard thresholding (BIHT)*, for parameter estimation in sparse GLMs with binary outcomes. We establish that BIHT is statistically efficient and converges to the correct solution for parameter estimation in a general class of sparse binary GLMs. Unlike many other methods for learning GLMs, including maximum likelihood estimation, generalized approximate message passing, and GLM-tron (Kakade et al., 2011; Bahmani et al., 2016), BIHT does not require knowledge of the GLM's link function, offering flexibility and generality in allowing the algorithm to learn arbitrary binary GLMs. As two applications, logistic and probit regression are additionally studied. In this regard, it is shown that in logistic regression, the algorithm is in fact statistically optimal in the sense that the order-wise sample complexity matches (up to logarithmic factors) the lower bound obtained previously. To the best of our knowledge, this is the first work achieving statistical optimality for logistic regression in all noise regimes with a computationally efficient algorithm. Moreover, for probit regression, our sample complexity is on the same order as that obtained for logistic regression.

1. Introduction

1.1. Generalized Linear Models

Generalized linear models (GLMs) are a popular statistical paradigm that has been extensively studied since their introduction by Nelder and Wedderburn (1972) as a generalization and unifying framework encompassing several common statistical models. In a GLM, each response (random) variable, $y \in \mathbb{R}$, has distribution with a parameter, $\boldsymbol{\theta}^* \in \Theta$, taken from a parameter space, $\Theta \subseteq \mathbb{R}^d$, and dependent on a covariate, $\mathbf{x} \in \mathbb{R}^d$, such that, $\mathbb{E}[y \mid \mathbf{x}] = g^{-1}(\langle \mathbf{x}, \boldsymbol{\theta}^* \rangle)$, where g is the *link function* that "links" the linear combination, $\langle \mathbf{x}, \boldsymbol{\theta}^* \rangle$, to the conditional expectation of the response, $y \mid \mathbf{x}$. This framework offers a flexible extension of the popular linear regression model, $\mathbb{E}[y \mid \mathbf{x}] = \langle \mathbf{x}, \boldsymbol{\theta}^* \rangle$—to allow for nonlinearities. The reader is referred to McCullagh (2019); Dobson and Barnett (2018); Fahrmeir and Tutz; Hardin and Hilbe (2007) for background on GLMs.

A fundamental problem in GLMs is parameter estimation—that is, the estimation of the parameter, $\boldsymbol{\theta}^* \in \Theta$, when given n i.i.d. samples, $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$, where the observed responses, $y_1, \ldots, y_n$, and known covariates, $\mathbf{x}_1, \ldots, \mathbf{x}_n$, are related in the manner stated earlier.

Maximum likelihood estimation (MLE) (Myung, 2003; Richards, 1961; Gallant and Nychka, 1987; Wald, 1949; Aitchison and Silvey, 1960) is a predominant approach to parameter estimation in GLMs (McCulloch, 1997; Hardin and Hilbe, 2007), where the estimates can be obtained through techniques such as iterative weighted least-squares methods (Nelder and Wedderburn, 1972; Firth, 1992; Hardin and Hilbe, 2007), the Newton-Raphson method (Jin et al., 2022; Hardin and Hilbe, 2007), and the Gauss-Newton method (Wedderburn, 1974). Gradient descent can also compute maximum likelihood estimates for parameter estimation in GLMs. One such line of work has studied gradient descent on the objective, $\mathcal{J}(\boldsymbol{\theta}) = \sum_i G(\langle \mathbf{x}_i, \boldsymbol{\theta} \rangle) - y_i \langle \mathbf{x}_i, \boldsymbol{\theta} \rangle$, where G is defined such that $g^{-1} = \partial G$. Kakade et al. (2011) propose a perceptron-like algorithm, *GLM-tron*, for learning GLMs by performing gradient descent on this loss function. Bahmani et al. (2016) study a similar gradient decent algorithm which incorporates (sparse) projections. When the outcomes, y_i, are i.i.d. and follow an exponential distribution, $\mathcal{J}$ is the negative log-likelihood function, and therefore, under these conditions, the minimizer of $\mathcal{J}$ is the maximum likelihood estimate. Note, however, that Kakade et al. (2011); Bahmani et al. (2016) consider more general classes of GLMs, meaning that their results extend beyond MLE in some cases. These works will be discussed further and compared to the contributions of this work in Appendix A. Of note, the *Sparsitron* algorithm of Klivans and Meka (2017), a multiplicative-weights update method, improves on the error of GLM-tron.

One consideration in learning GLMs is whether the link function is known. In contrast to many other works, e.g., Nelder and Wedderburn (1972); Kakade et al. (2011); Bahmani et al. (2016); Barbier et al. (2019), the algorithm studied in this work does not require knowledge of the specific link function. Some other works do consider learning a class of single-index models (SIMs) agnostically, without access to the link function, e.g. Gollakota et al. (2024) via omnipredictors, which are predictors that optimize over all loss functions in a collection. The setup and guarantees of this line of work is quite different than ours.

While the discussion thus far has not constrained the outcomes, y, GLMs with binary outcomes are popular for classification and particularly useful when data is non-separable. For brevity, this class of GLMs will be referred to as *binary GLMs* throughout this manuscript. They contain several important families of models, such as a subset of the exponential family that includes the ubiquitous logistic and probit regression models. This work is concerned with binary GLMs with a mild assumption on their link functions, which indeed holds for logistic and probit regression. Section 1.4 briefly surveys some relevant prior works on parameter estimation in binary GLMs.

In treating binary GLMs as classifiers, where $\boldsymbol{\theta}$ becomes a feature vector, the assumption of sparsity in a high-dimensional parameter space—or in this analogy the feature space—is commonplace in machine learning. Moreover, interpreting the parameter $\boldsymbol{\theta}$ as a signal or data vector, parameter estimation in GLMs can be framed as the inverse problem of signal reconstruction from noisy measurements. In this regard, requiring high-dimensional parameters—or signals—to be contained in a sparse subspace is again usual. In fact, this present work is motivated by the connection of binary GLMs to 1-bit compressed sensing, a topic within compressed sensing where the entries of the compressed signal representations are quantized to single bits: the $\pm$ signs of the unquantized values. Next, in Section 1.2, we briefly introduce 1-bit compressed sensing and explores this connection.

1.2. 1-Bit Compressed Sensing and Binary Iterative Hard Thresholding

The 1-bit compressed sensing problem Boufounos and Baraniuk (2008) seeks to recover an unknown vector $\boldsymbol{\theta}^*$ when given n responses, $y_i = \text{sign}(\langle \mathbf{x}_i, \boldsymbol{\theta}^* \rangle)$, where $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ are measurement vectors, and only $\pm$ signs of the linear measurements are being kept. Letting $\mathbf{X} = (\mathbf{x}_1 \ \cdots \ \mathbf{x}_n)^T$ and extending the notation of signs to vectors by applying it coordinate-wise, this representation can be written concisely as $\mathbf{y} = (y_1, \ldots, y_n) = \mathrm{sign}(\mathbf{X}\boldsymbol{\theta}^*)$. Typically, it is assumed that $n \ll d$, and thus, $\mathbf{y}$ is a compressed representation of the original signal vector, $\boldsymbol{\theta}^*$. The compressibility of $\boldsymbol{\theta}^*$ is often incorporated by a notion of sparsity. In this work, it is assumed that $\boldsymbol{\theta}^*$ is k-*sparse*—that is, the vector is supported on at most $k \leq d$ nonzero entries. Additionally, one of the most common choices of measurements is i.i.d. standard Gaussian random vectors, as is the design studied in this work. Most often, in 1-bit compressed sensing, the unknown vector, $\boldsymbol{\theta}^*$, is assumed to have unit norm since information about the norm is lost by the binarization of the responses. In relating 1-bit compressed sensing to binary classification, this noiseless model corresponds with classification of separable data. Alternatively, in connecting 1-bit compressed sensing to binary GLMs, the sign function can be replaced by a random function f, to incorporate noise, defined such that each i^{th} response takes the value 1 with probability $p(\langle \mathbf{x}_i, \boldsymbol{\theta}^* \rangle)$ and the value -1 with probability $1 - p(\langle \mathbf{x}_i, \boldsymbol{\theta}^* \rangle)$ for some function, $p : \mathbb{R} \to [0, 1]$. This setting, which can also be interpreted as binary classification of non-separable data, is studied in this work.

For reconstruction in 1-bit compressed sensing, Jacques et al. (2013b) propose the *binary iterative hard thresholding (BIHT)* algorithm, inspired by the existing *iterative hard thresholding* algorithm for compressed sensing Blumensath and Davies (2009). BIHT is a projected, (sub)gradient descent algorithm on the (negative) ReLU loss, given for $\boldsymbol{\theta} \in \mathbb{R}^d$ by

$$\mathcal{J}(\boldsymbol{\theta}) = \|[\mathrm{diag}(\mathbf{y})\mathbf{X}\boldsymbol{\theta}]_-\|_1, \tag{1}$$

where $\mathbf{y} \triangleq \mathrm{sign}(\mathbf{X}\boldsymbol{\theta}^*) \in \{-1, 1\}^n$ is the vector of binary responses and where $([\mathbf{v}]_-)_i = [v_i]_- = \min\{v_i, 0\}$, $i \in [n]$, for a vector $\mathbf{v} \in \mathbb{R}^n$. Jacques et al. (2013b) shows that

$$\mathbf{X}^T \ \frac{1}{2}(\mathrm{sign}\,(\mathbf{X}\boldsymbol{\theta}) - \mathbf{y}) \in \nabla_{\boldsymbol{\theta}} \mathcal{J}(\boldsymbol{\theta}), \tag{2}$$

leading to their BIHT algorithm. This present work studies the normalized variant of the BIHT algorithm of Jacques et al. (2013b), presented in Algorithm 1, which iteratively performs (i) first a (sub)gradient descent step on the objective function $\mathcal{J}$ in Equation (1), followed by (ii) a projection onto Θ, by performing the top-k thresholding operation and then normalizing. Note that in the dense regime, when $k = d$, the algorithm applies no thresholding in the second step but still normalizes the approximation.

For this noiseless response setting in 1-bit compressed sensing, Friedlander et al. (2021) shows that under the Gaussian design, the BIHT algorithm converges to the correct solutions with a suboptimal sample complexity. Subsequently, Matsumoto and Mazumdar (2024a) improves the sample complexity to the theoretically order-wise optimal (up to logarithmic factors) sample complexity: $\tilde{O}(\frac{k}{\epsilon}\log(\frac{d}{k})\sqrt{\log(\frac{1}{\epsilon})} + \frac{k}{\epsilon}\log^{3/2}(\frac{1}{\epsilon}))$, matching lower bound on the sample complexity for recovery in 1-bit compressed sensing established by Jacques et al. (2013b). One limitation of Friedlander et al. (2021); Matsumoto and Mazumdar (2024a) is the inability to immediately extend the convergence result to non-separable data and settings with noise. This was partially addressed in later work, in which Matsumoto and Mazumdar (2024b) shows robustness properties of BIHT when f incorporates adversarial noise: convergence of BIHT is possible with a similar sample complexity to that stated for the noiseless (or non-separable) setting. This begs the question: what convergence guarantees are possible in other models of noise or non-separable data? This work seeks an answer to this question for one such class of models: binary GLMs.

Algorithm 1: Normalized binary iterative hard thresholding (BIHT)

Given: $\mathbf{y}, \mathbf{X}$

$\hat{\boldsymbol{\theta}}^{(0)} \sim S^{d-1} \cap \Sigma_k^d$

for $t = 1, 2, 3, \ldots$ **do**

$\quad \tilde{\boldsymbol{\theta}}^{(t)} \leftarrow \hat{\boldsymbol{\theta}}^{(t-1)} + \frac{\sqrt{2\pi}}{n} \mathbf{X}^T \frac{1}{2} \left(\mathbf{y} - \mathrm{sign}\big(\mathbf{X}\hat{\boldsymbol{\theta}}^{(t-1)}\big) \right)$

$\quad \hat{\boldsymbol{\theta}}^{(t)} \leftarrow \frac{T_k(\tilde{\boldsymbol{\theta}}^{(t)})}{\|T_k(\tilde{\boldsymbol{\theta}}^{(t)})\|_2}$

end

1.3. Main Contributions

One realization that leads to this work is that the BIHT algorithm (Algorithm 1) can be applicable to any possibly randomized 1-bit quantization function, f, without any change, and not restricted to the sign function. Therefore, in particular, Algorithm 1 is a perfect candidate for efficient parameter estimation in binary GLMs, and can be advantageous over some existing estimation methods as it is oblivious of the link function.

This work establishes that, when applied to binary GLMs where the link function satisfies a reasonable property (i.e., Assumption 2), the BIHT algorithm iteratively produces approximations that converge to the ϵ-ball around the true parameter, $\boldsymbol{\theta}^*$, with high probability under the Gaussian covariates design with a sample complexity of

$$n = \tilde{O}\left(\max\left[\frac{\max(\alpha, \epsilon)}{\gamma^2 \epsilon^2} k \log\left(\frac{d}{\epsilon k}\right), \frac{k}{\epsilon} \log^{3/2}\left(\frac{1}{\epsilon}\right) \right] \right),$$

where the big-O term hides some less-significant factors for the sake of readability, and α and γ are properties of the link function defined in Equations (8) and (9) which can be explicitly computed for standard GLMs (leading to tighter sample complexity $\tilde{O}(k/\epsilon)$ in some regimes - equivalent to *optimistic rate* in learning theory). In addition the estimation error decays at an exponential rate with respect to the number of iterations, $t \in \mathbb{Z}_{\geq 0}$, of the algorithm, bounded from above by

$$\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}^{(t)}\|_2 = O(\epsilon^{1-2^{-t}}),$$

and hence, asymptotically approaches the error-rate of ϵ as $t \to \infty$.

When specialized to two of the most popular binary GLMs, logistic and probit regressions, convergence of BIHT can be shown and explicit expressions can derived for the sample complexities (via closed-form bounds on α and γ). For logistic and probit regressions these lead to the optimal scaling with the "signal-to-noise ratio" SNR, β, and error-rate, ϵ, as β is varied. Notably, as $\beta \to \infty$, the dependence of the sample complexity on the error-rate, reduces to $\frac{1}{\epsilon}$. Due to lower bounds of Hsu and Mazumdar (2024), the resultant sample complexity for logistic regression is optimal up to logarithmic factors in all regimes. As discussed in Hsu and Mazumdar (2024), for logistic regression with Gaussian covariates, computationally efficient algorithms with optimal sample complexity were known only for the high noise regime, e.g., Plan et al. (2017), and the noiseless regime Matsumoto and Mazumdar (2024a), while no such optimal algorithms were known for the intermediate regimes[1]. Thus, to the best of our knowledge, BIHT is the first computationally efficient algorithm with the optimal sample complexity in all noise regimes for logistic regression under the Gaussian covariate design.

1. However, for $k = d$, MLE achieves optimal samples for low noise case, by recent work of Chardon et al. (2024).

1.4. Comparison to Prior Work

A generalized version of the popular "LASSO" algorithm has been proposed for GLMs in Plan and Vershynin (2016). Our results for BIHT are analogous to their results for parameter estimation with LASSO; in particular their sample complexity is $\tilde{O}(\frac{k\log(\frac{d}{k})}{\epsilon^2})$; and their results also depends on properties of the link function, namely a scaling factor and noise variance, the former being same as the quantity γ as defined in Equation (9). Subsequently, a very precise error-analysis for generalized LASSO has been performed in Thrampoulidis et al. (2015). To compare with Theorem 5, their sample complexity is $\tilde{O}(\frac{(1-\gamma^2)k\log(\frac{d}{k})}{\epsilon^2})$ (see, (Thrampoulidis et al., 2015, Equation (8))); however, due to the differences in assumption and applicability as explained above, one should exercise caution in such comparisons.

A few other prior works are worth remarking on. In the following discussion, the parameter, $\boldsymbol{\theta}^*$, is assumed to have unit norm, but the models incorporate SNR denoted by $\beta > 0$. Formulating the (sparse) estimation problem as a convex program, Plan and Vershynin (2012) shows that the estimation of $\boldsymbol{\theta}^*$ from binary responses is possible with $\tilde{O}(\frac{k\log(\frac{d}{k})}{\min\{\beta^2,1\}\epsilon^4})$ samples under the Gaussian covariate design. Plan et al. (2017) improves this sample complexity to $\tilde{O}(\frac{k\log(\frac{d}{k})}{\min\{\beta^2,1\}\epsilon^2})$ using a method that effectively amounts to the "Average" algorithm of Servedio (1999).

Subsequently, in the case of logistic regression with Gaussian covariates, maximum likelihood estimators and their regularized versions have recently received renewed attention. In this regard, Sur and Candès (2019) and Salehi et al. (2019) are notable; however their precise asymptotic results are given in terms of solutions of a system of equations, and are not directly comparable to our sample complexity results. On the other hand, recently Hsu and Mazumdar (2024) established a lower bound on sample complexity for this case via a variant of Fano's inequality that is order-wise tight (up to logarithmic factors) when $\beta \leq 1$. Hsu and Mazumdar (2024) additionally obtains order-wise tight (again, up to logarithmic factors) bounds on the sample complexity in logistic regression with the dense parameter space (when $k = d$) for any β, summarized in Equation (17). Very recently, Chardon et al. (2024) show that for the $k = d$ case MLE achieves optimal sample complexity for $\beta = \Omega(1)$. While Plan and Vershynin (2012); Plan et al. (2017) pair their sample complexity bounds with efficient algorithms for the Gaussian design, polynomial time algorithms achieving the optimal sample complexity for logistic regression in the $\beta > 1$ regimes were not known in general. This work settles this question by proving that BIHT is a computationally efficient algorithm that in fact simultaneously achieves the order-wise optimal sample complexity (up to logarithmic factors) for all choices of β even with the sparsity constraint. Here, it is worth remarking that, although we are not aware of a lower bound on the sample complexity for probit regression in the literature, our result for probit model shares the same order sample complexity (which we believe to be tight) as our result for logistic regression. It should be noted that, for the probit model in the the non-sparse $k = d$ case, when restricted to $\beta > 1$, the same sample complexity (up to log factors) is also achieved by Kuchelmeister and van de Geer (2024).

1.4.1. COMPARISON TO MATSUMOTO AND MAZUMDAR (2024A,B)

Although numerous prior works have studied BIHT, e.g., Friedlander et al. (2021); Jacques et al. (2013a,b); Liu et al. (2019); Plan et al. (2017), the works most closely aligned with the analysis in this manuscript are Matsumoto and Mazumdar (2024a,b), and indeed, some elements of the approach in this work are analogous to components of the analyses in Matsumoto and Mazumdar (2024a,b). However, handling the GLM's randomness—introduced into the model through the function f—requires a novel approach. It turns out that the normalization step in each iteration of BIHT is crucial to obtain our bound on the approximation error, which distinguishes this work from Matsumoto and Mazumdar (2024a,b), even despite the fact that Matsumoto and Mazumdar (2024b) considers an alternative (adversarially) noisy setting. As discussed in Section 2, the analysis in this work largely centers around an invertibility condition that uniformly bounds an expression of the form

$$\left\| \boldsymbol{\theta}^* - \frac{\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})}{\|\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})\|_2} \right\|_2 \tag{3}$$

for all $\hat{\boldsymbol{\theta}} \in \Theta$ and all $J \subseteq [d]$, $|J| \leq k$. In contrast, Matsumoto and Mazumdar (2024a,b) consider invertibility conditions that bounds expressions of the respective forms

$$\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}} - h_{\mathrm{sign};J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})\|_2, \qquad \|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}} - h_{f_{\mathrm{adv}};J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})\|_2, \tag{4}$$

where the parameterizations by the sign and f_{adv} functions can be thought of as the noiseless and adversarially noisy analogs, respectively, to f in our setting for GLMs. Notice that both expressions in (4) omit the sort of normalization that appears in (3). But following Matsumoto and Mazumdar (2024a,b) in this way turns out to be problematic when applying the analysis to GLMs: in "lower" SNR regimes, it would effectively lead to an $\Omega(1)$ additive term in the bound, meaning that the bound on the error-rate for BIHT would become $\Omega(1)$, rather than ϵ, regardless of the number of covariates (i.e., the sample complexity). However, accounting for the normalization mitigates this issue to give the desired ϵ-error-rate.

The intuition behind this is the following. In expectation, the vector $\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})$ is aligned with $\boldsymbol{\theta}^*$, as is the noise introduced by the GLM's randomness. Therefore, the normalization essentially eliminates (or at least reduces) this noise, leading to the desired error-rate. Note that, on the other hand, if the noise was instead adversarial, as in Matsumoto and Mazumdar (2024b), it is unlikely that accounting for such normalization in this way would help as the noise can be (adversarially) chosen to be in more or less any direction.

In fact, an empirical study with logistic regression (see, Figure 1) corroborates these observations and suggests that BIHT may exhibit different convergence and stability behaviors when it does or does not normalize its approximations, at least at "higher" noise levels, which is a notable distinction from observations made in the "noiseless" setting, where BIHT has been empirically seen to converge well (and potentially even less brittlely) when the algorithm's approximations are not normalized (see, Figure 2). Similar empirical behavior is exhibited with probit regression, as well, but such empirical results have been omitted to avoid redundancy.

All in all, this key observation and distinguishing approach are essential for our analysis of and convergence guarantees for learning GLMs with BIHT, and may potentially even be less an artifact of our analysis and more an inherent feature of the algorithm itself.

Organization In Section 2 we give an overview of how our results are obtained. In Section 3, we provide the notations used throughout the paper, and formally define binary GLMs and related quantities and assumptions. In Section 4, the main theorems of this work, which establish the convergence of BIHT to the correct solution for parameter estimation in binary GLMs, are formally stated. Section 4.1 outlines the key steps in the proof of the main results. Section 5 presents the main technical theorem: the invertibility condition for Gaussian covariates. In Appendix A, we further compare our techniques with existing literature, notably with Bahmani et al. (2016) and Kakade et al. (2011). The main theorems are formally proved in Appendix B, while formal proofs of the main technical theorems are in Appendix C. Lastly, Appendix D derives some concentration inequalities that are needed in the proof of the main technical theorems.

2. Overview of Techniques

The proof of the convergence of the BIHT approximations for any GLM satisfying Assumption 2 adapts the following approach. It consists of two primary bounds on the approximation error: (a) a deterministic bound that relates the approximation error to an *invertibility condition* satisfied by Gaussian matrices, and (b) a probabilistic bound that describes this invertibility condition for Gaussian matrices. The former of these bounds, (a), is relatively straightforward to establish using standard techniques, including bounding of a recurrence relation. On the other hand, the derivation of the latter bound, (b), which is the primary technical contribution of this work, entails extensive analysis that constitutes the majority of this manuscript. The establishment of the invertibility condition for Gaussian covariate matrices in regard to this latter bound follows a similar approach to Matsumoto and Mazumdar (2024a) (i.e., the noiseless case) to prove an analogous invertibility condition for Gaussian matrices therein, though there are some major technical differences, which are highlighted in Section 1.4.1.

The intuition behind the argument for the probabilistic bound, (b), is as follows. The aforementioned invertibility condition (see Theorem 9) upper bounds

$$\left\| \boldsymbol{\theta}^* - \frac{\mathbf{w}_{\hat{\boldsymbol{\theta}}}}{\|\mathbf{w}_{\hat{\boldsymbol{\theta}}}\|_2} \right\|_2, \tag{5}$$

where $\mathbf{w}_{\hat{\boldsymbol{\theta}}} \triangleq \hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})$, uniformly for every $\hat{\boldsymbol{\theta}} \in \Theta$ and every $J \subseteq [d]$, $|J| \leq k$, and for a particular random function, $h_{f;J} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$, parameterized by the coordinate subset, J, and dependent on the covariate matrix, $\mathbf{X}$. The specification of the function, $h_{f;J}$, which is determined by the GLM function f, is deferred to Section 4.1 as this informal overview can be understood without its formal definition. One can view an upper bound on the quantity of Equation (5) to be a *single-step progress* towards estimating $\boldsymbol{\theta}^*$ via the BIHT algorithm. One salient characteristic of the invertibility condition is that stronger guarantees are provided for points, $\hat{\boldsymbol{\theta}}$, which are closer to $\boldsymbol{\theta}^*$. The invertibility condition will be proved to hold for Gaussian covariate matrices with high probability. Towards this, the quantity in Equation (5) will be shown to describe a notion of deviation of the random vector $\mathbf{w}_{\hat{\boldsymbol{\theta}}}$ around its mean in the sense that $\left\| \boldsymbol{\theta}^* - \frac{\mathbf{w}_{\hat{\boldsymbol{\theta}}}}{\|\mathbf{w}_{\hat{\boldsymbol{\theta}}}\|_2} \right\|_2 = \left\| \frac{\mathbf{w}_{\hat{\boldsymbol{\theta}}}}{\|\mathbf{w}_{\hat{\boldsymbol{\theta}}}\|_2} - \frac{\mathbb{E}[\mathbf{w}_{\hat{\boldsymbol{\theta}}}]}{\|\mathbb{E}[\mathbf{w}_{\hat{\boldsymbol{\theta}}}]\|_2} \right\|_2$. Furthermore, it can be shown that this deviation is roughly proportional to the deviation of the random function $h_{f;J}$ around its mean:

$$\left\| \frac{\mathbf{w}_{\hat{\boldsymbol{\theta}}}}{\|\mathbf{w}_{\hat{\boldsymbol{\theta}}}\|_2} - \frac{\mathbb{E}[\mathbf{w}_{\hat{\boldsymbol{\theta}}}]}{\|\mathbb{E}[\mathbf{w}_{\hat{\boldsymbol{\theta}}}]\|_2} \right\|_2 \propto \|h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) - \mathbb{E}[h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})]\|_2. \tag{6}$$

The deviation of $h_{f;J}$ is then decomposed into (and upper bounded by) three components of deviation: (i) a global component for each point in a cover over the parameter space, Θ, that is sufficiently far from $\boldsymbol{\theta}^*$; (ii) a local component for each point in the cover and every point in a small region surrounding it; and (iii) a component which arises from the randomness of the GLM—specifically, through the function f. Upper bounds on the first and third components, (i) and (iii), can be established following Gaussian concentration. Meanwhile, the second component, (ii), relies on the local binary embeddings of Oymak and Recht (2015).

Notably, the two components, (i) and (ii), do not depend on the random function f: they replace the random function f with the deterministic sign function. Thus, the third component, (iii), entirely captures the deviation associated with the randomness induced by f. The upper bounding of all three components exploits the statistical "niceness" of the Gaussian covariates. In particular, for the first component, (i), the angular uniformity of i.i.d. Gaussian random vectors is crucial because it controls the number of covariates involved in the computation of the random function, $h_{f;J}$. However, this breaks down when points are too close together as the number of samples involved in the computation of $h_{f;J}$ cannot be guaranteed to further decrease beyond a certain threshold (a distance on the order of ϵ). Once this occurs, the local guarantees provided by the local binary embeddings of Oymak and Recht (2015) take over through (ii) to ensure that the randomness of $h_{f;J}$ is well-controlled to provide sufficient guarantees within these local regions. The (sub)gaussianity of the covariates is also critical for bounding the third component of deviation, (iii), though not strictly through angular uniformity. In effect, it limits the amount of deviation that can be introduced by the randomness of the GLM, which comes from standard knowledge.

Taking a step back to examine the relationship between this invertibility condition and the convergence of BIHT, there are a couple key ideas underlying the behavior exhibited by the algorithm. First, the last two components of deviation, (ii) and (iii), introduce error to the algorithm's approximations which is more or less "baked in" once the model and its covariates are fixed—that is, these contributions to the approximation error will not decay as the algorithm continues to iterate. In contrast, the first component of deviation, (i), contributes error into the approximation which indeed decays. Simply put, this is the consequence of the invertibility condition imposing a stronger bound for points which are closer to the true parameter, $\boldsymbol{\theta}^*$. In essence, since the number of covariates—and hence also the variance—involved in the evaluation of $h_{f;J}$ at a pair of points decreases as the distance between the points decreases, the improvement of the approximation in one iteration of BIHT leads to even better control over $h_{f;J}$, and thus a better approximation, in the next iteration. However, because the invertibility condition only guarantees this improvement up to but not within small, local regions, the error cannot be guaranteed to reduce once the approximations reach the ϵ-ball around $\boldsymbol{\theta}^*$. At the same time, the error will remain within a small threshold due to the local result, i.e., (ii).

3. Notations and Key Properties of GLMs

For a set of real numbers, $\mathcal{S} \subseteq \mathbb{R}$, let $\mathcal{S}_{\geq 0}, \mathcal{S}_+ \subseteq \mathcal{S}$ denote the sets of nonnegative elements and, respectively, positive elements in $\mathcal{S}$—formally, $\mathcal{S}_{\geq 0} \triangleq \{s \in \mathcal{S} : s \geq 0\}$ and $\mathcal{S}_+ \triangleq \{s \in \mathcal{S} : s > 0\}$. For $t \in \mathbb{Z}_+$, let $[t] \triangleq \{1, \dots, t\}$. Vectors and matrices are denoted in lowercase and uppercase bold typeface, respectively, e.g., $\mathbf{v} \in \mathbb{R}^n$ and $\mathbf{M} \in \mathbb{R}^{m \times n}$, with their entries in italic font such that, e.g., $\mathbf{v} = (v_j)_{j \in [d]}$ and $\mathbf{M} = (M_{i,j})_{(i,j) \in [n] \times [d]}$. The all-zeros vector is written in boldface: $\mathbf{0} = (0, \dots, 0)$. For a coordinate subset $J \subseteq [d]$, the vector with entries indexed by J taking value 1 and with all other entries set to 0 is written as $\mathbf{1}^J \in \mathbb{R}^d$. The set of all s-sparse, real-valued vectors is denoted by $\Sigma_s^d \triangleq \{\mathbf{v} \in \mathbb{R}^d : \|\mathbf{v}\|_0 \leq s\}$. For $r > 0$ and $\mathbf{v} \in \mathbb{R}^d$, specify the radius-$r$ ball around $\mathbf{v}$ by $\mathcal{B}_r(\mathbf{v}) \triangleq \{\mathbf{u} \in \mathbb{R}^d : \|\mathbf{u} - \mathbf{v}\|_2 \leq r\}$, and let $\mathcal{B}'_r(\mathbf{v}) \triangleq \mathcal{B}_r(\mathbf{v}) \cap \mathcal{S}^{d-1} \cap \{\mathbf{u} \in \mathbb{R}^d : \operatorname{supp}(\mathbf{u}) = \operatorname{supp}(\mathbf{v})\}$. Let $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{U} \subseteq \mathcal{X}$ be sets, and let $f : \mathcal{X} \to \mathcal{Y}$. The image of $\mathcal{U}$ under f

is denoted by $f[\mathcal{U}] \subseteq \mathcal{Y}$. The natural logarithm is denoted by $\log : \mathbb{R} \to \mathbb{R}$. The indicator function, $\mathbb{I}$, is given for a true/false condition, C, by $\mathbb{I}(C) = 0$ if C is false and $\mathbb{I}(C) = 1$ if C is true, where this notation extends to vectors by applying it entry-wise. The sign function, $\mathrm{sign} : \mathbb{R} \to \{-1, 1\}$, simply returns the $\pm$ sign of the input, i.e., $\mathrm{sign}(a) = +1$ if and only if $a \geq 0$, for $a \in \mathbb{R}$ with this notation extending to vectors by applying it entry-wise. For $s \in \mathbb{Z}_+$, *the top-s thresholding operation*, written $T_s : \mathbb{R}^d \to \mathbb{R}^d$, maps $\mathbf{v} \mapsto T_s(\mathbf{v})$, where $T_s(\mathbf{v})$ retains the largest magnitude entries in $\mathbf{v}$ and sets all other entries to 0 with ties broken arbitrarily. Similarly, for a set $J \subseteq [d]$, define the *subset thresholding operation*, denoted by $T_J : \mathbb{R}^d \to \mathbb{R}^d$, to be the map which takes a vector $\mathbf{v} \in \mathbb{R}^d$ to a vector with j^{th} entries $T_J(\mathbf{v})_j = v_j \mathbb{I}(j \in J)$, $j \in [d]$. Note that the latter thresholding operation is a linear transformation given by $T_J(\mathbf{v}) = \mathrm{diag}(\mathbf{1}^J)\mathbf{v}$ for $\mathbf{v} \in \mathbb{R}^d$.

Denote by $X \sim \mathcal{D}$ a random variable X, which follows a distribution $\mathcal{D}$. If $\mathcal{S}$ is a set then $X \sim \mathcal{S}$ means X follows the uniform distribution over $\mathcal{S}$. Additionally, the density function and moment generating function (mgf) (when well-defined) of a random variable, X, are written f_X and ψ_X, respectively.

Binary GLMs Throughout this manuscript, $d, k, n \in \mathbb{Z}_+$ denote, in order, the dimension of the parameter space, the sparsity, and the number of samples (or measurements), and the error-rate is denoted by $\epsilon \in (0, 1)$. The parameter space is written as $\Theta = S^{d-1} \cap \Sigma_k^d \subseteq \mathbb{R}^d$. Note that the results in this manuscript extend to the dense regime by taking $k = d$ and $\Theta = S^{d-1}$. The covariates are d-variate i.i.d. standard Gaussian random vectors, written as $\mathbf{x}_1, \dots, \mathbf{x}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, which are stacked up into the covariate matrix, $\mathbf{X} = (\mathbf{x}_1 \cdots \mathbf{x}_n)^T \in \mathbb{R}^{n \times d}$. The unknown parameter vector which is being estimated is denoted by $\boldsymbol{\theta}^* \in \Theta$, and the t^{th} approximations produced by the t^{th} iteration of the recovery algorithms are written as $\hat{\boldsymbol{\theta}}^{(t)} \in \Theta$, where $t \in \mathbb{Z}_{\geq 0}$. There is assumed access to the covariates, $\mathbf{x}_i$, $i \in [n]$, as well as n binary measurement responses specified as the vector $\mathbf{y} \in \{-1, 1\}^n$.

For a function, $p : \mathbb{R} \to [0, 1]$, the i^{th} measurement responses, $y_i \in \{-1, 1\}$, $i \in [n]$, are obtained through a random function $f : \mathbb{R} \to \{-1, 1\}$, given by

$$f(z) = \begin{cases} -1, & \text{with probability } 1 - p(z), \\ 1, & \text{with probability } p(z), \end{cases} \tag{7}$$

for $z \in \mathbb{R}$, such that $y_i = f(\langle \mathbf{x}_i, \boldsymbol{\theta}^* \rangle)$, where the notation of f extends to vectors, i.e., $f : \mathbb{R}^n \to \{-1, 1\}^n$, by applying it entry-wise independently so that the response vector is given concisely by $\mathbf{y} = f(\mathbf{X}\boldsymbol{\theta}^*)$.

Definition 1 ("Noise" and "Slope") *There are two important quantities related to the function p, which are concisely represented as the variables $\alpha > 0$ that measures the amount of "noise" in the random function f compared to the* sign *function; and $\gamma > 0$ that measures the average "slope" of the function:*

$$\alpha \triangleq P(f(Z) \neq \mathrm{sign}\,(Z)), \tag{8}$$

$$\gamma \triangleq \mathbb{E}[Z f(Z)], \tag{9}$$

where $Z \sim \mathcal{N}(0, 1)$ is a standard univariate Gaussian random variable and the probabilities and expectations are with respect to Z and the randomness of f. Note that, $\gamma \leq \mathbb{E}[|Z|] = \sqrt{\frac{2}{\pi}}$. From

Stein's lemma, if the function p is differentiable then

$$\gamma = 2\,\mathbb{E}[p'(Z)],$$

where the expectation is now with respect to Z.

In addition, for $\epsilon > 0$, define

$$\alpha_0 \triangleq \max\{\alpha, \frac{\epsilon}{\frac{3}{2}(5+\sqrt{21})}\}. \tag{10}$$

Note that when $p(-z) = 1 - p(z)$—as is the case in logistic and probit regression (see, the formal definitions of these models in Definitions 3 and 4 below)—the expression for α simplifies to

$$\alpha = \frac{1}{\sqrt{2\pi}} \int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2}(1 - p(z) + p(-z))dz = \sqrt{\frac{2}{\pi}} \int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2} p(-z)dz.$$

In addition, the function p must satisfy two assumptions, stated together in Assumption 2, below.

Assumption 2 *The following conditions are enforced on p: (i) p monotonically increases over the real line; and (ii) let $\nu(z) \equiv 1 - p(z) + p(-z)$; the function*

$$\frac{\nu(z+w)}{\nu(z)} \tag{11}$$

is non-increasing in $z \geq 0$, for any $w > 0$.

Intuitively, the second condition means that the "noise" of the GLM defined above decreases at a faster rate away from "margin" ($z = 0$). Indeed, $\nu(z) = P(f(z) = -1) + P(f(-z) = 1)$ for $z \geq 0$ can be thought of as a proxy for the noise with respect to the sign function, and the ratio in (11) quantifies the growth-rate of the function.

Here, it is worth noting—and later, it will be proved—that Assumption 2 is satisfied by two ubiquitous models in binary classification and statistical modeling with binary outcomes: logistic and probit regression. As these two models will be studied later on, the functions, p, corresponding with these models are formally defined below in Definition 3 and 4. To provide greater generality with these models, these definitions introduce an addition parameter: $\beta > 0$, which denotes the inverse temperature and signal-to-noise ratio (SNR) in logistic and probit regression, respectively.

Definition 3 *For* **logistic regression** *with inverse temperature $\beta > 0$, the function $p : \mathbb{R} \to [0, 1]$ is given at $z \in \mathbb{R}$ by*

$$p(z) = \frac{1}{1 + e^{-\beta z}}. \tag{12}$$

Definition 4 *For the* **probit model** *with signal-to-noise ratio (SNR) $\beta > 0$, the function $p : \mathbb{R} \to [0, 1]$ is given at $z \in \mathbb{R}$ by*

$$p(z) = \frac{1}{\sqrt{2\pi}} \int_{u=-\infty}^{u=\beta z} e^{-\frac{1}{2}u^2} du. \tag{13}$$

Note that, equivalently, p is simply the distribution function of a standard Gaussian random variable composed with multiplication by β.

4. Main Results

The main result for the convergence of Algorithm 1 to the correct solution is stated below as Theorem 5, whose proof is overviewed in Section 4.1 and presented in full in Appendix B. Its analog for the dense parameter regime, when $k = d$ and $\Theta = S^{d-1}$, is provided as Corollary 6. Additionally, the specializations of the main result to logistic and probit regression are presented below in Corollary 7 which is also proved in Appendix B. Essentially, these theorems say that for a sufficiently large number of samples, n, the approximations produced by Algorithm 1 converge to the ϵ-ball around the true parameter, $\boldsymbol{\theta}^* \in \Theta$.

Theorem 5 *Fix $d, k, n \in \mathbb{Z}_+$, $k \leq d$, and $\epsilon, \rho \in (0, 1)$. Write $\alpha_0 \triangleq \max\{\alpha, \frac{\epsilon}{\frac{3}{2}(5+\sqrt{21})}\}$ as in Equation (10). Let $\Theta = S^{d-1} \cap \Sigma_k^d$, and fix $\boldsymbol{\theta}^* \in \Theta$ as the unknown parameter. Fix n i.i.d. standard Gaussian covariates, $\mathbf{x}_1, \ldots, \mathbf{x}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, and let $\mathbf{X} = (\mathbf{x}_1 \cdots \mathbf{x}_n)^T$ be the covariate matrix. For a number of samples*

$$n = O\left(\max\left\{\frac{\alpha_0 k}{\gamma^2 \epsilon^2} \log\left(\frac{d}{\epsilon k}\right) + \frac{\alpha_0}{\gamma^2 \epsilon^2} \log\left(\frac{1}{\rho}\right), \frac{k}{\epsilon} \log^{3/2}\left(\frac{1}{\epsilon}\right), \frac{1}{\epsilon}\sqrt{\log\left(\frac{1}{\epsilon}\right)} \log\left(\frac{1}{\rho}\right)\right\}\right), \quad . \tag{14}$$

of the model specified in Equation (7) and under Assumption 2, with probability at least $1 - \rho$, the sequence of approximations, $\{\hat{\boldsymbol{\theta}}^{(t)} \in \Theta\}_{t \in \mathbb{Z}_{\geq 0}}$, produced by Algorithm 1 with the covariate matrix $\mathbf{X}$ converges to the ϵ-ball around $\boldsymbol{\theta}^*$ such that

$$\lim_{t \to \infty} \|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}^{(t)}\|_2 \leq \epsilon, \tag{15}$$

with the rate of convergence upper bounded at each t^{th} iteration, $t \in \mathbb{Z}_{\geq 0}$ by

$$\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}^{(t)}\|_2 \leq 2^{2^{-t}} \epsilon^{1-2^{-t}}. \tag{16}$$

There are a few special cases of interest, formalized below as Corollaries 6–7. Unless stated otherwise, the following corollaries to Theorem 5 use notations consistent with the theorem. As the first of these, Corollary 6 takes a look at the dense parameter regime.

Corollary 6 *Let $\Theta = S^{d-1}$. Then, under Assumption 2, the convergence guarantees for Algorithm 1 stated in Equations (15) and (16) of Theorem 5 hold for a number of samples*

$$n = O\left(\max\left\{\frac{\alpha_0 d}{\gamma^2 \epsilon^2} \log\left(\frac{1}{\epsilon}\right) + \frac{\alpha_0}{\gamma^2 \epsilon^2} \log\left(\frac{1}{\rho}\right), \frac{d}{\epsilon} \log^{3/2}\left(\frac{1}{\epsilon}\right), \frac{1}{\epsilon}\sqrt{\log\left(\frac{1}{\epsilon}\right)} \log\left(\frac{1}{\rho}\right)\right\}\right).$$

We now proceed to applications of the main result to two well-studied GLMs: logistic and probit regression models. For these models, not only can Theorem 5 be shown to be valid, but also closed-form bounds on the sample complexity in the theorem can be obtained, as per Corollary 7 below. For conciseness, only order-wise sample complexities are presented in the corollary's statement, while precise bounds on the sample complexity are specified in its proof in Appendix B.3.

Corollary 7 *When p is the logistic function with inverse temperature $\beta \geq 0$, as in Definition 3 (or the probit regression with SNR $\beta \geq 0$ defined as in Definition 4), there exist positive constants, $b_1, b_2 > 0$, such that the convergence guarantees for Algorithm 1 stated in Equations (15) and (16) of Theorem 5 hold if*

$$n = \begin{cases} \tilde{O}\left(\frac{k}{\beta^2\epsilon^2}\right), & \text{if } \beta \in (0, b_1), \\ \tilde{O}\left(\frac{k}{\beta\epsilon^2}\right), & \text{if } \beta \in [b_1, \frac{b_2}{\epsilon}), \\ \tilde{O}\left(\frac{k}{\epsilon}\right), & \text{if } \beta \in [\frac{b_2}{\epsilon}, \infty). \end{cases} \tag{17}$$

Note that the constants $b_1, b_2 > 0$ can be different for the logistic and probit cases.

Remark 8 *In Corollary 7, Algorithm 1 achieves the order-wise optimal sample complexity (up to logarithmic factors) for parameter estimation in logistic regression under the Gaussian design. See, Hsu and Mazumdar (2024) for the establishment of the optimal sample complexity.*

4.1. Overview of the Proof of the Main Result

While the formal proof of the main theorem, Theorem 5, is deferred to Appendix B.2, the arguments are outlined here. (Meanwhile, Corollary 7 is proved in Appendix B.3 but not outlined here.) This proof resembles the approach in Matsumoto and Mazumdar (2024a), but some important differences are necessary to handle the randomness introduced into the responses. In particular, the analysis in this work relies on the normalization in Step 1 of Algorithm 1 in order to reduce the error induced by f. However, this feature of the analysis will not be apparent until the technical proofs.

To facilitate this overview, as well as the upcoming formal analysis, the following notations are defined. For $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$ and $J \subseteq [d]$, let

$$h(\mathbf{u}, \mathbf{v}) \triangleq \frac{\sqrt{2\pi}}{n} \mathbf{X}^T \frac{1}{2} \left(\text{sign}\left(\mathbf{X}\mathbf{u}\right) - \text{sign}\left(\mathbf{X}\mathbf{v}\right)\right), \tag{18}$$

$$h_J(\mathbf{u}, \mathbf{v}) \triangleq T_{\text{supp}(\mathbf{u}) \cup \text{supp}(\mathbf{v}) \cup J}\left(h(\mathbf{u}, \mathbf{v})\right), \tag{19}$$

$$h_f(\mathbf{u}, \mathbf{v}) \triangleq \frac{\sqrt{2\pi}}{n} \mathbf{X}^T \frac{1}{2} \left(f(\mathbf{X}\mathbf{u}) - \text{sign}\left(\mathbf{X}\mathbf{v}\right)\right), \tag{20}$$

$$h_{f;J}(\mathbf{u}, \mathbf{v}) \triangleq T_{\text{supp}(\mathbf{u}) \cup \text{supp}(\mathbf{v}) \cup J}\left(h_f(\mathbf{u}, \mathbf{v})\right). \tag{21}$$

4.1.1. Outline of the Proof

The proof of Theorem 5 is outlined as follows.

1. The error of the 0^{th} approximation, $\hat{\boldsymbol{\theta}}^{(0)} \sim \Theta$, produced by BIHT is clearly bounded from above by the diameter of the unit sphere S^{d-1}, i.e., no more than 2.
2. The vast majority of the work thus falls onto analyzing any subsequent t^{th} approximation, $\hat{\boldsymbol{\theta}}^{(t)} \in \Theta$, $t \in \mathbb{Z}_+$. For this, the analysis is divided into establishing two main bounds: (i) a deterministic bound on the error of the t^{th} approximation obtained from BIHT (see, Lemma 11), and (ii) a probabilistic bound, which is amounts to a *restricted invertibility* property that holds for Gaussian matrices with high probability (see, Theorem 9). Then, these bounds are combined into the convergence guarantees for BIHT stated in the main theorem.

3. Regarding the first bound, (i), it can be shown that the error of the t^{th} approximation obtained via Algorithm 1 is bounded from above by (see, Lemma 11 and Algorithm 1)

$$\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}^{(t)}\|_2 = O\left(\left\|\boldsymbol{\theta}^* - \frac{T_{\mathrm{supp}(\boldsymbol{\theta}^*)\cup\mathrm{supp}(\hat{\boldsymbol{\theta}}^{(t-1)})\cup\mathrm{supp}(\hat{\boldsymbol{\theta}}^{(t)})}(\hat{\boldsymbol{\theta}}^{(t-1)} + h_f(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}^{(t-1)}))}{\|T_{\mathrm{supp}(\boldsymbol{\theta}^*)\cup\mathrm{supp}(\hat{\boldsymbol{\theta}}^{(t-1)})\cup\mathrm{supp}(\hat{\boldsymbol{\theta}}^{(t)})}(\hat{\boldsymbol{\theta}}^{(t-1)} + h_f(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}^{(t-1)}))\|_2}\right\|_2\right).$$

4. For the second bound, (ii), stated in Step 2, a variant of the restricted approximate invertibility condition that appeared in (Matsumoto and Mazumdar, 2024a, Definition 3.1) is established for Gaussian matrices when the number of rows in the covariate matrix (alternatively, the number of covariates or measurements), n, is sufficiently large. More precisely, Gaussian matrices are shown to have the property that

$$\left\|\boldsymbol{\theta}^* - \frac{\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})}{\|\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})\|_2}\right\|_2 = O\left(\sqrt{\epsilon\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\|_2} + \epsilon\right)$$

uniformly for all $\hat{\boldsymbol{\theta}} \in \Theta$ and $J \subseteq [d]$, $|J| \leq k$, with high probability when n is at least what is specified in Equation (14) (see, Theorem 9).

5. The results of Steps 2 and 3 are then combined in order to upper bound the error of the t^{th} approximation by the following recurrence relation, which holds with bounded probability dictated by Theorem 9, i.e., by the probability that the bound (ii) holds:

$$\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}^{(0)}\|_2 \leq 2,$$
$$\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}^{(t)}\|_2 = O\left(\sqrt{\epsilon\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}^{(t-1)}\|_2} + \epsilon\right), \quad t \in \mathbb{Z}_+.$$

6. Per Fact 12, the above recurrence relation is point-wise bounded from above to yield the rate of convergence and, consequently, the asymptotic convergence in the limit as $t \to \infty$ of the approximations produced by BIHT:

$$\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}^{(t)}\|_2 \leq 2^{2^{-t}}\epsilon^{1-2^{-t}}, \quad t \in \mathbb{Z}_{\geq 0},$$
$$\lim_{t\to\infty}\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}^{(t)}\|_2 \leq \epsilon,$$

completing the proof of the main theorem.

5. Restricted Approximate Invertibility of GLMs

The crux of the analysis for the convergence of Algorithm 1 is a variant of the restricted approximate invertibility conditions (RAICs) studied in Friedlander et al. (2021); Matsumoto and Mazumdar (2024a), which is established for Gaussian matrices in Theorem 9. The formal proofs of these technical results, which constitute the primary technical contributions of this work, are located in Appendix C and overviewed in Section C.1.1.

The main technical theorem will be formalized next.

Theorem 9 *Fix $d, k, n \in \mathbb{Z}_+$, $k \leq d$, and $\rho, \delta \in (0,1)$ where $\delta \triangleq \frac{\epsilon}{\frac{3}{2}(5+\sqrt{21})}$. Write $\alpha_0 = \alpha_0(\delta) \triangleq \max\{\alpha, \delta\}$ as in Equation (10). Let $\Theta = S^{d-1} \cap \Sigma_k^d$, and fix $\boldsymbol{\theta}^* \in \Theta$. Under Assumption 2, for a* *number of samples*

$$n = O\left(\max\left\{ \frac{\alpha_0 k}{\gamma^2 \delta^2} \log\left(\frac{d}{\delta k}\right) + \frac{\alpha_0}{\gamma^2 \delta^2} \log\left(\frac{1}{\rho}\right), \frac{k}{\delta} \log^{3/2}\left(\frac{1}{\delta}\right), \frac{1}{\delta} \sqrt{\log\left(\frac{1}{\delta}\right)} \log\left(\frac{1}{\rho}\right) \right\} \right), \tag{22}$$

with probability at least $1 - \rho$*, uniformly for all* $\hat{\boldsymbol{\theta}} \in \Theta$ *and all* $J \subseteq [d]$*,* $|J| \leq k$*,*

$$\left\| \boldsymbol{\theta}^* - \frac{\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})}{\|\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})\|_2} \right\|_2 \leq \sqrt{\delta \|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\|_2} + \delta. \tag{23}$$

The main technical theorem holds for logistic and probit regression—as formalized below in Corollary 10—because Assumption 2 is satisfied by both models. Moreover, closed-form bounds on the sample complexity in Theorem 9 can be derived for these two exemplary GLMs, which are stated as order-wise results in Corollary 10 with the specification of precise bounds and constants left to the proof of the corollary in Appendix C.5.

Corollary 10 *Let* p *be the logistic function with inverse temperature* $\beta > 0$*, as in Definition 3 (or the probit function with SNR* $\beta > 0$*, as in Definition 4). If there exist absolute constants* $b_1, b_2 > 0$ *then for a number of samples given by Equation (17), the bound stated as Equation (23) in Theorem 9 holds uniformly for all* $\hat{\boldsymbol{\theta}} \in \Theta$ *and all* $J \subseteq [d]$*,* $|J| \leq k$*, with probability at least* $1 - \rho$*.*

6. Conclusion

In this paper, we made a case for binary iterative hard thresholding, (projected) gradient descent on the ReLU loss, as a universal learning algorithm for classification tasks. Under very general models of nonseparable (and separable) data, that include logistic, probit, and random classification noise models, BIHT is statistically optimal in parameter estimation. We observe this in practice as well.

We have restricted ourselves to Gaussian covariates, for which our results are tight. However it will be worth exploring the performance of BIHT for more general classes of distributions. We also note an observation that contrasts the noiseless case from generalized linear models, as far as the dynamics of BIHT is concerned. It is known that the normalization step in BIHT, though essential for the convergence proof of Friedlander et al. (2021); Matsumoto and Mazumdar (2024a), can be redundant in practice for the noiseless case, see Figure 2 in the appendix. On the other hand, for the setting of this paper, the normalization seems to be crucial for the stability of the algorithm especially in the high-noise regime (as can be seen in Figure 1 in the appendix).

Finally, it will be interesting to analyze BIHT (and a stochastic perceptron-like version of it) from a learning theory perspective, especially in the agnostic setting, where data not necessarily comes from a GLM. Other noise models, such as Massart noise, can also be interesting.

References

J Aitchison and SD Silvey. Maximum-likelihood estimation procedures and associated tests of significance. *Journal of the Royal Statistical Society: Series B (Methodological)*, 22(1):154–171, 1960.

Sohail Bahmani, Petros T Boufounos, and Bhiksha Raj. Learning model-based sparsity via projected gradient descent. *IEEE Transactions on Information Theory*, 62(4):2092–2099, 2016.

Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019.

Thomas Blumensath and Mike E Davies. Iterative hard thresholding for compressed sensing. *Appl. Comput. Harmon. Anal*, 27:265–274, 2009.

Petros T Boufounos and Richard G Baraniuk. 1-bit compressive sensing. In *2008 42nd Annual Conference on Information Sciences and Systems*, pages 16–21. IEEE, 2008.

Hugo Chardon, Matthieu Lerasle, and Jaouad Mourtada. Finite-sample performance of the maximum likelihood estimator in logistic regression. *arXiv preprint arXiv:2411.02137*, 2024.

Moses S Charikar. Similarity estimation techniques from rounding algorithms. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pages 380–388. ACM, 2002.

Annette J Dobson and Adrian G Barnett. *An introduction to generalized linear models*. Chapman and Hall/CRC, 2018.

Ludwig Fahrmeir and Gerhard Tutz. *Multivariate statistical modelling based on generalized linear models*, volume 425. Springer.

D Firth. Generalized linear models and jeffreys priors: an iterative weighted least-squares approach. In *Computational Statistics: Volume 1: Proceedings of the 10th Symposium on Computational Statistics*, pages 553–557. Springer, 1992.

Michael P Friedlander, Halyun Jeong, Yaniv Plan, and Özgür Yılmaz. Nbiht: An efficient algorithm for 1-bit compressed sensing with optimal error decay rate. *IEEE Transactions on Information Theory*, 68(2):1157–1177, 2021.

A Ronald Gallant and Douglas W Nychka. Semi-nonparametric maximum likelihood estimation. *Econometrica: Journal of the econometric society*, pages 363–390, 1987.

Aravind Gollakota, Parikshit Gopalan, Adam Klivans, and Konstantinos Stavropoulos. Agnostically learning single-index models using omnipredictors. *Advances in Neural Information Processing Systems*, 36, 2024.

James W Hardin and Joseph M Hilbe. *Generalized linear models and extensions*. Stata press, 2007.

Daniel Hsu and Arya Mazumdar. On the sample complexity of parameter estimation in logistic regression with normal design. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 2418–2437. PMLR, 2024.

Laurent Jacques, Kévin Degraux, and Christophe De Vleeschouwer. Quantized iterative hard thresholding: Bridging 1-bit and high-resolution quantized compressed sensing. *arXiv preprint arXiv:1305.1786*, 2013a.

Laurent Jacques, Jason N Laska, Petros T Boufounos, and Richard G Baraniuk. Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors. *IEEE Transactions on Information Theory*, 59(4):2082–2102, 2013b.

Qiujiang Jin, Tongzheng Ren, Nhat Ho, and Aryan Mokhtari. Statistical and computational complexities of bfgs quasi-newton method for generalized linear models. *arXiv preprint arXiv:2206.00207*, 2022.

Sham M Kakade, Varun Kanade, Ohad Shamir, and Adam Kalai. Efficient learning of generalized linear and single index models with isotonic regression. *Advances in Neural Information Processing Systems*, 24, 2011.

Adam Klivans and Raghu Meka. Learning graphical models using multiplicative weights. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 343–354. IEEE, 2017.

Felix Kuchelmeister and Sara van de Geer. Finite sample rates for logistic regression with small noise or few samples. *Sankhya A*, pages 1–70, 2024.

Dekai Liu, Song Li, and Yi Shen. One-bit compressive sensing with projected subgradient method under sparsity constraints. *IEEE Transactions on Information Theory*, 65(10):6650–6663, 2019.

Namiko Matsumoto and Arya Mazumdar. Binary iterative hard thresholding converges with optimal number of measurements for 1-bit compressed sensing. *J. ACM*, jul 2024a. ISSN 0004-5411. doi: 10.1145/3680542. URL https://doi.org/10.1145/3680542.

Namiko Matsumoto and Arya Mazumdar. Robust 1-bit compressed sensing with iterative hard thresholding. In *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 2941–2979. SIAM, 2024b.

Peter McCullagh. *Generalized linear models*. Routledge, 2019.

Charles E McCulloch. Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American statistical Association*, 92(437):162–170, 1997.

Marco Mondelli and Ramji Venkataramanan. Approximate message passing with spectral initialization for generalized linear models. In *International Conference on Artificial Intelligence and Statistics*, pages 397–405. PMLR, 2021.

In Jae Myung. Tutorial on maximum likelihood estimation. *Journal of mathematical Psychology*, 47(1):90–100, 2003.

John Ashworth Nelder and Robert WM Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 135(3):370–384, 1972.

Samet Oymak and Ben Recht. Near-optimal bounds for binary embeddings of arbitrary sets. *arXiv preprint arXiv:1512.04433*, 2015.

Yaniv Plan and Roman Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Transactions on Information Theory*, 59(1):482–494, 2012.

Yaniv Plan and Roman Vershynin. The generalized lasso with non-linear observations. *IEEE Transactions on information theory*, 62(3):1528–1537, 2016.

Yaniv Plan, Roman Vershynin, and Elena Yudovina. High-dimensional estimation with geometric constraints. *Information and Inference: A Journal of the IMA*, 6(1):1–40, 2017.

Sundeep Rangan. Generalized approximate message passing for estimation with random linear mixing. In *2011 IEEE International Symposium on Information Theory Proceedings*, pages 2168–2172. IEEE, 2011.

Sundeep Rangan, Philip Schniter, and Alyson K Fletcher. Vector approximate message passing. *IEEE Transactions on Information Theory*, 65(10):6664–6684, 2019.

Francis SG Richards. A method of maximum-likelihood estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 23(2):469–475, 1961.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

David J Sakrison. Efficient recursive estimation; application to estimating the parameters of a covariance function. *International journal of engineering science*, 3(4):461–483, 1965.

Fariborz Salehi, Ehsan Abbasi, and Babak Hassibi. The impact of regularization on high-dimensional logistic regression. *Advances in Neural Information Processing Systems*, 32, 2019.

Philip Schniter, Sundeep Rangan, and Alyson K Fletcher. Vector approximate message passing for the generalized linear model. In *2016 50th Asilomar conference on signals, systems and computers*, pages 1525–1529. IEEE, 2016.

Rocco A Servedio. On pac learning using winnow, perceptron, and a perceptron-like algorithm. In *Proceedings of the Twelfth Annual Conference on Computational learning theory*, pages 296–307, 1999.

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

Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi. Lasso with non-linear measurements is equivalent to one with linear measurements. *Advances in Neural Information Processing Systems*, 28, 2015.

Panagiotis Toulis, Edoardo Airoldi, and Jason Rennie. Statistical analysis of stochastic gradient methods for generalized linear models. In *International Conference on Machine Learning*, pages 667–675. PMLR, 2014.

Ramji Venkataramanan, Kevin Kögler, and Marco Mondelli. Estimation in rotationally invariant generalized linear models via approximate message passing. In *International Conference on Machine Learning*, pages 22120–22144. PMLR, 2022.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Abraham Wald. Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics*, 20(4):595–601, 1949.

Robert WM Wedderburn. Quasi-likelihood functions, generalized linear models, and the gauss—newton method. *Biometrika*, 61(3):439–447, 1974.

Yihan Zhang, Hong Chang Ji, Ramji Venkataramanan, and Marco Mondelli. Spectral estimators for structured generalized linear models via approximate message passing. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 5224–5230. PMLR, 2024.

Zilu Zhao, Fangqing Xiao, and Dirk Slock. Vector approximate message passing for not so large niid generalized i/o linear models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13281–13285. IEEE, 2024.

Jiang Zhu, Lin Han, and Xiangming Meng. An amp-based low complexity generalized sparse bayesian learning algorithm. *IEEE Access*, 7:7965–7976, 2018.

Appendix A. Comparison to GLM-tron

A special case of the projected gradient decent algorithm studied by Bahmani et al. (2016) offers an alternative gradient-based method to BIHT for parameter estimation in some classes of GLMs, including those with nondecreasing, Lipschitz transfer functions, g^{-1}. Note that this class encompasses GLMs whose responses follow an exponential distribution, which is one of the most widely studied families of GLMs. Like BIHT, this algorithm projects its approximations onto the set of k-sparse vectors, where the sparsity can be taken as $k = d$ in the dense parameter regime so as to effectively eliminate the projection step of the algorithm. In the dense parameter regime, this algorithm becomes the perceptron-like *GLM-tron* algorithm of the earlier work, Kakade et al. (2011), which learns GLMs with (possibly nonstrictly) monotonically increasing, Lipschitz transfer functions. For concise nomenclature, we will borrow the name of "GLM-tron" to refer to the original GLM-tron algorithm with the addition of a sparse projection, as analyzed in Bahmani et al. (2016).

BIHT and GLM-tron have a few key differences. While other covariate designs are possible, the analysis in Bahmani et al. (2016) assumes that the norm of each covariate is almost surely at most 1, in contrast to the Gaussian covariate design considered in this present work. As another distinction between BIHT and GLM-tron, BIHT requires that the GLM has binary outcomes with a mild condition on the link function, g, but otherwise need not know the specific choice of link function, while GLM-tron can learn a larger class of GLMs that only necessitates that the transfer function, g^{-1}, satisfies a certain derivative condition (which indeed holds when the transfer function is nondecreasing and Lipschitz). However, unlike BIHT, GLM-tron requires knowledge of the specific choice of link function. (Note that Kakade et al. (2011) proposes a second algorithm for learning single-index models which estimates an unknown link function, but this is outside the scope of this work.) Most significantly, BIHT and GLM-tron "try to" minimize different objective functions. Whereas BIHT performs gradient descent on the (negative) ReLU loss,

$$\mathscr{J}_{\text{BIHT}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} |[y_i \langle \mathbf{x}_i, \boldsymbol{\theta} \rangle]_-|,$$

by taking gradient steps in the negated direction of

$$\nabla_{\boldsymbol{\theta}} \mathscr{J}_{\text{BIHT}}(\boldsymbol{\theta}) \ni -\mathbf{X}^T \frac{1}{2} \left(\mathbf{y} - \text{sign}(\mathbf{X}\boldsymbol{\theta}) \right),$$

GLM-tron is a gradient descent procedure on the loss

$$\mathscr{J}_{\text{GLM-tron}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} G(\langle \mathbf{x}_i, \boldsymbol{\theta} \rangle) - y_i \langle \mathbf{x}_i, \boldsymbol{\theta} \rangle$$

with gradient steps in the negated direction of

$$\nabla_{\boldsymbol{\theta}} \mathscr{J}_{\text{GLM-tron}}(\boldsymbol{\theta}) = -\mathbf{X}^T \left(\mathbf{y} - g^{-1}(\mathbf{X}\boldsymbol{\theta}) \right),$$

where the function, G, is defined such that $g^{-1} = \partial G$. When the responses, $y_i \,|\, \mathbf{x}_i$, $i \in [n]$, follow an exponential distribution, $\mathscr{J}_{\text{GLM-tron}}$ becomes the negative log-likelihood function, and hence, roughly speaking, GLM-tron essentially "tries to" compute the MLE in this case. The difference in objective functions is fundamental: it precludes the application of the analysis for BIHT in this work to GLM-tron. Conversely, adapting the approach in Bahmani et al. (2016) is insufficient to achieve the sample complexity established for BIHT here.

Assuming that the responses, y_i, $i \in [n]$, are bounded, as is the case for binary GLMs, Bahmani et al. (2016) shows that GLM-tron achieves an error-rate of ϵ provided the number of covariates, n, is at least

$$n = \tilde{O}\left(\max\left\{ \frac{1}{\epsilon^4}, k \log\left(\frac{d}{k}\right) \right\} \right),$$

where this hides some terms. Notice that the dependency on the error-rate, ϵ, is ϵ^{-4} compared to ϵ^{-2} obtained in this work for BIHT, though in fairness, neither result should be considered "superior" to the other since, although this work obtains a smaller dependence on ϵ and need not know the link function, the analysis Bahmani et al. (2016) applies to a larger class of GLMs.

A.1. Other Related Work

Stochastic gradient descent (SGD) (Robbins and Monro, 1951; Sakrison, 1965) offers an alternative gradient-based method for parameter estimation in GLMs. Toulis et al. (2014) studies the statistical properties of SGD estimates in GLMs when updates are both explicit and implicit. However, such SGD estimates are asymptotically sub-optimal compared to the maximum likelihood estimates, as noted by Toulis et al. (2014). As another common approach, approximate message passing (AMP) and its extensions have also been heavily applied to parameter estimation in GLMs, e.g., Mondelli and Venkataramanan (2021); Venkataramanan et al. (2022); Zhang et al. (2024); Barbier et al. (2019); Zhu et al. (2018); Schniter et al. (2016); Zhao et al. (2024). This includes generalized approximate message passing (GAMP), an algorithm first proposed by Rangan (2011). While the

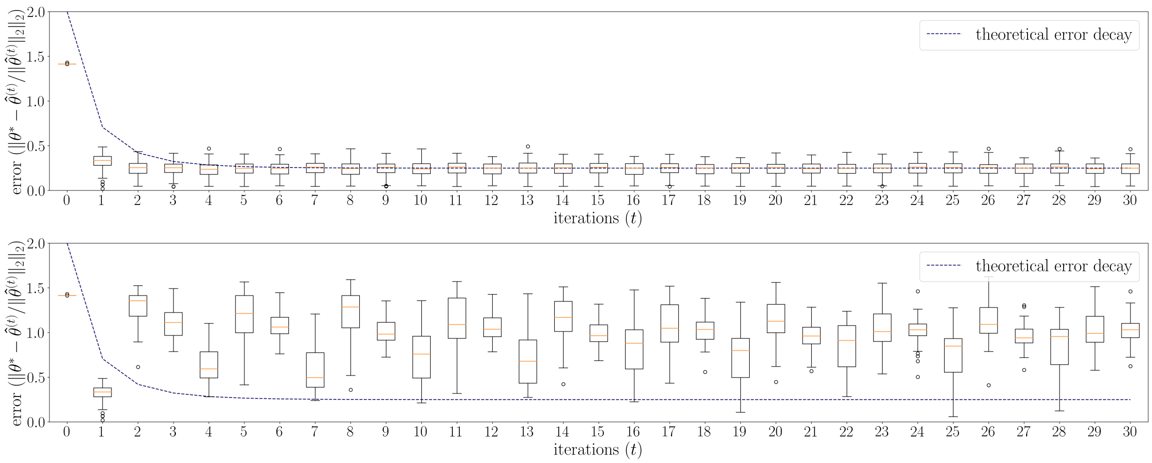

Figure 1: This experiment compares the iterative approximation errors for BIHT with (top plot) and without (bottom plot) the normalization step of the algorithm under logistic regression with inverse temperature $\beta = 1$. The error is the ℓ_2-distance between the normalized approximation and the true parameter. In both plots, the theoretical error decay for the normalized version of BIHT with logistic regression is displayed for reference. The experiment ran 100 trials of recovery for 30 iterations with parameters: $d = 2000$, $k = 5$, $n = 3000$, $\epsilon = 0.25$, and $\rho = 0.25$.

error-rate of GAMP is information theoretically optimal for some GLM's, it falls short of the information theoretical optimum for GLMs in some paradigms (Barbier et al., 2019). In fact, Barbier et al. (2019) characterizes the regions of the parameter space in which GAMP achieves the information theoretical optimal error-rate or is information theoretically sub-optimal for GLMs. As a relative to GAMP, another variant of approximation passing called vector approximate message passing (VAMP), introduced by Rangan et al. (2019), has been used for estimation in GLMs, initially by Schniter et al. (2016) and subsequently by, e.g., Zhao et al. (2024).

Appendix B. Proof of the Main Results

In this section, the main results—Theorem 5 and Corollary 7—are proved, contingent on the correctness of the main technical results—Theorem 9 and Corollary 10—and some auxiliary results, whose proofs are deferred to Appendices C and D.

B.1. Intermediate Results

Before Theorem 5 can be proved, two auxiliary results, stated below as Lemma 11 and Fact 12, are needed. The first of these intermediate results—whose proof is deferred to Appendix B.4—will allow the main technical result, Theorem 9, as well as its corollaries, to be related to the error of the approximations iteratively produced by BIHT (Algorithm 1).

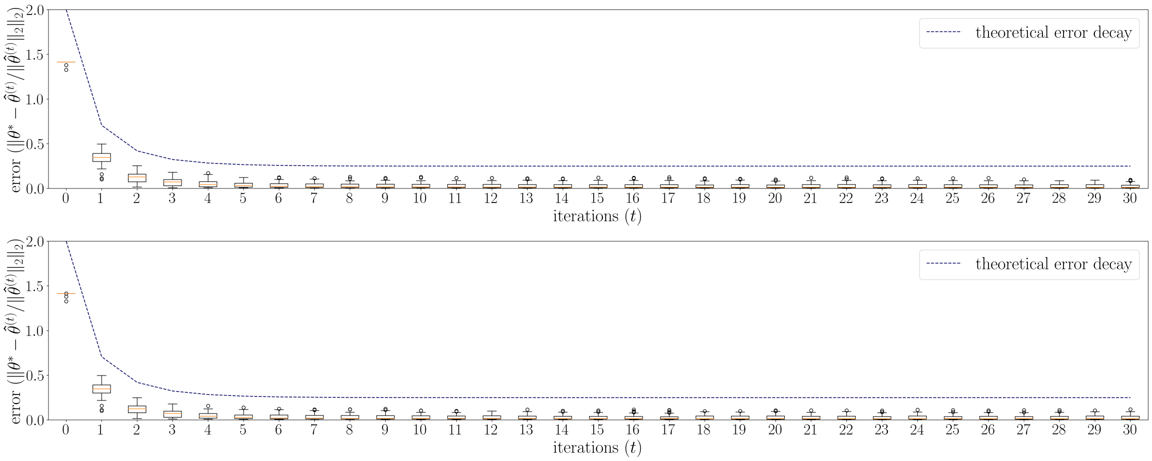

Figure 2: This experiment compares the iterative approximation errors for BIHT with (top plot) and without (bottom plot) the normalization step of the algorithm under the noiseless model. The error is the ℓ_2-distance between the normalized approximation and the true parameter. In both plots, the theoretical error decay for the normalized version of BIHT in the noiseless setting—established by Matsumoto and Mazumdar (2024a)—is displayed for reference. The experiment ran 100 trials of recovery for 30 iterations with parameters: $d = 2000$, $k = 5$, $n = 700$, $\epsilon = 0.25$, and $\rho = 0.25$.

Lemma 11 *Let $\mathbf{u} \in \mathbb{R}^d \cap \Sigma_k^d$ and $\mathbf{v} \in \mathbb{R}^d$, and let $J, J', J'' \subseteq [d]$, where $|J| \leq k$, $J' \triangleq \operatorname{supp}(\mathbf{u})$, and $J'' \triangleq \operatorname{supp}(T_k(\mathbf{v}))$. Then,*

$$\left\| \mathbf{u} - \frac{T_k(\mathbf{v})}{\|T_k(\mathbf{v})\|_2} \right\|_2 \leq 3 \left\| \mathbf{u} - \frac{T_{J \cup J' \cup J''}(\mathbf{v})}{\|T_{J \cup J' \cup J''}(\mathbf{v})\|_2} \right\|_2. \tag{24}$$

The proof of the main theorem will additionally utilize the following fact from Matsumoto and Mazumdar (2024a). The iterative approximation errors will turn out to be upper bounded by the functions in this fact, and thus, this fact will facilitate the calculation of a close-form bound on the iterative approximation errors, much like the approach in Matsumoto and Mazumdar (2024a).

Fact 12 ((Matsumoto and Mazumdar, 2024a, Fact 4.1)) *Let $u, v, w \in \mathbb{R}_+$, where $u \triangleq \frac{1}{2}(1 + \sqrt{1+4w})$ and $1 \leq u \leq \sqrt{\frac{2}{v}}$. Let $f_1, f_2 : \mathbb{Z}_{\geq 0} \to \mathbb{R}$ be functions given by*

$$\begin{aligned} f_1(0) &= 2, \\ f_1(t) &= \sqrt{v f_1(t-1)} + vw, \quad t \in \mathbb{Z}_+, \\ f_2(t) &= 2^{2^{-t}} (u^2 v)^{1-2^{-t}}, \quad t \in \mathbb{Z}_{\geq 0}. \end{aligned}$$

Then,

$$\begin{aligned} f_1(t) &> f_1(t'), \quad t < t' \in \mathbb{Z}_{\geq 0}, \\ f_2(t) &> f_2(t'), \quad t < t' \in \mathbb{Z}_{\geq 0}, \end{aligned}$$

$$f_1(t) \le f_2(t), \quad t \in \mathbb{Z}_{\ge 0},$$
$$\lim_{t\to\infty} f_1(t) \le \lim_{t\to\infty} f_2(t) = u^2 v.$$

B.2. Proof of Theorem 5

With the above results in Appendix B.1, the convergence of the BIHT approximations, as stated in the main theorem, can now be proved.

Proof Theorem 5 Setting

$$\delta = \frac{\epsilon}{\frac{3}{2}(5+\sqrt{21})} = \frac{\epsilon}{9\left(\frac{1}{2}(1+\sqrt{\frac{7}{3}})\right)^2}, \tag{25}$$

and taking

$$\begin{aligned} n \ge \max\Bigg\{ & \frac{c_1\alpha_0}{\gamma^2\delta^2}\log\left(\frac{a}{\rho}\sum_{\ell=0}^{k}\binom{d}{\ell}\sum_{\ell'=0}^{k}\binom{d}{\ell'}\left(\frac{b}{\tau}\right)^{\ell'}\right), \\ & \frac{c_2}{\gamma\delta\sqrt{\log\left(\frac{4e}{\eta}\right)}}\log\left(\frac{a}{\rho}\sum_{\ell=0}^{k}\binom{d}{\ell}\sum_{\ell'=0}^{k}\binom{d}{\ell'}\left(\frac{b}{\tau}\right)^{\ell'}\right), \\ & \frac{c_3}{\eta}\log\left(\frac{a}{\rho}\right), \frac{c_4 k}{\eta}\log\left(\frac{1}{\eta}\right)\Bigg\}, \end{aligned} \tag{26}$$

the following bound holds for all $\hat{\boldsymbol{\theta}} \in \Theta$ and all $J \subseteq [d]$, $|J| \le k$, uniformly with probability at least $1-\rho$ due to Theorem 9:

$$\left\| \boldsymbol{\theta}^* - \frac{\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})}{\|\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})\|_2} \right\|_2 \le \sqrt{\delta\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\|_2} + \delta. \tag{27}$$

The remainder of the proof will assume that the inequality in Equation (27) holds uniformly, which occurs with bounded probability, as just stated. Additionally, using the notations in Fact 12—wherein the variables are set as $u \triangleq \frac{1}{2}(1+\sqrt{\frac{7}{3}})$, $v \triangleq 9\delta$, and $w \triangleq \frac{1}{3}$ and satisfy the fact's requirement, $\sqrt{\frac{2}{v}} = \sqrt{\frac{2}{9\delta}} = \sqrt{\frac{2\cdot 9u^2}{9\epsilon}} = u\sqrt{\frac{2}{\epsilon}} > u$ —define the functions $f_1, f_2 : \mathbb{Z}_{\ge 0} \to \mathbb{R}$ by

$$f_1(0) = 2, \tag{28}$$
$$f_1(t) = \sqrt{v f_1(t-1)} + wv = \sqrt{9\delta f_1(t-1)} + 3\delta, \quad t \in \mathbb{Z}_+, \tag{29}$$
$$f_2(t) = 2^{2^{-t}}(u^2 v)^{1-2^{-t}} = 2^{2^{-t}}\left(\frac{3}{2}(5+\sqrt{21})\delta\right)^{1-2^{-t}} = 2^{2^{-t}}\epsilon^{1-2^{-t}}, \quad t \in \mathbb{Z}_{\ge 0}. \tag{30}$$

Then, by Fact 12, for all $t \in \mathbb{Z}_{\ge 0}$,

$$f_1(t) \le f_2(t) = 2^{2^{-t}}\epsilon^{1-2^{-t}}, \tag{31}$$

and asymptotically,

$$\lim_{t\to\infty} f_1(t) \le \lim_{t\to\infty} f_2(t) = \epsilon. \tag{32}$$

With these preliminaries laid out, we are ready to verify Equations (15) and (16) in Theorem 5, which can be argued inductively. Inducting on the iterations, $t = 0, 1, 2, 3, \ldots$, the following inductive claim will be shown:

$$C(t) \triangleq \text{``}\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}^{(t)}\|_2 \leq f_1(t).\text{''} \tag{33}$$

The base case, when $t = 0$, is trivial: since there is the membership of $\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}^{(0)} \in S^{d-1} \cap \Sigma_k^d \subseteq S^{d-1}$, the Euclidean distance between $\boldsymbol{\theta}^*$ and $\hat{\boldsymbol{\theta}}^{(0)}$ cannot exceed the diameter of the unit sphere (distance 2), i.e.,

$$\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}^{(0)}\|_2 \leq 2 = f_1(0),$$

where the rightmost equality is due to the definition of f_1 in Equation (28). Next, consider some arbitrary choice of $t \in \mathbb{Z}_+$, and suppose that for every $t' < t$, the t'^{th} inductive claim, $C(t')$, holds. Then, under this inductive assumption, the t^{th} inductive claim, $C(t)$, needs to be verified. Recall that for $t > 0$, Algorithm 1 sets

$$\tilde{\boldsymbol{\theta}}^{(t)} = \hat{\boldsymbol{\theta}}^{(t-1)} + \frac{\sqrt{2\pi}}{n} \mathbf{X}^T \frac{1}{2} \left(f(\mathbf{X}\boldsymbol{\theta}^*) - \operatorname{sign}(\mathbf{X}\hat{\boldsymbol{\theta}}^{(t-1)}) \right) = \hat{\boldsymbol{\theta}}^{(t-1)} + h_f(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}^{(t-1)}), \tag{34}$$

$$\hat{\boldsymbol{\theta}}^{(t)} = \frac{T_k(\tilde{\boldsymbol{\theta}}^{(t)})}{\|T_k(\tilde{\boldsymbol{\theta}}^{(t)})\|_2}. \tag{35}$$

Additionally, due to Lemma 11—where, in the context of this proof, the sets $J, J'', J' \subseteq [d]$ in the lemma are taken to be $J \triangleq \operatorname{supp}(\hat{\boldsymbol{\theta}}^{(t-1)})$, $J'' \triangleq \operatorname{supp}(\boldsymbol{\theta}^*)$, and $J' \triangleq \operatorname{supp}(\hat{\boldsymbol{\theta}}^{(t)})$—the following holds:

$$\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}^{(t)}\|_2 = \left\| \boldsymbol{\theta}^* - \frac{T_k(\tilde{\boldsymbol{\theta}}^{(t)})}{\|T_k(\tilde{\boldsymbol{\theta}}^{(t)})\|_2} \right\|_2 \leq 3 \left\| \boldsymbol{\theta}^* - \frac{T_{\operatorname{supp}(\boldsymbol{\theta}^*) \cup \operatorname{supp}(\hat{\boldsymbol{\theta}}^{(t-1)}) \cup \operatorname{supp}(\hat{\boldsymbol{\theta}}^{(t)})}(\tilde{\boldsymbol{\theta}}^{(t)})}{\|T_{\operatorname{supp}(\boldsymbol{\theta}^*) \cup \operatorname{supp}(\hat{\boldsymbol{\theta}}^{(t-1)}) \cup \operatorname{supp}(\hat{\boldsymbol{\theta}}^{(t)})}(\tilde{\boldsymbol{\theta}}^{(t)})\|_2} \right\|_2. \tag{36}$$

Then, the t^{th} inductive claim, $C(t)$, can now be established:

$$\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}^{(t)}\|_2 \leq 3 \left\| \boldsymbol{\theta}^* - \frac{T_{\operatorname{supp}(\boldsymbol{\theta}^*) \cup \operatorname{supp}(\hat{\boldsymbol{\theta}}^{(t-1)}) \cup \operatorname{supp}(\hat{\boldsymbol{\theta}}^{(t)})}(\tilde{\boldsymbol{\theta}}^{(t)})}{\|T_{\operatorname{supp}(\boldsymbol{\theta}^*) \cup \operatorname{supp}(\hat{\boldsymbol{\theta}}^{(t-1)}) \cup \operatorname{supp}(\hat{\boldsymbol{\theta}}^{(t)})}(\tilde{\boldsymbol{\theta}}^{(t)})\|_2} \right\|_2$$

► by Equation (36)

$$= 3 \left\| \boldsymbol{\theta}^* - \frac{T_{\operatorname{supp}(\boldsymbol{\theta}^*) \cup \operatorname{supp}(\hat{\boldsymbol{\theta}}^{(t-1)}) \cup \operatorname{supp}(\hat{\boldsymbol{\theta}}^{(t)})}(\hat{\boldsymbol{\theta}}^{(t-1)} + h_f(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}^{(t-1)}))}{\|T_{\operatorname{supp}(\boldsymbol{\theta}^*) \cup \operatorname{supp}(\hat{\boldsymbol{\theta}}^{(t-1)}) \cup \operatorname{supp}(\hat{\boldsymbol{\theta}}^{(t)})}(\hat{\boldsymbol{\theta}}^{(t-1)} + h_f(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}^{(t-1)}))\|_2} \right\|_2$$

► by Equation (34)

$$= 3 \left\| \boldsymbol{\theta}^* - \frac{\hat{\boldsymbol{\theta}}^{(t-1)} + h_{f;\operatorname{supp}(\hat{\boldsymbol{\theta}}^{(t)})}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}^{(t-1)})}{\|\hat{\boldsymbol{\theta}}^{(t-1)} + h_{f;\operatorname{supp}(\hat{\boldsymbol{\theta}}^{(t)})}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}^{(t-1)})\|_2} \right\|_2$$

► by the definitions of the subset thresholding operation and $h_{f;J}$ ($J \subseteq [d]$)

$$\leq 3 \left(\sqrt{\delta \|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}^{(t-1)}\|_2} + \delta \right)$$

$$
\begin{aligned}
&\qquad \blacktriangleright \text{ by Equation (27)} \\
&= \sqrt{9\delta \|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}^{(t-1)}\|_2} + 3\delta \\
&\leq \sqrt{9\delta f_1(t-1)} + 3\delta \\
&\qquad \blacktriangleright \text{ by the inductive hypothesis, i.e., the assumed correctness of } C(t-1) \\
&= f_1(t), \\
&\qquad \blacktriangleright \text{ by the definition of } f_1 \text{ in Equation (29)}
\end{aligned}
$$

as desired.

Having verified the t^{th} inductive claim, $C(t)$, under the inductive assumption, it follows by induction that for all $t \in \mathbb{Z}_{\geq 0}$, the t^{th} inductive claim, $C(t)$, holds:

$$
\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}^{(t)}\|_2 \leq f_1(t).
$$

Therefore, the assumption that Equation (27) holds uniformly—which occurs with probability at least $1-\rho$—and Equations (31) and (32) together imply that

$$
\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}^{(t)}\|_2 \leq f_1(t) \leq f_2(t) = 2^{2^{-t}} \epsilon^{1-2^{-t}}
$$

for every $t \in \mathbb{Z}_{\geq 0}$ and that

$$
\lim_{t\to\infty} \|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}^{(t)}\|_2 \leq \lim_{t\to\infty} f_1(t) \leq \lim_{t\to\infty} f_2(t) = \epsilon,
$$

concluding the theorem's proof. ■

B.3. Proof of Corollary 7

Proof Corollary 7 Under the presumed correctness of the main technical corollary, Corollary 10, Corollary 7 for the convergence of BIHT (Algorithm 1) in logistic and probit regressions, now follow along the same arguments as in the proof of Theorem 5. In this analogous proof, the use of Corollary 10 replaces Theorem 9 in order to establish the logistic and probit cases of Corollary 7. ■

B.4. Proof of the Intermediate Result, Lemma 11

This section verifies the intermediate result, Lemma 11, which was introduced in Appendix B.1. The proof of Lemma 11 will use the following fact.

Fact 13 *Let* $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$. *Then,*

$$
\left\| \frac{\mathbf{u}}{\|\mathbf{u}\|_2} - \frac{\mathbf{v}}{\|\mathbf{v}\|_2} \right\|_2 \leq 2 \min \left\{ \frac{\|\mathbf{u}-\mathbf{v}\|_2}{\|\mathbf{u}\|_2}, \frac{\|\mathbf{u}-\mathbf{v}\|_2}{\|\mathbf{v}\|_2} \right\}. \tag{37}
$$

Proof Fact 13 Before the fact is verified, the following easily verifiable claim is derived.

Claim 14 *Let* $z \geq 0$. *Then,* $|1 - |1-z|| \leq z$.

Now, returning to the proof of Fact 13, fix $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$ arbitrarily. Observe:

$$\begin{aligned}
\left\| \frac{\mathbf{u}}{\|\mathbf{u}\|_2} - \frac{\mathbf{v}}{\|\mathbf{v}\|_2} \right\|_2 &= \left\| \left(\frac{\mathbf{u}}{\|\mathbf{u}\|_2} - \frac{\mathbf{v}}{\|\mathbf{u}\|_2} \right) + \left(\frac{\mathbf{v}}{\|\mathbf{u}\|_2} - \frac{\mathbf{v}}{\|\mathbf{v}\|_2} \right) \right\|_2 \\
&\leq \left\| \frac{\mathbf{u}}{\|\mathbf{u}\|_2} - \frac{\mathbf{v}}{\|\mathbf{u}\|_2} \right\|_2 + \left\| \frac{\mathbf{v}}{\|\mathbf{u}\|_2} - \frac{\mathbf{v}}{\|\mathbf{v}\|_2} \right\|_2 \\
&\qquad \blacktriangleright \text{ by the triangle inequality} \\
&= \frac{\|\mathbf{u}-\mathbf{v}\|_2}{\|\mathbf{u}\|_2} + \left| 1 - \frac{\|\mathbf{u}-(\mathbf{u}-\mathbf{v})\|_2}{\|\mathbf{u}\|_2} \right| \\
&\leq \frac{\|\mathbf{u}-\mathbf{v}\|_2}{\|\mathbf{u}\|_2} + \max\left\{ \left| 1 - \frac{\|\mathbf{u}\|_2 + \|\mathbf{u}-\mathbf{v}\|_2}{\|\mathbf{u}\|_2} \right|, \left| 1 - \left| \frac{\|\mathbf{u}\|_2 - \|\mathbf{u}-\mathbf{v}\|_2}{\|\mathbf{u}\|_2} \right| \right| \right\} \\
&\qquad \blacktriangleright \text{ since } \big| \|\mathbf{u}\|_2 - \|\mathbf{u}-\mathbf{v}\|_2 \big| \leq \|\mathbf{u}-(\mathbf{u}-\mathbf{v})\|_2 \leq \|\mathbf{u}\|_2 + \|\mathbf{u}-\mathbf{v}\|_2 \\
&\qquad\ \ \text{by the triangle inequality} \\
&= \frac{\|\mathbf{u}-\mathbf{v}\|_2}{\|\mathbf{u}\|_2} + \max\left\{ \left| 1 - \left(1 + \frac{\|\mathbf{u}-\mathbf{v}\|_2}{\|\mathbf{u}\|_2} \right) \right|, \left| 1 - \left| 1 - \frac{\|\mathbf{u}-\mathbf{v}\|_2}{\|\mathbf{u}\|_2} \right| \right| \right\} \\
&= \frac{\|\mathbf{u}-\mathbf{v}\|_2}{\|\mathbf{u}\|_2} + \max\left\{ \frac{\|\mathbf{u}-\mathbf{v}\|_2}{\|\mathbf{u}\|_2}, \frac{\|\mathbf{u}-\mathbf{v}\|_2}{\|\mathbf{u}\|_2} \right\} \\
&\qquad \blacktriangleright \text{ by Claim 14} \\
&= \frac{2\|\mathbf{u}-\mathbf{v}\|_2}{\|\mathbf{u}\|_2}.
\end{aligned}$$

A nearly identical derivation obtains

$$\left\| \frac{\mathbf{u}}{\|\mathbf{u}\|_2} - \frac{\mathbf{v}}{\|\mathbf{v}\|_2} \right\|_2 \leq \frac{2\|\mathbf{u}-\mathbf{v}\|_2}{\|\mathbf{v}\|_2}.$$

Combining the two acquired bounds implies the fact:

$$\left\| \frac{\mathbf{u}}{\|\mathbf{u}\|_2} - \frac{\mathbf{v}}{\|\mathbf{v}\|_2} \right\|_2 \leq 2 \min\left\{ \frac{\|\mathbf{u}-\mathbf{v}\|_2}{\|\mathbf{u}\|_2}, \frac{\|\mathbf{u}-\mathbf{v}\|_2}{\|\mathbf{v}\|_2} \right\},$$

as desired. ■

We now proceed to the proof of Lemma 11.

Proof Lemma 11 Consider any $\mathbf{u} \in \mathbb{R}^d \cap \Sigma_k^d$, $\mathbf{v} \in \mathbb{R}^d$, and $J \subseteq [d]$, $|J| \leq k$. Recall the notations of the coordinate subsets $J', J'' \subseteq [d]$, where $J' \triangleq \mathrm{supp}(\mathbf{u})$ and $J'' \triangleq \mathrm{supp}(T_k(\mathbf{v}))$. Due to the definition of J'',

$$\left\| \mathbf{u} - \frac{T_k(\mathbf{v})}{\|T_k(\mathbf{v})\|_2} \right\|_2 = \left\| \mathbf{u} - \frac{T_{J''}(\mathbf{v})}{\|T_{J''}(\mathbf{v})\|_2} \right\|_2.$$

Then,

$$\left\| \mathbf{u} - \frac{T_k(\mathbf{v})}{\|T_k(\mathbf{v})\|_2} \right\|_2 = \left\| \mathbf{u} - \frac{T_{J''}(\mathbf{v})}{\|T_{J''}(\mathbf{v})\|_2} \right\|_2$$

$$= \left\| \left(\mathbf{u} - \frac{T_{J\cup J'\cup J''}(\mathbf{v})}{\|T_{J\cup J'\cup J''}(\mathbf{v})\|_2} \right) + \left(\frac{T_{J\cup J'\cup J''}(\mathbf{v})}{\|T_{J\cup J'\cup J''}(\mathbf{v})\|_2} - \frac{T_{J''}(\mathbf{v})}{\|T_{J''}(\mathbf{v})\|_2} \right) \right\|_2$$
$$\leq \left\| \mathbf{u} - \frac{T_{J\cup J'\cup J''}(\mathbf{v})}{\|T_{J\cup J'\cup J''}(\mathbf{v})\|_2} \right\|_2 + \left\| \frac{T_{J\cup J'\cup J''}(\mathbf{v})}{\|T_{J\cup J'\cup J''}(\mathbf{v})\|_2} - \frac{T_{J''}(\mathbf{v})}{\|T_{J''}(\mathbf{v})\|_2} \right\|_2, \quad (38)$$

where the last line follows from the triangle inequality. Focusing in on the second term in the last line above, it follows from Fact 13 that

$$\left\| \frac{T_{J\cup J'\cup J''}(\mathbf{v})}{\|T_{J\cup J'\cup J''}(\mathbf{v})\|_2} - \frac{T_{J''}(\mathbf{v})}{\|T_{J''}(\mathbf{v})\|_2} \right\|_2 \leq \frac{2\|T_{J\cup J'\cup J''}(\mathbf{v}) - T_{J''}(\mathbf{v})\|_2}{\|T_{J\cup J'\cup J''}(\mathbf{v})\|_2}.$$

Note that $T_{J\cup J'\cup J''}(\mathbf{v}) - T_{J''}(\mathbf{v}) = T_{(J\cup J')\setminus J''}(\mathbf{v})$, and hence,

$$\left\| \frac{T_{J\cup J'\cup J''}(\mathbf{v})}{\|T_{J\cup J'\cup J''}(\mathbf{v})\|_2} - \frac{T_{J''}(\mathbf{v})}{\|T_{J''}(\mathbf{v})\|_2} \right\|_2 \leq \frac{2\|T_{J\cup J'\cup J''}(\mathbf{v}) - T_{J''}(\mathbf{v})\|_2}{\|T_{J\cup J'\cup J''}(\mathbf{v})\|_2} = \frac{2\|T_{(J\cup J')\setminus J''}(\mathbf{v})\|_2}{\|T_{J\cup J'\cup J''}(\mathbf{v})\|_2}. \quad (39)$$

Since $|J'| = |\mathrm{supp}(\mathbf{u})| \leq k$, the definitions of J' and the top-k thresholding operation imply that $|\mathrm{supp}(T_{J'}(\mathbf{v}))| \leq |\mathrm{supp}(T_{J''}(\mathbf{v}))|$, as well as that

$$\|T_{(J\cup J')\setminus J''}(\mathbf{v})\|_2 = \|T_{(J\cup J'\cup J'')\setminus J''}(\mathbf{v})\|_2 \leq \|T_{(J\cup J'\cup J'')\setminus J'}(\mathbf{v})\|_2 = \|T_{(J\cup J'')\setminus J'}(\mathbf{v})\|_2. \quad (40)$$

Additionally, observe:

$$\left\| \mathbf{u} - \frac{T_{J\cup J'\cup J''}(\mathbf{v})}{\|T_{J\cup J'\cup J''}(\mathbf{v})\|_2} \right\|_2^2 = \left\| \mathbf{u} - \frac{T_{J'}(\mathbf{v})}{\|T_{J\cup J'\cup J''}(\mathbf{v})\|_2} - \frac{T_{(J\cup J'')\setminus J'}(\mathbf{v})}{\|T_{J\cup J'\cup J''}(\mathbf{v})\|_2} \right\|_2^2$$
$$= \left\| \mathbf{u} - \frac{T_{J'}(\mathbf{v})}{\|T_{J\cup J'\cup J''}(\mathbf{v})\|_2} \right\|_2^2 + \left\| \frac{T_{(J\cup J'')\setminus J'}(\mathbf{v})}{\|T_{J\cup J'\cup J''}(\mathbf{v})\|_2} \right\|_2^2, \quad (41)$$

where the first equality holds since $T_{J\cup J'\cup J''}(\mathbf{v}) = T_{J'}(\mathbf{v}) + T_{(J\cup J'\cup J'')\setminus J'}(\mathbf{v}) = T_{J'}(\mathbf{v}) + T_{(J\cup J')\setminus J'}(\mathbf{v})$, and where the second equality is due to the orthogonality of $\mathbf{u} + wT_{J'}(\mathbf{v})$ and $T_{(J\cup J'')\setminus J'}(\mathbf{v})$ for any scalar $w \in \mathbb{R}$. This orthogonality is the result of disjoint support sets: $\mathrm{supp}(\mathbf{u} + wT_{J'}(\mathbf{v})) \cap \mathrm{supp}(T_{(J\cup J'')\setminus J'}(\mathbf{v})) \subseteq J' \cap ((J \cup J'') \setminus J') = \emptyset$. Rearranging the terms in Equation (41) and taking the square root yields:

$$\left\| \frac{T_{(J\cup J'')\setminus J'}(\mathbf{v})}{\|T_{J\cup J'\cup J''}(\mathbf{v})\|_2} \right\|_2 = \sqrt{\left\| \mathbf{u} - \frac{T_{J\cup J'\cup J''}(\mathbf{v})}{\|T_{J\cup J'\cup J''}(\mathbf{v})\|_2} \right\|_2^2 - \left\| \mathbf{u} - \frac{T_{J'}(\mathbf{v})}{\|T_{J\cup J'\cup J''}(\mathbf{v})\|_2} \right\|_2^2}. \quad (42)$$

From Equation (42), it follows that

$$\frac{\|T_{(J\cup J'')\setminus J'}(\mathbf{v})\|_2}{\|T_{J\cup J'\cup J''}(\mathbf{v})\|_2} = \sqrt{\left\| \mathbf{u} - \frac{T_{J\cup J'\cup J''}(\mathbf{v})}{\|T_{J\cup J'\cup J''}(\mathbf{v})\|_2} \right\|_2^2 - \left\| \mathbf{u} - \frac{T_{J'}(\mathbf{v})}{\|T_{J\cup J'\cup J''}(\mathbf{v})\|_2} \right\|_2^2} \quad (43)$$
$$\leq \left\| \mathbf{u} - \frac{T_{J\cup J'\cup J''}(\mathbf{v})}{\|T_{J\cup J'\cup J''}(\mathbf{v})\|_2} \right\|_2. \quad (44)$$

Combining Equations (39), (40), and (44),

$$\left\| \frac{T_{J\cup J'\cup J''}(\mathbf{v})}{\|T_{J\cup J'\cup J''}(\mathbf{v})\|_2} - \frac{T_{J''}(\mathbf{v})}{\|T_{J''}(\mathbf{v})\|_2} \right\|_2 \leq \frac{2\|T_{(J\cup J')\setminus J''}(\mathbf{v})\|_2}{\|T_{J\cup J'\cup J''}(\mathbf{v})\|_2}$$

$$
\begin{aligned}
&\qquad \blacktriangleright \text{ by Equation (39)} \\
&\le \frac{2\|T_{(J\cup J'')\setminus J'}(\mathbf{v})\|_2}{\|T_{J\cup J'\cup J''}(\mathbf{v})\|_2} \\
&\qquad \blacktriangleright \text{ by Equation (40)} \\
&\le 2\left\|\mathbf{u} - \frac{T_{J\cup J'\cup J''}(\mathbf{v})}{\|T_{J\cup J'\cup J''}(\mathbf{v})\|_2}\right\|_2 . \\
&\qquad \blacktriangleright \text{ by Equation (44)}
\end{aligned}
\tag{45}
$$

Now, returning to Equation (38), the proof is completed as follows:

$$
\begin{aligned}
\left\|\mathbf{u} - \frac{T_k(\mathbf{v})}{\|T_k(\mathbf{v})\|_2}\right\|_2 &\le \left\|\mathbf{u} - \frac{T_{J\cup J'\cup J''}(\mathbf{v})}{\|T_{J\cup J'\cup J''}(\mathbf{v})\|_2}\right\|_2 + \left\|\frac{T_{J\cup J'\cup J''}(\mathbf{v})}{\|T_{J\cup J'\cup J''}(\mathbf{v})\|_2} - \frac{T_{J''}(\mathbf{v})}{\|T_{J''}(\mathbf{v})\|_2}\right\|_2 \\
&\qquad \blacktriangleright \text{ by Equation (38)} \\
&\le \left\|\mathbf{u} - \frac{T_{J\cup J'\cup J''}(\mathbf{v})}{\|T_{J\cup J'\cup J''}(\mathbf{v})\|_2}\right\|_2 + 2\left\|\mathbf{u} - \frac{T_{J\cup J'\cup J''}(\mathbf{v})}{\|T_{J\cup J'\cup J''}(\mathbf{v})\|_2}\right\|_2 \\
&\qquad \blacktriangleright \text{ by Equation (45)} \\
&= 3\left\|\mathbf{u} - \frac{T_{J\cup J'\cup J''}(\mathbf{v})}{\|T_{J\cup J'\cup J''}(\mathbf{v})\|_2}\right\|_2 ,
\end{aligned}
$$

as desired. ∎

Appendix C. Proof of the Main Technical Results

C.1. Overview of the Proof of Theorem 9

The proof of the main technical theorem, Theorem 9, takes up the majority of the work in this manuscript. This section provides an overview of the proof. The proof in full is located in Appendix C with some auxiliary results therein proved in Appendix D. Before outlining the arguments, recall the definitions of Equations (18)–(21) from Appendix 4.1. Additionally, define the following related notations for $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$ and $J \subseteq [d]$:

$$
\bar{h}(\mathbf{u},\mathbf{v}) \triangleq h(\mathbf{u},\mathbf{v}) - \left\langle h(\mathbf{u},\mathbf{v}), \frac{\mathbf{u}-\mathbf{v}}{\|\mathbf{u}-\mathbf{v}\|_2}\right\rangle \frac{\mathbf{u}-\mathbf{v}}{\|\mathbf{u}-\mathbf{v}\|_2} - \left\langle h(\mathbf{u},\mathbf{v}), \frac{\mathbf{u}+\mathbf{v}}{\|\mathbf{u}+\mathbf{v}\|_2}\right\rangle \frac{\mathbf{u}+\mathbf{v}}{\|\mathbf{u}+\mathbf{v}\|_2}, \tag{46}
$$

$$
\bar{h}_J(\mathbf{u},\mathbf{v}) \triangleq T_{\mathrm{supp}(\mathbf{u})\cup\mathrm{supp}(\mathbf{v})\cup J}\left(\bar{h}(\mathbf{u},\mathbf{v})\right), \tag{47}
$$

$$
\bar{h}_f(\mathbf{u},\mathbf{u}) \triangleq h_f(\mathbf{u},\mathbf{u}) - \langle h_f(\mathbf{u},\mathbf{u}), \mathbf{u}\rangle \mathbf{u}, \tag{48}
$$

$$
\bar{h}_{f;J}(\mathbf{u},\mathbf{u}) \triangleq T_{\mathrm{supp}(\mathbf{u})\cup\mathrm{supp}(\mathbf{v})\cup J}\left(\bar{h}_f(\mathbf{u},\mathbf{u})\right). \tag{49}
$$

Note that

$$
\bar{h}_J(\mathbf{u},\mathbf{v}) = h_J(\mathbf{u},\mathbf{v}) - \left\langle h_J(\mathbf{u},\mathbf{v}), \frac{\mathbf{u}-\mathbf{v}}{\|\mathbf{u}-\mathbf{v}\|_2}\right\rangle \frac{\mathbf{u}-\mathbf{v}}{\|\mathbf{u}-\mathbf{v}\|_2} - \left\langle h_J(\mathbf{u},\mathbf{v}), \frac{\mathbf{u}+\mathbf{v}}{\|\mathbf{u}+\mathbf{v}\|_2}\right\rangle \frac{\mathbf{u}+\mathbf{v}}{\|\mathbf{u}+\mathbf{v}\|_2}
$$

and that

$$
\bar{h}_{f;J}(\mathbf{u},\mathbf{u}) = h_{f;J}(\mathbf{u},\mathbf{u}) - \langle h_{f;J}(\mathbf{u},\mathbf{u}), \mathbf{u}\rangle \mathbf{u}.
$$

C.1.1. KEY STEPS OF THE PROOF

The proof of Theorem 9 are sketched as follows.

1. Recall that the aim is to bound
$$\left\| \boldsymbol{\theta}^* - \frac{\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})}{\|\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})\|_2} \right\|_2 \tag{50}$$
from above with high probability uniformly for all $\hat{\boldsymbol{\theta}} \in \Theta$ and all $J \subseteq [d]$, $|J| \leq k$.

2. To obtain a uniform result, a τ-net, $\mathcal{C} \subset \Theta$, over the parameter space, Θ, is constructed with a particular design, the details of which are left to the formal proof of Theorem 9. For the purpose of this overview, its suffices to say that, crucially, the design of $\mathcal{C}$ ensures that for each $\hat{\boldsymbol{\theta}} \in \Theta$, there exists an element, $\boldsymbol{\theta} \in \mathcal{C}$, such that both $\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_2 \leq \tau$ and $\mathrm{supp}(\boldsymbol{\theta}) = \mathrm{supp}(\hat{\boldsymbol{\theta}})$. This cover, $\mathcal{C}$, will allow the establishment of a global result for points, $\boldsymbol{\theta} \in \mathcal{C}$, within it, which can subsequently be extended to arbitrary points, $\hat{\boldsymbol{\theta}} \in \Theta$, in the entire parameter space via a local analysis.

3. As another preliminary step, it will be shown that for any $\hat{\boldsymbol{\theta}} \in \Theta$ and $J \subseteq [d]$,
$$\frac{\mathbb{E}[\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})]}{\| \mathbb{E}[\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})]\|_2} = \boldsymbol{\theta}^*.$$
In other words, the quantity in (50)—which we seek to bound—describes a notion of deviation of $\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})$ from its mean (after normalization):
$$\left\| \boldsymbol{\theta}^* - \frac{\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})}{\|\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})\|_2} \right\|_2 = \left\| \frac{\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})}{\|\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})\|_2} - \frac{\mathbb{E}[\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})]}{\| \mathbb{E}[\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})]\|_2} \right\|_2.$$
In fact, this deviation turns out to roughly scale with the deviation of the random function $h_{f;J}$ around its mean:
$$\left\| \frac{\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})}{\|\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})\|_2} - \frac{\mathbb{E}[\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})]}{\| \mathbb{E}[\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})]\|_2} \right\|_2 \propto \|h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) - \mathbb{E}[h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})]\|_2. \tag{51}$$
Analyzing (a decomposition of) the deviation of $h_{f;J}$ will be at the core of the proof.

4. Letting $\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}} \in \Theta$ be arbitrary, and using the observations in Step 3, the triangle inequality, algebraic manipulations, and other standard techniques, the expression in (50) is bounded by the sum of three terms which will admit an easier analysis than directly handling (50):
$$\left\| \boldsymbol{\theta}^* - \frac{\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})}{\|\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})\|_2} \right\|_2 \leq \frac{2\|h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}) - \mathbb{E}[h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})]\|_2}{\| \mathbb{E}[\boldsymbol{\theta} + h_f(\boldsymbol{\theta}^*, \boldsymbol{\theta})]\|_2} \tag{52a}$$
$$+ \frac{2\|h_{\mathrm{supp}(\boldsymbol{\theta}^*) \cup J}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) - \mathbb{E}[h_{\mathrm{supp}(\boldsymbol{\theta}^*) \cup J}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})]\|_2}{\| \mathbb{E}[\boldsymbol{\theta} + h_f(\boldsymbol{\theta}^*, \boldsymbol{\theta})]\|_2} \tag{52b}$$

$$+\frac{2\|h_{f;\operatorname{supp}(\boldsymbol{\theta})\cup J}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)-\mathbb{E}[h_{f;\operatorname{supp}(\boldsymbol{\theta})\cup J}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)]\|_2}{\|\mathbb{E}[\boldsymbol{\theta}+h_f(\boldsymbol{\theta}^*,\boldsymbol{\theta})]\|_2}, \tag{52c}$$

where $J \subseteq [d]$, $|J| \le k$, is arbitrary, and where $\boldsymbol{\theta} \in \mathcal{C} \setminus \mathcal{B}_\tau(\boldsymbol{\theta}^*)$ such that $\|\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\|_2 \le 2\tau$ and $\operatorname{supp}(\boldsymbol{\theta}) \cup J = \operatorname{supp}(\hat{\boldsymbol{\theta}}) \cup J$ (see, Lemma 18). Per the design of the τ-net, $\mathcal{C} \subset \Theta$, in Step 2, such a point $\boldsymbol{\theta} \in \mathcal{C}$ exists for any choice of $\hat{\boldsymbol{\theta}} \in \Theta$.

5. The three terms on the right-hand-side of Equation (52) can be viewed as bounding (50) by relating it (with appropriate scaling) to the deviation of h_f specified on the right-hand-side of (51), and then controlling the right-hand-side of (51) by decomposing the deviation of h_f into three components of deviation, in order: (52a), a component handling points in the cover, $\mathcal{C}$, over Θ that are sufficiently far from $\boldsymbol{\theta}^*$—a "global" result; (52b), a component reconciling the discrepancy between the original point, $\hat{\boldsymbol{\theta}} \in \Theta$, which may be outside the cover, and a nearby neighbor in the cover, $\boldsymbol{\theta} \in \mathcal{C} \setminus \mathcal{B}_\tau(\boldsymbol{\theta}^*)$—a "local" result; and (52c), a component handling the "noise" introduced into the GLM through the randomness of f.

6. The technical work in this manuscript then lies largely with bounding the three terms on the right-hand-side of Equation (52). While most of the details of this analysis are left to the formal proofs (see, Appendices C.6–D), a few salient ideas in the approach are mentioned here.

7. For the three terms, (52a)–(52c), the (shared) denominator can be calculated directly.

8. On the other hand, the numerators in (52a)–(52c) are upper bounded with bounded probability through concentration inequalities derived with standard techniques. To do so, each numerator is orthogonally decomposed into two to three components for which derivations of concentration inequalities are easier. Subsequently, for each numerator, the concentration inequalities for its associated components are combined via the triangle inequality. Then, these are extended into uniform results by appropriate union bounds.

9. For the first and last terms, (52a) and (52c), the union bounds are straightforward: simply taken over the coordinate subsets of cardinality at most k, as well as, in the case of (52a), over the cover, $\mathcal{C}$.

10. In contrast, the second term, (52b), requires a more careful—and somewhat indirect—argument. In this case, the union bound is taken over the set $\{h_{\operatorname{supp}(\boldsymbol{\theta}^*)\cup J}(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}) : \hat{\boldsymbol{\theta}} \in \mathcal{B}'_{2\tau}(\boldsymbol{\theta})\}$, which has a sufficiently small cardinality due to the local binary embeddings of (Oymak and Recht, 2015, Corollary 3.3).

11. Using the uniform results obtained in Steps 7–10, the number of covariates, n, can then be determined such that desired bounds on the terms (52a)–(52c), and hence also the desired bound on (50), hold uniformly with high probability. This will establish the invertibility condition for Gaussian covariate matrices claimed in Theorem 9.

C.2. Detailed Proofs

Several constants will appear throughout this section. For convenient reference later, they are specified in the following definition.

Definition 15 *Let $a_1, a_2, a_3, c, c' > 0$ be (absolute) constants such that $a_1 \triangleq \frac{a_2}{a_3^2} \geq \frac{1}{50}$, $\sqrt{8}a_2 < \frac{c'}{2}$, $a_3 \triangleq \frac{c'}{2} - \sqrt{8}a_2 = \Omega(1)$, $c < 1 - \sqrt{\frac{2a_2}{c_4}}$, and $c' \leq 1 - c - \sqrt{\frac{2a_2}{c_4}}$. Additionally, let $a, b, c_1, c_2, c_3, c_4 > 0$ be the (absolute) constants given by: $a = 24$, $b = 3$, $c_1 = \frac{192}{c}$, $c_2 = \sqrt{\frac{800}{\pi}} a_1$, $c_3 = 64$, and $c_4 = 256$.*

Two additional notations will be used in this manuscript, which are introduced in Definition 16, below.

Definition 16 *For $\delta > 0$, let $\eta(\delta), \tau(\delta) > 0$ be given by*

$$\eta(\delta) = \frac{\gamma a_2 \delta}{\sqrt{\frac{2}{\pi} \log\left(\frac{4e}{\eta(\delta)}\right)}}, \tag{53}$$

and

$$\tau(\delta) \triangleq \frac{\eta(\delta)}{c_4 \log\left(\frac{2e}{\eta(\delta)}\right)}, \tag{54}$$

where $c_4 > 0$ is given in Definition 15. To condense notation, the explicit parameterization by δ will in general be dropped and left implicit in this manuscript, i.e., $\eta = \eta(\delta)$ and $\tau = \tau(\delta)$, where the specific choice of δ may vary but will be clear from the context.

We restate Theorem 9 below with the sample complexity more specific with the above-defined constants.

Theorem 17 (Restatement of Theorem 9) *Let $a, b, c_1, c_1, c_2, c_3, c_4 > 0$ be absolute constants as specified in Definition 15, and fix $d, k, n \in \mathbb{Z}_+$, $k \leq d$, and $\rho, \delta \in (0,1)$ where*

$$\delta \triangleq \frac{\epsilon}{\frac{3}{2}(5 + \sqrt{21})}. \tag{55}$$

Set $\eta = \eta(\delta) > 0$ and let $\tau = \tau(\delta) > 0$ as in Definition 16. Write $\alpha_0 = \alpha_0(\delta) \triangleq \max\{\alpha, \delta\}$ as in Equation (10). Let $\Theta = S^{d-1} \cap \Sigma_k^d$, and fix $\boldsymbol{\theta}^* \in \Theta$. Under Assumption 2, if

$$\begin{aligned}
n \geq \max \Bigg\{ & \frac{c_1 \alpha_0}{\gamma^2 \delta^2} \log\left(\frac{a}{\rho} \sum_{\ell=0}^{k} \binom{d}{\ell} \sum_{\ell'=0}^{k} \binom{d}{\ell'} \left(\frac{b}{\tau}\right)^{\ell'}\right), \\
& \frac{c_2}{\gamma \delta \sqrt{\log\left(\frac{4e}{\eta}\right)}} \log\left(\frac{a}{\rho} \sum_{\ell=0}^{k} \binom{d}{\ell} \sum_{\ell'=0}^{k} \binom{d}{\ell'} \left(\frac{b}{\tau}\right)^{\ell'}\right), \\
& \frac{c_3}{\eta} \log\left(\frac{a}{\rho}\right), \frac{c_4 k}{\eta} \log\left(\frac{1}{\eta}\right) \Bigg\} \\
= O\Bigg(& \max\left\{ \frac{\alpha_0 k}{\gamma^2 \delta^2} \log\left(\frac{d}{\delta k}\right) + \frac{\alpha_0}{\gamma^2 \delta^2} \log\left(\frac{1}{\rho}\right), \frac{k}{\delta} \log^{3/2}\left(\frac{1}{\delta}\right), \frac{1}{\delta} \sqrt{\log\left(\frac{1}{\delta}\right)} \log\left(\frac{1}{\rho}\right) \right\} \Bigg),
\end{aligned} \tag{56}$$

then with probability at least $1-\rho$*, uniformly for all* $\hat{\boldsymbol{\theta}} \in \Theta$ *and all* $J \subseteq [d]$, $|J| \leq k$,

$$\left\| \boldsymbol{\theta}^* - \frac{\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})}{\|\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})\|_2} \right\|_2 \leq \sqrt{\delta \|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\|_2} + \delta. \tag{57}$$

C.3. Intermediate Results for the Proof of Theorem 9

Lemmas 18–21, stated below in this section, lay the groundwork for proving the main technical theorem, Theorem 9. The proofs of these intermediate results can be found in Appendix C.6. Recall that the ultimate goal is to uniformly bound

$$\left\| \boldsymbol{\theta}^* - \frac{\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})}{\|\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})\|_2} \right\|_2 \tag{58}$$

from above. Lemma 18 starts off by upper bounding (58) by the sum of three terms (with some scaling), each of which describes how much the functions h and h_f (with thresholding) deviate from their means. Subsequently, Lemmas 19–21 provide bounds on these deviations.

Lemma 18 *Let* $J \subseteq [d]$*, and fix* $\boldsymbol{\theta}^*, \boldsymbol{\theta}, \hat{\boldsymbol{\theta}} \in \Theta$ *such that* $\mathrm{supp}(\hat{\boldsymbol{\theta}}) \cup J = \mathrm{supp}(\boldsymbol{\theta}) \cup J$*. Then,*

$$\begin{aligned}\left\| \boldsymbol{\theta}^* - \frac{\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})}{\|\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})\|_2} \right\|_2 \leq{} & \frac{2\|h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}) - \mathbb{E}[h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})]\|_2}{\|\mathbb{E}[\boldsymbol{\theta} + h_f(\boldsymbol{\theta}^*, \boldsymbol{\theta})]\|_2} \\ & + \frac{2\|h_{\mathrm{supp}(\boldsymbol{\theta}^*)\cup J}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) - \mathbb{E}[h_{\mathrm{supp}(\boldsymbol{\theta}^*)\cup J}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})]\|_2}{\|\mathbb{E}[\boldsymbol{\theta} + h_f(\boldsymbol{\theta}^*, \boldsymbol{\theta})]\|_2} \\ & + \frac{2\|h_{f;\mathrm{supp}(\boldsymbol{\theta})\cup J}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) - \mathbb{E}[h_{f;\mathrm{supp}(\boldsymbol{\theta})\cup J}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)]\|_2}{\|\mathbb{E}[\boldsymbol{\theta} + h_f(\boldsymbol{\theta}^*, \boldsymbol{\theta})]\|_2}. \end{aligned} \tag{59}$$

Lemma 18 motivates three additional results, presented next in Lemmas 19–21. Note that whereas Lemma 18 holds deterministically, Lemmas 19–21 are probabilistic results.

Lemma 19 *Let* $\rho_1, \delta \in (0,1)$*, and define* $\tau = \tau(\delta)$ *according to Definition 16. Fix* $\boldsymbol{\theta}^* \in \Theta$*, and let* $\mathcal{J} \subseteq 2^{[d]}$ *and* $\mathcal{C} \subset \Theta$ *be finite sets. Define* $k_0 \triangleq \min\{2k + \max_{J\in\mathcal{J}} |J|, d\}$*. If*

$$n \geq \frac{16}{\gamma^2 \delta} \max\left\{ 27\pi \log\left(\frac{12}{\rho_1} |\mathcal{J}||\mathcal{C}| \right), 4(k_0 - 2) \right\}, \tag{60}$$

then with probability at least $1 - \rho_1$*, uniformly for all* $J \in \mathcal{J}$ *and all* $\boldsymbol{\theta} \in \mathcal{C} \setminus \mathcal{B}_\tau(\boldsymbol{\theta}^*)$*,*

$$\frac{2\|h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}) - \mathbb{E}[h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})]\|_2}{\|\mathbb{E}[\boldsymbol{\theta} + h_f(\boldsymbol{\theta}^*, \boldsymbol{\theta})]\|_2} \leq \sqrt{\delta \|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2}. \tag{61}$$

Lemma 20 *Let* $c', c_4 > 0$ *be constants specified in Definition 15. Let* $\rho_2, \delta \in (0,1)$*, and define* $\eta = \eta(\delta)$ *and* $\tau = \tau(\delta)$ *according to Definition 16. Let* $\mathcal{C} \subset \Theta$ *be a finite set, and fix* $\boldsymbol{\theta}^* \in \Theta$*. Let* $\mathcal{J}, \mathcal{J}' \subseteq 2^{[d]}$*, where* $\mathcal{J}' \triangleq \{\mathrm{supp}(\boldsymbol{\theta}^*) \cup J : J \in \mathcal{J}\}$*. Set* $k_0' \triangleq \min\{\max\{2k, \max_{J''\in\mathcal{J}'} |J''|\}, d\}$*. If*

$$n \geq \max\left\{ \frac{200\eta \log\left(\frac{6}{\rho_2}|\mathcal{J}||\mathcal{C}|\right)}{\left(\sqrt{\frac{\pi}{8}}\gamma c' \delta - \eta\sqrt{8\log\left(\frac{e}{\eta}\right)}\right)^2}, \frac{200\eta k_0'}{\pi \gamma^2 c'^2 \delta^2}, \frac{64}{\eta} \log\left(\frac{6}{\rho_2}\binom{d}{k}\right), \frac{c_4 k}{\eta}\log\left(\frac{1}{\eta}\right) \right\}, \tag{62}$$

then with probability at least $1-\rho_2$*, uniformly for all* $J' \in \mathcal{J}'$, $\boldsymbol{\theta} \in \mathcal{C} \setminus \mathcal{B}_\tau(\boldsymbol{\theta}^*)$*, and* $\hat{\boldsymbol{\theta}} \in \mathcal{B}'_{2\tau}(\boldsymbol{\theta})$*,*

$$\frac{2\|h_{J'}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) - \mathbb{E}[h_{J'}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})]\|_2}{\|\,\mathbb{E}[\boldsymbol{\theta} + h_f(\boldsymbol{\theta}^*, \boldsymbol{\theta})]\|_2} \leq c'\delta. \tag{63}$$

Lemma 21 *Let* $c > 0$ *be a constant specified in Definition 15. Let* $\rho_3, \delta \in (0,1)$*, and define* $\alpha_0 = \alpha_0(\delta) \triangleq \max\{\alpha, \delta\}$*. Fix* $\boldsymbol{\theta}^* \in \Theta$*. Let* $\mathcal{J} \subseteq 2^{[d]}$ *and* $\mathcal{C} \subset \Theta$ *be finite sets, and let* $\mathcal{J}'' \triangleq \{\mathrm{supp}(\boldsymbol{\theta}) \cup J : \boldsymbol{\theta} \in \mathcal{C}, J \in \mathcal{J}\}$*. Define* $k''_0 \triangleq \min\{k + \max_{J'' \in \mathcal{J}''} |J''|, d\}$*. If*

$$n \geq \max\left\{ \frac{64\alpha_0}{\gamma^2 c^2 \delta^2} \max\left\{ 3\log\left(\frac{6}{\rho_3}|\mathcal{J}||\mathcal{C}|\right), 2(k''_0 - 1)\right\}, \frac{4}{\alpha_0}\log\left(\frac{6}{\rho_3}|\mathcal{J}||\mathcal{C}|\right)\right\}, \tag{64}$$

then with probability at least $1-\rho_3$*, uniformly for all* $J'' \in \mathcal{J}''$*,*

$$\frac{2\|h_{f;J''}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) - \mathbb{E}[h_{f;J''}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)]\|_2}{\|\,\mathbb{E}[\boldsymbol{\theta} + h_f(\boldsymbol{\theta}^*, \boldsymbol{\theta})]\|_2} \leq c\delta, \tag{65}$$

for every $\boldsymbol{\theta} \in \Theta$*.*

C.4. Proof of Theorem 9

An important (and standard) construct for the analysis in this work is a τ-net. Fix $\tau > 0$. Let $(\mathcal{S}, d_{\mathcal{S}})$ be a metric space. A subset, $\mathcal{T} \subseteq \mathcal{S}$, is a τ*-net* over $\mathcal{S}$ if $\inf_{t \in \mathcal{T}} d_{\mathcal{S}}(s,t) \leq \tau$ for all $s \in \mathcal{S}$. We will use the following upper bound on the minimal cardinality of a τ-net of a sphere.

Lemma 22 (see, e.g., Vershynin (2018)) *Fix* $\tau > 0$*, and let* $m \in \mathbb{Z}_+$*. There exists an* ℓ_2 τ*-net,* $\mathcal{T} \subset S^{m-1}$*, over* S^{m-1} *of cardinality not exceeding* $|\mathcal{T}| \leq (\frac{3}{\tau})^m$*.*

Proof Theorem 9 Fix $\boldsymbol{\theta}^* \in \Theta$. Let $\mathcal{J} \subseteq 2^{[d]}$ be the set of coordinate subsets with cardinality at most k—that is, the set given by $\mathcal{J} \triangleq \{J \subseteq [d] : |J| \leq k\}$. Construct a τ-net, $\mathcal{C} \subset \Theta$, over Θ with the following design. For each $J \in \mathcal{J}$, let $\mathcal{C}_J \subset \Theta$ be a τ-net over the set of points in Θ whose support is a subset of J—formally, over the set $\{\mathbf{v} \in \Theta : \mathrm{supp}(\mathbf{v}) \subseteq J\}$—such that each vector in the cover, $\mathcal{C}_J$, has support exactly J, i.e., $\mathrm{supp}(\mathbf{v}) = J$ for all $\mathbf{v} \in \mathcal{C}_J$. (This last condition on the support of elements in the τ-net is possible through a rotation.) Then, let $\mathcal{C} \triangleq \bigcup_{J \in \mathcal{J}} \mathcal{C}_J$. Note that this construction ensures that for every point in the parameter space, Θ, the cover, $\mathcal{C}$, contains at least one point within distance τ of it and with precisely the same support. Additionally, the cardinalities of $\mathcal{J}$ and $\mathcal{C}$ satisfy $|\mathcal{J}| = \sum_{\ell=0}^{k} \binom{d}{\ell}$ and $|\mathcal{C}| \leq \sum_{J \in \mathcal{J}} (\frac{3}{\tau})^{|J|} = \sum_{\ell=0}^{k} \binom{d}{\ell} \left(\frac{3}{\tau}\right)^{\ell}$, where the bound on $|\mathcal{C}|$ is due to Lemma 22 combined with a union bound.

Consider an arbitrary choice of $\hat{\boldsymbol{\theta}} \in \Theta$, to later be varied over the entire parameter space, Θ, and let $\boldsymbol{\theta} \in \mathcal{C}$ satisfy $\boldsymbol{\theta} \notin \mathcal{B}_\tau(\boldsymbol{\theta}^*)$ and $\hat{\boldsymbol{\theta}} \in \mathcal{B}'_{2\tau}(\boldsymbol{\theta})$, where such a point, $\boldsymbol{\theta}$, exists in the τ-net, $\mathcal{C}$, by its design. Note that this ensures that $\mathrm{supp}(\boldsymbol{\theta}) \cup J = \mathrm{supp}(\hat{\boldsymbol{\theta}}) \cup J$ for all $J \in \mathcal{J}$, and hence, by Lemma 18,

$$\left\| \boldsymbol{\theta}^* - \frac{\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})}{\|\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})\|_2} \right\|_2 \leq \frac{2\|h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}) - \mathbb{E}[h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})]\|_2}{\|\,\mathbb{E}[\boldsymbol{\theta} + h_f(\boldsymbol{\theta}^*, \boldsymbol{\theta})]\|_2}$$
$$+ \frac{2\|h_{\mathrm{supp}(\boldsymbol{\theta}^*) \cup J}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) - \mathbb{E}[h_{\mathrm{supp}(\boldsymbol{\theta}^*) \cup J}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})]\|_2}{\|\,\mathbb{E}[\boldsymbol{\theta} + h_f(\boldsymbol{\theta}^*, \boldsymbol{\theta})]\|_2}$$

$$+\frac{2\|h_{f;\mathrm{supp}(\boldsymbol{\theta})\cup J}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)-\mathbb{E}[h_{f;\mathrm{supp}(\boldsymbol{\theta})\cup J}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)]\|_2}{\|\,\mathbb{E}[\boldsymbol{\theta}+h_f(\boldsymbol{\theta}^*,\boldsymbol{\theta})]\|_2}, \tag{66}$$

where this bound holds deterministically. Now, suppose

$$\begin{aligned} n \geq \max\Bigg\{ & \frac{c_1\alpha_0}{\gamma^2\delta^2}\log\left(\frac{a}{\rho}\sum_{\ell=0}^{k}\binom{d}{\ell}\sum_{\ell'=0}^{k}\binom{d}{\ell'}\left(\frac{b}{\tau}\right)^{\ell'}\right), \\ & \frac{c_2}{\gamma\delta\sqrt{\log\left(\frac{4e}{\eta}\right)}}\log\left(\frac{a}{\rho}\sum_{\ell=0}^{k}\binom{d}{\ell}\sum_{\ell'=0}^{k}\binom{d}{\ell'}\left(\frac{b}{\tau}\right)^{\ell'}\right), \\ & \frac{c_3}{\eta}\log\left(\frac{a}{\rho}\right), \frac{c_4 k}{\eta}\log\left(\frac{1}{\eta}\right)\Bigg\}. \end{aligned}$$

This choice of n is sufficiently large so that taking $\rho_1=\frac{\rho}{2}$, $\rho_2=\frac{\rho}{4}$, and $\rho_3=\frac{\rho}{4}$—such that $\rho_1+\rho_2+\rho_3=\rho$—in Lemmas 19–21, respectively, and then combining the bounds in these lemmas with a union bound, the three terms on the right-hand-side of the inequality in Equation (66) are simultaneously bounded from above with probability at least $1-\rho_1-\rho_2-\rho_3=1-\rho$ by first,

$$\sup_{\substack{J\in\mathcal{J},\\ \boldsymbol{\theta}\in\mathcal{C}\setminus\mathcal{B}_\tau(\boldsymbol{\theta}^*)}}\frac{2\|h_J(\boldsymbol{\theta}^*,\boldsymbol{\theta})-\mathbb{E}[h_J(\boldsymbol{\theta}^*,\boldsymbol{\theta})]\|_2}{\|\,\mathbb{E}[\boldsymbol{\theta}+h_f(\boldsymbol{\theta}^*,\boldsymbol{\theta})]\|_2}\leq\sqrt{\delta\|\boldsymbol{\theta}^*-\boldsymbol{\theta}\|_2},$$

second,

$$\begin{aligned} &\sup_{\substack{J\in\mathcal{J},\\ \boldsymbol{\theta}\in\mathcal{C}\setminus\mathcal{B}_\tau(\boldsymbol{\theta}^*),\\ \hat{\boldsymbol{\theta}}\in\mathcal{B}'_{2\tau}(\boldsymbol{\theta})}}\frac{2\|h_{\mathrm{supp}(\boldsymbol{\theta}^*)\cup J}(\boldsymbol{\theta},\hat{\boldsymbol{\theta}})-\mathbb{E}[h_{\mathrm{supp}(\boldsymbol{\theta}^*)\cup J}(\boldsymbol{\theta},\hat{\boldsymbol{\theta}})]\|_2}{\|\,\mathbb{E}[\boldsymbol{\theta}+h_f(\boldsymbol{\theta}^*,\boldsymbol{\theta})]\|_2} \\ &\leq \sup_{\substack{J'\in\mathcal{J}',\\ \boldsymbol{\theta}\in\mathcal{C}\setminus\mathcal{B}_\tau(\boldsymbol{\theta}^*),\\ \hat{\boldsymbol{\theta}}\in\mathcal{B}'_{2\tau}(\boldsymbol{\theta})}}\frac{2\|h_{J'}(\boldsymbol{\theta},\hat{\boldsymbol{\theta}})-\mathbb{E}[h_{J'}(\boldsymbol{\theta},\hat{\boldsymbol{\theta}})]\|_2}{\|\,\mathbb{E}[\boldsymbol{\theta}+h_f(\boldsymbol{\theta}^*,\boldsymbol{\theta})]\|_2} \\ &\leq c'\delta \\ &\leq\left(1-c-\sqrt{\frac{2a_2}{c_4}}\right)\delta \\ &\qquad \blacktriangleright \text{ by Definition 15} \\ &\leq\left(1-c-\sqrt{\frac{2\tau}{\delta}}\right)\delta, \\ &\qquad \blacktriangleright \text{ by the definitions of } \delta,\tau \end{aligned}$$

and third,

$$\sup_{\substack{J\in\mathcal{J},\\ \boldsymbol{\theta}\in\mathcal{C}\setminus\mathcal{B}_\tau(\boldsymbol{\theta}^*)}}\frac{2\|h_{f;\mathrm{supp}(\boldsymbol{\theta})\cup J}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)-\mathbb{E}[h_{f;\mathrm{supp}(\boldsymbol{\theta})\cup J}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)]\|_2}{\|\,\mathbb{E}[\boldsymbol{\theta}+h_f(\boldsymbol{\theta}^*,\boldsymbol{\theta})]\|_2}\leq\sup_{J''\in\mathcal{J}''}\frac{2\|h_{f;J''}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)-\mathbb{E}[h_{f;J''}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)]\|_2}{\|\,\mathbb{E}[\boldsymbol{\theta}+h_f(\boldsymbol{\theta}^*,\boldsymbol{\theta})]\|_2}\leq c\delta,$$

where $\mathcal{J}' \triangleq \{\mathrm{supp}(\boldsymbol{\theta}^*) \cup J : J \in \mathcal{J}\}$ and $\mathcal{J}'' \triangleq \{\mathrm{supp}(\boldsymbol{\theta}) \cup J : \boldsymbol{\theta} \in \mathcal{C}, J \in \mathcal{J}\}$. It follows that under the stated condition on n, with probability at least $1-\rho$, for all $\hat{\boldsymbol{\theta}} \in \Theta$ and $J \in \mathcal{J}$,

$$\begin{aligned}
\left\| \boldsymbol{\theta}^* - \frac{\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})}{\|\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})\|_2} \right\|_2 &\leq \sqrt{\delta \|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2} + \left(1 - c - \sqrt{\frac{2\tau}{\delta}}\right)\delta + c\delta \\
&\qquad \blacktriangleright \text{ for some } \boldsymbol{\theta} \in (\mathcal{C} \cap \mathcal{B}'_{2\tau}(\hat{\boldsymbol{\theta}})) \setminus \mathcal{B}_\tau(\boldsymbol{\theta}^*) \\
&= \sqrt{\delta \|(\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}) - (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})\|_2} + \left(1 - c - \sqrt{\frac{2\tau}{\delta}}\right)\delta + c\delta \\
&\leq \sqrt{\delta \|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\|_2} + \sqrt{\delta \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_2} + \left(1 - c - \sqrt{\frac{2\tau}{\delta}}\right)\delta + c\delta \\
&\qquad \blacktriangleright \text{ by the triangle inequality} \\
&\leq \sqrt{\delta \|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\|_2} + \sqrt{2\delta\tau} + \left(1 - c - \sqrt{\frac{2\tau}{\delta}}\right)\delta + c\delta \\
&\qquad \blacktriangleright \because \boldsymbol{\theta} \in \mathcal{B}'_{2\tau}(\hat{\boldsymbol{\theta}}) \\
&= \sqrt{\delta \|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\|_2} + \sqrt{\frac{2\tau}{\delta}}\delta + \left(1 - c - \sqrt{\frac{2\tau}{\delta}}\right)\delta + c\delta \\
&= \sqrt{\delta \|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\|_2} + \delta,
\end{aligned}$$

as claimed. ■

C.5. Proof of Corollary 10

Proof Corollary 10 The specialization of the main technical result to logistic regression in Corollary 10 requires two arguments: (a) Assumption 2 needs to be shown to hold for logistic regression, i.e., when p is the logistic function with inverse temperature $\beta > 0$, as in Definition 3; and (b) explicit forms for the variables α (and α_0) and γ need specification. Once these are achieved, the corollary will follow from combining the bounds on α and γ obtained from Task (b) with Theorem 9. Throughout this proof, p is taken to be the logistic function, parameterized by the inverse temperature, $\beta > 0$, per Definition 3, which is recalled for convenience:

$$p(z) = \frac{1}{1 + e^{-\beta z}}.$$

For Task (a), recall that Assumption 2 imposes two conditions: (i) that p is nondecreasing over the entire real line, and (ii) that $\frac{\partial}{\partial z} \frac{1-p(z+w)+p(-(z+w))}{1-p(z)+p(-z)} \leq 0$. Let $z < z' \in \mathbb{R}$. Then, clearly,

$$p(z) = \frac{1}{1 + e^{-\beta z}} < \frac{1}{1 + e^{-\beta z'}} = p(z'),$$

and thus, Condition (i) holds. On the other hand, Condition (ii) can be established via basis calculus. First, note that for any $z \in \mathbb{R}$,

$$1 - p(z) = 1 - \frac{1}{1 + e^{-\beta z}} = \frac{e^{-\beta z}}{1 + e^{-\beta z}} = \frac{1}{1 + e^{\beta z}} = p(-z), \tag{67}$$

and hence, for $w, z \in \mathbb{R}$, $w > 0$,

$$\frac{1 - p(z+w) + p(-(z+w))}{1 - p(z) + p(-z)} = \frac{2p(-(z+w))}{2p(-z)} = \frac{p(-(z+w))}{p(-z)} = \frac{1 + e^{\beta z}}{1 + e^{\beta(z+w)}}.$$

Then,

$$\frac{\partial}{\partial z} \frac{1 - p(z+w) + p(-(z+w))}{1 - p(z) + p(-z)} = \frac{\partial}{\partial z} \frac{1 + e^{\beta z}}{1 + e^{\beta(z+w)}} = -\frac{\beta e^{\beta z}(e^{\beta w} - 1)}{(1 + e^{\beta(z+w)})^2} \leq 0,$$

as desired. Thus, Condition (ii) also holds when p is the logistic function. This complete Task (a).

Proceeding to Task (b), the aim now is to derive closed-form bounds on α and γ. Looking first at α, recall its definition from Equation (8):

$$\alpha \triangleq \underset{Z \sim \mathcal{N}(0,1)}{P} (f(Z) \neq \operatorname{sign}(Z)) = \frac{1}{\sqrt{2\pi}} \int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2} (1 - p(z) + p(-z)) dz.$$

By the earlier observation in (67), when p is the logistic function,

$$\begin{aligned}
\alpha &= \frac{1}{\sqrt{2\pi}} \int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2} (1 - p(z) + p(-z)) dz \\
&\qquad \blacktriangleright \text{ by Equation (8)} \\
&= \sqrt{\frac{2}{\pi}} \int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2} p(-z) dz \\
&\qquad \blacktriangleright \text{ by Equation (67)} \\
&= \mathbb{E}\left[p(-|Z|)\right] \\
&\qquad \blacktriangleright \text{ by the law of the lazy statistician and the density function} \\
&\qquad\quad \text{for standard half-normal random variables} \\
&= \mathbb{E}\left[\frac{1}{1 + e^{\beta|Z|}}\right], \qquad (68)
\end{aligned}$$

where $Z \sim \mathcal{N}(0,1)$ is a standard univariate Gaussian random variable. Note that

$$p(0) = \frac{1}{1 + e^0} = \frac{1}{2}. \qquad (69)$$

Hence, when $\beta = 0$, Equation (68) trivially evaluates to $\alpha = \frac{1}{2}$. To bound Equation (68) when $\beta > 0$, we can directly apply the following result from Hsu and Mazumdar (2024).

Lemma 23 ((Hsu and Mazumdar, 2024, Lemma 13)) *Fix $r > 0$, and let $Z \sim \mathcal{N}(0,1)$ be a standard univariate Gaussian random variable. Then,*

$$\mathbb{E}\left[\frac{1}{1 + e^{r|Z|}}\right] \leq \min\left\{\frac{1}{2}, \frac{1}{2} - \sqrt{\frac{2}{\pi}}\left(1 - \frac{r^2}{6}\right)\frac{r}{4}, \sqrt{\frac{2}{\pi}}\frac{1}{r}\right\}. \qquad (70)$$

It immediately follows from Equation (68) and Lemma 23 that for $\beta > 0$,

$$\alpha = \mathbb{E}\left[\frac{1}{1 + e^{\beta|Z|}}\right] \leq \min\left\{\frac{1}{2}, \frac{1}{2} - \sqrt{\frac{2}{\pi}}\left(1 - \frac{\beta^2}{6}\right)\frac{\beta}{4}, \sqrt{\frac{2}{\pi}}\frac{1}{\beta}\right\}. \qquad (71)$$

Using the above bound on α in Equation (71), an upper bound on α_0 can also be obtained. Noting that $\delta \leq \frac{1}{2}$, and letting $b_2 \triangleq \frac{3}{\sqrt{2\pi}}(5+\sqrt{21})$, if $\beta < \frac{b_2}{\epsilon} = \sqrt{\frac{2}{\pi}}\frac{1}{\delta}$, then $\alpha_0 = \max\{\alpha, \delta\} \leq \min\{\frac{1}{2}, \sqrt{\frac{2}{\pi}}\frac{1}{\beta}\}$, whereas if $\beta \geq \frac{b_2}{\epsilon} = \sqrt{\frac{2}{\pi}}\frac{1}{\delta}$, then $\alpha_0 = \max\{\alpha, \delta\} = \delta$.

Next, an explicit form for an lower bound on γ will be derived. This will largely hinge on showing that $\gamma \geq \sqrt{\frac{2}{\pi}}(1-2\alpha)$, from where the above bound on α can subsequently provide a closed-form bound on γ. Towards this, define $\zeta \triangleq 1 - \sqrt{\frac{\pi}{2}}\gamma$. Then, the inequality $\gamma \geq \sqrt{\frac{2}{\pi}}(1-2\alpha)$ is equivalent to $\zeta \leq 2\alpha$, the latter of which will be our focus. Note that ζ can be calculated by the following expression:

$$\zeta = \int_{z=0}^{z=\infty} z e^{-\frac{1}{2}z^2}(1 - p(z) + p(-z))dz.$$

It is convenient to view α and ζ as being parameterized by β, and hence, the following argument will use the notations: $\alpha(\beta)$ and $\zeta(\beta)$. Note that the inverse temperature, $\beta > 0$, is left as implicit in p to simplify the notation. Then, when p is taken as in Definition 3 for logistic regression, $\zeta(\beta)$ has the form:

$$\zeta(\beta) = \int_{z=0}^{z=\infty} z e^{-\frac{1}{2}z^2}(1 - p(z) + p(-z))dz = 2\int_{z=0}^{z=\infty} z e^{-\frac{1}{2}z^2} p(-z)dz,$$

where the second equality applies Equation (67). Notice that due to Equation (69), when $\beta = 0$,

$$\zeta(0) = 2\int_{z=0}^{z=\infty} z e^{-\frac{1}{2}z^2} p(0)dz = \int_{z=0}^{z=\infty} z e^{-\frac{1}{2}z^2} dz = 1,$$

where the last equality is obtained by scaling the expected value of a standard half-normal random variable by $\sqrt{\frac{\pi}{2}}$. Thus, in this scenario, $\zeta(0) = 1 = 2 \cdot \frac{1}{2} = 2\alpha(0)$, where the last equality follows from the earlier observation that $\alpha(0) = \frac{1}{2}$. Given this, it suffices to show that the ratio $\frac{\zeta(\beta)}{\alpha(\beta)}$ is maximized when $\beta = 0$, i.e., that $\sup_{\beta>0} \frac{\zeta(\beta)}{\alpha(\beta)} = \frac{\zeta(0)}{\alpha(0)} = 2$. This would indeed be true if $\frac{\partial}{\partial\beta}\frac{\zeta(\beta)}{\alpha(\beta)} \leq 0$ for all $\beta > 0$. As scaling this by a positive constant will not affect the inequality, it will be more convenient to establish that $\frac{\partial}{\partial\beta}\frac{\zeta(\beta)}{\sqrt{2\pi}\alpha(\beta)} \leq 0$, which will be done next.

To begin, note the following partial derivatives.

$$\frac{\partial}{\partial\beta}\sqrt{2\pi}\alpha(\beta) = \frac{\partial}{\partial\beta}2\int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2} p(-z)dz = 2\int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2}\frac{\partial}{\partial\beta}p(-z)dz, \tag{72}$$

$$\frac{\partial}{\partial\beta}\zeta(\beta) = \frac{\partial}{\partial\beta}2\int_{z=0}^{z=\infty} z e^{-\frac{1}{2}z^2} p(-z)dz = 2\int_{z=0}^{z=\infty} z e^{-\frac{1}{2}z^2}\frac{\partial}{\partial\beta}p(-z)dz, \tag{73}$$

where due to the quotient rule,

$$\frac{\partial}{\partial\beta}p(-z) = \frac{\partial}{\partial\beta}\frac{1}{1+e^{\beta z}} = -\frac{ze^{\beta z}}{(1+e^{\beta z})^2} = -\frac{z}{(1+e^{-\beta z})(1+e^{\beta z})} = -zp(z)p(-z). \tag{74}$$

Plugging (74) into (72) and (73) and scaling by a factor of $\frac{1}{2}$ yields

$$\frac{1}{2}\frac{\partial}{\partial\beta}\sqrt{2\pi}\alpha(\beta) = \int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2}\frac{\partial}{\partial\beta}p(-z)dz = -\int_{z=0}^{z=\infty} z e^{-\frac{1}{2}z^2} p(z)p(-z)dz, \tag{75}$$

$$\frac{1}{2}\frac{\partial}{\partial\beta}\zeta(\beta) = \int_{z=0}^{z=\infty} ze^{-\frac{1}{2}z^2}\frac{\partial}{\partial\beta}p(-z)dz = -\int_{z=0}^{z=\infty} z^2e^{-\frac{1}{2}z^2}p(z)p(-z)dz. \tag{76}$$

Then, by applying the quotient rule and plugging in (75) and (76),

$$\begin{aligned}
&\frac{\partial}{\partial\beta}\frac{\zeta(\beta)}{\sqrt{2\pi}\alpha(\beta)} \\
&= \frac{\sqrt{2\pi}\alpha(\beta)\frac{\partial}{\partial\beta}\zeta(\beta) - \zeta(\beta)\frac{\partial}{\partial\beta}\sqrt{2\pi}\alpha(\beta)}{2\pi\alpha(\beta)^2} \\
&= \frac{(-\zeta(\beta)\frac{\partial}{\partial\beta}\sqrt{2\pi}\alpha(\beta)) - (-\sqrt{2\pi}\alpha(\beta)\frac{\partial}{\partial\beta}\zeta(\beta))}{2\pi\alpha(\beta)^2} \\
&= \frac{(-\frac{1}{2}\zeta(\beta)\frac{1}{2}\frac{\partial}{\partial\beta}\sqrt{2\pi}\alpha(\beta)) - (-\frac{1}{2}\sqrt{2\pi}\alpha(\beta)\frac{1}{2}\frac{\partial}{\partial\beta}\zeta(\beta))}{2\pi(\frac{1}{2}\alpha(\beta))^2} \\
&= \frac{\left(\int_{z=0}^{z=\infty} ze^{-\frac{1}{2}z^2}p(-z)dz\right)\left(\int_{z=0}^{z=\infty} ze^{-\frac{1}{2}z^2}p(z)p(-z)dz\right) - \left(\int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2}p(-z)dz\right)\left(\int_{z=0}^{z=\infty} z^2e^{-\frac{1}{2}z^2}p(z)p(-z)dz\right)}{\left(\int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2}p(-z)dz\right)^2}.
\end{aligned}$$

In the last line, the numerator clearly determines the sign of $\frac{\partial}{\partial\beta}\frac{\zeta(\beta)}{\sqrt{2\pi}\alpha(\beta)}$. Focusing in on this expression, the following claim provides an upper bound. Its verification is deferred to the end of this proof of the corollary.

Claim 24 *Using the notations of this proof,*

$$\begin{aligned}
&\left(\int_{z=0}^{z=\infty} ze^{-\frac{1}{2}z^2}p(-z)dz\right)\left(\int_{z=0}^{z=\infty} ze^{-\frac{1}{2}z^2}p(z)p(-z)dz\right) \\
&-\left(\int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2}p(-z)dz\right)\left(\int_{z=0}^{z=\infty} z^2e^{-\frac{1}{2}z^2}p(z)p(-z)dz\right) \\
&\quad\leq \left(\int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2}p(-z)dz\right)\left(\int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2}p(z)p(-z)dz\right) \\
&\quad\quad\left(\left(\frac{\int_{z=0}^{z=\infty} ze^{-\frac{1}{2}z^2}p(z)p(-z)dz}{\int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2}p(z)p(-z)dz}\right)^2 - \frac{\int_{z=0}^{z=\infty} z^2e^{-\frac{1}{2}z^2}p(z)p(-z)dz}{\int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2}p(z)p(-z)dz}\right).
\end{aligned} \tag{77}$$

Under the assumed correctness of Claim 24, the proof of Corollary 10 can be completed. The right-hand-side of the inequality in Equation (77) has the same sign as

$$y \triangleq \left(\frac{\int_{z=0}^{z=\infty} ze^{-\frac{1}{2}z^2}p(z)p(-z)dz}{\int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2}p(z)p(-z)dz}\right)^2 - \frac{\int_{z=0}^{z=\infty} z^2e^{-\frac{1}{2}z^2}p(z)p(-z)dz}{\int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2}p(z)p(-z)dz} \tag{78}$$

since the product of the first two integrals is positive, i.e.,

$$\left(\int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2}p(-z)dz\right)\left(\int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2}p(z)p(-z)dz\right) > 0.$$

Hence, if $y \leq 0$, then due to Claim 24 and the earlier discussion, it must also happen that $\frac{\partial}{\partial\zeta}\frac{\zeta(\beta)}{\sqrt{2\pi}\alpha(\beta)} \leq 0$.

To establish the nonpositivity of y, consider a univariate standard Gaussian random variable, $Z \sim \mathcal{N}(0,1)$, and a random variable U which takes values in $\{0,1\}$ such that for $z \geq 0$,

$$(U \,|\, |Z| = z) = \begin{cases} 0, & \text{with probability } 1 - p(z)p(-z), \\ 1, & \text{with probability } p(z)p(-z). \end{cases}$$

The mass function of this conditioned random variable, $U \,|\, |Z|$, is given for $z \geq 0$ by

$$f_{U||Z|}(u \,|\, z) = \begin{cases} 1 - p(z)p(-z), & \text{if } u = 0, \\ p(z)p(-z), & \text{if } u = 1. \end{cases}$$

In addition, by the law of the total probability and the definition of conditional probabilities,

$$f_U(1) = \int_{z=-\infty}^{z=\infty} f_{U||Z|}(1 \,|\, z) f_{|Z|}(z) dz = \sqrt{\frac{2}{\pi}} \int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2} p(z)p(-z) dz.$$

Then, by Bayes' theorem, the density function of the conditioned random variable $|Z| \,|\, U = 1$ is given for $z \geq 0$ by

$$f_{|Z||U}(z \,|\, 1) = \frac{f_{U||Z|}(1 \,|\, z) f_{|Z|}(z)}{f_U(1)} = \frac{\sqrt{\frac{2}{\pi}} e^{-\frac{1}{2}z^2} p(z)p(-z) dz}{\sqrt{\frac{2}{\pi}} \int_{z'=0}^{z'=\infty} e^{-\frac{1}{2}z'^2} p(z')p(-z') dz'},$$

while for $z < 0$, $f_{|Z||U}(z \,|\, 1) = 0$. The variance of $|Z| \,|\, U = 1$ is therefore:

$$\begin{aligned} \operatorname{Var}(|Z| \,|\, U = 1) &= \mathbb{E}[|Z|^2 \,|\, U = 1] - \mathbb{E}[|Z| \,|\, U = 1]^2 \\ &= \frac{\sqrt{\frac{2}{\pi}} \int_{z=0}^{z\infty} z^2 e^{-\frac{1}{2}z^2} p(z)p(-z) dz}{\sqrt{\frac{2}{\pi}} \int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2} p(z)p(-z) dz} - \left(\frac{\sqrt{\frac{2}{\pi}} \int_{z=0}^{z\infty} z e^{-\frac{1}{2}z^2} p(z)p(-z) dz}{\sqrt{\frac{2}{\pi}} \int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2} p(z)p(-z) dz} \right)^2 \\ &= \frac{\int_{z=0}^{z\infty} z^2 e^{-\frac{1}{2}z^2} p(z)p(-z) dz}{\int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2} p(z)p(-z) dz} - \left(\frac{\int_{z=0}^{z\infty} z e^{-\frac{1}{2}z^2} p(z)p(-z) dz}{\int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2} p(z)p(-z) dz} \right)^2 \\ &= -y. \end{aligned}$$

The variance of a random variable is always nonnegative, which implies that

$$y = -\operatorname{Var}(|Z| \,|\, U = 1) \leq 0,$$

and thus, combining this with some previous remarks, it follows that $\frac{\partial}{\partial \beta} \frac{\zeta(\beta)}{\sqrt{2\pi}\alpha(\beta)} \leq 0$ when $\beta > 0$, as claimed.

As noted earlier, this further implies that

$$\sup_{\beta > 0} \frac{\zeta(\beta)}{\alpha(\beta)} = \frac{\zeta(0)}{\alpha(0)} = 2.$$

Then, by Lemma 23,

$$\zeta \leq 2\alpha \leq \min \left\{ 1, 1 - \sqrt{\frac{2}{\pi}} \left(1 - \frac{\beta^2}{6} \right) \frac{\beta}{2}, \sqrt{\frac{2}{\pi}} \frac{2}{\beta} \right\}. \tag{79}$$

This upper bound on ζ now gives the following lower bound on γ:

$$\gamma = \sqrt{\frac{2}{\pi}}(1-\zeta) \geq \sqrt{\frac{2}{\pi}}(1-2\alpha) \geq \sqrt{\frac{2}{\pi}}\left(1 - \min\left\{1, 1 - \sqrt{\frac{2}{\pi}}\left(1-\frac{\beta^2}{6}\right)\frac{\beta}{2}, \sqrt{\frac{2}{\pi}}\frac{2}{\beta}\right\}\right). \tag{80}$$

This completes Task (b).

The above work sets up the realization of the corollary's proof. With the explicit bounds on α and γ, note the following:

$$\gamma \geq \begin{cases} \left(1-\frac{\beta^2}{6}\right)\frac{\beta}{\pi}, & \text{if } \beta \in (0, b_1), \\ \sqrt{\frac{2}{\pi}}\left(1-\sqrt{\frac{2}{\pi}}\frac{2}{\beta}\right), & \text{if } \beta \in [b_1, \frac{b_2}{\epsilon}), \\ \sqrt{\frac{2}{\pi}}\left(1-\sqrt{\frac{2}{\pi}}\frac{2}{\beta}\right), & \text{if } \beta \in [\frac{b_2}{\epsilon}, \infty), \end{cases} \geq \begin{cases} \left(1-\frac{b_1^2}{6}\right)\frac{\beta}{\pi}, & \text{if } \beta \in (0, b_1), \\ b_0, & \text{if } \beta \in [b_1, \frac{b_2}{\epsilon}), \\ b_0, & \text{if } \beta \in [\frac{b_2}{\epsilon}, \infty). \end{cases} \tag{81}$$

where $b_0, b_1, b_2 > 0$ are constants such that $b_1 = \frac{\sqrt{8/\pi}}{1+b_0} \geq 1$ and $b_2 \triangleq \frac{3}{\sqrt{2\pi}}(5+\sqrt{21})$. Recall from an earlier discussion that

$$\alpha_0 \leq \begin{cases} \min\left\{\frac{1}{2}, \sqrt{\frac{2}{\pi}}\frac{1}{\beta}\right\}, & \text{if } \beta < \frac{b_2}{\delta}, \\ \delta, & \text{if } \beta \geq \frac{b_2}{\delta}. \end{cases}$$

As a result,

$$\frac{\alpha_0}{\gamma^2} \leq \begin{cases} \frac{\pi^2}{2\left(1-\frac{\beta^2}{6}\right)^2\beta^2}, & \text{if } \beta \in (0, b_1), \\ \frac{\sqrt{\pi/2}}{\left(1-\sqrt{\frac{2}{\pi}}\frac{2}{\beta}\right)^2\beta}, & \text{if } \beta \in [b_1, \frac{b_2}{\epsilon}), \\ \frac{\pi\delta}{2\left(1-\sqrt{\frac{2}{\pi}}\frac{2}{\beta}\right)^2}, & \text{if } \beta \in [\frac{b_2}{\epsilon}, \infty), \end{cases} \leq \begin{cases} \frac{\pi^2}{2\left(1-\frac{b_1^2}{6}\right)^2\beta^2}, & \text{if } \beta \in (0, b_1), \\ \frac{\sqrt{\pi/2}}{b_0^2\beta}, & \text{if } \beta \in [b_1, \frac{b_2}{\epsilon}), \\ \frac{\pi\delta}{2b_0^2}, & \text{if } \beta \in [\frac{b_2}{\epsilon}, \infty). \end{cases} \tag{82}$$

Now, the corollary's result for logistic regression can be established by substituting the bounds in Equations (81)–(82) into Equation (22) of Theorem 9 and the definitions of $\eta(\delta)$ and $\tau(\delta)$. Take n to be at least

$$n \geq \max\{n_1, n_2, n_3, n_4, n_5\} = \begin{cases} \tilde{O}\left(\frac{k}{\beta^2\epsilon^2}\right), & \text{if } \beta \in (0, b_1), \\ \tilde{O}\left(\frac{k}{\beta\epsilon^2}\right), & \text{if } \beta \in [b_1, \frac{b_2}{\epsilon}), \\ \tilde{O}\left(\frac{k}{\epsilon}\right), & \text{if } \beta \in [\frac{b_2}{\epsilon}, \infty), \end{cases}$$

where

$$n_1 = \begin{cases} \frac{432\pi^3}{(1-\frac{b_1^2}{6})^2\beta^2\delta}\log\left(\frac{24}{\rho}\sum_{\ell=0}^{k}\binom{d}{\ell}\sum_{\ell'=0}^{k}\binom{d}{\ell'}\left(\frac{b}{\tau}\right)^{\ell'}\right), & \text{if } \beta \in (0, b_1), \\ \frac{216\pi^2}{b_0^2\delta}\log\left(\frac{24}{\rho}\sum_{\ell=0}^{k}\binom{d}{\ell}\sum_{\ell'=0}^{k}\binom{d}{\ell'}\left(\frac{b}{\tau}\right)^{\ell'}\right), & \text{if } \beta \in [b_1, \frac{b_2}{\epsilon}), \\ \frac{216\pi^2}{b_0^2\delta}\log\left(\frac{24}{\rho}\sum_{\ell=0}^{k}\binom{d}{\ell}\sum_{\ell'=0}^{k}\binom{d}{\ell'}\left(\frac{b}{\tau}\right)^{\ell'}\right), & \text{if } \beta \in [\frac{b_2}{\epsilon}, \infty), \end{cases}$$

$$n_2 = \begin{cases} \dfrac{96\pi^2}{\left(1-\frac{b_1^2}{6}\right)^2 c^2\beta^2\delta^2} \log\left(\dfrac{24}{\rho}\displaystyle\sum_{\ell=0}^{k}\binom{d}{\ell}\sum_{\ell'=0}^{k}\binom{d}{\ell'}\left(\dfrac{b}{\tau}\right)^{\ell'}\right), & \text{if } \beta\in(0,b_1),\\ \dfrac{\sqrt{2\pi}96}{b_0^2c^2\beta\delta^2}\log\left(\dfrac{24}{\rho}\displaystyle\sum_{\ell=0}^{k}\binom{d}{\ell}\sum_{\ell'=0}^{k}\binom{d}{\ell'}\left(\dfrac{b}{\tau}\right)^{\ell'}\right), & \text{if } \beta\in[b_1,\frac{b_2}{\epsilon}),\\ \dfrac{96\pi}{b_0^2c^2\delta}\log\left(\dfrac{24}{\rho}\displaystyle\sum_{\ell=0}^{k}\binom{d}{\ell}\sum_{\ell'=0}^{k}\binom{d}{\ell'}\left(\dfrac{b}{\tau}\right)^{\ell'}\right), & \text{if } \beta\in[\frac{b_2}{\epsilon},\infty), \end{cases}$$

$$n_3 = \begin{cases} \dfrac{200\sqrt{2\pi}a_2}{(1-\frac{b_1^2}{6})a_3^2\beta\delta}\dfrac{\log\left(\frac{24}{\rho}\sum_{\ell=0}^{k}\binom{d}{\ell}\sum_{\ell'=0}^{k}\binom{d}{\ell'}\left(\frac{b}{\tau}\right)^{\ell'}\right)}{\sqrt{\log\left(\frac{4e}{\eta}\right)}}, & \text{if } \beta\in(0,b_1),\\ \dfrac{200a_2}{b_0a_3^2\delta}\dfrac{\log\left(\frac{24}{\rho}\sum_{\ell=0}^{k}\binom{d}{\ell}\sum_{\ell'=0}^{k}\binom{d}{\ell'}\left(\frac{b}{\tau}\right)^{\ell'}\right)}{\sqrt{\log\left(\frac{4e}{\eta}\right)}}, & \text{if } \beta\in[b_1,\frac{b_2}{\epsilon}),\\ \dfrac{200a_2}{b_0a_3^2\delta}\dfrac{\log\left(\frac{24}{\rho}\sum_{\ell=0}^{k}\binom{d}{\ell}\sum_{\ell'=0}^{k}\binom{d}{\ell'}\left(\frac{b}{\tau}\right)^{\ell'}\right)}{\sqrt{\log\left(\frac{4e}{\eta}\right)}}, & \text{if } \beta\in[\frac{b_2}{\epsilon},\infty), \end{cases}$$

$$n_4 = \begin{cases} \dfrac{64\sqrt{2\pi}}{(1-\frac{b_1^2}{6})a_2\beta\delta}\sqrt{\log\left(\dfrac{4e}{\eta}\right)}\log\left(\dfrac{24}{\rho}\right), & \text{if } \beta\in(0,b_1),\\ \dfrac{64}{b_0a_2\delta}\sqrt{\log\left(\dfrac{4e}{\eta}\right)}\log\left(\dfrac{24}{\rho}\right), & \text{if } \beta\in[b_1,\frac{b_2}{\epsilon}),\\ \dfrac{64}{b_0a_2\delta}\sqrt{\log\left(\dfrac{4e}{\eta}\right)}\log\left(\dfrac{24}{\rho}\right), & \text{if } \beta\in[\frac{b_2}{\epsilon},\infty), \end{cases}$$

$$n_5 = \begin{cases} \dfrac{\sqrt{2\pi}c_4k}{(1-\frac{b_1^2}{6})a_2\beta\delta}\sqrt{\log\left(\dfrac{4e}{\eta}\right)}\log\left(\dfrac{1}{\eta}\right), & \text{if } \beta\in(0,b_1),\\ \dfrac{c_4k}{b_0a_2\delta}\sqrt{\log\left(\dfrac{4e}{\eta}\right)}\log\left(\dfrac{1}{\eta}\right), & \text{if } \beta\in[b_1,\frac{b_2}{\epsilon}),\\ \dfrac{c_4k}{b_0a_2\delta}\sqrt{\log\left(\dfrac{4e}{\eta}\right)}\log\left(\dfrac{1}{\eta}\right), & \text{if } \beta\in[\frac{b_2}{\epsilon},\infty), \end{cases}$$

and where, as a result of Equation (81), η is bounded from below by

$$\eta = \frac{\gamma a_2 \delta}{\sqrt{\frac{2}{\pi}\log\left(\frac{4e}{\eta}\right)}} \geq \begin{cases} \frac{(1-\frac{b_1^2}{6})a_2\beta\delta}{\sqrt{2\pi\log\left(\frac{4e}{\eta}\right)}}, & \text{if } \beta \in (0, b_1), \\ \frac{b_0 a_2\delta}{\sqrt{\log\left(\frac{4e}{\eta}\right)}}, & \text{if } \beta \in [b_1, \frac{b_2}{\epsilon}), \\ \frac{b_0 a_2\delta}{\sqrt{\log\left(\frac{4e}{\eta}\right)}}, & \text{if } \beta \in [\frac{b_2}{\epsilon}, \infty). \end{cases}$$

Then, due to Equations (81) and (82),

$$\begin{aligned} n \geq \max\Bigg\{ & \frac{c_1\alpha_0}{\gamma^2\delta^2}\log\left(\frac{a}{\rho}\sum_{\ell=0}^{k}\binom{d}{\ell}\sum_{\ell'=0}^{k}\binom{d}{\ell'}\left(\frac{b}{\tau}\right)^{\ell'}\right), \\ & \frac{c_2}{\gamma\delta\sqrt{\log\left(\frac{4e}{\eta}\right)}}\log\left(\frac{a}{\rho}\sum_{\ell=0}^{k}\binom{d}{\ell}\sum_{\ell'=0}^{k}\binom{d}{\ell'}\left(\frac{b}{\tau}\right)^{\ell'}\right), \\ & \frac{c_3}{\eta}\log\left(\frac{a}{\rho}\right), \frac{c_4 k}{\eta}\log\left(\frac{1}{\eta}\right)\Bigg\}. \end{aligned}$$

Therefore, by Theorem 9 and this choice of n, with probability at least $1-\rho$, uniformly for all $\hat{\boldsymbol{\theta}} \in \Theta$ and $J \subseteq [d]$, $|J| \leq k$, Equation (23) holds:

$$\left\| \boldsymbol{\theta}^* - \frac{\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})}{\|\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})\|_2} \right\|_2 \leq \sqrt{\delta\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\|_2} + \delta.$$

Barring the verification of Claim 24, this concludes the proof of Corollary 10 for logistic regression. The last remaining task is returning to and proving Claim 24.

Proof Claim 24 Looking at the left-hand-side of Equation (77), observe:

$$\begin{aligned} &\left(\int_{z=0}^{z=\infty} z e^{-\frac{1}{2}z^2} p(-z)dz\right)\left(\int_{z=0}^{z=\infty} z e^{-\frac{1}{2}z^2} p(z)p(-z)dz\right) \\ &-\left(\int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2} p(-z)dz\right)\left(\int_{z=0}^{z=\infty} z^2 e^{-\frac{1}{2}z^2} p(z)p(-z)dz\right) \\ &= \left(\int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2} p(-z)dz\right)\left(\int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2} p(z)p(-z)dz\right) \\ &\quad \left(\left(\frac{\int_{z=0}^{z=\infty} z e^{-\frac{1}{2}z^2} p(-z)dz}{\int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2} p(-z)dz}\right)\left(\frac{\int_{z=0}^{z=\infty} z e^{-\frac{1}{2}z^2} p(z)p(-z)dz}{\int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2} p(z)p(-z)dz}\right) - \left(\frac{\int_{z=0}^{z=\infty} z^2 e^{-\frac{1}{2}z^2} p(z)p(-z)dz}{\int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2} p(z)p(-z)dz}\right)\right). \quad (83) \end{aligned}$$

In the above equation, (83), the term

$$\frac{\int_{z=0}^{z=\infty} z e^{-\frac{1}{2}z^2} p(-z)dz}{\int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2} p(-z)dz}$$

can be bounded from above as follows. Similarly to earlier in the proof, let $Z \sim \mathcal{N}(0,1)$ be a univariate standard Gaussian random variable such that $|Z|$ is a standard half-normal random variable with density

$$f_Z(z) = \begin{cases} 0, & \text{if } z < 0, \\ \sqrt{\frac{2}{\pi}} e^{-\frac{1}{2}z^2}, & \text{if } z \geq 0, \end{cases} \tag{84}$$

and let U and V be random variables taking values in $\{0,1\}$, where for $z \geq 0$,

$$(U \,|\, |Z| = z) = \begin{cases} 0, & \text{with probability } 1 - p(z)p(-z), \\ 1, & \text{with probability } p(z)p(-z), \end{cases}$$

$$(V \,|\, |Z| = z) = \begin{cases} 0, & \text{with probability } 1 - p(-z), \\ 1, & \text{with probability } p(-z). \end{cases}$$

These conditioned random variables, $U \,|\, |Z|$ and $V \,|\, |Z|$, have mass functions:

$$f_{U||Z|}(u \,|\, z) = \begin{cases} 1 - p(z)p(-z), & \text{if } u = 0, \\ p(z)p(-z), & \text{if } u = 1, \end{cases} \tag{85}$$

$$f_{V||Z|}(v \,|\, z) = \begin{cases} 1 - p(-z), & \text{if } v = 0, \\ p(-z), & \text{if } v = 1, \end{cases} \tag{86}$$

for $u, v \in \{0,1\}$ and $z \geq 0$. Applying, in order twice, the law of total probability, the definition of conditional probabilities, and Equations (84)-(86) obtains:

$$f_U(1) = \int_{z=-\infty}^{z=\infty} f_{U,|Z|}(1,z)dz = \int_{z=0}^{z=\infty} f_{U||Z|}(1 \,|\, z) f_{|Z|}(z)dz = \sqrt{\frac{2}{\pi}} \int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2} p(z)p(-z)dz,$$

$$f_V(1) = \int_{z=-\infty}^{z=\infty} f_{V,|Z|}(1,z)dz = \int_{z=0}^{z=\infty} f_{V||Z|}(1 \,|\, z) f_{|Z|}(z)dz = \sqrt{\frac{2}{\pi}} \int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2} p(-z)dz.$$

Using Bayes' theorem, for $z \geq 0$, the density functions of the conditioned random variables $|Z| \,|\, U = 1$ and $|Z| \,|\, V = 1$ are respectively given by

$$f_{|Z||U}(z \,|\, 1) = \frac{f_{U||Z|}(1 \,|\, z) f_{|Z|}(z)}{f_U(1)} = \frac{\sqrt{\frac{2}{\pi}} e^{-\frac{1}{2}z^2} p(z)p(-z)dz}{\sqrt{\frac{2}{\pi}} \int_{z'=0}^{z'=\infty} e^{-\frac{1}{2}z'^2} p(z')p(-z')dz'},$$

$$f_{|Z||V}(z \,|\, 1) = \frac{f_{V||Z|}(1 \,|\, z) f_{|Z|}(z)}{f_V(1)} = \frac{\sqrt{\frac{2}{\pi}} e^{-\frac{1}{2}z^2} p(-z)dz}{\sqrt{\frac{2}{\pi}} \int_{z'=0}^{z'=\infty} e^{-\frac{1}{2}z'^2} p(-z')dz'},$$

while for $z < 0$, $f_{|Z||U}(z \,|\, 1) = 0$ and $f_{|Z||V}(z \,|\, 1) = 0$. Then, in expectation,

$$\mathbb{E}[|Z| \,|\, U = 1] = \frac{\sqrt{\frac{2}{\pi}} \int_{z=0}^{z=\infty} z e^{-\frac{1}{2}z^2} p(z)p(-z)dz}{\sqrt{\frac{2}{\pi}} \int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2} p(z)p(-z)dz} = \frac{\int_{z=0}^{z=\infty} z e^{-\frac{1}{2}z^2} p(z)p(-z)dz}{\int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2} p(z)p(-z)dz},$$

$$\mathbb{E}[|Z|\,|\,V=1] = \frac{\sqrt{\frac{2}{\pi}}\int_{z=0}^{z=\infty} ze^{-\frac{1}{2}z^2}p(-z)dz}{\sqrt{\frac{2}{\pi}}\int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2}p(-z)dz} = \frac{\int_{z=0}^{z=\infty} ze^{-\frac{1}{2}z^2}p(-z)dz}{\int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2}p(-z)dz}.$$

Since $p(z) = \frac{1}{1+e^{-\beta z}}$ is a nondecreasing function and the support of $f_{|Z||U=1}$ and $f_{|Z||V=1}$ lies on the nonnegative real line, it follows that $\mathbb{E}[|Z|\,|\,U=1] \geq \mathbb{E}[|Z|\,|\,V=1]$, and hence, by the above pair of equations,

$$\frac{\int_{z=0}^{z=\infty} ze^{-\frac{1}{2}z^2}p(-z)dz}{\int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2}p(-z)dz} = \mathbb{E}[|Z|\,|\,V=1] \leq \mathbb{E}[|Z|\,|\,U=1] = \frac{\int_{z=0}^{z=\infty} ze^{-\frac{1}{2}z^2}p(z)p(-z)dz}{\int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2}p(z)p(-z)dz}. \quad (87)$$

Therefore, returning to Equation (83) and applying the above inequality in Equation (87), Claim 24 follows:

$$\begin{aligned}
&\left(\int_{z=0}^{z=\infty} ze^{-\frac{1}{2}z^2}p(-z)dz\right)\left(\int_{z=0}^{z=\infty} ze^{-\frac{1}{2}z^2}p(z)p(-z)dz\right) \\
&-\left(\int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2}p(-z)dz\right)\left(\int_{z=0}^{z=\infty} z^2e^{-\frac{1}{2}z^2}p(z)p(-z)dz\right) \\
&\quad= \left(\int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2}p(-z)dz\right)\left(\int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2}p(z)p(-z)dz\right) \\
&\qquad\left(\left(\frac{\int_{z=0}^{z=\infty} ze^{-\frac{1}{2}z^2}p(-z)dz}{\int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2}p(-z)dz}\right)\left(\frac{\int_{z=0}^{z=\infty} ze^{-\frac{1}{2}z^2}p(z)p(-z)dz}{\int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2}p(z)p(-z)dz}\right) - \left(\frac{\int_{z=0}^{z=\infty} z^2e^{-\frac{1}{2}z^2}p(z)p(-z)dz}{\int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2}p(z)p(-z)dz}\right)\right) \\
&\quad\leq \left(\int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2}p(-z)dz\right)\left(\int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2}p(z)p(-z)dz\right) \\
&\qquad\left(\left(\frac{\int_{z=0}^{z=\infty} ze^{-\frac{1}{2}z^2}p(z)p(-z)dz}{\int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2}p(z)p(-z)dz}\right)^2 - \frac{\int_{z=0}^{z=\infty} z^2e^{-\frac{1}{2}z^2}p(z)p(-z)dz}{\int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2}p(z)p(-z)dz}\right),
\end{aligned}$$

as desired. ■

Having established Claim 24, Corollary 10 for logistic regression is thus proved.

Next, the specialization to **probit regression** in Corollary 10 is addressed.

As in the proof of the logistic case, the probit case follows from a two-step argument—the first being Step (a), verifying that Assumption 2 holds when p is parameterized by the SNR, $\beta > 0$, and defined at $z \in \mathbb{R}$ as

$$p(z) = \frac{1}{\sqrt{2\pi}}\int_{u=-\infty}^{u=\beta z} e^{-\frac{1}{2}u^2}du \quad (88)$$

for probit regression, as per Definition 4; and the second being Step (b), deriving α and γ.

For (a), beginning with Condition (i) of Assumption 2—that p nondecreasing over the real line—notice that when $\beta = 0$, the function p is constant: $p(z) = \frac{1}{2}$ for all $z \in \mathbb{R}$, and therefore, in this scenario, p is nondecreasing. Otherwise, when $\beta > 0$, p can be rewritten as follows for $z \in \mathbb{R}$:

$$p(z) = \frac{1}{\sqrt{2\pi}}\int_{u=-\infty}^{u=\beta z} e^{-\frac{1}{2}u^2}du = \frac{\beta}{\sqrt{2\pi}}\int_{u=-\infty}^{u=z} e^{-\frac{1}{2}\beta^2u^2}du,$$

where the second equality applies a change of variables. Hence, p is the distribution function of a mean-0, variance-$\frac{1}{\beta^2}$ Gaussian random variable. Since distribution functions are nondecreasing, clearly the first condition of Assumption 2 holds under this model. Proceeding to the second requirement, Condition (ii), of Assumption 2—that

$$\frac{\partial}{\partial z} \frac{1 - p(z+w) + p(-(z+w))}{1 - p(z) + p(-z)} \leq 0 \tag{89}$$

for all $z \in \mathbb{R}$ and $w > 0$—note the following properties of p:

$$1 - p(z) = 1 - \frac{1}{\sqrt{2\pi}} \int_{u=-\infty}^{u=\beta z} e^{-\frac{1}{2}u^2} du = \frac{1}{\sqrt{2\pi}} \int_{u=\beta z}^{u=\infty} e^{-\frac{1}{2}u^2} du = \frac{1}{\sqrt{2\pi}} \int_{u=-\infty}^{u=-\beta z} e^{-\frac{1}{2}u^2} du = p(-z), \tag{90}$$

$$1 - p(z) = 1 - \frac{1}{\sqrt{2\pi}} \int_{u=-\infty}^{u=\beta z} e^{-\frac{1}{2}u^2} du = \frac{1}{\sqrt{2\pi}} \int_{u=\beta z}^{u=\infty} e^{-\frac{1}{2}u^2} du = \frac{1}{\sqrt{2\pi}} \int_{u=0}^{u=\infty} e^{-\frac{1}{2}(u+\beta z)^2} du, \tag{91}$$

where $z \in \mathbb{R}$. Additionally, it suffices to verify Assumption (ii) for $\beta = 1$. Thus, for $w > 0$,

$$\begin{aligned}
\frac{1 - p(z+w) + p(-(z+w))}{1 - p(z) + p(-z)} &= \frac{2(1 - p(z+w))}{2(1 - p(z))} \\
&= \frac{1 - p(z+w)}{1 - p(z)} \\
&= \frac{\frac{1}{\sqrt{2\pi}} \int_{u=0}^{u=\infty} e^{-\frac{1}{2}(u+z+w)^2} du}{\frac{1}{\sqrt{2\pi}} \int_{u=0}^{u=\infty} e^{-\frac{1}{2}(u+z)^2} du} \\
&= \frac{\int_{u=0}^{u=\infty} e^{-\frac{1}{2}(u+z+w)^2} du}{\int_{u=0}^{u=\infty} e^{-\frac{1}{2}(u+z)^2} du},
\end{aligned}$$

and hence, the desired inequality in Equation (89) is equivalently stated as:

$$\frac{\partial}{\partial z} \frac{1 - p(z+w) + p(-(z+w))}{1 - p(z) + p(-z)} = \frac{\partial}{\partial z} \frac{\int_{u=0}^{u=\infty} e^{-\frac{1}{2}(u+z+w)^2} du}{\int_{u=0}^{u=\infty} e^{-\frac{1}{2}(u+z)^2} du} \leq 0.$$

To evaluate this partial derivative, observe:

$$\begin{aligned}
&\frac{\partial}{\partial z} \frac{1 - p(z+w) + p(-(z+w))}{1 - p(z) + p(-z)} \\
&= \frac{\partial}{\partial z} \frac{\int_{u=0}^{u=\infty} e^{-\frac{1}{2}(u+z+w)^2} du}{\int_{u=0}^{u=\infty} e^{-\frac{1}{2}(u+z)^2} du} \\
&= \frac{\left(\int_{u=0}^{u=\infty} e^{-\frac{1}{2}(u+z)^2} du\right) \left(\frac{\partial}{\partial z} \int_{u=0}^{u=\infty} e^{-\frac{1}{2}(u+z+w)^2} du\right)}{\left(\int_{u=0}^{u=\infty} e^{-\frac{1}{2}(u+z)^2} du\right)^2} \\
&\quad - \frac{\left(\int_{u=0}^{u=\infty} e^{-\frac{1}{2}(u+z+w)^2} du\right) \left(\frac{\partial}{\partial z} \int_{u=0}^{u=\infty} e^{-\frac{1}{2}(u+z)^2} du\right)}{\left(\int_{u=0}^{u=\infty} e^{-\frac{1}{2}(u+z)^2} du\right)^2}
\end{aligned}$$

▶ by the quotient rule

$$= \frac{1}{\left(\int_{u=0}^{u=\infty} e^{-\frac{1}{2}(u+z)^2} du\right)^2} \left(\left(\int_{u=0}^{u=\infty} e^{-\frac{1}{2}(u+z)^2} du \right) \left(\frac{\partial}{\partial z} \int_{u=0}^{u=\infty} e^{-\frac{1}{2}(u+z+w)^2} du \right) \right.$$
$$\left. - \left(\int_{u=0}^{u=\infty} e^{-\frac{1}{2}(u+z+w)^2} du \right) \left(\frac{\partial}{\partial z} \int_{u=0}^{u=\infty} e^{-\frac{1}{2}(u+z)^2} du \right) \right)$$

$$= \frac{1}{\left(\int_{u=0}^{u=\infty} e^{-\frac{1}{2}(u+z)^2} du\right)^2} \left(\left(\int_{u=0}^{u=\infty} e^{-\frac{1}{2}(u+z)^2} du \right) \left(\int_{u=0}^{u=\infty} \frac{\partial}{\partial z} e^{-\frac{1}{2}(u+z+w)^2} du \right) \right.$$
$$\left. - \left(\int_{u=0}^{u=\infty} e^{-\frac{1}{2}(u+z+w)^2} du \right) \left(\int_{u=0}^{u=\infty} \frac{\partial}{\partial z} e^{-\frac{1}{2}(u+z)^2} du \right) \right)$$

▶ z does not depend on the variable of integration

$$= \frac{1}{\left(\int_{u=0}^{u=\infty} e^{-\frac{1}{2}(u+z)^2} du\right)^2} \left(\left(\int_{u=0}^{u=\infty} e^{-\frac{1}{2}(u+z)^2} du \right) \left(- \int_{u=0}^{u=\infty} (u+z+w) e^{-\frac{1}{2}(u+z+w)^2} du \right) \right.$$
$$\left. - \left(\int_{u=0}^{u=\infty} e^{-\frac{1}{2}(u+z+w)^2} du \right) \left(- \int_{u=0}^{u=\infty} (u+z) e^{-\frac{1}{2}(u+z)^2} du \right) \right)$$

▶ via the chain rule

$$= \frac{\left(\int_{u=0}^{u=\infty} e^{-\frac{1}{2}(u+z)^2} du\right)\left(\int_{u=0}^{u=\infty} e^{-\frac{1}{2}(u+z+w)^2} du\right)}{\left(\int_{u=0}^{u=\infty} e^{-\frac{1}{2}(u+z)^2} du\right)^2}$$
$$\left(\frac{\int_{u=0}^{u=\infty} (u+z) e^{-\frac{1}{2}(u+z)^2} du}{\int_{u=0}^{u=\infty} e^{-\frac{1}{2}(u+z)^2} du} - \frac{\int_{u=0}^{u=\infty} (u+z+w) e^{-\frac{1}{2}(u+z+w)^2} du}{\int_{u=0}^{u=\infty} e^{-\frac{1}{2}(u+z+w)^2} du} \right)$$

▶ by distributivity and commutativity

$$= \frac{\left(\int_{u=z}^{u=\infty} e^{-\frac{1}{2}u^2} du\right)\left(\int_{u=z+w}^{u=\infty} e^{-\frac{1}{2}u^2} du\right)}{\left(\int_{u=z}^{u=\infty} e^{-\frac{1}{2}u^2} du\right)^2} \left(\frac{\int_{u=z}^{u=\infty} u e^{-\frac{1}{2}u^2} du}{\int_{u=z}^{u=\infty} e^{-\frac{1}{2}u^2} du} - \frac{\int_{u=z+w}^{u=\infty} u e^{-\frac{1}{2}u^2} du}{\int_{u=z+w}^{u=\infty} e^{-\frac{1}{2}u^2} du} \right).$$

▶ by a change of variables (applied to each of the four integrals)

Notice in the last line that the sign of $\frac{\partial}{\partial z} \frac{1-p(z+w)+p(-(z+w))}{1-p(z)+p(-z)}$ is entirely determined by the sign of the rightmost multiplicand,

$$\frac{\int_{u=z}^{u=\infty} u e^{-\frac{1}{2}u^2} du}{\int_{u=z}^{u=\infty} e^{-\frac{1}{2}u^2} du} - \frac{\int_{u=z+w}^{u=\infty} u e^{-\frac{1}{2}u^2} du}{\int_{u=z+w}^{u=\infty} e^{-\frac{1}{2}u^2} du},$$

and hence, it suffices to show nonpositivity of this term, i.e., that

$$\frac{\int_{u=z}^{u=\infty} ue^{-\frac{1}{2}u^2}du}{\int_{u=z}^{u=\infty} e^{-\frac{1}{2}u^2}du} - \frac{\int_{u=z+w}^{u=\infty} ue^{-\frac{1}{2}u^2}du}{\int_{u=z+w}^{u=\infty} e^{-\frac{1}{2}u^2}du} \leq 0. \tag{92}$$

Towards verifying this last inequality, (92), let $U \sim \mathcal{N}(0,1)$ be a standard univariate Gaussian random variable, and let V and W be indicator random variables given by $V \triangleq \mathbb{I}(|U| \geq z)$ and $W \triangleq \mathbb{I}(|U| \geq z+w)$, respectively. The random variable $|U|$ is standard half-normal with density

$$f_{|U|}(u) = \sqrt{\frac{2}{\pi}}e^{-\frac{1}{2}u^2}. \tag{93}$$

The masses of the conditioned random variables $V = 1 \,|\, |U|$ and $W = 1 \,|\, |U|$ are given by

$$f_{V\,||U|}(1\,|\,u) = \begin{cases} 0, & \text{if } u < z, \\ 1, & \text{if } u \geq z, \end{cases} \tag{94}$$

$$f_{W\,||U|}(1\,|\,u) = \begin{cases} 0, & \text{if } u < z+w, \\ 1, & \text{if } u \geq z+w. \end{cases} \tag{95}$$

Observe:

$$\begin{aligned}
f_V(1) &= \int_{u=0}^{u=\infty} f_{|U|}(u) f_{V\,||U|}(1\,|\,u)du \\
&\qquad \blacktriangleright \text{ by the law of total probability, definition of} \\
&\qquad\quad \text{conditional probabilities, and support of } f_{|U|} \\
&= \int_{u=0}^{u=z} f_{|U|}(u)\cdot 0du + \int_{u=z}^{u=\infty} f_{|U|}(u)\cdot 1du \\
&\qquad \blacktriangleright \text{ by Equation (94)} \\
&= \sqrt{\frac{2}{\pi}}\int_{u=z}^{u=\infty} e^{-\frac{1}{2}u^2}du, \\
&\qquad \blacktriangleright \text{ by Equation (93)}
\end{aligned}$$

and likewise,

$$\begin{aligned}
f_W(1) &= \int_{u=0}^{u=\infty} f_{|U|}(u) f_{W\,||U|}(1\,|\,u)du \\
&\qquad \blacktriangleright \text{ by the law of total probability, definition of} \\
&\qquad\quad \text{conditional probabilities, and support of } f_{|U|} \\
&= \int_{u=0}^{u=z+w} f_{|U|}(u)\cdot 0du + \int_{u=z+w}^{u=\infty} f_{|U|}(u)\cdot 1du \\
&\qquad \blacktriangleright \text{ by Equation (95)} \\
&= \sqrt{\frac{2}{\pi}}\int_{u=z+w}^{u=\infty} e^{-\frac{1}{2}u^2}du,
\end{aligned}$$

► by Equation (93)

By Bayes' theorem,

$$\begin{aligned} f_{|U||V}(u\,|\,1) &= \frac{f_{|U|}(u) f_{V||U|}(1\,|\,u)}{f_V(1)} \\ &= \begin{cases} 0, & \text{if } u < z, \\ \frac{\sqrt{\frac{2}{\pi}} e^{-\frac{1}{2}u^2} du}{\sqrt{\frac{2}{\pi}} \int_{y=z}^{y=\infty} e^{-\frac{1}{2}y^2} dy}, & \text{if } u \geq z, \end{cases} \\ &= \begin{cases} 0, & \text{if } u < z, \\ \frac{e^{-\frac{1}{2}u^2} du}{\int_{y=z}^{y=\infty} e^{-\frac{1}{2}y^2} dy}, & \text{if } u \geq z, \end{cases} \end{aligned}$$

and

$$\begin{aligned} f_{|U||W}(u\,|\,1) &= \frac{f_{|U|}(u) f_{W||U|}(1\,|\,u)}{f_W(1)} \\ &= \begin{cases} 0, & \text{if } u < z+w, \\ \frac{\sqrt{\frac{2}{\pi}} e^{-\frac{1}{2}u^2} du}{\sqrt{\frac{2}{\pi}} \int_{y=z+w}^{y=\infty} e^{-\frac{1}{2}y^2} dy}, & \text{if } u \geq z+w, \end{cases} \\ &= \begin{cases} 0, & \text{if } u < z+w, \\ \frac{e^{-\frac{1}{2}u^2} du}{\int_{y=z+w}^{y=\infty} e^{-\frac{1}{2}y^2} dy}, & \text{if } u \geq z+w. \end{cases} \end{aligned}$$

Therefore, in expectation,

$$\begin{aligned} \mathbb{E}[|U|\,|\,|U| \geq z] = \mathbb{E}[|U|\,|\,V=1] &= \frac{\int_{u=z}^{u=\infty} ue^{-\frac{1}{2}u^2} du}{\int_{u=z}^{u=\infty} e^{-\frac{1}{2}u^2} du}, \\ \mathbb{E}[|U|\,|\,|U| \geq z+w] = \mathbb{E}[|U|\,|\,W=1] &= \frac{\int_{u=z+w}^{u=\infty} ue^{-\frac{1}{2}u^2} du}{\int_{u=z+w}^{u=\infty} e^{-\frac{1}{2}u^2} du}. \end{aligned}$$

Note that $\mathbb{E}[|U|\,|\,|U| \geq z] \leq \mathbb{E}[|U|\,|\,|U| \geq z+w]$ when $w > 0$, implying that

$$\frac{\int_{u=z}^{u=\infty} ue^{-\frac{1}{2}u^2} du}{\int_{u=z}^{u=\infty} e^{-\frac{1}{2}u^2} du} = \mathbb{E}[|U|\,|\,|U| \geq z] \leq \mathbb{E}[|U|\,|\,|U| \geq z+w] = \frac{\int_{u=z+w}^{u=\infty} ue^{-\frac{1}{2}u^2} du}{\int_{u=z+w}^{u=\infty} e^{-\frac{1}{2}u^2} du},$$

and hence also that

$$\frac{\int_{u=z}^{u=\infty} ue^{-\frac{1}{2}u^2} du}{\int_{u=z}^{u=\infty} e^{-\frac{1}{2}u^2} du} - \frac{\int_{u=z+w}^{u=\infty} ue^{-\frac{1}{2}u^2} du}{\int_{u=z+w}^{u=\infty} e^{-\frac{1}{2}u^2} du} \leq 0,$$

as desired. It follows from this and the earlier discussion that

$$\frac{\partial}{\partial z} \frac{1 - p(z+w) + p(-(z+w))}{1 - p(z) + p(-z)} = \frac{\partial}{\partial z} \frac{\int_{u=0}^{u=\infty} e^{-\frac{1}{2}(u+z+w)^2} du}{\int_{u=0}^{u=\infty} e^{-\frac{1}{2}(u+z)^2} du} \leq 0,$$

which verifies that Condition (ii) of Assumption 2 is upheld when p is as defined for probit regression. Having shown that both conditions of Assumption 2 are satisfied for probit regression, Step (a) is completed.

Moving ahead with Step (b), recall that the aim here is to derive α and γ. For α, observe:

$$\begin{aligned}
\alpha &= \frac{1}{\sqrt{2\pi}}\int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2}(1-p(z)+p(-z))dz \\
&= \frac{2}{\sqrt{2\pi}}\int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2}p(-z)dz \\
&= \frac{2}{\sqrt{2\pi}}\int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2}\frac{1}{\sqrt{2\pi}}\int_{u=-\infty}^{u=-\beta z} e^{-\frac{1}{2}u^2}dudz \\
&= \frac{2}{\sqrt{2\pi}}\int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2}\left(\frac{1}{\sqrt{2\pi}}\int_{u=-\infty}^{u=0} e^{-\frac{1}{2}u^2}du - \frac{1}{\sqrt{2\pi}}\int_{u=-\beta z}^{u=0} e^{-\frac{1}{2}u^2}du\right)dz \\
&= \frac{2}{\sqrt{2\pi}}\int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2}\left(\frac{1}{2} - \frac{1}{\sqrt{2\pi}}\int_{u=-\beta z}^{u=0} e^{-\frac{1}{2}u^2}du\right)dz \\
&= \frac{1}{\sqrt{2\pi}}\int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2}\left(1 - \frac{1}{\sqrt{2\pi}}\int_{u=-\beta z}^{u=\beta z} e^{-\frac{1}{2}u^2}du\right)dz \\
&= \frac{1}{\sqrt{2\pi}}\int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2}dz - \frac{1}{\sqrt{2\pi}}\int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2}\frac{1}{\sqrt{2\pi}}\int_{u=-\beta z}^{u=\beta z} e^{-\frac{1}{2}u^2}dudz \\
&= \frac{1}{2} - \frac{1}{\pi}\arctan(\beta) \\
&= \frac{1}{\pi}\arctan\left(\frac{1}{\beta}\right),
\end{aligned}$$

where the last equality holds for $\beta > 0$. Towards deriving a closed form expression for γ, define $\zeta = 1 - \sqrt{\frac{\pi}{2}}\gamma$. Then, ζ is similarly obtained as follows:

$$\begin{aligned}
\zeta &= \int_{z=0}^{z=\infty} ze^{-\frac{1}{2}z^2}(1-p(z)+p(-z))dz \\
&= 2\int_{z=0}^{z=\infty} ze^{-\frac{1}{2}z^2}p(-z)dz \\
&= 2\int_{z=0}^{z=\infty} ze^{-\frac{1}{2}z^2}\frac{1}{\sqrt{2\pi}}\int_{u=-\infty}^{u=-\beta z} e^{-\frac{1}{2}u^2}dudz \\
&= 2\int_{z=0}^{z=\infty} ze^{-\frac{1}{2}z^2}\left(\frac{1}{\sqrt{2\pi}}\int_{u=-\infty}^{u=0} e^{-\frac{1}{2}u^2}du - \frac{1}{\sqrt{2\pi}}\int_{u=-\beta z}^{u=0} e^{-\frac{1}{2}u^2}du\right)dz \\
&= 2\int_{z=0}^{z=\infty} ze^{-\frac{1}{2}z^2}\left(\frac{1}{2} - \frac{1}{\sqrt{2\pi}}\int_{u=-\beta z}^{u=0} e^{-\frac{1}{2}u^2}du\right)dz \\
&= \int_{z=0}^{z=\infty} ze^{-\frac{1}{2}z^2}\left(1 - \frac{1}{\sqrt{2\pi}}\int_{u=-\beta z}^{u=\beta z} e^{-\frac{1}{2}u^2}du\right)dz \\
&= \int_{z=0}^{z=\infty} ze^{-\frac{1}{2}z^2}dz - \int_{z=0}^{z=\infty} ze^{-\frac{1}{2}z^2}\frac{1}{\sqrt{2\pi}}\int_{u=-\beta z}^{u=\beta z} e^{-\frac{1}{2}u^2}dudz \\
&= 1 - \sin(\arctan(\beta))
\end{aligned}$$

$$= 1 - \frac{\beta}{\sqrt{\beta^2+1}}.$$

It follows that $1-\zeta = \frac{\beta}{\sqrt{\beta^2+1}}$, and hence, $\gamma = \frac{\sqrt{2/\pi}\beta}{\sqrt{\beta^2+1}}$. Thus,

$$\frac{1}{\gamma} = \sqrt{\frac{\pi}{2}\left(1+\frac{1}{\beta^2}\right)} = \begin{cases} O\left(\frac{1}{\beta}\right), & \text{if } \beta \in (0, b_3), \\ O(1), & \text{if } \beta \in [b_3, \infty), \end{cases},$$

$$\frac{1}{\gamma^2} = \frac{\pi}{2}\left(1+\frac{1}{\beta^2}\right) = \begin{cases} O\left(\frac{1}{\beta^2}\right), & \text{if } \beta \in (0, b_3), \\ O(1), & \text{if } \beta \in [b_3, \infty), \end{cases}$$

where $b_3 \triangleq 1$. Additionally, notice that $\alpha = \frac{1}{\pi}\arctan\left(\frac{1}{\beta}\right) \leq \min\left\{\frac{1}{2}, \frac{1}{\pi\beta}\right\}$, and therefore,

$$\alpha_0 \leq \max\left\{\min\left\{\frac{1}{2}, \frac{1}{\pi\beta}\right\}, \delta\right\} = \begin{cases} O(1), & \text{if } \beta \in (0, b_3), \\ O(\frac{1}{\beta}), & \text{if } \beta \in [b_3, \frac{b_4}{\epsilon}), \\ O(\epsilon), & \text{if } \beta \in [\frac{b_4}{\epsilon}, \infty), \end{cases}$$

where $b_4 \triangleq \frac{3}{2}(5+\sqrt{21})$. Incorporating these expressions for α_0 and γ into the sample complexity in Equation (22) of Theorem 9, and into the definitions of $\eta = \eta(\delta)$ and $\tau = \tau(\delta)$, the corollary's bound holds due to Theorem 9:

$$\left\| \boldsymbol{\theta}^* - \frac{\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})}{\|\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})\|_2} \right\|_2 \leq \sqrt{\delta\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\|_2} + \delta,$$

for every $\hat{\boldsymbol{\theta}} \in \Theta$ and every $J \subseteq [d]$, $|J| \leq k$, uniformly with probability at least $1-\rho$ as long as

$$\begin{aligned} n \geq \max \Bigg\{ & \frac{108\pi^3(1+\frac{1}{\beta^2})}{\delta} \log\left(\frac{24}{\rho}\sum_{\ell=0}^{k}\binom{d}{\ell}\sum_{\ell'=0}^{k}\binom{d}{\ell'}\left(\frac{b}{\tau}\right)^{\ell'}\right), \\ & \frac{48\pi^2\alpha_0(1+\frac{1}{\beta^2})}{c^2\delta^2} \log\left(\frac{24}{\rho}\sum_{\ell=0}^{k}\binom{d}{\ell}\sum_{\ell'=0}^{k}\binom{d}{\ell'}\left(\frac{b}{\tau}\right)^{\ell'}\right), \\ & \frac{200a_2\sqrt{\frac{\pi}{2}\left(1+\frac{1}{\beta^2}\right)}}{a_3^2\delta} \frac{\log\left(\frac{24}{\rho}\sum_{\ell=0}^{k}\binom{d}{\ell}\sum_{\ell'=0}^{k}\binom{d}{\ell'}\left(\frac{b}{\tau}\right)^{\ell'}\right)}{\sqrt{\log\left(\frac{4e}{\eta}\right)}}, \\ & \frac{64}{a_2\delta}\sqrt{\frac{\pi}{2}\left(1+\frac{1}{\beta^2}\right)\log\left(\frac{4e}{\eta}\right)} \log\left(\frac{24}{\rho}\right), \\ & \frac{c_4 k}{a_2\delta}\sqrt{\frac{\pi}{2}\left(1+\frac{1}{\beta^2}\right)\log\left(\frac{4e}{\eta}\right)} \log\left(\frac{1}{\eta}\right) \Bigg\} \end{aligned}$$

$$= \begin{cases} \tilde{O}\left(\frac{k}{\beta^2\epsilon^2}\right), & \text{if } \beta \in (0, b_3), \\ \tilde{O}\left(\frac{k}{\beta\epsilon^2}\right), & \text{if } \beta \in [b_3, \frac{b_4}{\epsilon}), \\ \tilde{O}\left(\frac{k}{\epsilon}\right), & \text{if } \beta \in [\frac{b_4}{\epsilon}, \infty), \end{cases}$$

where $\eta = \eta(\delta) = \frac{a_2 \delta}{\sqrt{\frac{\pi}{2}(1+\frac{1}{\beta^2}) \log\left(\frac{4e}{\eta}\right)}}$. ■

C.6. Proof of the Intermediate Results

Having completed the proofs of the main technical results, Theorem 9 and Corollary 10, the auxiliary results used therein, Lemmas 18–21—which were introduced in Appendix C.3—are proved next.

C.6.1. CONCENTRATION INEQUALITIES, EXPECTATIONS, AND A DETERMINISTIC BOUND

The concentration inequalities and expectations in Lemma 25, below, will be crucial to the proofs of the intermediate results, Lemmas 18–21. The proof this lemma is deferred to Appendix D.

Lemma 25 *Fix $s, s', t, t', \delta \in (0,1)$. Fix $\boldsymbol{\theta}^* \in \Theta$, and let $\mathcal{J}, \mathcal{J}'' \subseteq 2^{[d]}$ and $\mathcal{C} \subset \Theta$ be finite sets, and define $\tilde{\mathcal{C}} \triangleq \mathcal{C} \setminus \mathcal{B}_\tau(\boldsymbol{\theta}^*)$. Let $k_0 \triangleq \min\{2k + \max_{J \in \mathcal{J}} |J|, d\}$ and $k_0'' \triangleq \min\{k + \max_{J'' \in \mathcal{J}''} |J''|, d\}$, and let $\alpha_0 = \alpha_0(\delta) = \max\{\alpha, \delta\}$. Then,*

$$P\left(\forall J \in \mathcal{J}, \boldsymbol{\theta} \in \tilde{\mathcal{C}} \quad \left\| \frac{h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})}{\sqrt{2\pi}} - \mathbb{E}\left[\frac{h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})}{\sqrt{2\pi}}\right] \right\|_2 \leq \sqrt{\frac{(1+s)(k_0-2)\arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle)}{\pi n}} + \frac{t \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle)}{\pi}\right)$$
$$\geq 1 - 4|\mathcal{J}||\tilde{\mathcal{C}}| e^{-\frac{1}{27\pi} n t^2 \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle)} - |\mathcal{J}||\tilde{\mathcal{C}}| e^{-\frac{1}{18\pi(1+s)} n t^2 \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle)} - |\tilde{\mathcal{C}}| e^{-\frac{1}{3\pi} n s^2 \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle)}, \tag{96}$$

$$P\left(\forall J'' \in \mathcal{J}'' \quad \left\| \frac{h_{f;J''}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)}{\sqrt{2\pi}} - \mathbb{E}\left[\frac{h_{f;J''}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)}{\sqrt{2\pi}}\right] \right\|_2 \leq \sqrt{\frac{\alpha_0(1+s')(k_0''-1)}{n}} + \alpha_0 t'\right)$$
$$\geq 1 - 2|\mathcal{J}''| e^{-\frac{1}{12}\alpha_0 n t'^2} - |\mathcal{J}''| e^{-\frac{1}{8}\frac{\alpha_0 n t'^2}{1+s'}} - e^{-\frac{1}{3}\alpha_0 n s'^2}, \tag{97}$$

where in expectation, for any $J \subseteq [d]$ and $\boldsymbol{\theta} \in \Theta$,

$$\mathbb{E}\left[h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})\right] = \boldsymbol{\theta}^* - \boldsymbol{\theta}, \tag{98}$$

$$\mathbb{E}\left[h_{f;J}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)\right] = \mathbb{E}\left[\langle h_{f;J}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*), \boldsymbol{\theta}^* \rangle\right] \boldsymbol{\theta}^* = -\left(1 - \sqrt{\frac{\pi}{2}}\gamma\right) \boldsymbol{\theta}^*, \tag{99}$$

$$\| \mathbb{E}[\boldsymbol{\theta} + h_f(\boldsymbol{\theta}^*, \boldsymbol{\theta})] \|_2 = \| \mathbb{E}\left[\boldsymbol{\theta} + h_{f;J}(\boldsymbol{\theta}^*, \boldsymbol{\theta})\right] \|_2 = \sqrt{\frac{\pi}{2}}\gamma. \tag{100}$$

In addition to the above lemma, the following fact will facilitate the analysis in this section.

Fact 26 *Let $\mathbf{u}, \mathbf{v} \in S^{d-1}$. Then,*

$$\|\mathbf{u} - \mathbf{v}\|_2 \leq \arccos(\langle \mathbf{u}, \mathbf{v} \rangle) \leq \frac{\pi}{2} \|\mathbf{u} - \mathbf{v}\|_2. \tag{101}$$

Proof Fact 26 Fix $\mathbf{u}, \mathbf{v} \in S^{d-1}$ arbitrarily. To verify the first inequality in Equation (101)—that $\|\mathbf{u} - \mathbf{v}\|_2 \leq \arccos(\langle \mathbf{u}, \mathbf{v} \rangle)$—observe:

$$\|\mathbf{u} - \mathbf{v}\|_2 = \sqrt{2 - 2\langle \mathbf{u}, \mathbf{v} \rangle}$$

$$\leq \sqrt{2\left(1-\left(1-\frac{\arccos^2(\langle \mathbf{u},\mathbf{v}\rangle)}{2}\right)\right)}$$

► by the Taylor series for the cosine function, $\cos(x) \geq 1 - \dfrac{x^2}{2}$, $x \in \mathbb{R}$

$$= \arccos(\langle \mathbf{u},\mathbf{v}\rangle),$$

as desired. For the second inequality in Equation (101)—that $\arccos(\langle \mathbf{u},\mathbf{v}\rangle) \leq \frac{\pi}{2}\|\mathbf{u}-\mathbf{v}\|_2$—note that by standard trigonometric properties, $\|\mathbf{u}-\mathbf{v}\|_2 = 2\sin(\frac{\arccos(\langle \mathbf{u},\mathbf{v}\rangle)}{2})$. To proceed, some basic calculus is needed to examine the function $\frac{\sin(x)}{x}$ on the interval $x \in (0, \frac{\pi}{2}]$. Using the quotient rule,

$$\frac{d}{dx}\frac{\sin(x)}{x} = \frac{x\cos(x)-\sin(x)}{x^2},$$

where the numerator determines the sign of the above expression and has a Taylor series given by

$$x\cos(x)-\sin(x) = \sum_{z=0}^{\infty}\frac{(-1)^z x^{2z+1}}{(2z)!} - \sum_{z=0}^{\infty}\frac{(-1)^z x^{2z+1}}{(2z+1)!} = \sum_{z=1}^{\infty}\frac{(-1)^z x^{2z+1}}{(2z)!}\left(1-\frac{1}{2z+1}\right).$$

Now, it can be seen that for $x \in (0, \frac{\pi}{2}]$,

$$\frac{d}{dx}\frac{\sin(x)}{x} = \frac{x\cos(x)-\sin(x)}{x^2} = \frac{1}{x^2}\sum_{z=1}^{\infty}\frac{(-1)^z x^{2z+1}}{(2z)!}\left(1-\frac{1}{2z+1}\right) < 0,$$

which implies that $\frac{\sin(x)}{x}$ decreases over the interval $x \in (0, \frac{\pi}{2}]$. Hence,

$$\inf_{x\in(0,\frac{\pi}{2}]}\frac{\sin(x)}{x} = \left.\frac{\sin(x)}{x}\right|_{x=\frac{\pi}{2}} = \frac{\sin\left(\frac{\pi}{2}\right)}{\frac{\pi}{2}} = \frac{2}{\pi}.$$

Then,

$$\frac{\|\mathbf{u}-\mathbf{v}\|_2}{\arccos(\langle \mathbf{u},\mathbf{v}\rangle)} = \frac{2\sin\left(\frac{\arccos(\langle \mathbf{u},\mathbf{v}\rangle)}{2}\right)}{\arccos(\langle \mathbf{u},\mathbf{v}\rangle)} \geq \frac{2}{\pi},$$

implying that

$$\arccos(\langle \mathbf{u},\mathbf{v}\rangle) \leq \frac{\pi}{2}\|\mathbf{u}-\mathbf{v}\|_2,$$

as claimed. ■

C.6.2. PROOFS OF LEMMAS 18–21

With the above auxiliary results, Lemmas 18–21 can now be established. We begin with the proof of Lemma 18.

Proof Lemma 18 The first step towards proving the lemma will be showing that

$$\frac{\mathbb{E}[\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*,\hat{\boldsymbol{\theta}})]}{\|\,\mathbb{E}[\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*,\hat{\boldsymbol{\theta}})]\|_2} = \boldsymbol{\theta}^*,$$

where $\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}} \in \Theta$ and $J \subseteq [d]$ are arbitrary. Notice that for any $\mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathbb{R}^d$ such that $\mathrm{supp}(\mathbf{w}) \cup J = \mathrm{supp}(\mathbf{v}) \cup J$, the following pair of equations holds:

$$h_{f;J}(\mathbf{u}, \mathbf{v}) = h_J(\mathbf{u}, \mathbf{v}) + h_{f;\mathrm{supp}(\mathbf{v}) \cup J}(\mathbf{u}, \mathbf{u}) = h_J(\mathbf{u}, \mathbf{v}) + h_{f;\mathrm{supp}(\mathbf{w}) \cup J}(\mathbf{u}, \mathbf{u}), \tag{102}$$

$$h_J(\mathbf{u}, \mathbf{v}) = h_J(\mathbf{u}, \mathbf{w}) + h_{\mathrm{supp}(\mathbf{u}) \cup J}(\mathbf{w}, \mathbf{v}). \tag{103}$$

To justify these equations, observe:

$$\begin{aligned}
h_{f;J}(\mathbf{u}, \mathbf{v}) &= T_{\mathrm{supp}(\mathbf{u}) \cup \mathrm{supp}(\mathbf{v}) \cup J} \left(\frac{\sqrt{2\pi}}{n} \, \mathbf{X}^T \, \frac{1}{2} \left(f(\mathbf{X}\mathbf{u}) - \mathrm{sign}\left(\mathbf{X}\mathbf{v}\right) \right) \right) \\
&= T_{\mathrm{supp}(\mathbf{u}) \cup \mathrm{supp}(\mathbf{v}) \cup J} \left(\frac{\sqrt{2\pi}}{n} \, \mathbf{X}^T \, \frac{1}{2} \left(\mathrm{sign}\left(\mathbf{X}\mathbf{u}\right) - \mathrm{sign}\left(\mathbf{X}\mathbf{v}\right) \right) \right) \\
&\quad + T_{\mathrm{supp}(\mathbf{u}) \cup (\mathrm{supp}(\mathbf{v}) \cup J)} \left(\frac{\sqrt{2\pi}}{n} \, \mathbf{X}^T \, \frac{1}{2} \left(f(\mathbf{X}\mathbf{u}) - \mathrm{sign}\left(\mathbf{X}\mathbf{u}\right) \right) \right) \\
&= h_J(\mathbf{u}, \mathbf{v}) + h_{f;\mathrm{supp}(\mathbf{v}) \cup J}(\mathbf{u}, \mathbf{u}) \\
&= h_J(\mathbf{u}, \mathbf{v}) + h_{f;\mathrm{supp}(\mathbf{w}) \cup J}(\mathbf{u}, \mathbf{u})
\end{aligned}$$

and

$$\begin{aligned}
h_J(\mathbf{u}, \mathbf{v}) &= T_{\mathrm{supp}(\mathbf{u}) \cup \mathrm{supp}(\mathbf{v}) \cup J} \left(\frac{\sqrt{2\pi}}{n} \, \mathbf{X}^T \, \frac{1}{2} \left(\mathrm{sign}\left(\mathbf{X}\mathbf{u}\right) - \mathrm{sign}\left(\mathbf{X}\mathbf{v}\right) \right) \right) \\
&= T_{\mathrm{supp}(\mathbf{u}) \cup \mathrm{supp}(\mathbf{v}) \cup J} \left(\frac{\sqrt{2\pi}}{n} \, \mathbf{X}^T \, \frac{1}{2} \left(\mathrm{sign}\left(\mathbf{X}\mathbf{u}\right) - \mathrm{sign}\left(\mathbf{X}\mathbf{w}\right) \right) \right) \\
&\quad + T_{\mathrm{supp}(\mathbf{u}) \cup \mathrm{supp}(\mathbf{v}) \cup J} \left(\frac{\sqrt{2\pi}}{n} \, \mathbf{X}^T \, \frac{1}{2} \left(\mathrm{sign}\left(\mathbf{X}\mathbf{w}\right) - \mathrm{sign}\left(\mathbf{X}\mathbf{v}\right) \right) \right) \\
&= T_{\mathrm{supp}(\mathbf{u}) \cup \mathrm{supp}(\mathbf{w}) \cup J} \left(\frac{\sqrt{2\pi}}{n} \, \mathbf{X}^T \, \frac{1}{2} \left(\mathrm{sign}\left(\mathbf{X}\mathbf{u}\right) - \mathrm{sign}\left(\mathbf{X}\mathbf{w}\right) \right) \right) \\
&\quad + T_{\mathrm{supp}(\mathbf{w}) \cup \mathrm{supp}(\mathbf{v}) \cup (\mathrm{supp}(\mathbf{u}) \cup J)} \left(\frac{\sqrt{2\pi}}{n} \, \mathbf{X}^T \, \frac{1}{2} \left(\mathrm{sign}\left(\mathbf{X}\mathbf{w}\right) - \mathrm{sign}\left(\mathbf{X}\mathbf{v}\right) \right) \right) \\
&= h_J(\mathbf{u}, \mathbf{w}) + h_{\mathrm{supp}(\mathbf{u}) \cup J}(\mathbf{w}, \mathbf{v}).
\end{aligned}$$

Additionally, by Lemma 25,

$$\mathbb{E}[h_J(\mathbf{u}, \mathbf{v})] = \mathbf{u} - \mathbf{v}, \tag{104}$$

$$\mathbb{E}[h_{f;J}(\mathbf{u}, \mathbf{u})] = -\left(1 - \sqrt{\frac{\pi}{2}} \gamma \right) \mathbf{u}. \tag{105}$$

Thus,

$$\begin{aligned}
\mathbb{E}[\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})] &= \hat{\boldsymbol{\theta}} + \mathbb{E}[h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})] \\
&= \hat{\boldsymbol{\theta}} + \mathbb{E}[h_J(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})] + \mathbb{E}[h_{f;J}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)]
\end{aligned}$$

$$
\begin{aligned}
&\quad \blacktriangleright \text{ by Equation (102) and the linearity of expectation} \\
&= \hat{\boldsymbol{\theta}} + (\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}) - \left(1 - \sqrt{\frac{\pi}{2}}\gamma\right)\boldsymbol{\theta}^* \\
&\quad \blacktriangleright \text{ by Equations (104) and (105)} \\
&= \sqrt{\frac{\pi}{2}}\gamma\boldsymbol{\theta}^*.
\end{aligned} \tag{106}
$$

It follows that

$$
\frac{\mathbb{E}[\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})]}{\|\,\mathbb{E}[\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})]\|_2} = \frac{\sqrt{\pi/2}\gamma\boldsymbol{\theta}^*}{\sqrt{\pi/2}\gamma} = \boldsymbol{\theta}^*, \tag{107}
$$

as claimed. Having achieved the first task, the left-hand-side of the inequality in Equation (66) can now be bounded from above as follows:

$$
\begin{aligned}
&\left\| \boldsymbol{\theta}^* - \frac{\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})}{\|\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})\|_2} \right\|_2 \\
&\quad = \left\| \frac{\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})}{\|\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})\|_2} - \frac{\mathbb{E}[\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})]}{\mathbb{E}[\|\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})\|_2]} \right\|_2 \\
&\qquad \blacktriangleright \text{ by Equation (107)} \\
&\quad \leq \frac{2\|\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) - \mathbb{E}[\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})]\|_2}{\|\,\mathbb{E}[\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})]\|_2} \\
&\qquad \blacktriangleright \text{ by Fact 13} \\
&\quad = \frac{2\|h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) - \mathbb{E}[h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})]\|_2}{\|\,\mathbb{E}[\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})]\|_2} \\
&\quad = \frac{2\|h_J(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) + h_{f;\mathrm{supp}(\boldsymbol{\theta})\cup J}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) - \mathbb{E}[h_J(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}) + h_{f;\mathrm{supp}(\boldsymbol{\theta})\cup J}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)]\|_2}{\|\,\mathbb{E}[\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})]\|_2} \\
&\qquad \blacktriangleright \text{ by Equation (102) and the lemma's condition} \\
&\quad = \frac{2\|h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}) - \mathbb{E}[h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})]\|_2}{\|\,\mathbb{E}[\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})]\|_2} + \frac{2\|h_{\mathrm{supp}(\boldsymbol{\theta}^*)\cup J}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) - \mathbb{E}[h_{\mathrm{supp}(\boldsymbol{\theta}^*)\cup J}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})]\|_2}{\|\,\mathbb{E}[\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})]\|_2} \\
&\qquad + \frac{2\|h_{f;\mathrm{supp}(\boldsymbol{\theta})\cup J}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) - \mathbb{E}[h_{f;\mathrm{supp}(\boldsymbol{\theta})\cup J}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)]\|_2}{\|\,\mathbb{E}[\hat{\boldsymbol{\theta}} + h_{f;J}(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}})]\|_2} \\
&\qquad \blacktriangleright \text{ by the linearity of expectation and the triangle inequality} \\
&\quad = \frac{2\|h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}) - \mathbb{E}[h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})]\|_2}{\|\,\mathbb{E}[\boldsymbol{\theta} + h_f(\boldsymbol{\theta}^*, \boldsymbol{\theta})]\|_2} + \frac{2\|h_{\mathrm{supp}(\boldsymbol{\theta}^*)\cup J}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) - \mathbb{E}[h_{\mathrm{supp}(\boldsymbol{\theta}^*)\cup J}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})]\|_2}{\|\,\mathbb{E}[\boldsymbol{\theta} + h_f(\boldsymbol{\theta}^*, \boldsymbol{\theta})]\|_2} \\
&\qquad + \frac{2\|h_{f;\mathrm{supp}(\boldsymbol{\theta})\cup J}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) - \mathbb{E}[h_{f;\mathrm{supp}(\boldsymbol{\theta})\cup J}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)]\|_2}{\|\,\mathbb{E}[\boldsymbol{\theta} + h_f(\boldsymbol{\theta}^*, \boldsymbol{\theta})]\|_2}. \\
&\qquad \blacktriangleright \text{ by the first equality in Equation (100)}
\end{aligned}
$$

This completes the proof of Lemma 18. ■

Proof Lemma 19 Let $\boldsymbol{\theta}^* \in \Theta$ be arbitrary. Write $\tilde{\mathcal{C}} \triangleq \mathcal{C} \setminus \mathcal{B}_\tau(\boldsymbol{\theta}^*)$. For the time being, fix $J \in \mathcal{J}$ and $\boldsymbol{\theta} \in \tilde{\mathcal{C}}$ arbitrarily—to be varied over all possible choices later—and let $J'' \triangleq \operatorname{supp}(\boldsymbol{\theta}) \cup J$. By Equation (100) in Lemma 25,

$$\| \mathbb{E}[\boldsymbol{\theta} + h_f(\boldsymbol{\theta}^*, \boldsymbol{\theta})]\|_2 = \sqrt{\frac{\pi}{2}}\gamma, \tag{108}$$

and thus, substituting Equation (108) into the right-hand-side of Equation (61) in Lemma 19 yields

$$\frac{2\|h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}) - \mathbb{E}[h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})]\|_2}{\| \mathbb{E}[\boldsymbol{\theta} + h_f(\boldsymbol{\theta}^*, \boldsymbol{\theta})]\|_2} = \frac{2\|h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}) - \mathbb{E}[h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})]\|_2}{\sqrt{\pi/2}\gamma}. \tag{109}$$

Towards bounding the term $\|h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}) - \mathbb{E}[h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})]\|_2$ in the numerator on the right-hand-side of Equation (109), consider Equation (96) in Lemma 25, where $s, t \in (0,1)$ are taken to be

$$s \triangleq \sqrt{\frac{3\pi \log\left(\frac{3}{\rho_1}|\mathcal{C}|\right)}{n\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2}}, \tag{110}$$

and

$$t \triangleq \sqrt{\frac{27\pi \log\left(\frac{12}{\rho_1}|\mathcal{J}||\mathcal{C}|\right)}{n\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2}}, \tag{111}$$

and where n is at least

$$n \geq \frac{16}{\gamma^2\delta} \max\left\{27\pi \log\left(\frac{12}{\rho_1}|\mathcal{J}||\mathcal{C}|\right), 4(k_0 - 2)\right\}. \tag{112}$$

This bound on n suffices to ensure that under the lemma's condition that $\boldsymbol{\theta} \in \tilde{\mathcal{C}} = \mathcal{C} \setminus \mathcal{B}_\tau(\boldsymbol{\theta}^*)$, the variable s satisfies the requirement that $s < 1$, which further implies that $1 + s < 2$. Hence also,

$$\begin{aligned}
t &= \sqrt{\frac{27\pi \log\left(\frac{12}{\rho_1}|\mathcal{J}||\mathcal{C}|\right)}{n\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2}} \\
&= \max\left\{\sqrt{\frac{27\pi \log\left(\frac{12}{\rho_1}|\mathcal{J}||\mathcal{C}|\right)}{n\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2}}, \sqrt{\frac{4\pi \log\left(\frac{3}{\rho_1}|\mathcal{J}||\mathcal{C}|\right)}{n\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2}}\right\} \\
&= \max\left\{\sqrt{\frac{27\pi \log\left(\frac{12}{\rho_1}|\mathcal{J}||\mathcal{C}|\right)}{n\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2}}, \sqrt{\frac{2\pi(1+s) \log\left(\frac{3}{\rho_1}|\mathcal{J}||\mathcal{C}|\right)}{n\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2}}\right\}.
\end{aligned} \tag{113}$$

Moreover, with these choices,

$$\sqrt{\frac{\pi(1+s)(k_0-2)\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2}{n}} + \sqrt{\frac{\pi}{2}} t\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2$$

$$\leq \frac{1}{2} \cdot \sqrt{\frac{\pi}{8}} \gamma \sqrt{\delta \|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2} + \frac{1}{2} \cdot \sqrt{\frac{\pi}{8}} \gamma \sqrt{\delta \|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2}$$
$$= \sqrt{\frac{\pi}{8}} \gamma \sqrt{\delta \|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2}. \tag{114}$$

For any random variable U taking values in $\mathcal{S} \subseteq \mathbb{R}$, and for values $u \leq u' \in \mathcal{S}$, the event that $U \leq u$ implies $U \leq u'$, and therefore,

$$P(U \leq u) \leq P(U \leq u'), \quad u \leq u'. \tag{115}$$

Combining these observations with Equation (96) in Lemma 25 yields:

$$P\left(\forall J \in \mathcal{J}, \boldsymbol{\theta} \in \tilde{\mathcal{C}} \;\; \|h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}) - \mathbb{E}\left[h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})\right]\|_2 \leq \sqrt{\tfrac{\pi}{8}} \gamma \sqrt{\delta \|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2}\right)$$

$$\geq P\left(\forall J \in \mathcal{J}, \boldsymbol{\theta} \in \tilde{\mathcal{C}} \;\; \|h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}) - \mathbb{E}\left[h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})\right]\|_2 \leq \sqrt{\tfrac{\pi(1+s)(k_0-2)\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2}{n}} + \sqrt{\tfrac{\pi}{2}} t \|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2\right)$$

▶ by Equations (114) and (115)

$$\geq P\left(\forall J \in \mathcal{J}, \boldsymbol{\theta} \in \tilde{\mathcal{C}} \;\; \|h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}) - \mathbb{E}\left[h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})\right]\|_2 \leq \sqrt{\tfrac{2(1+s)(k_0-2)\arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta}\rangle)}{n}} + \sqrt{\tfrac{2}{\pi}} t \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta}\rangle)\right)$$

▶ $\arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta}\rangle) \leq \frac{\pi}{2} \|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2$ by Fact 26

$$\geq 1 - 4|\mathcal{J}||\tilde{\mathcal{C}}|e^{-\frac{1}{27\pi} n t^2 \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta}\rangle)} - |\mathcal{J}||\tilde{\mathcal{C}}|e^{-\frac{1}{2\pi(1+s)} n t^2 \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta}\rangle)} - |\tilde{\mathcal{C}}|e^{-\frac{1}{3\pi} n s^2 \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta}\rangle)}$$

▶ by Equation (96)

$$\geq 1 - 4|\mathcal{J}||\tilde{\mathcal{C}}|e^{-\frac{1}{27\pi} n t^2 \|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2} - |\mathcal{J}||\tilde{\mathcal{C}}|e^{-\frac{1}{2\pi(1+s)} n t^2 \|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2} - |\tilde{\mathcal{C}}|e^{-\frac{1}{3\pi} n s^2 \|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2}$$

▶ $\arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta}\rangle) \geq \|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2$ due to Fact 26

$$\geq 1 - \frac{\rho_1}{3} - \frac{\rho_1}{3} - \frac{\rho_1}{3}$$

▶ by the choice of s, t in Equations (110) and (111), respectively, and by Equation (113)

$$= 1 - \rho_1. \tag{116}$$

Returning to Equation (109) and inserting Equation (116), it follows that if n satisfies Equation (112), then with probability at least $1 - \rho_1$, for all $J \in \mathcal{J}$ and $\boldsymbol{\theta} \in \tilde{\mathcal{C}}$,

$$\begin{aligned}
\frac{2\|h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}) - \mathbb{E}[h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})]\|_2}{\|\mathbb{E}[\boldsymbol{\theta} + h_f(\boldsymbol{\theta}^*, \boldsymbol{\theta})]\|_2} &= \frac{2\|h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}) - \mathbb{E}[h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})]\|_2}{\sqrt{\pi/2}\gamma} \\
&\leq \sqrt{\frac{8}{\pi}} \frac{1}{\gamma} \sqrt{\frac{\pi}{8}} \gamma \sqrt{\delta \|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2} \\
&= \sqrt{\delta \|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2},
\end{aligned}$$

as desired. ■

Proof Lemma 20 Take any $\boldsymbol{\theta}^* \in \Theta$, and write $\tilde{\mathcal{C}} \triangleq \mathcal{C} \setminus \mathcal{B}_\tau(\boldsymbol{\theta}^*)$ and $\mathcal{J}' \triangleq \{\mathrm{supp}(\boldsymbol{\theta}^*) \cup J : J \in \mathcal{J}\}$, where $\mathcal{J} \subseteq 2^{[d]}$ is arbitrary. Let $\boldsymbol{\theta} \in \tilde{\mathcal{C}}$, $\hat{\boldsymbol{\theta}} \in \mathcal{B}'_{2\tau}(\boldsymbol{\theta})$, and $J' \in \mathcal{J}'$ be arbitrary. Write

$\tilde{\mathbf{x}}_i \triangleq T_{\operatorname{supp}(\boldsymbol{\theta}) \cup J'}(\mathbf{x}_i) \in \mathbb{R}^d$, $i \in [n]$, and let $\tilde{\mathbf{X}} \triangleq (\tilde{\mathbf{x}}_1 \cdots \tilde{\mathbf{x}}_n)^T \in \mathbb{R}^{n \times d}$. Using these notations, $\frac{1}{\sqrt{2\pi}} h_{J'}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ can be expressed as follows:

$$\begin{aligned}
\frac{1}{\sqrt{2\pi}} h_{J'}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) &= T_{\operatorname{supp}(\boldsymbol{\theta}) \cup J'} \left(\frac{1}{m} \sum_{i=1}^{n} \mathbf{x}_i \, \frac{1}{2} \left(\operatorname{sign}(\langle \mathbf{x}_i, \boldsymbol{\theta} \rangle) - \operatorname{sign}(\langle \mathbf{x}_i, \hat{\boldsymbol{\theta}} \rangle) \right) \right) \\
&= \frac{1}{m} \sum_{i=1}^{n} \tilde{\mathbf{x}}_i \, \frac{1}{2} \left(\operatorname{sign}(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta} \rangle) - \operatorname{sign}(\langle \tilde{\mathbf{x}}_i, \hat{\boldsymbol{\theta}} \rangle) \right) \\
&= \frac{1}{m} \sum_{i=1}^{n} \tilde{\mathbf{x}}_i \operatorname{sign}(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta} \rangle) \, \mathbb{I}\big(\operatorname{sign}(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta} \rangle) \neq \operatorname{sign}(\langle \tilde{\mathbf{x}}_i, \hat{\boldsymbol{\theta}} \rangle) \big) \\
&= \frac{1}{m} \tilde{\mathbf{X}}^T \operatorname{diag}(\operatorname{sign}(\tilde{\mathbf{X}} \boldsymbol{\theta})) \, \mathbb{I}\big(\operatorname{sign}(\tilde{\mathbf{X}} \boldsymbol{\theta}) \neq \operatorname{sign}(\tilde{\mathbf{X}} \hat{\boldsymbol{\theta}}) \big).
\end{aligned}$$

It is clear from the last line that after fixing the covariates, $\tilde{\mathbf{x}}_i$, $i \in [n]$, the function $h_{J'}$ can only take finitely many values. Moreover, upon additionally fixing $\boldsymbol{\theta} \in \Theta$, the finitely many values that can be taken by the function $h_{J'}(\boldsymbol{\theta}, \cdot)$ is determined by the number of values that can be taken by $\mathbb{I}(\operatorname{sign}(\tilde{\mathbf{X}} \boldsymbol{\theta}) \neq \operatorname{sign}(\tilde{\mathbf{X}} \hat{\boldsymbol{\theta}})) \in \{0,1\}^n$ over all choices of $\hat{\boldsymbol{\theta}} \in \mathcal{B}'_{2\tau}(\boldsymbol{\theta})$. As such, write $\mathcal{W}(\boldsymbol{\theta}) \triangleq \{\mathbb{I}(\operatorname{sign}(\tilde{\mathbf{X}} \boldsymbol{\theta}) \neq \operatorname{sign}(\tilde{\mathbf{X}} \hat{\boldsymbol{\theta}})) : \hat{\boldsymbol{\theta}} \in \mathcal{B}'_{2\tau}(\boldsymbol{\theta})\}$ for $\boldsymbol{\theta} \in \Theta$. In addition, for $\boldsymbol{\theta}, \hat{\boldsymbol{\theta}} \in \Theta$, define $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) \triangleq \|\mathbb{I}(\operatorname{sign}(\mathbf{X} \boldsymbol{\theta}) \neq \operatorname{sign}(\mathbf{X} \hat{\boldsymbol{\theta}}))\|_0$, and let $L(\boldsymbol{\theta}) \triangleq \sup_{\boldsymbol{\theta}' \in \mathcal{B}'_{2\tau}(\boldsymbol{\theta})} L(\boldsymbol{\theta}, \boldsymbol{\theta}')$. While there is a naïve upper bound of $|\mathcal{W}(\boldsymbol{\theta})| \leq 2^n$, it turns out that a little more nuance admits a tighter bound on the cardinality of $\mathcal{W}(\boldsymbol{\theta})$ by means of the random variable $L(\boldsymbol{\theta})$ and the following lemma.

Lemma 27 (Corollary to (Oymak and Recht, 2015, Corollary 3.3)) *Let $c_4 > 0$ be the constant defined in Definition 15. If*

$$n \geq \frac{c_4 k}{\eta} \log\left(\frac{1}{\eta}\right), \tag{117}$$

then with probability at least $1 - \binom{d}{k} e^{-\frac{1}{64}\eta n}$, the random variable $L(\boldsymbol{\theta})$ is bounded from above by $L(\boldsymbol{\theta}) \leq \eta n$ uniformly for all $\boldsymbol{\theta} \in S^{d-1} \cap \Sigma_k^d$.

Proof Lemma 27 Lemma 27 is a corollary to (Oymak and Recht, 2015, Corollary 3.3), which is presented below as Lemma 28.

Lemma 28 ((part of) (Oymak and Recht, 2015, Corollary 3.3)) *Let $\mathcal{U} \subseteq \mathbb{R}^d$. If the set $\hat{\mathcal{U}} \triangleq \{w\mathbf{u} : \mathbf{u} \in \mathcal{U}, w \in \mathbb{R}\}$ is a subspace with dimension $\dim \hat{\mathcal{U}} = t$ and*

$$n \geq \frac{c_4 t}{\eta} \log\left(\frac{1}{\eta}\right), \tag{118}$$

then $L(\mathbf{u}, \mathbf{v}) \leq \eta n$ for each pair $\mathbf{u}, \mathbf{v} \in \mathcal{U}$ such that $\|\mathbf{u} - \mathbf{v}\|_2 \leq \frac{\eta}{c_4 \sqrt{\log\left(\frac{1}{\eta}\right)}}$, uniformly with probability at least $1 - e^{-\frac{1}{64}\eta n}$.

Resuming the verification of Lemma 27, let $\mathcal{J}''' \triangleq \{J''' \subseteq [d] : |J'''| = k\}$, where $|\mathcal{J}'''| = \binom{d}{k}$. Note that $\bigcup_{J''' \in \mathcal{J}'''} \{\mathbf{u} \in S^{d-1} : \operatorname{supp}(\mathbf{u}) \subseteq J'''\} = S^{d-1} \cap \Sigma_k^d$. Fixing $J''' \in \mathcal{J}'''$ arbitrarily, and writing $\mathcal{U} \triangleq \{\mathbf{u} \in S^{d-1} : \operatorname{supp}(\mathbf{u}) \subseteq J'''\}$ and $\hat{\mathcal{U}} \triangleq \{w\mathbf{u} : \mathbf{u} \in \mathcal{U}, w \in \mathbb{R}\}$—where $\hat{\mathcal{U}}$ has dimension $\dim \hat{\mathcal{U}} = k$—consider any $\boldsymbol{\theta} \in \mathcal{U}$. Notice that because $\mathcal{B}'_{2\tau}(\boldsymbol{\theta}) \subseteq \mathcal{U}$, it happens that $L(\boldsymbol{\theta}) = \sup_{\hat{\boldsymbol{\theta}} \in \mathcal{B}'_{2\tau}(\boldsymbol{\theta})} L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) \leq \sup_{\hat{\boldsymbol{\theta}} \in \mathcal{U}} L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) \leq \sup_{\boldsymbol{\theta}', \hat{\boldsymbol{\theta}} \in \mathcal{U}} L(\boldsymbol{\theta}', \hat{\boldsymbol{\theta}})$. Hence, it immediately follows from Lemma 28 that with probability at least $1 - e^{-\frac{1}{64}\eta n}$, the desired upper bound holds: $L(\boldsymbol{\theta}) \leq \sup_{\boldsymbol{\theta}', \hat{\boldsymbol{\theta}} \in \mathcal{U}} L(\boldsymbol{\theta}', \hat{\boldsymbol{\theta}}) \leq \eta n$. By a union bound over $\mathcal{J}'''$, this bound on $L(\boldsymbol{\theta})$ holds uniformly over all $\boldsymbol{\theta} \in S^{d-1} \cap \Sigma^d_k$ with probability at least $1 - |\mathcal{J}'''| e^{-\frac{1}{64}\eta n} = 1 - \binom{d}{k} e^{-\frac{1}{64}\eta n}$, as claimed. ■

Returning to the proof of Lemma 20, recall that $\Theta = S^{d-1} \cap \Sigma^d_k$. Thus, due to Lemma 27 and the sufficiently large choice of

$$n \geq \frac{c_4 k}{\eta} \log\left(\frac{1}{\eta}\right) \tag{119}$$

in Lemma 20, with probability no less than $1 - \binom{d}{k} e^{-\frac{1}{64}\eta n}$, for every $\boldsymbol{\theta} \in S^{d-1} \cap \Sigma^d_k = \Theta$ and every $\hat{\boldsymbol{\theta}} \in \mathcal{B}'_{2\tau}(\boldsymbol{\theta})$, the indicator random vector $\mathbb{I}(\operatorname{sign}(\tilde{\mathbf{X}}\boldsymbol{\theta}) \neq \operatorname{sign}(\tilde{\mathbf{X}}\hat{\boldsymbol{\theta}})) \in \{0,1\}^n$ contains at most ηn-many nonzero entries. Therefore, with probability at least $1 - \binom{d}{k} e^{-\frac{1}{64}\eta n}$, for every $\boldsymbol{\theta} \in \Theta$,

$$|\mathcal{W}(\boldsymbol{\theta})| \leq \sum_{\ell=0}^{\eta n} \binom{n}{\ell} \leq \left(\frac{en}{\eta n}\right)^{\eta n} = \left(\frac{e}{\eta}\right)^{\eta n}, \tag{120}$$

where the second inequality is due to a well-known bound for sums of binomial coefficients. In light of this, for each $\boldsymbol{\theta} \in \Theta$, construct the following cover, $\mathcal{D}(\boldsymbol{\theta}) \subset \mathcal{B}'_{2\tau}(\boldsymbol{\theta})$, over $\mathcal{B}'_{2\tau}(\boldsymbol{\theta})$: for each $\mathbf{w} \in \mathcal{W}(\boldsymbol{\theta})$, insert into $\mathcal{D}(\boldsymbol{\theta})$ exactly one $\hat{\boldsymbol{\theta}} \in \mathcal{B}'_{2\tau}(\boldsymbol{\theta})$ for which $\mathbb{I}(\operatorname{sign}(\tilde{\mathbf{X}}\boldsymbol{\theta}) \neq \operatorname{sign}(\tilde{\mathbf{X}}\hat{\boldsymbol{\theta}})) = \mathbf{w}$. Note that $|\mathcal{D}(\boldsymbol{\theta})| = |\mathcal{W}(\boldsymbol{\theta})|$. Define the random variable $Q \triangleq \max_{\boldsymbol{\theta}' \in \Theta} |\mathcal{D}(\boldsymbol{\theta}')| \equiv \max_{\boldsymbol{\theta}' \in \Theta} |\mathcal{W}(\boldsymbol{\theta}')|$, and write $q \triangleq \sum_{\ell=0}^{\eta n} \binom{n}{\ell} \leq (\frac{e}{\eta})^{\eta n}$. Due to Lemma 27 and Equation (120), as well as the sufficient choice of n in Equation (119), the random variable Q is bounded from above by $Q \leq q$ with, once again, probability at least $1 - \binom{d}{k} e^{-\frac{1}{64}\eta n}$.

An additional helpful technique is an orthogonal decomposition of $h_{J'}$—in this case:

$$h_{J'}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \left\langle h_{J'}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}), \frac{\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}}{\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_2} \right\rangle \frac{\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}}{\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_2} + \left\langle h_{J'}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}), \frac{\boldsymbol{\theta} + \hat{\boldsymbol{\theta}}}{\|\boldsymbol{\theta} + \hat{\boldsymbol{\theta}}\|_2} \right\rangle \frac{\boldsymbol{\theta} + \hat{\boldsymbol{\theta}}}{\|\boldsymbol{\theta} + \hat{\boldsymbol{\theta}}\|_2} + \bar{h}_{J'}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}), \tag{121}$$

where

$$\bar{h}_{J'}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = h_{J'}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) - \left\langle h_{J'}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}), \frac{\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}}{\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_2} \right\rangle \frac{\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}}{\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_2} - \left\langle h_{J'}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}), \frac{\boldsymbol{\theta} + \hat{\boldsymbol{\theta}}}{\|\boldsymbol{\theta} + \hat{\boldsymbol{\theta}}\|_2} \right\rangle \frac{\boldsymbol{\theta} + \hat{\boldsymbol{\theta}}}{\|\boldsymbol{\theta} + \hat{\boldsymbol{\theta}}\|_2}$$

per Equation (47). Note that similar orthogonal decompositions appear in, e.g., Plan et al. (2017); Friedlander et al. (2021); Matsumoto and Mazumdar (2024a,b). Due to Equation (121) in combination with the linearity of expectation,

$$\begin{aligned}
&h_{J'}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) - \mathbb{E}[h_{J'}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})] \\
&= \left(\left\langle h_{J'}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}), \frac{\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}}{\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_2} \right\rangle - \mathbb{E}\left[\left\langle h_{J'}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}), \frac{\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}}{\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_2} \right\rangle \right] \right) \frac{\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}}{\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_2} \\
&\quad + \left(\left\langle h_{J'}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}), \frac{\boldsymbol{\theta} + \hat{\boldsymbol{\theta}}}{\|\boldsymbol{\theta} + \hat{\boldsymbol{\theta}}\|_2} \right\rangle - \mathbb{E}\left[\left\langle h_{J'}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}), \frac{\boldsymbol{\theta} + \hat{\boldsymbol{\theta}}}{\|\boldsymbol{\theta} + \hat{\boldsymbol{\theta}}\|_2} \right\rangle \right] \right) \frac{\boldsymbol{\theta} + \hat{\boldsymbol{\theta}}}{\|\boldsymbol{\theta} + \hat{\boldsymbol{\theta}}\|_2}
\end{aligned}$$

$$+ (\bar{h}_{J'}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) - \mathbb{E}[\bar{h}_{J'}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})]).$$

Combining this orthogonal decomposition with the triangle inequality yields the following upper bound on the ℓ_2-distance of $h_{J'}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ from its mean:

$$\begin{aligned}\|h_{J'}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) - \mathbb{E}[h_{J'}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})]\|_2 \leq & \left| \left\langle h_{J'}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}), \frac{\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}}{\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_2} \right\rangle - \mathbb{E}\left[\left\langle h_{J'}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}), \frac{\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}}{\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_2} \right\rangle \right] \right| \\ & + \left| \left\langle h_{J'}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}), \frac{\boldsymbol{\theta} + \hat{\boldsymbol{\theta}}}{\|\boldsymbol{\theta} + \hat{\boldsymbol{\theta}}\|_2} \right\rangle - \mathbb{E}\left[\left\langle h_{J'}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}), \frac{\boldsymbol{\theta} + \hat{\boldsymbol{\theta}}}{\|\boldsymbol{\theta} + \hat{\boldsymbol{\theta}}\|_2} \right\rangle \right] \right| \\ & + \|\bar{h}_{J'}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) - \mathbb{E}[\bar{h}_{J'}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})]\|_2. \end{aligned} \tag{122}$$

Most of the remaining arguments in this proof are towards bounding the three terms on the right-hand-side of Equation (122). The following lemma from Matsumoto and Mazumdar (2024a) will facilitate the bound on Equation (122). Note that Lemma 29 is not an exact restatement of, but is implied by, (Matsumoto and Mazumdar, 2024a, Lemma A.1) and its proof.

Lemma 29 (due to (Matsumoto and Mazumdar, 2024a, Lemma A.1)) *Let $t > 0$ and $u \in (0, 1)$, and let $\boldsymbol{\theta}, \hat{\boldsymbol{\theta}} \in S^{d-1} \cap \Sigma_k^d$ and $J' \in \mathcal{J}'$. Write $k_0' \triangleq \min\{\max\{2k, \max_{J'' \in \mathcal{J}'} |J''|\}, d\} \geq \min\{\max\{|\mathrm{supp}(\boldsymbol{\theta}) \cup \mathrm{supp}(\hat{\boldsymbol{\theta}})|, |J'|\}, d\}$. Then,*

$$P\left(\left| \left\langle \tfrac{h_{J'}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})}{\sqrt{2\pi}}, \tfrac{\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}}{\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_2} \right\rangle - \mathbb{E}\left[\left\langle \tfrac{h_{J'}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})}{\sqrt{2\pi}}, \tfrac{\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}}{\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_2} \right\rangle \right] \right| > ut \middle| L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) \leq un \right) \leq 2e^{-\frac{1}{2}unt^2}, \tag{123}$$

$$P\left(\left| \left\langle \tfrac{h_{J'}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})}{\sqrt{2\pi}}, \tfrac{\boldsymbol{\theta} + \hat{\boldsymbol{\theta}}}{\|\boldsymbol{\theta} + \hat{\boldsymbol{\theta}}\|_2} \right\rangle - \mathbb{E}\left[\left\langle \tfrac{h_{J'}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})}{\sqrt{2\pi}}, \tfrac{\boldsymbol{\theta} + \hat{\boldsymbol{\theta}}}{\|\boldsymbol{\theta} + \hat{\boldsymbol{\theta}}\|_2} \right\rangle \right] \right| > ut \middle| L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) \leq un \right) \leq 2e^{-\frac{1}{2}unt^2}, \tag{124}$$

$$P\left(\left\| \tfrac{\bar{h}_{J'}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})}{\sqrt{2\pi}} - \mathbb{E}\left[\tfrac{\bar{h}_{J'}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})}{\sqrt{2\pi}} \right] \right\|_2 > 2\sqrt{\tfrac{k_0' u}{n}} + ut \middle| L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) \leq un \right) \leq e^{-\frac{1}{8}unt^2}. \tag{125}$$

By symmetry, this implies that for any $\boldsymbol{\theta} \in \Theta$ and $\hat{\boldsymbol{\theta}} \in \mathcal{B}'_{2\tau}(\boldsymbol{\theta})$,

$$P\left(\left| \left\langle \tfrac{h_{J'}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})}{\sqrt{2\pi}}, \tfrac{\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}}{\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_2} \right\rangle - \mathbb{E}\left[\left\langle \tfrac{h_{J'}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})}{\sqrt{2\pi}}, \tfrac{\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}}{\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_2} \right\rangle \right] \right| > ut \middle| \forall \boldsymbol{\theta}' \in \Theta \;\; L(\boldsymbol{\theta}') \leq un \right) \leq 2e^{-\frac{1}{2}unt^2}, \tag{126}$$

$$P\left(\left| \left\langle \tfrac{h_{J'}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})}{\sqrt{2\pi}}, \tfrac{\boldsymbol{\theta} + \hat{\boldsymbol{\theta}}}{\|\boldsymbol{\theta} + \hat{\boldsymbol{\theta}}\|_2} \right\rangle - \mathbb{E}\left[\left\langle \tfrac{h_{J'}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})}{\sqrt{2\pi}}, \tfrac{\boldsymbol{\theta} + \hat{\boldsymbol{\theta}}}{\|\boldsymbol{\theta} + \hat{\boldsymbol{\theta}}\|_2} \right\rangle \right] \right| > ut \middle| \forall \boldsymbol{\theta}' \in \Theta \;\; L(\boldsymbol{\theta}') \leq un \right) \leq 2e^{-\frac{1}{2}unt^2}, \tag{127}$$

$$P\left(\left\| \tfrac{\bar{h}_{J'}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})}{\sqrt{2\pi}} - \mathbb{E}\left[\tfrac{\bar{h}_{J'}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})}{\sqrt{2\pi}} \right] \right\|_2 > 2\sqrt{\tfrac{k_0' u}{n}} + ut \middle| \forall \boldsymbol{\theta}' \in \Theta \;\; L(\boldsymbol{\theta}') \leq un \right) \leq e^{-\frac{1}{8}unt^2}, \tag{128}$$

where notations are taken from the above lemma.

To help condense notations in the upcoming analysis, define the following indicator random variables:

$$\begin{aligned} A_1(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}, J') &\triangleq \mathbb{I}\left(\left| \left\langle \tfrac{h_{J'}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})}{\sqrt{2\pi}}, \tfrac{\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}}{\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_2} \right\rangle - \mathbb{E}\left[\left\langle \tfrac{h_{J'}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})}{\sqrt{2\pi}}, \tfrac{\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}}{\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_2} \right\rangle \right] \right| > \eta t \right), \\ A_2(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}, J') &\triangleq \mathbb{I}\left(\left| \left\langle \tfrac{h_{J'}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})}{\sqrt{2\pi}}, \tfrac{\boldsymbol{\theta} + \hat{\boldsymbol{\theta}}}{\|\boldsymbol{\theta} + \hat{\boldsymbol{\theta}}\|_2} \right\rangle - \mathbb{E}\left[\left\langle \tfrac{h_{J'}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})}{\sqrt{2\pi}}, \tfrac{\boldsymbol{\theta} + \hat{\boldsymbol{\theta}}}{\|\boldsymbol{\theta} + \hat{\boldsymbol{\theta}}\|_2} \right\rangle \right] \right| > \eta t \right), \\ A_3(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}, J') &\triangleq \mathbb{I}\left(\left\| \tfrac{\bar{h}_{J'}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})}{\sqrt{2\pi}} - \mathbb{E}\left[\tfrac{\bar{h}_{J'}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})}{\sqrt{2\pi}} \right] \right\|_2 > 2\sqrt{\tfrac{k_0' \eta}{n}} + \eta t \right). \end{aligned}$$

Now, observe:

$$
\begin{aligned}
&P(A_1(\boldsymbol{\theta},\hat{\boldsymbol{\theta}},J')=1 \text{ or } A_2(\boldsymbol{\theta},\hat{\boldsymbol{\theta}},J')=1 \text{ or } A_3(\boldsymbol{\theta},\hat{\boldsymbol{\theta}},J')=1 \,|\, \forall \boldsymbol{\theta}' \in \Theta \;\; L(\boldsymbol{\theta}') \le \eta n) \\
&\quad \le P(A_1(\boldsymbol{\theta},\hat{\boldsymbol{\theta}},J')=1 \,|\, \forall \boldsymbol{\theta}' \in \Theta \;\; L(\boldsymbol{\theta}') \le \eta n) \\
&\qquad + P(A_2(\boldsymbol{\theta},\hat{\boldsymbol{\theta}},J')=1 \,|\, \forall \boldsymbol{\theta}' \in \Theta \;\; L(\boldsymbol{\theta}') \le \eta n) \\
&\qquad + P(A_3(\boldsymbol{\theta},\hat{\boldsymbol{\theta}},J')=1 \,|\, \forall \boldsymbol{\theta}' \in \Theta \;\; L(\boldsymbol{\theta}') \le \eta n) \\
&\qquad\quad \blacktriangleright \text{ by a union bound} \\
&\quad \le 2e^{-\frac{1}{2}\eta n t^2} + 2e^{-\frac{1}{2}\eta n t^2} + e^{-\frac{1}{8}\eta n t^2} \\
&\qquad\quad \blacktriangleright \text{ by Equations (126)–(128)} \\
&\quad \le 5e^{-\frac{1}{8}\eta n t^2}.
\end{aligned}
\tag{129}
$$

A uniform bound over all $J' \in \mathcal{J}'$, $\boldsymbol{\theta} \in \tilde{\mathcal{C}}$, and $\hat{\boldsymbol{\theta}} \in \mathcal{D}(\boldsymbol{\theta})$ is then obtained as follows:

$$
\begin{aligned}
&P(\exists J' \in \mathcal{J}', \boldsymbol{\theta} \in \tilde{\mathcal{C}}, \hat{\boldsymbol{\theta}} \in \mathcal{D}(\boldsymbol{\theta}) \;\; A_1(\boldsymbol{\theta},\hat{\boldsymbol{\theta}},J')=1 \text{ or } A_2(\boldsymbol{\theta},\hat{\boldsymbol{\theta}},J')=1 \text{ or } A_3(\boldsymbol{\theta},\hat{\boldsymbol{\theta}},J')=1) \\
&\quad = P(\exists J' \in \mathcal{J}', \boldsymbol{\theta} \in \tilde{\mathcal{C}}, \hat{\boldsymbol{\theta}} \in \mathcal{D}(\boldsymbol{\theta}) \;\; A_1(\boldsymbol{\theta},\hat{\boldsymbol{\theta}},J')=1 \text{ or } A_2(\boldsymbol{\theta},\hat{\boldsymbol{\theta}},J')=1 \text{ or } A_3(\boldsymbol{\theta},\hat{\boldsymbol{\theta}},J')=1 \\
&\qquad\quad | \, \forall \boldsymbol{\theta}' \in \Theta \;\; L(\boldsymbol{\theta}') \le \eta n) \\
&\qquad \cdot P(\forall \boldsymbol{\theta}' \in \Theta \;\; L(\boldsymbol{\theta}') \le \eta n) \\
&\qquad + P(\exists J' \in \mathcal{J}', \boldsymbol{\theta} \in \tilde{\mathcal{C}}, \hat{\boldsymbol{\theta}} \in \mathcal{D}(\boldsymbol{\theta}) \;\; A_1(\boldsymbol{\theta},\hat{\boldsymbol{\theta}},J')=1 \text{ or } A_2(\boldsymbol{\theta},\hat{\boldsymbol{\theta}},J')=1 \text{ or } A_3(\boldsymbol{\theta},\hat{\boldsymbol{\theta}},J')=1 \\
&\qquad\qquad | \, \exists \boldsymbol{\theta}' \in \Theta \;\; L(\boldsymbol{\theta}') > \eta n) \\
&\qquad\quad \cdot P(\exists \boldsymbol{\theta}' \in \Theta \;\; L(\boldsymbol{\theta}') > \eta n) \\
&\qquad\quad \blacktriangleright \text{ by the law of total probability and the definition of conditional probabilities} \\
&\quad \le P(\exists J' \in \mathcal{J}', \boldsymbol{\theta} \in \tilde{\mathcal{C}}, \hat{\boldsymbol{\theta}} \in \mathcal{D}(\boldsymbol{\theta}) \;\; A_1(\boldsymbol{\theta},\hat{\boldsymbol{\theta}},J')=1 \text{ or } A_2(\boldsymbol{\theta},\hat{\boldsymbol{\theta}},J')=1 \text{ or } A_3(\boldsymbol{\theta},\hat{\boldsymbol{\theta}},J')=1 \\
&\qquad\quad | \, \forall \boldsymbol{\theta}' \in \Theta \;\; L(\boldsymbol{\theta}') \le \eta n) \\
&\qquad + P(\exists \boldsymbol{\theta}' \in \Theta \;\; L(\boldsymbol{\theta}') > \eta n) \\
&\quad \le |\mathcal{J}'||\tilde{\mathcal{C}}|q \, P(A_1(\boldsymbol{\theta},\hat{\boldsymbol{\theta}},J')=1 \text{ or } A_2(\boldsymbol{\theta},\hat{\boldsymbol{\theta}},J')=1 \text{ or } A_3(\boldsymbol{\theta},\hat{\boldsymbol{\theta}},J')=1 \,|\, \forall \boldsymbol{\theta}' \in \Theta \;\; L(\boldsymbol{\theta}') \le \eta n) \\
&\qquad + P(\exists \boldsymbol{\theta}' \in \Theta \;\; L(\boldsymbol{\theta}') > \eta n) \\
&\qquad\quad \blacktriangleright \text{ for an arbitrary choice of } J' \in \mathcal{J}', \boldsymbol{\theta} \in \tilde{\mathcal{C}}, \hat{\boldsymbol{\theta}} \in \mathcal{D}(\boldsymbol{\theta}); \\
&\qquad\quad \blacktriangleright \text{ by a union bound and an earlier discussion about the cardinality of } \mathcal{D}(\cdot) \\
&\quad \le 5|\mathcal{J}'||\tilde{\mathcal{C}}|q e^{-\frac{1}{8}\eta n t^2} + \binom{d}{k} e^{-\frac{1}{64}\eta n} \\
&\qquad\quad \blacktriangleright \text{ by Equation (129) and Lemma 27} \\
&\quad \le 5|\mathcal{J}||\mathcal{C}|q e^{-\frac{1}{8}\eta n t^2} + \binom{d}{k} e^{-\frac{1}{64}\eta n}. \\
&\qquad\quad \blacktriangleright \because |\mathcal{J}'| \le |\mathcal{J}| \text{ and } |\tilde{\mathcal{C}}| \le |\mathcal{C}|
\end{aligned}
$$

Note that the last line follows from recalling the definition of $\mathcal{J}'$: $\mathcal{J}' \triangleq \{J \cup \operatorname{supp}(\boldsymbol{\theta}^*) : J \in \mathcal{J}\}$, which inserts at most one coordinate subset into $\mathcal{J}'$ for each coordinate subset in $\mathcal{J}$.

Returning to Equation (122) and now applying the results just derived above, the following bound holds for all $J' \in \mathcal{J}'$, $\boldsymbol{\theta} \in \tilde{\mathcal{C}}$, and $\hat{\boldsymbol{\theta}} \in \mathcal{D}(\boldsymbol{\theta})$ with probability at least $1 - 5|\mathcal{J}||\mathcal{C}|qe^{-\frac{1}{8}\eta n t^2} -$

$\binom{d}{k}e^{-\frac{1}{64}\eta n}$:

$$\begin{aligned}
\|h_{J'}(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}) - \mathbb{E}[h_{J'}(\boldsymbol{\theta},\hat{\boldsymbol{\theta}})]\|_2 &\le \left|\left\langle h_{J'}(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}), \frac{\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}}{\|\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\|_2}\right\rangle - \mathbb{E}\left[\left\langle h_{J'}(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}), \frac{\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}}{\|\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\|_2}\right\rangle\right]\right| \\
&\quad + \left|\left\langle h_{J'}(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}), \frac{\boldsymbol{\theta}+\hat{\boldsymbol{\theta}}}{\|\boldsymbol{\theta}+\hat{\boldsymbol{\theta}}\|_2}\right\rangle - \mathbb{E}\left[\left\langle h_{J'}(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}), \frac{\boldsymbol{\theta}+\hat{\boldsymbol{\theta}}}{\|\boldsymbol{\theta}+\hat{\boldsymbol{\theta}}\|_2}\right\rangle\right]\right| \\
&\quad + \|\bar{h}_{J'}(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}) - \mathbb{E}[\bar{h}_{J'}(\boldsymbol{\theta},\hat{\boldsymbol{\theta}})]\|_2 \\
&\le 2\sqrt{\frac{k_0'\eta}{n}} + 3\eta t \\
&\le 5\max\left\{\sqrt{\frac{k_0'\eta}{n}}, \eta t\right\}. \qquad (130)
\end{aligned}$$

Set

$$\begin{aligned}
t &= \sqrt{\frac{8\log\left(\frac{6}{\rho_2}|\mathcal{J}||\mathcal{C}|q\right)}{\eta n}} \\
&= \sqrt{\frac{8\log\left(\frac{6}{\rho_2}|\mathcal{J}||\mathcal{C}|\right)}{\eta n} + \frac{8\log\left(\sum_{\ell=0}^{\eta n}\binom{n}{\ell}\right)}{\eta n}} \\
&\le \sqrt{\frac{8\log\left(\frac{6}{\rho_2}|\mathcal{J}||\mathcal{C}|\right)}{\eta n} + 8\log\left(\frac{e}{\eta}\right)}
\end{aligned}$$

▶ by an earlier remark

$$\le \sqrt{\frac{8\log\left(\frac{6}{\rho_2}|\mathcal{J}||\mathcal{C}|\right)}{\eta n}} + \sqrt{8\log\left(\frac{e}{\eta}\right)}. \qquad (131)$$

▶ by the triangle inequality (in one dimension)

In accordance with Equation (62) of Lemma 20, let

$$n \ge \max\left\{\frac{200\eta\log\left(\frac{6}{\rho_2}|\mathcal{J}||\mathcal{C}|\right)}{\left(\sqrt{\frac{\pi}{8}}\gamma c'\delta - \eta\sqrt{8\log\left(\frac{e}{\eta}\right)}\right)^2}, \frac{200\eta k_0'}{\pi\gamma^2 c'^2\delta^2}, \frac{64}{\eta}\log\left(\frac{6}{\rho_2}\binom{d}{k}\right), \frac{c_4 k}{\eta}\log\left(\frac{1}{\eta}\right)\right\}, \qquad (132)$$

where $c', c_4 > 0$ are constants as per Definition 15 and η is as defined in Definition 16. Then, with probability at least

$$\begin{aligned}
&1 - 5|\mathcal{J}||\mathcal{C}|qe^{-\frac{1}{8}\eta n t^2} - \binom{d}{k}e^{-\frac{1}{64}\eta n} \\
&\ge 1 - 5|\mathcal{J}||\mathcal{C}|qe^{-\frac{1}{8}\eta n\cdot\frac{8}{\eta n}\log(\frac{6}{\rho_2}|\mathcal{J}||\mathcal{C}|q)} - \binom{d}{k}e^{-\frac{1}{64}\eta\cdot\frac{64}{\eta}\log(\frac{6}{\rho_2}\binom{d}{k})}
\end{aligned}$$

$$
\begin{aligned}
&\qquad \blacktriangleright \text{ by the choices of } t, n \text{ in Equations (131) and (132)} \\
&\geq 1 - \rho_2,
\end{aligned}
$$

for all $J' \in \mathcal{J}'$, $\boldsymbol{\theta} \in \tilde{\mathcal{C}}$, and $\hat{\boldsymbol{\theta}} \in \mathcal{D}(\boldsymbol{\theta})$,

$$
\begin{aligned}
\|h_{J'}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) - \mathbb{E}[h_{J'}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})]\|_2 &\leq 5 \max\left\{\sqrt{\frac{k_0'\eta}{n}}, \eta t\right\} \\
&\qquad \blacktriangleright \text{ by Equation (130)} \\
&= 5 \max\left\{\sqrt{\frac{k_0'\eta}{n}}, \sqrt{\frac{8\eta \log\left(\frac{6}{\rho_2}|\mathcal{J}||\mathcal{C}|\right)}{n}} + \eta\sqrt{8\log\left(\frac{e}{\eta}\right)}\right\} \\
&\qquad \blacktriangleright \text{ by the choice of } t \text{ specified in Equation (131)} \\
&\leq \sqrt{\frac{\pi}{8}}\gamma c' \delta. \\
&\qquad \blacktriangleright \text{ by the choice of } n \text{ specified in Equation (132)} \\
&\qquad \text{ and the definition of } \eta \text{ in Equation (53)}
\end{aligned}
$$

The bound in Equation (63) of Lemma 20 now follows: with probability at least $1 - \rho_2$, uniformly for every $J' \in \mathcal{J}'$, $\boldsymbol{\theta} \in \tilde{\mathcal{C}}$, and $\hat{\boldsymbol{\theta}} \in \mathcal{D}(\boldsymbol{\theta})$,

$$
\frac{2\|h_{J'}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) - \mathbb{E}[h_{J'}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})]\|_2}{\|\mathbb{E}[\boldsymbol{\theta} + h_f(\boldsymbol{\theta}^*, \boldsymbol{\theta})]\|_2} \leq \sqrt{\frac{8}{\pi}}\frac{1}{\gamma}\sqrt{\frac{\pi}{8}}\gamma c'\delta = c'\delta.
$$

■

Proof Lemma 21 Fix any $\boldsymbol{\theta}^* \in \Theta$. Let $J'' \in \mathcal{J}''$ be arbitrary. Once again, due to Equation (100) in Lemma 25,

$$
\|\mathbb{E}[\boldsymbol{\theta} + h_{f;J''}(\boldsymbol{\theta}^*, \boldsymbol{\theta})]\|_2 = \sqrt{\frac{\pi}{2}}\gamma. \tag{133}
$$

Then, inserting (133) into the right-hand-side Equation (65) in Lemma 21,

$$
\frac{2\|h_{f;J''}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) - \mathbb{E}[h_{f;J''}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)]\|_2}{\|\mathbb{E}[\boldsymbol{\theta} + h_f(\boldsymbol{\theta}^*, \boldsymbol{\theta})]\|_2} = \frac{2\|h_{f;J''}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) - \mathbb{E}[h_{f;J''}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)]\|_2}{\sqrt{\pi/2}\gamma}. \tag{134}
$$

The ℓ_2-distance between $h_{f;J''}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)$ and its mean, i.e., the term $\|h_{f;J''}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) - \mathbb{E}[h_{f;J''}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)]\|_2$ on the right-hand-side of Equation (134), is bound from above as follows. Take $s', t' > 0$ in Equation (97) in Lemma 25 to be

$$
s' \triangleq \sqrt{\frac{3\log\left(\frac{3}{\rho_3}\right)}{\alpha_0 n}} \tag{135}
$$

and

$$t' \triangleq \max\left\{\sqrt{\frac{3\log\left(\frac{6}{\rho_3}|\mathcal{J}''|\right)}{\alpha_0 n}}, \sqrt{\frac{2(1+s')\log\left(\frac{3}{\rho_3}|\mathcal{J}''|\right)}{\alpha_0 n}}\right\}, \tag{136}$$

and take n to be bounded from below by

$$n \geq \max\left\{\frac{64\alpha_0}{\gamma^2 c^2 \delta^2}\max\left\{3\log\left(\frac{6}{\rho_3}|\mathcal{J}||\mathcal{C}|\right), 2(k_0''-1)\right\}, \frac{4}{\alpha_0}\log\left(\frac{6}{\rho_3}|\mathcal{J}||\mathcal{C}|\right)\right\}, \tag{137}$$

where this choice of n satisfies

$$n \geq \max\left\{\frac{64\alpha_0}{\gamma^2 c^2 \delta^2}\max\left\{3\log\left(\frac{6}{\rho_3}|\mathcal{J}''|\right), (1+s')(k_0''-1)\right\}, \frac{4}{\alpha_0}\log\left(\frac{6}{\rho_3}|\mathcal{J}''|\right)\right\}$$

because $|\mathcal{J}''| \leq |\mathcal{J}||\mathcal{C}|$. This condition on n also ensures that $s', t' < 1$, as required to utilize Lemma 25. Then, observe:

$$\begin{aligned}
\sqrt{\frac{2\pi\alpha_0(1+s')(k_0''-1)}{n}} + \sqrt{2\pi}\alpha_0 t' &\leq \sqrt{\frac{2\pi\alpha_0(1+s')(k_0''-1)}{n}} + \sqrt{\frac{6\pi\alpha_0\log\left(\frac{6}{\rho_3}|\mathcal{J}''|\right)}{n}} \\
&\leq \frac{1}{2}\cdot\sqrt{\frac{\pi}{8}}\gamma c\delta + \frac{1}{2}\cdot\sqrt{\frac{\pi}{8}}\gamma c\delta \\
&= \sqrt{\frac{\pi}{8}}\gamma c\delta,
\end{aligned} \tag{138}$$

where $c > 0$ is a constant as per Definition 15. Together with Equation (115) from the proof of Lemma 19, Equation (138) gives way to the following bound:

$$\begin{aligned}
&P\left(\forall J'' \in \mathcal{J}'' \quad \|h_{J''}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) - \mathbb{E}\left[h_{J''}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)\right]\|_2 \leq \sqrt{\frac{\pi}{8}}\gamma c\delta\right) \\
&\geq P\left(\forall J'' \in \mathcal{J}'' \quad \|h_{J''}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) - \mathbb{E}\left[h_{J''}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)\right]\|_2 \leq \sqrt{\frac{2\pi\alpha_0(1+s')(k_0''-1)}{n}} + \sqrt{2\pi}\alpha_0 t'\right) \\
&\qquad \blacktriangleright \text{ due to Equations (115) and (138)} \\
&\geq 1 - 2|\mathcal{J}''|e^{-\frac{1}{3}\alpha_0 n t'^2} - |\mathcal{J}''|e^{-\frac{1}{2}\frac{\alpha_0 n t'^2}{1+s'}} - e^{-\frac{1}{3}\alpha_0 n s'^2} \\
&\qquad \blacktriangleright \text{ by Equation (97) in Lemma 25} \\
&\geq 1 - \frac{\rho_3}{3} - \frac{\rho_3}{3} - \frac{\rho_3}{3} \\
&\qquad \blacktriangleright \text{ by the choice of } s', t' \text{ in Equations (135) and (136)} \\
&= 1 - \rho_3.
\end{aligned} \tag{139}$$

Therefore, taken together, Equations (133) and (139) imply that if n is bounded from below as in Equation (137), then with probability at least $1 - \rho_3$, for all $J'' \in \mathcal{J}''$,

$$\frac{2\|h_{f;J''}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) - \mathbb{E}[h_{f;J''}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)]\|_2}{\|\mathbb{E}[\boldsymbol{\theta} + h_f(\boldsymbol{\theta}^*, \boldsymbol{\theta})]\|_2} = \frac{2\|h_{f;J''}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) - \mathbb{E}[h_{f;J''}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)]\|_2}{\sqrt{\pi/2}\gamma}$$

$$
\begin{aligned}
&\leq \sqrt{\frac{8}{\pi}}\frac{1}{\gamma}\sqrt{\frac{\pi}{8}}\gamma c\delta \\
&= c\delta,
\end{aligned}
$$

thus establishing Lemma 21. ■

Appendix D. Proof of the Concentration Inequalities, Lemma 25

D.1. Intermediate Results

We return to Lemma 25 to present its proof. Towards this, the following pair of intermediate lemmas, whose proofs can be found in Appendices D.3 and D.4, are provided below.

Lemma 30 *Fix $s, t, \tau \in (0,1)$, and let $\boldsymbol{\theta}^* \in \Theta$, $\mathcal{J} \subseteq 2^{[d]}$, and $\mathcal{C}, \tilde{\mathcal{C}} \subset \Theta$, where $\mathcal{J}$, $\mathcal{C}$, and $\tilde{\mathcal{C}}$ are finite sets, and where $\tilde{\mathcal{C}} \triangleq \mathcal{C} \setminus \mathcal{B}_\tau(\boldsymbol{\theta}^*)$. Let $k_0 \triangleq \min\{2k + \max_{J\in\mathcal{J}} |J|, d\}$. Then,*

$$
\begin{aligned}
&P\left(\exists J \in \mathcal{J}, \boldsymbol{\theta} \in \tilde{\mathcal{C}} \quad \left|\left\langle \frac{h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})}{\sqrt{2\pi}}, \frac{\boldsymbol{\theta}^* - \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2} \right\rangle - \mathbb{E}\left[\left\langle \frac{h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})}{\sqrt{2\pi}}, \frac{\boldsymbol{\theta}^* - \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2} \right\rangle\right]\right| > \frac{t \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta}\rangle)}{\pi}\right) \\
&\qquad \leq 2|\mathcal{J}||\tilde{\mathcal{C}}|e^{-\frac{1}{3\pi}nt^2 \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta}\rangle)}, \qquad (140) \\
&P\left(\exists J \in \mathcal{J}, \boldsymbol{\theta} \in \tilde{\mathcal{C}} \quad \left|\left\langle \frac{h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})}{\sqrt{2\pi}}, \frac{\boldsymbol{\theta}^* + \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* + \boldsymbol{\theta}\|_2} \right\rangle - \mathbb{E}\left[\left\langle \frac{h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})}{\sqrt{2\pi}}, \frac{\boldsymbol{\theta}^* + \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* + \boldsymbol{\theta}\|_2} \right\rangle\right]\right| > \frac{t \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta}\rangle)}{\pi}\right) \\
&\qquad \leq 2|\mathcal{J}||\tilde{\mathcal{C}}|e^{-\frac{1}{3\pi}nt^2 \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta}\rangle)}, \qquad (141) \\
&P\left(\exists J \in \mathcal{J}, \boldsymbol{\theta} \in \tilde{\mathcal{C}} \quad \left\|\frac{\bar{h}_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})}{\sqrt{2\pi}} - \mathbb{E}\left[\frac{\bar{h}_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})}{\sqrt{2\pi}}\right]\right\|_2 > \sqrt{\frac{2(1+s)(k_0-2)\arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta}\rangle)}{n}} + \frac{t \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta}\rangle)}{\pi}\right) \\
&\qquad \leq |\mathcal{J}||\tilde{\mathcal{C}}|e^{-\frac{1}{2\pi(1+s)}nt^2 \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta}\rangle)} + |\tilde{\mathcal{C}}|e^{-\frac{1}{3\pi}ns^2 \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta}\rangle)}, \qquad (142)
\end{aligned}
$$

where for any $J \subseteq [d]$ and $\boldsymbol{\theta} \in \Theta$,

$$
\mathbb{E}\left[\left\langle h(\boldsymbol{\theta}^*, \boldsymbol{\theta}), \frac{\boldsymbol{\theta}^* - \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2}\right\rangle\right] = \mathbb{E}\left[\left\langle h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}), \frac{\boldsymbol{\theta}^* - \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2}\right\rangle\right] = \|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2, \qquad (143)
$$

$$
\mathbb{E}\left[\left\langle h(\boldsymbol{\theta}^*, \boldsymbol{\theta}), \frac{\boldsymbol{\theta}^* + \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* + \boldsymbol{\theta}\|_2}\right\rangle\right] = \mathbb{E}\left[\left\langle h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}), \frac{\boldsymbol{\theta}^* + \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* + \boldsymbol{\theta}\|_2}\right\rangle\right] = 0, \qquad (144)
$$

$$
\mathbb{E}[\bar{h}(\boldsymbol{\theta}^*, \boldsymbol{\theta})] = \mathbb{E}[\bar{h}_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})] = \mathbf{0}. \qquad (145)
$$

Lemma 31 *Fix $t > 0$, $s' \in (0,1)$, and $\delta \in (0,1)$, and let $\mathcal{J}'' \subseteq 2^{[d]}$. Let $k_0'' \triangleq \min\{k + \max_{J''\in\mathcal{J}''} |J''|, d\}$, and define $\alpha_0 = \alpha_0(\delta) = \max\{\alpha, \delta\}$. Then, for $\boldsymbol{\theta}^* \in \Theta$,*

$$
P\left(\exists J'' \in \mathcal{J}'' \quad \left|\left\langle \frac{h_{f;J''}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)}{\sqrt{2\pi}}, \boldsymbol{\theta}^* \right\rangle - \mathbb{E}\left[\left\langle \frac{h_{f;J''}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)}{\sqrt{2\pi}}, \boldsymbol{\theta}^* \right\rangle\right]\right| > \alpha t\right) \leq 2|\mathcal{J}''|e^{-\frac{1}{3}\alpha n t^2}, \qquad (146)
$$

$$
\begin{aligned}
&P\left(\exists J'' \in \mathcal{J}'' \quad \left\|\frac{\bar{h}_{f;J''}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)}{\sqrt{2\pi}} - \mathbb{E}\left[\frac{\bar{h}_{f;J''}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)}{\sqrt{2\pi}}\right]\right\|_2 > \sqrt{\frac{\alpha_0(1+s')(k_0''-1)}{n}} + \alpha_0 t\right) \\
&\qquad \leq |\mathcal{J}''|e^{-\frac{1}{2(1+s')}\alpha_0 n t^2} + e^{-\frac{1}{3}\alpha_0 n s'^2}, \qquad (147)
\end{aligned}
$$

where for any $J'' \subseteq [d]$,

$$\mathbb{E}\left[\langle h_f(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*), \boldsymbol{\theta}^* \rangle\right] = \mathbb{E}\left[\left\langle h_{f;J''}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*), \boldsymbol{\theta}^* \right\rangle\right] = -\left(1 - \sqrt{\frac{\pi}{2}}\gamma\right), \tag{148}$$

$$\mathbb{E}[\bar{h}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)] = \mathbb{E}[\bar{h}_{J''}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)] = \mathbf{0}. \tag{149}$$

D.2. Proof of Lemma 25

We are ready to prove Lemma 25 by means of the intermediate lemmas in Appendix D.1.
Proof Lemma 25 The proof of the lemma is split across Appendices D.2.1 and D.2.2, where the former derives Equations (96) and (97) and the latter establishes Equations (98)–(100).

D.2.1. PROOF OF EQUATIONS (96) AND (97)

Verification of Equation (96) Fix $\boldsymbol{\theta}^* \in \Theta$, $\boldsymbol{\theta} \in \tilde{\mathcal{C}}$, and $J \in \mathcal{J}$ arbitrarily. Towards bounding the concentration of $h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})$ around its mean, consider the following orthogonal decomposition:

$$\begin{aligned} h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = &\left\langle h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}), \frac{\boldsymbol{\theta}^* - \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2} \right\rangle \frac{\boldsymbol{\theta}^* - \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2} \\ &+ \left\langle h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}), \frac{\boldsymbol{\theta}^* + \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* + \boldsymbol{\theta}\|_2} \right\rangle \frac{\boldsymbol{\theta}^* + \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* + \boldsymbol{\theta}\|_2} + \bar{h}_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}), \end{aligned} \tag{150}$$

where, recalling Equation (47),

$$\begin{aligned} \bar{h}_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}) &- \left\langle h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}), \frac{\boldsymbol{\theta}^* - \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2} \right\rangle \frac{\boldsymbol{\theta}^* - \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2} \\ &- \left\langle h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}), \frac{\boldsymbol{\theta}^* + \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* + \boldsymbol{\theta}\|_2} \right\rangle \frac{\boldsymbol{\theta}^* + \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* + \boldsymbol{\theta}\|_2}. \end{aligned} \tag{151}$$

Due to Equation (150) and the linearity of expectation, the centered random vector $h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}) - \mathbb{E}[h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})]$ has the following orthogonal decomposition:

$$\begin{aligned} &h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}) - \mathbb{E}[h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})] \\ &= \left(\left\langle h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}), \frac{\boldsymbol{\theta}^* - \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2} \right\rangle - \mathbb{E}\left[\left\langle h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}), \frac{\boldsymbol{\theta}^* - \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2} \right\rangle \right] \right) \frac{\boldsymbol{\theta}^* - \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2} \\ &\quad + \left(\left\langle h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}), \frac{\boldsymbol{\theta}^* + \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* + \boldsymbol{\theta}\|_2} \right\rangle - \mathbb{E}\left[\left\langle h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}), \frac{\boldsymbol{\theta}^* + \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* + \boldsymbol{\theta}\|_2} \right\rangle \right] \right) \frac{\boldsymbol{\theta}^* + \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* + \boldsymbol{\theta}\|_2} \\ &\quad + \left(\bar{h}_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}) - \mathbb{E}[\bar{h}_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})] \right) \end{aligned} \tag{152}$$

due to Equation (150) and the linearity of expectation. Applying the triangle inequality to the ℓ_2-norm of the orthogonal decomposition in Equation (152) and scaling it by a factor of $\frac{1}{\sqrt{2\pi}}$ yields:

$$\begin{aligned} &\left\| \frac{1}{\sqrt{2\pi}} h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}) - \mathbb{E}\left[\frac{1}{\sqrt{2\pi}} h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}) \right] \right\|_2 \\ &\quad \leq \left| \left\langle \frac{1}{\sqrt{2\pi}} h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}), \frac{\boldsymbol{\theta}^* - \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2} \right\rangle - \mathbb{E}\left[\left\langle \frac{1}{\sqrt{2\pi}} h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}), \frac{\boldsymbol{\theta}^* - \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2} \right\rangle \right] \right| \end{aligned}$$

$$+\left|\left\langle \frac{1}{\sqrt{2\pi}}h_J(\boldsymbol{\theta}^*,\boldsymbol{\theta}), \frac{\boldsymbol{\theta}^*+\boldsymbol{\theta}}{\|\boldsymbol{\theta}^*+\boldsymbol{\theta}\|_2}\right\rangle - \mathbb{E}\left[\left\langle \frac{1}{\sqrt{2\pi}}h_J(\boldsymbol{\theta}^*,\boldsymbol{\theta}), \frac{\boldsymbol{\theta}^*+\boldsymbol{\theta}}{\|\boldsymbol{\theta}^*+\boldsymbol{\theta}\|_2}\right\rangle\right]\right|$$
$$+\left\|\frac{1}{\sqrt{2\pi}}\bar{h}_J(\boldsymbol{\theta}^*,\boldsymbol{\theta}) - \mathbb{E}\left[\frac{1}{\sqrt{2\pi}}\bar{h}_J(\boldsymbol{\theta}^*,\boldsymbol{\theta})\right]\right\|_2. \tag{153}$$

Due to Lemma 30, the three terms in the last expression in Equation (153) are individually controlled with bounded probability as follows:

$$P\left(\exists J\in\mathcal{J},\boldsymbol{\theta}\in\tilde{\mathcal{C}}\quad \left|\left\langle \frac{h_J(\boldsymbol{\theta}^*,\boldsymbol{\theta})}{\sqrt{2\pi}}, \frac{\boldsymbol{\theta}^*-\boldsymbol{\theta}}{\|\boldsymbol{\theta}^*-\boldsymbol{\theta}\|_2}\right\rangle - \mathbb{E}\left[\left\langle \frac{h_J(\boldsymbol{\theta}^*,\boldsymbol{\theta})}{\sqrt{2\pi}}, \frac{\boldsymbol{\theta}^*-\boldsymbol{\theta}}{\|\boldsymbol{\theta}^*-\boldsymbol{\theta}\|_2}\right\rangle\right]\right| > \frac{t\arccos(\langle\boldsymbol{\theta}^*,\boldsymbol{\theta}\rangle)}{3\pi}\right)$$
$$\leq 2|\mathcal{J}||\tilde{\mathcal{C}}|e^{-\frac{1}{27\pi}nt^2\arccos(\langle\boldsymbol{\theta}^*,\boldsymbol{\theta}\rangle)},$$
$$P\left(\exists J\in\mathcal{J},\boldsymbol{\theta}\in\tilde{\mathcal{C}}\quad \left|\left\langle \frac{h_J(\boldsymbol{\theta}^*,\boldsymbol{\theta})}{\sqrt{2\pi}}, \frac{\boldsymbol{\theta}^*+\boldsymbol{\theta}}{\|\boldsymbol{\theta}^*+\boldsymbol{\theta}\|_2}\right\rangle - \mathbb{E}\left[\left\langle \frac{h_J(\boldsymbol{\theta}^*,\boldsymbol{\theta})}{\sqrt{2\pi}}, \frac{\boldsymbol{\theta}^*+\boldsymbol{\theta}}{\|\boldsymbol{\theta}^*+\boldsymbol{\theta}\|_2}\right\rangle\right]\right| > \frac{t\arccos(\langle\boldsymbol{\theta}^*,\boldsymbol{\theta}\rangle)}{3\pi}\right)$$
$$\leq 2|\mathcal{J}||\tilde{\mathcal{C}}|e^{-\frac{1}{27\pi}nt^2\arccos(\langle\boldsymbol{\theta}^*,\boldsymbol{\theta}\rangle)},$$
$$P\Big(\exists J\in\mathcal{J},\boldsymbol{\theta}\in\tilde{\mathcal{C}}\quad \left\|\frac{\bar{h}_J(\boldsymbol{\theta}^*,\boldsymbol{\theta})}{\sqrt{2\pi}} - \mathbb{E}\left[\frac{\bar{h}_J(\boldsymbol{\theta}^*,\boldsymbol{\theta})}{\sqrt{2\pi}}\right]\right\|_2 > \sqrt{\frac{2(1+s)(k_0-2)\arccos(\langle\boldsymbol{\theta}^*,\boldsymbol{\theta}\rangle)}{n}} + \frac{t\arccos(\langle\boldsymbol{\theta}^*,\boldsymbol{\theta}\rangle)}{3\pi}\Big)$$
$$\leq |\mathcal{J}||\tilde{\mathcal{C}}|e^{-\frac{1}{18\pi(1+s)}nt^2\arccos(\langle\boldsymbol{\theta}^*,\boldsymbol{\theta}\rangle)} + |\tilde{\mathcal{C}}|e^{-\frac{1}{3\pi}ns^2\arccos(\langle\boldsymbol{\theta}^*,\boldsymbol{\theta}\rangle)}.$$

Combining the three above concentration inequalities via a union bound and complementing, it follows that with probability at least

$$1-2|\mathcal{J}||\tilde{\mathcal{C}}|e^{-\frac{1}{27\pi}nt^2\arccos(\langle\boldsymbol{\theta}^*,\boldsymbol{\theta}\rangle)} - 2|\mathcal{J}||\tilde{\mathcal{C}}|e^{-\frac{1}{27\pi}nt^2\arccos(\langle\boldsymbol{\theta}^*,\boldsymbol{\theta}\rangle)} - |\mathcal{J}||\tilde{\mathcal{C}}|e^{-\frac{1}{18\pi(1+s)}nt^2\arccos(\langle\boldsymbol{\theta}^*,\boldsymbol{\theta}\rangle)}$$
$$-|\tilde{\mathcal{C}}|e^{-\frac{1}{3\pi}ns^2\arccos(\langle\boldsymbol{\theta}^*,\boldsymbol{\theta}\rangle)}$$
$$=1-4|\mathcal{J}||\tilde{\mathcal{C}}|e^{-\frac{1}{27\pi}nt^2\arccos(\langle\boldsymbol{\theta}^*,\boldsymbol{\theta}\rangle)} - |\mathcal{J}||\tilde{\mathcal{C}}|e^{-\frac{1}{18\pi(1+s)}nt^2\arccos(\langle\boldsymbol{\theta}^*,\boldsymbol{\theta}\rangle)} - |\tilde{\mathcal{C}}|e^{-\frac{1}{3\pi}ns^2\arccos(\langle\boldsymbol{\theta}^*,\boldsymbol{\theta}\rangle)},$$

for all $J\in\mathcal{J}$ and all $\boldsymbol{\theta}\in\tilde{\mathcal{C}}$, the following three inequalities hold simultaneously:

$$\left|\left\langle \frac{h_J(\boldsymbol{\theta}^*,\boldsymbol{\theta})}{\sqrt{2\pi}}, \frac{\boldsymbol{\theta}^*-\boldsymbol{\theta}}{\|\boldsymbol{\theta}^*-\boldsymbol{\theta}\|_2}\right\rangle - \mathbb{E}\left[\left\langle \frac{h_J(\boldsymbol{\theta}^*,\boldsymbol{\theta})}{\sqrt{2\pi}}, \frac{\boldsymbol{\theta}^*-\boldsymbol{\theta}}{\|\boldsymbol{\theta}^*-\boldsymbol{\theta}\|_2}\right\rangle\right]\right| \leq \frac{t\arccos(\langle\boldsymbol{\theta}^*,\boldsymbol{\theta}\rangle)}{3\pi},$$
$$\left|\left\langle \frac{h_J(\boldsymbol{\theta}^*,\boldsymbol{\theta})}{\sqrt{2\pi}}, \frac{\boldsymbol{\theta}^*+\boldsymbol{\theta}}{\|\boldsymbol{\theta}^*+\boldsymbol{\theta}\|_2}\right\rangle - \mathbb{E}\left[\left\langle \frac{h_J(\boldsymbol{\theta}^*,\boldsymbol{\theta})}{\sqrt{2\pi}}, \frac{\boldsymbol{\theta}^*+\boldsymbol{\theta}}{\|\boldsymbol{\theta}^*+\boldsymbol{\theta}\|_2}\right\rangle\right]\right| \leq \frac{t\arccos(\langle\boldsymbol{\theta}^*,\boldsymbol{\theta}\rangle)}{3\pi},$$
$$\left\|\frac{\bar{h}_J(\boldsymbol{\theta}^*,\boldsymbol{\theta})}{\sqrt{2\pi}} - \mathbb{E}\left[\frac{\bar{h}_J(\boldsymbol{\theta}^*,\boldsymbol{\theta})}{\sqrt{2\pi}}\right]\right\|_2 \leq \sqrt{\frac{2(1+s)(k_0-2)\arccos(\langle\boldsymbol{\theta}^*,\boldsymbol{\theta}\rangle)}{n}} + \frac{t\arccos(\langle\boldsymbol{\theta}^*,\boldsymbol{\theta}\rangle)}{3\pi}.$$

Then, taking this with Equation (153), this implies that with probability at least

$$1-4|\mathcal{J}||\tilde{\mathcal{C}}|e^{-\frac{1}{27\pi}nt^2\arccos(\langle\boldsymbol{\theta}^*,\boldsymbol{\theta}\rangle)} - |\mathcal{J}||\tilde{\mathcal{C}}|e^{-\frac{1}{18\pi(1+s)}nt^2\arccos(\langle\boldsymbol{\theta}^*,\boldsymbol{\theta}\rangle)} - |\tilde{\mathcal{C}}|e^{-\frac{1}{3\pi}ns^2\arccos(\langle\boldsymbol{\theta}^*,\boldsymbol{\theta}\rangle)},$$

uniformly for all $J\in\mathcal{J}$ and all $\boldsymbol{\theta}\in\tilde{\mathcal{C}}$,

$$\left\|\frac{h_J(\boldsymbol{\theta}^*,\boldsymbol{\theta})}{\sqrt{2\pi}} - \mathbb{E}\left[\frac{h_J(\boldsymbol{\theta}^*,\boldsymbol{\theta})}{\sqrt{2\pi}}\right]\right\|_2 \leq \sqrt{\frac{2(1+s)(k_0-2)\arccos(\langle\boldsymbol{\theta}^*,\boldsymbol{\theta}\rangle)}{n}} + \frac{t\arccos(\langle\boldsymbol{\theta}^*,\boldsymbol{\theta}\rangle)}{\pi},$$

as desired. This establishes Equation (96).

Verification of Equation (97) Next, Equation (97) is derived via an analogous technique. Again, an orthogonal decomposition will facilitate the proof, this time into just two components:

$$h_{f;J}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*) = \langle h_{f;J}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*),\boldsymbol{\theta}^*\rangle\boldsymbol{\theta}^* + \bar{h}_{f;J}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*), \tag{154}$$

where, as defined in Equation (49),

$$\bar{h}_{f;J}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*) = h_{f;J}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*) - \langle h_{f;J}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*),\boldsymbol{\theta}^*\rangle\boldsymbol{\theta}^*. \tag{155}$$

By the above orthogonal decomposition in (154) and the linearity of expectation,

$$\begin{aligned}
&h_{f;J}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*) - \mathbb{E}[h_{f;J}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)] \\
&\quad= \langle h_{f;J}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*),\boldsymbol{\theta}^*\rangle\boldsymbol{\theta}^* + \bar{h}_{f;J}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*) - \mathbb{E}[\langle h_{f;J}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*),\boldsymbol{\theta}^*\rangle\boldsymbol{\theta}^* + \bar{h}_{f;J}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)] \\
&\quad= (\langle h_{f;J}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*),\boldsymbol{\theta}^*\rangle\boldsymbol{\theta}^* - \mathbb{E}[\langle h_{f;J}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*),\boldsymbol{\theta}^*\rangle\boldsymbol{\theta}^*]) + \left(\bar{h}_{f;J}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*) - \mathbb{E}[\bar{h}_{f;J}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)]\right) \\
&\quad= (\langle h_{f;J}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*),\boldsymbol{\theta}^*\rangle - \mathbb{E}[\langle h_{f;J}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*),\boldsymbol{\theta}^*\rangle])\,\boldsymbol{\theta}^* + \left(\bar{h}_{f;J}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*) - \mathbb{E}[\bar{h}_{f;J}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)]\right).
\end{aligned}$$

Then, taking the norm of the above expressions and applying the triangle inequality to the last line,

$$\begin{aligned}
&\|h_{f;J}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*) - \mathbb{E}[h_{f;J}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)]\|_2 \\
&\quad\le \|(\langle h_{f;J}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*),\boldsymbol{\theta}^*\rangle - \mathbb{E}[\langle h_{f;J}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*),\boldsymbol{\theta}^*\rangle])\,\boldsymbol{\theta}^*\|_2 + \left\|\bar{h}_{f;J}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*) - \mathbb{E}[\bar{h}_{f;J}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)]\right\|_2 \\
&\quad= |\langle h_{f;J}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*),\boldsymbol{\theta}^*\rangle - \mathbb{E}[\langle h_{f;J}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*),\boldsymbol{\theta}^*\rangle]| + \left\|\bar{h}_{f;J}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*) - \mathbb{E}[\bar{h}_{f;J}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)]\right\|_2.
\end{aligned}$$

Scaling this by a factor of $\frac{1}{\sqrt{2\pi}}$ yields:

$$\begin{aligned}
&\left\|\frac{h_{f;J}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)}{\sqrt{2\pi}} - \mathbb{E}\left[\frac{h_{f;J}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)}{\sqrt{2\pi}}\right]\right\|_2 \\
&\quad\le \left|\left\langle \frac{h_{f;J}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)}{\sqrt{2\pi}},\boldsymbol{\theta}^*\right\rangle - \mathbb{E}\left[\left\langle \frac{h_{f;J}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)}{\sqrt{2\pi}},\boldsymbol{\theta}^*\right\rangle\right]\right| \\
&\qquad+ \left\|\frac{\bar{h}_{f;J}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)}{\sqrt{2\pi}} - \mathbb{E}\left[\frac{\bar{h}_{f;J}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)}{\sqrt{2\pi}}\right]\right\|_2.
\end{aligned}$$

By Lemma 31, for $s', t' \in (0,1)$,

$$\begin{aligned}
&P\left(\exists J'' \in \mathcal{J}'' \quad \left|\left\langle \frac{h_{f;J''}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)}{\sqrt{2\pi}},\boldsymbol{\theta}^*\right\rangle - \mathbb{E}\left[\left\langle \frac{h_{f;J''}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)}{\sqrt{2\pi}},\boldsymbol{\theta}^*\right\rangle\right]\right| > \frac{\alpha_0 t'}{2}\right) \\
&\quad= P\left(\exists J'' \in \mathcal{J}'' \quad \left|\left\langle \frac{h_{f;J''}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)}{\sqrt{2\pi}},\boldsymbol{\theta}^*\right\rangle - \mathbb{E}\left[\left\langle \frac{h_{f;J''}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)}{\sqrt{2\pi}},\boldsymbol{\theta}^*\right\rangle\right]\right| > \alpha\left(\frac{\alpha_0 t'}{2\alpha}\right)\right) \\
&\quad= 2|\mathcal{J}''|e^{-\frac{1}{12}\frac{\alpha_0}{\alpha}\alpha_0 n t'^2} \\
&\qquad\qquad \blacktriangleright \text{ due to Lemma 31} \\
&\quad\le 2|\mathcal{J}''|e^{-\frac{1}{12}\alpha_0 n t'^2}, \\
&\qquad\qquad \blacktriangleright\ \alpha_0 = \max\{\alpha,\delta\} \ge \alpha \text{ implies } \tfrac{\alpha_0}{\alpha} \ge 1
\end{aligned}$$

and

$$P\left(\exists J'' \in \mathcal{J}'' \quad \left\|\frac{\bar{h}_{f;J''}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)}{\sqrt{2\pi}} - \mathbb{E}\left[\frac{\bar{h}_{f;J''}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)}{\sqrt{2\pi}}\right]\right\|_2 > \sqrt{\frac{\alpha_0(1+s')(k_0''-1)}{n}} + \frac{\alpha_0 t'}{2}\right)$$

$$\le |\mathcal{J}''| e^{-\frac{1}{8(1+s')}\alpha_0 n t'^2} + e^{-\frac{1}{3}\alpha_0 n s'^2},$$

and hence, by a union bound over the above two probabilities, with probability at least

$$1 - 2|\mathcal{J}''| e^{-\frac{1}{12}\alpha_0 n t'^2} - |\mathcal{J}''| e^{-\frac{1}{8(1+s')}\alpha_0 n t'^2} - e^{-\frac{1}{3}\alpha_0 n s'^2},$$

the norm of the centered random vector $\frac{1}{\sqrt{2\pi}} h_{f;J}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) - \mathbb{E}[\frac{1}{\sqrt{2\pi}} h_{f;J}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)]$ is bounded from above by

$$\begin{aligned}
\left\| \frac{h_{f;J}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)}{\sqrt{2\pi}} - \mathbb{E}\left[\frac{h_{f;J}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)}{\sqrt{2\pi}} \right] \right\|_2 &\le \left| \left\langle \frac{h_{f;J}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)}{\sqrt{2\pi}}, \boldsymbol{\theta}^* \right\rangle - \mathbb{E}\left[\left\langle \frac{h_{f;J}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)}{\sqrt{2\pi}}, \boldsymbol{\theta}^* \right\rangle \right] \right| \\
&\quad + \left\| \frac{\bar{h}_{f;J}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)}{\sqrt{2\pi}} - \mathbb{E}\left[\frac{\bar{h}_{f;J}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)}{\sqrt{2\pi}} \right] \right\|_2 \\
&\le \frac{\alpha_0 t'}{2} + \sqrt{\frac{\alpha_0 (1+s')(k_0''-1)}{n}} + \frac{\alpha_0 t'}{2} \\
&= \sqrt{\frac{\alpha_0 (1+s')(k_0''-1)}{n}} + \alpha_0 t'.
\end{aligned}$$

Thus, Equation (97) holds.

D.2.2. PROOF OF EQUATIONS (98)–(100)

Next, the four expectations, Equations (98)–(100), in Lemma 31 are verified. Let $J \subseteq [d]$ be an arbitrary coordinate subset. Note that it suffices to establish the results for h_J as those for h immediately follow by taking $J = [d]$.

Verification of Equation (98) Towards establishing the first expectation, Equation (98), recall the orthogonal decomposition in Equation (154) in the proof of Equation (96):

$$\begin{aligned}
h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = \left\langle h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}), \frac{\boldsymbol{\theta}^* - \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2} \right\rangle \frac{\boldsymbol{\theta}^* - \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2} + \left\langle h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}), \frac{\boldsymbol{\theta}^* + \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* + \boldsymbol{\theta}\|_2} \right\rangle \frac{\boldsymbol{\theta}^* + \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* + \boldsymbol{\theta}\|_2} \\
+ \bar{h}_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}),
\end{aligned} \tag{156}$$

where $\bar{h}_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})$ is given in Equation (47) or (151). Hence, in expectation,

$$\begin{aligned}
\mathbb{E}[h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})] &= \mathbb{E}\left[\left\langle h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}), \frac{\boldsymbol{\theta}^* - \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2} \right\rangle \right] \frac{\boldsymbol{\theta}^* - \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2} \\
&\quad + \mathbb{E}\left[\left\langle h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}), \frac{\boldsymbol{\theta}^* + \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* + \boldsymbol{\theta}\|_2} \right\rangle \right] \frac{\boldsymbol{\theta}^* + \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* + \boldsymbol{\theta}\|_2} + \mathbb{E}\left[\bar{h}_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}) \right]
\end{aligned} \tag{157}$$

by Equation (156) and the linearity of expectation. Due to Lemma 30,

$$\mathbb{E}\left[\left\langle h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}), \frac{\boldsymbol{\theta}^* - \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2} \right\rangle \right] \frac{\boldsymbol{\theta}^* - \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2} = \|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2 \frac{\boldsymbol{\theta}^* - \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2} = \boldsymbol{\theta}^* - \boldsymbol{\theta}, \tag{158}$$

$$\mathbb{E}\left[\left\langle h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}), \frac{\boldsymbol{\theta}^* + \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* + \boldsymbol{\theta}\|_2} \right\rangle \right] \frac{\boldsymbol{\theta}^* + \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* + \boldsymbol{\theta}\|_2} = 0 \cdot \frac{\boldsymbol{\theta}^* + \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* + \boldsymbol{\theta}\|_2} = \mathbf{0}, \tag{159}$$

$$\mathbb{E}[\bar{h}_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})] = \mathbf{0}. \tag{160}$$

Then, by Equations (157)–(160),

$$
\begin{aligned}
&\mathbb{E}[h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})] \\
&\quad = \mathbb{E}\left[\left\langle h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}), \frac{\boldsymbol{\theta}^* - \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2} \right\rangle\right] \frac{\boldsymbol{\theta}^* - \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2} + \mathbb{E}\left[\left\langle h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}), \frac{\boldsymbol{\theta}^* + \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* + \boldsymbol{\theta}\|_2} \right\rangle\right] \frac{\boldsymbol{\theta}^* + \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* + \boldsymbol{\theta}\|_2} \\
&\qquad + \mathbb{E}\left[\bar{h}_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})\right] \\
&\qquad\qquad \blacktriangleright \text{ by Equation (157)} \\
&\quad = \boldsymbol{\theta}^* - \boldsymbol{\theta} + \mathbf{0} + \mathbf{0} \\
&\qquad\qquad \blacktriangleright \text{ by Equations (158)–(160)} \\
&\quad = \boldsymbol{\theta}^* - \boldsymbol{\theta},
\end{aligned}
$$

as claimed.

Verification of Equation (99) To verify Equation (99), we turn to the orthogonal decomposition in Equation (154) used to prove Equation (97):

$$
h_{f;J}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) = \langle h_{f;J}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*), \boldsymbol{\theta}^* \rangle \boldsymbol{\theta}^* + \bar{h}_{f;J}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*),
$$

where $h_{f;J}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)$ is as stated in Equation (49) or (155). With this decomposition, due to the linearity of expectation,

$$
\mathbb{E}[h_{f;J}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)] = \mathbb{E}[\langle h_{f;J}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*), \boldsymbol{\theta}^* \rangle] \boldsymbol{\theta}^* + \mathbb{E}[\bar{h}_{f;J}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)], \tag{161}
$$

where the last line uses the fact that $\boldsymbol{\theta}^*$ is nonrandom. Recall from Lemma 31 that

$$
\begin{aligned}
\mathbb{E}[\langle h_{f;J}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*), \boldsymbol{\theta}^* \rangle] &= -\left(1 - \sqrt{\frac{\pi}{2}}\gamma\right), \\
\mathbb{E}[\bar{h}_{f;J}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)] &= \mathbf{0}.
\end{aligned}
$$

Thus, continuing the above derivation in (161), Equation (99) follows:

$$
\mathbb{E}[h_{f;J}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)] = -\left(1 - \sqrt{\frac{\pi}{2}}\gamma\right) \boldsymbol{\theta}^*.
$$

Verification of Equation (100) Observe:

$$
\begin{aligned}
&h_{f;J}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) \\
&\quad = T_{\operatorname{supp}(\boldsymbol{\theta}^*) \cup \operatorname{supp}(\boldsymbol{\theta}) \cup J}\left(\frac{\sqrt{2\pi}}{n} \mathbf{X}^T \frac{1}{2}(f(\mathbf{X}\boldsymbol{\theta}^*) - \operatorname{sign}(\mathbf{X}\boldsymbol{\theta}))\right) \\
&\qquad\qquad \blacktriangleright \text{ by the definition of } h_{f;J} \text{ in Equation (21)} \\
&\quad = T_{\operatorname{supp}(\boldsymbol{\theta}^*) \cup \operatorname{supp}(\boldsymbol{\theta}) \cup J}\left(\frac{\sqrt{2\pi}}{n} \mathbf{X}^T \frac{1}{2}(\operatorname{sign}(\mathbf{X}\boldsymbol{\theta}^*) - \operatorname{sign}(\mathbf{X}\boldsymbol{\theta}))\right) \\
&\qquad + T_{\operatorname{supp}(\boldsymbol{\theta}^*) \cup \operatorname{supp}(\boldsymbol{\theta}) \cup J}\left(\frac{\sqrt{2\pi}}{n} \mathbf{X}^T \frac{1}{2}(f(\mathbf{X}\boldsymbol{\theta}^*) - \operatorname{sign}(\mathbf{X}\boldsymbol{\theta}^*))\right)
\end{aligned}
$$

$$\blacktriangleright \text{ the subset thresholding operation is a linear transformation (see, Section 3)}$$
$$= h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}) + h_{f;\mathrm{supp}(\boldsymbol{\theta})\cup J}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*). \tag{162}$$
$$\blacktriangleright \text{ by the definitions of } h_J \text{ and } h_{f;J} \text{ in Equations (19) and (21), respectively}$$

Thus,

$$\mathbb{E}[h_{f;J}(\boldsymbol{\theta}^*, \boldsymbol{\theta})] = \mathbb{E}[h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}) + h_{f;\mathrm{supp}(\boldsymbol{\theta})\cup J}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)] = \mathbb{E}[h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})] + \mathbb{E}[h_{f;\mathrm{supp}(\boldsymbol{\theta})\cup J}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)], \tag{163}$$

where the first equality applies Equation (162) and the second equality follows from the linearity of expectation. By Equations (98) and (99), respectively,

$$\mathbb{E}[h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})] = \boldsymbol{\theta}^* - \boldsymbol{\theta}, \tag{164}$$
$$\mathbb{E}[h_{f;\mathrm{supp}(\boldsymbol{\theta})\cup J}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)] = -\left(1 - \sqrt{\frac{\pi}{2}}\gamma\right)\boldsymbol{\theta}^*, \tag{165}$$

and therefore,

$$\begin{aligned}
\mathbb{E}[\boldsymbol{\theta} + h_{f;J}(\boldsymbol{\theta}^*, \boldsymbol{\theta})] &= \boldsymbol{\theta} + \mathbb{E}[h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})] + \mathbb{E}[h_{f;\mathrm{supp}(\boldsymbol{\theta})\cup J}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)] \\
&\qquad \blacktriangleright \text{ by Equation (163)} \\
&= \boldsymbol{\theta} + \boldsymbol{\theta}^* - \boldsymbol{\theta} - \left(1 - \sqrt{\frac{\pi}{2}}\gamma\right)\boldsymbol{\theta}^* \\
&\qquad \blacktriangleright \text{ by Equations (164) and (165)} \\
&= \sqrt{\frac{\pi}{2}}\gamma\boldsymbol{\theta}^*. \qquad (166)
\end{aligned}$$

Then,

$$\| \, \mathbb{E}\left[\boldsymbol{\theta} + h_{f;J}(\boldsymbol{\theta}^*, \boldsymbol{\theta})\right] \|_2 = \sqrt{\frac{\pi}{2}}\gamma\|\boldsymbol{\theta}^*\|_2 = \sqrt{\frac{\pi}{2}}\gamma,$$

where the first equality applies Equation (166), the second follows from the homogeneity of the (ℓ_2-)norm, and the third equality recalls that $\boldsymbol{\theta}^*$ has unit ℓ_2-norm. This completes the proof of Lemma 25. ■

D.3. Proof of Lemma 30

Proof Lemma 30 The proof of the expectations, Equations (143)–(145), in Lemma 30 are presented in Appendix D.3.1, while the concentration inequalities in Equations (140)–(142) are proved in Appendix D.3.2.

D.3.1. Proof the Expectations, Equations (143)–(145)

The expectations, Equations (143)–(145), follow largely from work already done by Matsumoto and Mazumdar (2024a), which is summarized below as Lemma 32.

Lemma 32 (due to (Matsumoto and Mazumdar, 2024a, Appendix B)) *Fix $\boldsymbol{\theta}^*, \boldsymbol{\theta} \in \Theta$ and $\ell \in \{0, \ldots, n\}$. Let $L \triangleq \|\mathbb{I}(\mathrm{sign}(\mathbf{X}\boldsymbol{\theta}^*) \neq \mathrm{sign}(\mathbf{X}\boldsymbol{\theta}))\|_0$. Then,*

$$\mathop{\mathbb{E}}_{\mathbf{X}|L}\left[\left\langle h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}), \frac{\boldsymbol{\theta}^* - \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2}\right\rangle \middle| L = \ell\right] = \frac{\pi \ell \|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2}{n \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta}\rangle)}, \tag{167}$$

$$\mathop{\mathbb{E}}_{\mathbf{X}|L}\left[\left\langle h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}), \frac{\boldsymbol{\theta}^* + \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* + \boldsymbol{\theta}\|_2}\right\rangle \middle| L = \ell\right] = 0, \tag{168}$$

$$\mathop{\mathbb{E}}_{\mathbf{X}|L}[\bar{h}_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}) \mid L = \ell] = \mathbf{0}. \tag{169}$$

Taking $\boldsymbol{\theta}^*, \boldsymbol{\theta} \in \Theta$ arbitrarily, via the law of total expectation, Equations (144) and (145) follow from Equations (168) and (169), respectively:

$$\mathop{\mathbb{E}}_{\mathbf{X}}\left[\left\langle h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}), \frac{\boldsymbol{\theta}^* + \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* + \boldsymbol{\theta}\|_2}\right\rangle\right] = \mathop{\mathbb{E}}_{L}\left[\mathop{\mathbb{E}}_{\mathbf{X}|L}\left[\left\langle h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}), \frac{\boldsymbol{\theta}^* + \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* + \boldsymbol{\theta}\|_2}\right\rangle \middle| L\right]\right] = \mathop{\mathbb{E}}_{L}[0] = 0, \tag{170}$$

$$\mathop{\mathbb{E}}_{\mathbf{X}}[\bar{h}_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})] = \mathop{\mathbb{E}}_{L}\left[\mathop{\mathbb{E}}_{\mathbf{X}|L}\left[\bar{h}_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}) \middle| L\right]\right] = \mathop{\mathbb{E}}_{L}[\mathbf{0}] = \mathbf{0}. \tag{171}$$

Note that because $J \subseteq [d]$ is arbitrary, Equations (170) and (171) further imply that

$$\mathop{\mathbb{E}}_{\mathbf{X}}\left[\left\langle h(\boldsymbol{\theta}^*, \boldsymbol{\theta}), \frac{\boldsymbol{\theta}^* + \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* + \boldsymbol{\theta}\|_2}\right\rangle\right] = 0,$$
$$\mathop{\mathbb{E}}_{\mathbf{X}}[\bar{h}(\boldsymbol{\theta}^*, \boldsymbol{\theta})] = \mathbf{0}$$

by taking $J = [d]$.

Proceeding to Equation (143), define the random variable $L \triangleq \|\mathbb{I}(\mathrm{sign}(\mathbf{X}\boldsymbol{\theta}^*) \neq \mathrm{sign}(\mathbf{X}\boldsymbol{\theta}))\|_0$ as in Lemma 32. To derive the result, first note that L follows a binomial distribution: $L \sim \mathrm{Binomial}(n, \frac{1}{\pi} \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta}\rangle))$, where the characterization of this random variable, L, is folklore (see, e.g., Charikar (2002)). Then, observe:

$$\begin{aligned}
&\mathbb{E}\left[\left\langle h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}), \frac{\boldsymbol{\theta}^* - \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2}\right\rangle\right] \\
&\quad = \mathop{\mathbb{E}}_{L}\left[\mathop{\mathbb{E}}_{\mathbf{X}|L}\left[\left\langle h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}), \frac{\boldsymbol{\theta}^* - \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2}\right\rangle \middle| L\right]\right] \\
&\qquad \blacktriangleright \text{ by the law of total expectation} \\
&\quad = \sum_{\ell=0}^{n} \binom{n}{\ell} \left(\frac{\arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta}\rangle)}{\pi}\right)^{\ell} \left(1 - \frac{\arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta}\rangle)}{\pi}\right)^{n-\ell} \mathop{\mathbb{E}}_{\mathbf{X}|L}\left[\left\langle h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}), \frac{\boldsymbol{\theta}^* - \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2}\right\rangle \middle| L = \ell\right] \\
&\qquad \blacktriangleright \text{ by the law of the lazy statistician and the mass function of } L \sim \mathrm{Binomial}(n, \tfrac{1}{\pi} \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta}\rangle)) \\
&\quad = \frac{\pi \|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2}{n \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta}\rangle)} \sum_{\ell=0}^{n} \binom{n}{\ell} \left(\frac{\arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta}\rangle)}{\pi}\right)^{\ell} \left(1 - \frac{\arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta}\rangle)}{\pi}\right)^{n-\ell} \ell \\
&\qquad \blacktriangleright \text{ by Equation (167)} \\
&\quad = \frac{\pi \|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2}{n \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta}\rangle)} \mathbb{E}[L] \\
&\qquad \blacktriangleright \text{ by the definition of expectation and the mass function of a binomial random variable}
\end{aligned}$$

$$= \|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2.$$

▶ by the expectation of a binomial random variable, $\mathbb{E}[L] = \frac{1}{\pi} n \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle)$

Again, because $J \subseteq [d]$ is arbitrary, it directly follows from the above derivation that

$$\mathbb{E}\left[\left\langle h(\boldsymbol{\theta}^*, \boldsymbol{\theta}), \frac{\boldsymbol{\theta}^* - \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2} \right\rangle\right] = \|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2.$$

This verifies Equation (143).

D.3.2. Proof of the Concentration Inequalities, Equations (140)–(142)

Next, we turn our attention to Equations (140)–(142). We begin with some preliminary analysis that will facilitate the derivations of these equations. Initially, fix $\boldsymbol{\theta} \in \tilde{\mathcal{C}}$ and $J \in \mathcal{J}$ arbitrarily, where later $\boldsymbol{\theta}$ and J will be varied over the entire sets $\tilde{\mathcal{C}}$ and $\mathcal{J}$, respectively, in union bounds. Write $\tilde{\mathbf{x}}_i \triangleq T_{\mathrm{supp}(\boldsymbol{\theta}^*) \cup \mathrm{supp}(\boldsymbol{\theta}) \cup J}(\mathbf{x}_i) \in \mathbb{R}^d$, $i \in [n]$. The definition of $\frac{1}{\sqrt{2\pi}} h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})$ in Equation (19) can be rewritten as follows:

$$\frac{h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})}{\sqrt{2\pi}} = T_{\mathrm{supp}(\boldsymbol{\theta}^*) \cup \mathrm{supp}(\boldsymbol{\theta}) \cup J}\left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \, \frac{1}{2}\left(\mathrm{sign}\left(\langle \mathbf{x}, \boldsymbol{\theta}^* \rangle\right) - \mathrm{sign}\left(\langle \mathbf{x}, \boldsymbol{\theta} \rangle\right)\right)\right)$$

▶ by the definition of h_J in Equation (19)

$$= \frac{1}{n}\sum_{i=1}^{n} \tilde{\mathbf{x}}_i \, \frac{1}{2}\left(\mathrm{sign}\left(\langle \mathbf{x}_i, \boldsymbol{\theta}^* \rangle\right) - \mathrm{sign}\left(\langle \mathbf{x}_i, \boldsymbol{\theta} \rangle\right)\right)$$

▶ by the linearity of the map $T_{\mathrm{supp}(\boldsymbol{\theta}^*) \cup \mathrm{supp}(\boldsymbol{\theta}) \cup J}$
(see, Section 3), and by the definition of $\tilde{\mathbf{x}}_i$, $i \in [n]$

$$= \frac{1}{n}\sum_{i=1}^{n} \tilde{\mathbf{x}}_i \, \frac{1}{2}\left(\mathrm{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle\right) - \mathrm{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta} \rangle\right)\right)$$

▶ $\mathrm{supp}(\boldsymbol{\theta}^*), \mathrm{supp}(\boldsymbol{\theta}) \subseteq \mathrm{supp}(\boldsymbol{\theta}^*) \cup \mathrm{supp}(\boldsymbol{\theta}) \cup J$

$$= \frac{1}{n}\sum_{i=1}^{n} \tilde{\mathbf{x}}_i \, \mathrm{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle\right) \mathbb{I}\big(\mathrm{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle\right) \neq \mathrm{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta} \rangle\right)\big). \tag{172}$$

▶ see, justification below

The last line can be verified by checking the value taken by $\frac{1}{2}(\mathrm{sign}(u) - \mathrm{sign}(v))$ at $u, v \in \mathbb{R}$ for each possible pair values of $\mathrm{sign}(u), \mathrm{sign}(v) \in \{-1, 1\}$:

$$\frac{1}{2}(\mathrm{sign}\,(u) - \mathrm{sign}\,(v)) = \begin{cases} 0, & \text{if } \mathrm{sign}(u) = \mathrm{sign}(v) = 1, \\ 0, & \text{if } \mathrm{sign}(u) = \mathrm{sign}(v) = -1, \\ 1, & \text{if } \mathrm{sign}(u) = 1 \neq 1 = \mathrm{sign}(v), \\ -1, & \text{if } \mathrm{sign}(u) = -1 \neq 1 = \mathrm{sign}(v), \end{cases}$$

$$= \mathrm{sign}\,(u)\, \mathbb{I}\big(\mathrm{sign}\,(u) \neq \mathrm{sign}\,(v)\big).$$

Therefore,

$$\left\langle \frac{1}{\sqrt{2\pi}} h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}), \frac{\boldsymbol{\theta}^* - \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2} \right\rangle$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left\langle \tilde{\mathbf{x}}_i, \frac{\boldsymbol{\theta}^* - \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2} \right\rangle \operatorname{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle\right) \mathbb{I}\big(\operatorname{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle\right) \neq \operatorname{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta} \rangle\right)\big)$$

▶ by Equation (172) and the linearity of inner products

$$= \frac{1}{n} \sum_{i=1}^{n} \left\langle \tilde{\mathbf{x}}_i, \frac{\boldsymbol{\theta}^* - \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2} \right\rangle \operatorname{sign}\left(\left\langle \tilde{\mathbf{x}}_i, \frac{\boldsymbol{\theta}^* - \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2} \right\rangle \right) \mathbb{I}\big(\operatorname{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle\right) \neq \operatorname{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta} \rangle\right)\big) \tag{173}$$

▶ see, justification below

$$= \frac{1}{n} \sum_{i=1}^{n} \left| \left\langle \tilde{\mathbf{x}}_i, \frac{\boldsymbol{\theta}^* - \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2} \right\rangle \right| \mathbb{I}\big(\operatorname{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle\right) \neq \operatorname{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta} \rangle\right)\big), \tag{174}$$

▶ $u \operatorname{sign}(u) = |u|$ for any $u \in \mathbb{R}$

where the second to last equality, (173), follows from the observation that either the indicator term takes the value 0, or otherwise, if $\operatorname{sign}(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle) \neq \operatorname{sign}(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta} \rangle)$, then $\operatorname{sign}(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta} \rangle) = -\operatorname{sign}(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle)$, and hence,

$$\begin{aligned}
\operatorname{sign}\left(\left\langle \tilde{\mathbf{x}}_i, \frac{\boldsymbol{\theta}^* - \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2} \right\rangle \right) &= \operatorname{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle - \langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta} \rangle\right) \\
&= \operatorname{sign}\left(\operatorname{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle\right) |\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle| - \operatorname{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta} \rangle\right) |\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta} \rangle|\right) \\
&= \operatorname{sign}\left(\operatorname{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle\right) |\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle| + \operatorname{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle\right) |\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta} \rangle|\right) \\
&= \operatorname{sign}\left(\operatorname{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle\right) \left(|\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle| + |\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta} \rangle|\right)\right) \\
&= \operatorname{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle\right).
\end{aligned}$$

With the above work out of the way, we are ready to derive Equations (140)–(142).

Verification of Equation (140) For $i \in [n]$, let

$$U_i \triangleq |\langle \tilde{\mathbf{x}}_i, \frac{\boldsymbol{\theta}^* - \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2} \rangle| \, \mathbb{I}\big(\operatorname{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle\right) \neq \operatorname{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta} \rangle\right)\big),$$

and let

$$R_i \triangleq \mathbb{I}\big(\operatorname{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle\right) \neq \operatorname{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta} \rangle\right)\big).$$

Note that although this definition of R_i differs slightly from a similar random variable analyzed in (Matsumoto and Mazumdar, 2024a, Appendix B), nearly the same arguments as those in Matsumoto and Mazumdar (2024a) apply here, and hence we omit the analysis here. Due to the analysis in (Matsumoto and Mazumdar, 2024a, Appendix B.1.1), the mass function of the random variable R_i is given by

$$f_{R_i}(r) = \begin{cases} 1 - \frac{1}{\pi} \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle), & \text{if } r = 0, \\ \frac{1}{\pi} \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle), & \text{if } r = 1, \end{cases} \tag{175}$$

for $r \in \{0,1\}$. For $z \in \mathbb{R}$ and $r \in \{0,1\}$, the density function of the conditioned random variable $U_i \mid R_i$ is given by

$$f_{U_i|R_i}(z \mid r) = \begin{cases} 0, & \text{if } r = 0, z \neq 0, \\ 1, & \text{if } r = 0, z = 0, \\ 0, & \text{if } r = 1, z < 0, \\ \frac{\pi}{\arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle)} \sqrt{\frac{2}{\pi}} e^{-\frac{1}{2}z^2} \frac{1}{\sqrt{2\pi}} \int_{y=-z\tan(\frac{1}{2}\arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle))}^{y=z\tan(\frac{1}{2}\arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle))} e^{-\frac{1}{2}y^2} dy, & \text{if } r = 1, z \geq 0. \end{cases} \tag{176}$$

Having specified the density function of the conditioned random variable $U_i \mid R_i$, the next step is obtaining the mgfs of the centered random variables $(U_i \mid R_i) - \mathbb{E}[U_i \mid R_i]$ and $(-U_i \mid R_i) - \mathbb{E}[-U_i \mid R_i]$. To simplify notation, write $\mu_1 \triangleq \mathbb{E}[U_i \mid R_i = 1]$ and $\mu_0 \triangleq \mathbb{E}[U_i \mid R_i = 0]$, where in the latter case,

$$\mu_0 = \mathbb{E}[U_i \mid R_i = 0] = \int_{z=-\infty}^{z=\infty} z f_{U_i|R_i}(z \mid 0) dz = 0 f_{U_i|R_i}(0 \mid 0) = 0. \tag{177}$$

Due to (Matsumoto and Mazumdar, 2024a, Appendix B.1), the mgf of $(U_i \mid R_i = 1) - \mathbb{E}[U_i \mid R_i = 1]$, denoted by $\psi_{(U_i|R_i=1)-\mathbb{E}[U_i|R_i=1]}$, is given and upper bounded at $s \in [0, \infty)$ by

$$\begin{aligned} &\psi_{(U_i|R_i=1)-\mathbb{E}[U_i|R_i=1]}(s) \\ &= \mathbb{E}\left[e^{s(U_i - \mathbb{E}[U_i])} \middle| R_i = 1 \right] \\ &= e^{\frac{1}{2}s^2} e^{-s\mu_1} \frac{\pi}{\arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle)} \sqrt{\frac{2}{\pi}} \int_{z=0}^{z=\infty} e^{-\frac{1}{2}(z-s)^2} \frac{1}{\sqrt{2\pi}} \int_{y=-z\tan(\frac{1}{2}\arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle))}^{y=z\tan(\frac{1}{2}\arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle))} e^{-\frac{1}{2}y^2} dy \, dz \\ &\qquad \blacktriangleright \text{ by the law of the lazy statistician and Equation (176)} \\ &\leq e^{\frac{1}{2}s^2}. \end{aligned} \tag{178}$$

Likewise, the mgf of $(-U_i \mid R_i = 1) - \mathbb{E}[-U_i \mid R_i = 1]$, denoted by $\psi_{(-U_i|R_i=1)-\mathbb{E}[-U_i|R_i=1]}$, is upper bounded at $s \in [0, \infty)$ by

$$\psi_{(-U_i|R_i=1)-\mathbb{E}[-U_i|R_i=1]}(s) \leq e^{\frac{1}{2}s^2}. \tag{179}$$

On the other hand, when conditioning on $R_i = 0$, the mgf of the centered conditioned random variable $(U_i \mid R_i = 0) - \mathbb{E}[U_i \mid R_i = 0]$, written $\psi_{(U_i|R_i=0)-\mathbb{E}[U_i|R_i=0]}$, is given at $s \in [0, \infty)$ by

$$\begin{aligned} \psi_{(U_i|R_i=0)-\mathbb{E}[U_i|R_i=0]}(s) &= \mathbb{E}\left[e^{sU_i} \middle| R_i = 0 \right] \\ &\qquad \blacktriangleright \text{ using Equation (177)} \\ &= e^{s \cdot 0} f_{U_i|R_i}(0 \mid 0) \\ &\qquad \blacktriangleright \text{ by the law of the lazy statistician and Equation (176), and} \\ &\qquad\quad \text{since the mass of } U_i \mid R_i = 0 \text{ is entirely concentrated at } 0 \\ &= 1, \end{aligned} \tag{180}$$

and the mgf of the centered conditioned random variable $(-U_i \mid R_i = 0) - \mathbb{E}[-U_i \mid R_i = 0]$, written $\psi_{(-U_i|R_i=0)-\mathbb{E}[-U_i|R_i=0]}$, is similarly given at $s \in [0, \infty)$ by

$$\psi_{(-U_i|R_i=0)-\mathbb{E}[-U_i|R_i=0]}(s) = 1. \tag{181}$$

Taking together the two cases for $R_i = 1$ and $R_i = 0$, the mgf of $U_i - \mathbb{E}[U_i]$, written $\psi_{U_i-\mathbb{E}[U_i]}$, is given and bounded from above at $s \in [0, \infty)$ by

$$\begin{aligned}
\psi_{U_i-\mathbb{E}[U_i]}(s) &= \mathbb{E}\left[e^{s(U_i-\mathbb{E}[U_i])}\right] \\
&= f_{R_i}(1)\,\mathbb{E}\left[e^{s(U_i-\mathbb{E}[U_i])}\middle| R_i = 1\right] + f_{R_i}(0)\,\mathbb{E}\left[e^{s(U_i-\mathbb{E}[U_i])}\middle| R_i = 0\right] \\
&\qquad \blacktriangleright \text{ by the law of total expectation} \\
&= f_{R_i}(1)\psi_{(U_i|R_i=1)-\mathbb{E}[U_i|R_i=1]}(s) + f_{R_i}(0)\psi_{(U_i|R_i=0)-\mathbb{E}[U_i|R_i=0]}(s) \\
&\qquad \blacktriangleright \text{ by the definition of mgfs} \\
&= \frac{1}{\pi}\arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta}\rangle)\psi_{(U_i|R_i=1)-\mathbb{E}[U_i|R_i=1]}(s) \\
&\quad + \left(1 - \frac{1}{\pi}\arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta}\rangle)\right)\psi_{(U_i|R_i=0)-\mathbb{E}[U_i|R_i=0]}(s) \\
&\qquad \blacktriangleright \text{ by Equation (175)} \\
&\le \frac{1}{\pi}\arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta}\rangle)e^{\frac{1}{2}s^2} + \left(1 - \frac{1}{\pi}\arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta}\rangle)\right) \\
&\qquad \blacktriangleright \text{ by Equations (175), (178), and (180)} \\
&= 1 + \frac{1}{\pi}\arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta}\rangle)\left(e^{\frac{1}{2}s^2} - 1\right),
\end{aligned} \tag{182}$$

The mgf of $-U_i - \mathbb{E}[-U_i]$, denoted by $\psi_{-U_i-\mathbb{E}[-U_i]}$, is similarly bounded from above by

$$\psi_{-U_i-\mathbb{E}[-U_i]}(s) \le 1 + \frac{1}{\pi}\arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta}\rangle)\left(e^{\frac{1}{2}s^2} - 1\right). \tag{183}$$

Let $U \triangleq \sum_{i=1}^{n} U_i$. Using the above bound on the mgfs of $U_i - \mathbb{E}[U_i]$, $i \in [n]$, in Equation (182), the mgf of $U - \mathbb{E}[U]$, written $\psi_{U-\mathbb{E}[U]}$, is given and upper bounded at $s \in [0, \infty)$ as follows:

$$\begin{aligned}
\psi_{U-\mathbb{E}[U]}(s) &= \mathbb{E}\left[e^{s(U-\mathbb{E}[U])}\right] \\
&\qquad \blacktriangleright \text{ by the definition of the mgf } \psi_{U-\mathbb{E}[U]} \\
&= \mathbb{E}\left[e^{s\sum_{i=1}^{n} U_i-\mathbb{E}[U_i]}\right] \\
&\qquad \blacktriangleright \text{ by the definition of } U \\
&= \prod_{i=1}^{n}\mathbb{E}\left[e^{s(U_i-\mathbb{E}[U_i])}\right] \\
&\qquad \blacktriangleright \text{ since } U_1, \dots, U_n \text{ are mutually independent} \\
&= (\psi_{U_i-\mathbb{E}[U_i]}(s))^n \\
&\qquad \blacktriangleright \text{ for any } i \in [n];
\end{aligned}$$

$$
\begin{aligned}
&\qquad \blacktriangleright \text{ by the definition of } \psi_{U_i - \mathbb{E}[U_i]},\ i \in [n], \text{ and} \\
&\qquad\quad \text{since } U_1, \ldots, U_n \text{ are identically distributed} \\
&\leq \left(1 + \frac{1}{\pi} \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle) \left(e^{\frac{1}{2}s^2} - 1\right)\right)^n \\
&\qquad \blacktriangleright \text{ by Equation (182)} \\
&\leq e^{\frac{1}{\pi} n \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle)(e^{\frac{1}{2}s^2} - 1)}, \qquad (184) \\
&\qquad \blacktriangleright \text{ by a well-known inequality, } \log(1+u) \leq u \text{ for } u > -1
\end{aligned}
$$

Likewise, applying Equation (183) for the negated random variable, $-U - \mathbb{E}[-U]$, obtains:

$$
\psi_{-U - \mathbb{E}[-U]}(s) \leq e^{\frac{1}{\pi} n \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle)(e^{\frac{1}{2}s^2} - 1)}. \tag{185}
$$

Then, due to Bernstein (see, e.g., Vershynin (2018)),

$$
\begin{aligned}
P\left(\frac{U}{n} - \mathbb{E}\left[\frac{U}{n}\right] > \frac{1}{\pi} t \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle)\right) &\leq \inf_{s \geq 0} e^{-\frac{1}{\pi} nst \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle)} \psi_{U - \mathbb{E}[U]}(s) \\
&\qquad \blacktriangleright \text{ due to Bernstein} \\
&\leq \inf_{s \geq 0} e^{-\frac{1}{\pi} nst \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle)} e^{\frac{1}{\pi} n \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle)(e^{\frac{1}{2}s^2} - 1)} \\
&\qquad \blacktriangleright \text{ by Equation (184)} \\
&= \inf_{s \geq 0} e^{-\frac{1}{\pi} n \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle)(st - e^{\frac{1}{2}s^2} + 1)}, \qquad (186)
\end{aligned}
$$

and for the concentration on the other side, an analogous derivation gives

$$
P\left(\frac{U}{n} - \mathbb{E}\left[\frac{U}{n}\right] < -\frac{1}{\pi} t \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle)\right) \leq \inf_{s \geq 0} e^{-\frac{1}{\pi} n \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle)(st - e^{\frac{1}{2}s^2} + 1)}. \tag{187}
$$

To minimize the last expressions in Equations (186) and (187), observe that when $s, t > 0$ are small,

$$
\left.\frac{\partial}{\partial s} st - e^{\frac{1}{2}s^2} + 1\right|_{s=t} = \left. t - s e^{\frac{1}{2}s^2}\right|_{s=t} \approx 0,
$$

where

$$
\frac{\partial^2}{\partial s^2} st - e^{\frac{1}{2}s^2} + 1 = -(1 + s^2) e^{\frac{1}{2}s^2} < 0
$$

for any $s \in \mathbb{R}$. Hence, for small $s, t \in (0, 1]$, the expression $st - e^{\frac{1}{2}s^2} + 1$ is maximized with respect to s at approximately $s \approx t$ (which minimizes Equations (186) and (187)), and moreover, when $t \in (0, 1]$,

$$
\left. st - e^{\frac{1}{2}s^2} + 1 \right|_{s=t} = t^2 - e^{\frac{1}{2}t^2} + 1 \geq \frac{t^2}{3}.
$$

Therefore, taking $s = t$ in Equations (186) and (187) yields the following concentration inequalities for $t \in (0, 1]$:

$$
P\left(\frac{U}{n} - \mathbb{E}\left[\frac{U}{n}\right] > \frac{1}{\pi} t \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle)\right) \leq e^{-\frac{1}{3\pi} nt^2 \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle)},
$$

$$P\left(\frac{U}{n}-\mathbb{E}\left[\frac{U}{n}\right]<-\frac{1}{\pi}t\arccos(\langle\boldsymbol{\theta}^*,\boldsymbol{\theta}\rangle)\right)\leq e^{-\frac{1}{3\pi}nt^2\arccos(\langle\boldsymbol{\theta}^*,\boldsymbol{\theta}\rangle)},$$

and combining these via a union bound subsequently obtains:

$$P\left(\left|\frac{U}{n}-\mathbb{E}\left[\frac{U}{n}\right]\right|>\frac{1}{\pi}t\arccos(\langle\boldsymbol{\theta}^*,\boldsymbol{\theta}\rangle)\right)\leq 2e^{-\frac{1}{3\pi}nt^2\arccos(\langle\boldsymbol{\theta}^*,\boldsymbol{\theta}\rangle)}.$$

This, along with Equation (174) and the definition of the random variable U, immediately implies that

$$P\left(\left|\left\langle\frac{h_J(\boldsymbol{\theta}^*,\boldsymbol{\theta})}{\sqrt{2\pi}},\frac{\boldsymbol{\theta}^*-\boldsymbol{\theta}}{\|\boldsymbol{\theta}^*-\boldsymbol{\theta}\|_2}\right\rangle-\mathbb{E}\left[\left\langle\frac{h_J(\boldsymbol{\theta}^*,\boldsymbol{\theta})}{\sqrt{2\pi}},\frac{\boldsymbol{\theta}^*-\boldsymbol{\theta}}{\|\boldsymbol{\theta}^*-\boldsymbol{\theta}\|_2}\right\rangle\right]\right|>\frac{t\arccos(\langle\boldsymbol{\theta}^*,\boldsymbol{\theta}\rangle)}{\pi}\right)$$
$$\leq 2e^{-\frac{1}{3\pi}nt^2\arccos(\langle\boldsymbol{\theta}^*,\boldsymbol{\theta}\rangle)}.$$

Then, Equation (140) follows from union bounds over all $J\in\mathcal{J}$ and $\boldsymbol{\theta}\in\tilde{\mathcal{C}}$:

$$P\left(\exists J\in\mathcal{J},\boldsymbol{\theta}\in\tilde{\mathcal{C}}\;\left|\left\langle\frac{h_J(\boldsymbol{\theta}^*,\boldsymbol{\theta})}{\sqrt{2\pi}},\frac{\boldsymbol{\theta}^*-\boldsymbol{\theta}}{\|\boldsymbol{\theta}^*-\boldsymbol{\theta}\|_2}\right\rangle-\mathbb{E}\left[\left\langle\frac{h_J(\boldsymbol{\theta}^*,\boldsymbol{\theta})}{\sqrt{2\pi}},\frac{\boldsymbol{\theta}^*-\boldsymbol{\theta}}{\|\boldsymbol{\theta}^*-\boldsymbol{\theta}\|_2}\right\rangle\right]\right|>\frac{t\arccos(\langle\boldsymbol{\theta}^*,\boldsymbol{\theta}\rangle)}{\pi}\right)$$
$$\leq 2|\mathcal{J}||\tilde{\mathcal{C}}|e^{-\frac{1}{3\pi}nt^2\arccos(\langle\boldsymbol{\theta}^*,\boldsymbol{\theta}\rangle)},$$

as desired.

Verification of Equation (141) Next, the concentration inequality in Equation (141) will be verified. Again, take any $J\in\mathcal{J}$ and $\boldsymbol{\theta}\in\tilde{\mathcal{C}}$. Write the random variables $V_i\triangleq\langle\tilde{\mathbf{x}}_i,\frac{\boldsymbol{\theta}^*+\boldsymbol{\theta}}{\|\boldsymbol{\theta}^*+\boldsymbol{\theta}\|_2}\rangle$, $i\in[n]$, and carry over the notation of the random variables $R_i\triangleq\mathbb{I}(\mathrm{sign}(\langle\tilde{\mathbf{x}}_i,\boldsymbol{\theta}^*\rangle)\neq\mathrm{sign}(\langle\tilde{\mathbf{x}}_i,\boldsymbol{\theta}\rangle))$, $i\in[n]$, from the verification of Equation (140), whose mass functions are given in Equation (175). Due to (Matsumoto and Mazumdar, 2024a, Lemma B.6), the conditioned random variable $V_i\,|\,R_i=1$ is standard Gaussian, i.e., $(V_i\,|\,R_i=1)\sim\mathcal{N}(0,1)$, and thus, the density function of $V_i\,|\,R_i$ is given for $z\in\mathbb{R}$ and $r\in\{0,1\}$ by:

$$f_{V_i|R_i}(z\,|\,r)=\begin{cases}0, & \text{if } r=0, z\neq 0,\\ 1, & \text{if } r=0, z=0,\\ \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}, & \text{if } r=1.\end{cases}$$

Since $V_i\,|\,R_i=1$ is standard Gaussian, its expectation is $\mathbb{E}[V_i\,|\,R_i=1]=0$, while also,

$$\mathbb{E}[V_i\,|\,R_i=0]=\int_{z=-\infty}^{z=\infty}zf_{V_i|R_i}(z\,|\,0)dz=0f_{V_i|R_i}(0\,|\,0)=0.$$

With this, the mgfs of the (centered) conditioned random variables $V_i\,|\,R_i$ and $-V_i\,|\,R_i$ are obtained for $s\in[0,\infty)$ as follows:

$$\begin{aligned}\psi_{V_i|R_i=1}(s)&=\mathbb{E}\left[e^{sV_i}\middle|R_i=1\right]\\&=\frac{1}{\sqrt{2\pi}}\int_{z=-\infty}^{z=\infty}e^{sz}e^{-\frac{1}{2}z^2}dz\\&=e^{\frac{1}{2}s^2}\frac{1}{\sqrt{2\pi}}\int_{z=-\infty}^{z=\infty}e^{-\frac{1}{2}(z-s)^2}dz\end{aligned}$$

$$= e^{\frac{1}{2}s^2}, \tag{188}$$

▶ by evaluating the density of a mean-s, variance-1 Gaussian random variable over its entire support

and by an analogous derivation,

$$\psi_{-V_i|R_i=1}(s) = e^{\frac{1}{2}s^2}. \tag{189}$$

Additionally,

$$\psi_{V_i|R_i=0}(s) = \mathbb{E}\left[e^{sV_i} \middle| R_i = 0\right] = \int_{z=-\infty}^{z=\infty} e^{sz} f_{V_i|R_i}(z \,|\, 0) dz = e^{s \cdot 0} = 1, \tag{190}$$

$$\psi_{-V_i|R_i=0}(s) = \mathbb{E}\left[e^{-sV_i} \middle| R_i = 0\right] = \int_{z=-\infty}^{z=\infty} e^{-sz} f_{V_i|R_i}(z \,|\, 0) dz = e^{-s \cdot 0} = 1. \tag{191}$$

It follows that the mgf of V_i is bounded from above by

$$\begin{aligned}
\psi_{V_i}(s) &= \mathbb{E}\left[e^{sV_i}\right] \\
&= f_{R_i}(1)\, \mathbb{E}\left[e^{sV_i} \middle| R_i = 1\right] + f_{R_i}(0)\, \mathbb{E}\left[e^{sV_i} \middle| R_i = 0\right] \\
&\qquad \blacktriangleright \text{ by the law of total expectation} \\
&= f_{R_i}(1)\psi_{V_i|R_i=1}(s) + f_{R_i}(0)\psi_{V_i|R_i=0}(s) \\
&\qquad \blacktriangleright \text{ by the definitions of } \psi_{V_i|R_i=1}, \psi_{V_i|R_i=0} \\
&\leq \frac{1}{\pi} \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle) e^{\frac{1}{2}s^2} + \left(1 - \frac{1}{\pi} \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle)\right) \\
&\qquad \blacktriangleright \text{ by Equations (175), (188), and (190)} \\
&= 1 + \frac{1}{\pi} \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle) \left(e^{\frac{1}{2}s^2} - 1\right),
\end{aligned} \tag{192}$$

and likewise, the mgf of $-V_i$ is bounded by

$$\psi_{-V_i}(s) \leq 1 + \frac{1}{\pi} \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle) \left(e^{\frac{1}{2}s^2} - 1\right). \tag{193}$$

Let $V \triangleq \sum_{i=1}^{n} V_i$, where in expectation,

$$\begin{aligned}
\mathbb{E}[V] &= \sum_{i=1}^{n} \mathbb{E}[V_i] \\
&\qquad \blacktriangleright \text{ by by the definition of } V \text{ and the linearity of expectation} \\
&= \sum_{i=1}^{n} f_{R_i}(0)\, \mathbb{E}[V_i \,|\, R_i = 0] + f_{R_i}(1)\, \mathbb{E}[V_i \,|\, R_i = 1] \\
&\qquad \blacktriangleright \text{ by the law of total expectation} \\
&= \sum_{i=1}^{n} f_{R_i}(0) \cdot 0 + f_{R_i}(1) \cdot 0
\end{aligned}$$

$$\begin{aligned}
&\qquad \blacktriangleright \text{ as argued earlier} \\
&= 0.
\end{aligned}$$

Then, the mgfs of the (centered) random variables V and $-V$ are given and upper bounded at $s \in [0, \infty)$ as follows:

$$\begin{aligned}
\psi_V(s) &= (\psi_{V_i}(s))^n \\
&\qquad \blacktriangleright \text{ for any } i \in [n]; \\
&\qquad \blacktriangleright \text{ since } V_1, \dots, V_n \text{ are identically distributed} \\
&\leq \left(1 + \frac{1}{\pi} \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle) \left(e^{\frac{1}{2}s^2} - 1\right)\right)^n \\
&\qquad \blacktriangleright \text{ by Equation (192)} \\
&\leq e^{\frac{1}{\pi} n \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle)(e^{\frac{1}{2}s^2} - 1)}, \qquad (194) \\
&\qquad \blacktriangleright \text{ by a well-known inequality, } \log(1+u) \leq u \text{ for } u > -1
\end{aligned}$$

and likewise,

$$\psi_{-V}(s) \leq e^{\frac{1}{\pi} n \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle)(e^{\frac{1}{2}s^2} - 1)}. \tag{195}$$

Now, observe:

$$\begin{aligned}
&P\left(\frac{V}{n} - \mathbb{E}\left[\frac{V}{n}\right] > \frac{1}{\pi} t \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle) \middle| R_i = 1\right) \\
&\quad \leq \inf_{s \geq 0} e^{-\frac{1}{\pi} nst \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle)} \psi_V(s) \\
&\qquad \blacktriangleright \text{ due to Bernstein (see, e.g., Vershynin (2018))} \\
&\quad \leq \inf_{s \geq 0} e^{-\frac{1}{\pi} nst \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle)} e^{\frac{1}{\pi} n \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle)(e^{\frac{1}{2}s^2} - 1)} \\
&\qquad \blacktriangleright \text{ by Equation (194)} \\
&\quad \leq e^{\frac{1}{3\pi} nt^2 \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle)}, \qquad (196) \\
&\qquad \blacktriangleright \text{ as argued earlier in the proof of Equation (140)}
\end{aligned}$$

and on the other side:

$$P\left(\frac{V}{n} - \mathbb{E}\left[\frac{V}{n}\right] < -\frac{1}{\pi} t \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle) \middle| R_i = 1\right) \leq e^{\frac{1}{3\pi} nt^2 \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle)}. \tag{197}$$

By a union bound over the above pair of inequalities in Equations (196) and (197),

$$P\left(\left|\frac{V}{n} - \mathbb{E}\left[\frac{V}{n}\right]\right| > \frac{1}{\pi} t \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle)\right) \leq 2e^{-\frac{1}{3\pi} nt^2 \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle)},$$

and therefore, recalling the definitions of the random variables V_i, $i \in [n]$, and their relationship to $h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})$, the above concentration inequality further implies that

$$P\left(\left|\left\langle \frac{h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})}{\sqrt{2\pi}}, \frac{\boldsymbol{\theta}^* + \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* + \boldsymbol{\theta}\|_2} \right\rangle - \mathbb{E}\left[\left\langle \frac{h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})}{\sqrt{2\pi}}, \frac{\boldsymbol{\theta}^* + \boldsymbol{\theta}}{\|\boldsymbol{\theta}^* + \boldsymbol{\theta}\|_2} \right\rangle\right]\right| > \frac{1}{\pi} t \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle)\right)$$

$$\leq 2e^{-\frac{1}{3\pi}nt^2 \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle)}.$$

Lastly, by union bounding over all $J \in \mathcal{J}$ and $\boldsymbol{\theta} \in \tilde{\mathcal{C}}$, Equation (141) follows:

$$P\left(\exists J \in \mathcal{J}, \boldsymbol{\theta} \in \tilde{\mathcal{C}} \quad \left| \left\langle \frac{h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})}{\sqrt{2\pi}}, \frac{\boldsymbol{\theta}^*+\boldsymbol{\theta}}{\|\boldsymbol{\theta}^*+\boldsymbol{\theta}\|_2} \right\rangle - \mathbb{E}\left[\left\langle \frac{h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})}{\sqrt{2\pi}}, \frac{\boldsymbol{\theta}^*+\boldsymbol{\theta}}{\|\boldsymbol{\theta}^*+\boldsymbol{\theta}\|_2} \right\rangle \right] \right| > \frac{1}{\pi} t \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle) \right)$$
$$\leq 2|\mathcal{J}||\tilde{\mathcal{C}}| e^{-\frac{1}{3\pi}nt^2 \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle)}.$$

Verification of Equation (142) Towards deriving the third concentration inequality, Equation (142), consider an orthonormal basis, $\{\mathbf{v}_1, \ldots, \mathbf{v}_{k'}\} \subset \mathbb{R}^d$, for the subspace $\mathcal{V} \triangleq \{\mathbf{v} \in \mathbb{R}^d : \mathrm{supp}(\mathbf{v}) \subseteq \mathrm{supp}(\boldsymbol{\theta}^*) \cup \mathrm{supp}(\boldsymbol{\theta}) \cup J\}$, where $k' \triangleq |\mathrm{supp}(\boldsymbol{\theta}^*) \cup \mathrm{supp}(\boldsymbol{\theta}) \cup J|$, and where $\mathbf{v}_{k'-1} \triangleq \frac{\boldsymbol{\theta}^*-\boldsymbol{\theta}}{\|\boldsymbol{\theta}^*-\boldsymbol{\theta}\|_2}$ and $\mathbf{v}_{k'} \triangleq \frac{\boldsymbol{\theta}^*+\boldsymbol{\theta}}{\|\boldsymbol{\theta}^*+\boldsymbol{\theta}\|_2}$. Then, the orthogonal decomposition of $\frac{1}{\sqrt{2\pi}}\bar{h}_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})$ using this basis is given and subsequently rewritten as follows:

$$\begin{aligned}
\bar{h}_J(\boldsymbol{\theta}^*, \boldsymbol{\theta}) &= \sum_{j=1}^{k'} \left\langle \frac{\bar{h}_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})}{\sqrt{2\pi}}, \mathbf{v}_j \right\rangle \mathbf{v}_j \\
&= \sum_{j=1}^{k'} \left\langle \frac{h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})}{\sqrt{2\pi}} - \left\langle \frac{h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})}{\sqrt{2\pi}}, \mathbf{v}_{k'-1} \right\rangle \mathbf{v}_{k'-1} - \left\langle \frac{h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})}{\sqrt{2\pi}}, \mathbf{v}_{k'} \right\rangle \mathbf{v}_{k'}, \mathbf{v}_j \right\rangle \mathbf{v}_j \\
&\qquad \blacktriangleright \text{ by the choice of } \mathbf{v}_{k'-1} = \tfrac{\boldsymbol{\theta}^*-\boldsymbol{\theta}}{\|\boldsymbol{\theta}^*-\boldsymbol{\theta}\|_2}, \mathbf{v}_{k'} = \tfrac{\boldsymbol{\theta}^*+\boldsymbol{\theta}}{\|\boldsymbol{\theta}^*+\boldsymbol{\theta}\|_2} \\
&= \sum_{j=1}^{k'} \left\langle \frac{h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})}{\sqrt{2\pi}}, \mathbf{v}_j \right\rangle \mathbf{v}_j - \left\langle \frac{h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})}{\sqrt{2\pi}}, \mathbf{v}_{k'-1} \right\rangle \mathbf{v}_{k'-1} - \left\langle \frac{h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})}{\sqrt{2\pi}}, \mathbf{v}_{k'} \right\rangle \mathbf{v}_{k'} \\
&\qquad \blacktriangleright \text{ due to the orthogonality of } \mathbf{v}_1, \ldots, \mathbf{v}_{k'} \\
&= \sum_{j=1}^{k'-2} \left\langle \frac{h_J(\boldsymbol{\theta}^*, \boldsymbol{\theta})}{\sqrt{2\pi}}, \mathbf{v}_j \right\rangle \mathbf{v}_j \\
&= \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k'-2} \langle \tilde{\mathbf{x}}_i, \mathbf{v}_j \rangle \mathbf{v}_j \,\mathrm{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle\right) \mathbb{I}\big(\mathrm{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle\right) \neq \mathrm{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta} \rangle\right)\big). \\
&\qquad \blacktriangleright \text{ by Equation (172)}
\end{aligned}$$

Note that $\boldsymbol{\theta}^*, \boldsymbol{\theta} \in \mathrm{span}(\{\mathbf{v}_{k'-1}, \mathbf{v}_{k'}\})$, which implies by the orthogonality of the set $\{\mathbf{v}_1, \ldots, \mathbf{v}_{k'}\}$ that $\boldsymbol{\theta}^*, \boldsymbol{\theta} \perp \mathbf{v}_j$ for every $j \in [k'-2]$. Thus, applying standard facts about Gaussians, for each $i \in [n]$, $j \in [k'-2]$, there is an equivalence in distribution:

$$\langle \tilde{\mathbf{x}}_i, \mathbf{v}_j \rangle \,\mathrm{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle\right) \mathbb{I}\big(\mathrm{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle\right) \neq \mathrm{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta} \rangle\right)\big) \sim W_{i,j} \triangleq Z_{i,j} Y_i,$$

where for each $i \in [n]$, $j \in [k']$, the random variable $Z_{i,j} \sim \mathcal{N}(0,1)$ is standard Gaussian and

$$Y_i \triangleq \mathrm{sign}\left(Z_{i,k'-1}\right) \mathbb{I}\big(\mathrm{sign}\left(Z_{i,k'-1}\right) \neq \mathrm{sign}\left(Z_{i,k'}\right)\big).$$

Notice that the random variables $\{Z_{i,j}\}_{i \in [n], j \in [k'-2]}$ are i.i.d. and also independent of $Z_{i,k'-1}$, $Z_{i,k'}$, and Y_i, $i \in [n]$. Moreover,

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k'-2} \langle \tilde{\mathbf{x}}_i, \mathbf{v}_j \rangle \mathbf{v}_j \,\mathrm{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle\right) \mathbb{I}\big(\mathrm{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle\right) \neq \mathrm{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta} \rangle\right)\big) \sim \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k'-2} W_{i,j} \mathbf{v}_j.$$

Due to the rotational invariance of Gaussians,

$$\begin{aligned}&\langle \tilde{\mathbf{x}}_i, \mathbf{v}_j \rangle \operatorname{sign}(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle) \, \mathbb{I}\big(\operatorname{sign}(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle) \neq \operatorname{sign}(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta} \rangle)\big) \\ &\quad \sim \langle \tilde{\mathbf{x}}_i, \mathbf{e}_j \rangle \operatorname{sign}(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle) \, \mathbb{I}\big(\operatorname{sign}(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle) \neq \operatorname{sign}(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta} \rangle)\big),\end{aligned}$$

where $\mathbf{e}_j \triangleq \mathbf{1}^{\{j\}} \in \mathbb{R}^d$ is the j^{th} standard basis vector for $\mathbb{R}^d$ in which the j^{th} entry is set to 1 and all other entries are set to 0. Hence, without loss of generality, the analysis will proceed under the assumption that the first $(k'-2)$-many j^{th} basis vectors are $\mathbf{v}_j = \mathbf{e}_j$, $j \in [k'-2]$. Under this assumption, the random vector, $\mathbf{U}$, which is given by

$$\mathbf{U} \triangleq \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k'-2} W_{i,j} \mathbf{v}_j = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k'-2} W_{i,j} \mathbf{e}_j,$$

has j^{th} entries, $j \in [d]$,

$$U_j = \begin{cases} 0, & \text{if } j \in [d] \setminus [k'-2], \\ \frac{1}{n} \sum_{i=1}^{n} W_{i,j}, & \text{if } j \in [k'-2]. \end{cases}$$

For $i \in [n]$, define the random variable $R_i \triangleq \mathbb{I}(\operatorname{sign}(Z_{i,k'-1}) \neq \operatorname{sign}(Z_{i,k'}))$, whose a mass function given at $r \in \{0,1\}$ by

$$f_{R_i}(r) = \begin{cases} 1 - \frac{1}{\pi} \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle), & \text{if } r = 0, \\ \frac{1}{\pi} \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle), & \text{if } r = 1. \end{cases}$$

As in the verification of Equation (140), this mass function can be derived by way of an approach similar to that which appears in (Matsumoto and Mazumdar, 2024a, Appendix B.1.1). Additionally, write the random vector $\mathbf{R} \triangleq (R_1, \dots, R_n)$, whose entries are i.i.d., and let $L \triangleq \|\mathbf{R}\|_0$. Because each random variable $Z_{i,j}$, $j \in [k'-2]$, is independent of $Z_{i,k'-1}$ and $Z_{i,k'}$, it is also independent of $\operatorname{sign}(Z_{i,k'-1})$ and R_i, where $\operatorname{sign}(Z_{i,k'-1})$ follows a Rademacher distribution. Since mean-0 Gaussian random variables have the same distribution as their negations, there are the following equivalences in distribution: $-Z_{i,j} \sim Z_{i,j} \sim \mathcal{N}(0,1)$ and $Z_{i,j} \operatorname{sign}(Z_{i,k'-1}) \sim Z_{i,j} \sim \mathcal{N}(0,1)$ (see, e.g., (Matsumoto and Mazumdar, 2024a, Appendix B) for a formal argument). Hence, $(W_{i,j} \,|\, R_i = 1) \sim \mathcal{N}(0,1)$. Since the random variables $W_{1,j}, \dots, W_{n,j}$ are i.i.d., it follows that $(U_j \,|\, L = \ell) \sim (U_j \,|\, \mathbf{R} = \mathbf{r}) \sim \mathcal{N}(0, \frac{\ell}{n^2})$ for each $j \in [k'-2]$ and an arbitrary choice of $\mathbf{r} \in \{0,1\}^n$, and where $\ell \triangleq \|\mathbf{r}\|_0$. (A more rigorous analysis can employ the law of total probability.) Therefore, $\mathbf{U}$ is a $\frac{\sqrt{\ell}}{n}$-subgaussian random vector with support of cardinality $\|\mathbf{U}\|_0 = k'-2$.

Before proceeding, two results are introduced to facilitate the proof.

Lemma 33 (Lemma (Matsumoto and Mazumdar, 2024a, Lemma A.2)) *Fix $s \in (0,1)$. Let $\mathbf{Z}_1, \dots, \mathbf{Z}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, and let $\mathbf{u}, \mathbf{v} \in S^{d-1}$. Define the random variable $L \triangleq |\{i \in [n] : \operatorname{sign}(\langle \mathbf{Z}_i, \mathbf{u} \rangle) \neq \operatorname{sign}(\langle \mathbf{Z}_i, \mathbf{v} \rangle)\}|$. Then,*

$$\mu_L \triangleq \mathbb{E}[L] = \frac{n \arccos(\langle \mathbf{u}, \mathbf{v} \rangle)}{\pi} \tag{198}$$

and

$$P(L > (1+s)\mu_L) \leq e^{-\frac{1}{3\pi} n s^2 \arccos(\langle \mathbf{u}, \mathbf{v} \rangle)}. \tag{199}$$

Lemma 34 *Fix $t'', \sigma > 0$ and $0 < m \le d$. Let $J'' \subseteq [d]$, $|J''| = m$, and $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \sum_{j \in J''} \mathbf{e}_j \mathbf{e}_j^T)$. Then,*

$$P\left(\|\mathbf{X} - \mathbb{E}[\mathbf{X}]\|_2 > \sqrt{m}\sigma + t''\right) \le P\left(\|\mathbf{X} - \mathbb{E}[\mathbf{X}]\|_2 > \mathbb{E}[\|\mathbf{X}\|_2] + t''\right) \le e^{-\frac{1}{2\sigma^2} t''^2}. \tag{200}$$

Proof Lemma 34 Note that $\|\mathbf{X} - \mathbb{E}[\mathbf{X}]\|_2 = \|\mathbf{X}\|_2$ due to the lemma's condition that $\mathbf{X}$ is zero-mean. By standard properties of Gaussians, the expected ℓ_2-norm of $\mathbf{X}$ is bound from above by $\mathbb{E}[\|\mathbf{X}\|_2] \le \sqrt{m}\sigma$. Due to a well-known concentration inequality for Lipschitz functions of sub-gaussian random vectors (see, e.g., Wainwright (2019)), and noting that the ℓ_2-norm is 1-Lipschitz, the claimed inequality holds:

$$\begin{aligned} P\left(\|\mathbf{X} - \mathbb{E}[\mathbf{X}]\|_2 > \sqrt{m}\sigma + t''\right) &= P(\|\mathbf{X}\|_2 > \sqrt{m}\sigma + t'') \\ &\le P(\|\mathbf{X}\|_2 > \mathbb{E}[\|\mathbf{X}\|_2] + t'') \\ &\le e^{-\frac{1}{2\sigma^2} t''^2}, \end{aligned}$$

as desired. ∎

Fixing $s \in (0,1)$, the random variable L exceeds $L > (1+s)\frac{1}{\pi} n \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle)$ with probability at most $e^{-\frac{1}{3\pi} n s^2 \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle)}$ by Lemma 33. Additionally, by an earlier observation, $\mathbb{E}[\mathbf{U} \mid L = \ell] = \mathbf{0}$, and thus, due to Lemma 34, for $t'' > 0$,

$$\begin{aligned} P\left(\|\mathbf{U} - \mathbb{E}[\mathbf{U}]\|_2 > \frac{\sqrt{(k'-2)\ell}}{n} + t'' \,\middle|\, L \le \ell\right) &\le P\left(\|\mathbf{U} - \mathbb{E}[\mathbf{U}]\|_2 > \frac{\sqrt{(k'-2)\ell}}{n} + t'' \,\middle|\, L = \ell\right) \\ &\le e^{-\frac{n^2 t''^2}{2\ell}}, \end{aligned} \tag{201}$$

where in particular, taking $\ell = (1+s)\frac{1}{\pi} n \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle)$ and $t'' = \frac{1}{\pi} t \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle)$ in Equation (201) and noting that

$$\begin{aligned} k' &= |\mathrm{supp}(\boldsymbol{\theta}^*) \cup \mathrm{supp}(\boldsymbol{\theta}) \cup J| \le \min\{|\mathrm{supp}(\boldsymbol{\theta}^*)| + |\mathrm{supp}(\boldsymbol{\theta})| + |J|, d\} \\ &\le \min\{2k + \max_{J' \in \mathcal{J}} |J'|, d\} = k_0 \end{aligned} \tag{202}$$

for any $J \in \mathcal{J}$, the following holds:

$$\begin{aligned} &P\Big(\|\mathbf{U} - \mathbb{E}[\mathbf{U}]\|_2 > \sqrt{\tfrac{1}{\pi n}(1+s)(k_0-2)\arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle)} + \tfrac{1}{\pi} t \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle) \\ &\qquad \Big| L \le (1+s)\tfrac{1}{\pi} n \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle)\Big) \\ &\le P\Big(\|\mathbf{U} - \mathbb{E}[\mathbf{U}]\|_2 > \sqrt{\tfrac{1}{\pi n}(1+s)(k'-2)\arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle)} + \tfrac{1}{\pi} t \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle) \\ &\qquad \Big| L \le (1+s)\tfrac{1}{\pi} n \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle)\Big) \\ &\qquad \blacktriangleright \text{ by Equation (202), } k' \le k_0 \\ &\le e^{-\frac{1}{2\pi(1+s)} n t^2 \arccos(\langle \boldsymbol{\theta}^*, \boldsymbol{\theta} \rangle)}. \\ &\qquad \blacktriangleright \text{ by Equation (201)} \end{aligned} \tag{203}$$

Combining the above arguments obtains:

$$\begin{aligned}
&P\Big(\|\mathbf{U}-\mathbb{E}[\mathbf{U}]\|_2 > \sqrt{\tfrac{1}{\pi n}(1+s)(k_0-2)\arccos(\langle\boldsymbol{\theta}^*,\boldsymbol{\theta}\rangle)+\tfrac{1}{\pi}t\arccos(\langle\boldsymbol{\theta}^*,\boldsymbol{\theta}\rangle)}\Big) \\
&\quad\le P\Big(\|\mathbf{U}-\mathbb{E}[\mathbf{U}]\|_2 > \sqrt{\tfrac{1}{\pi n}(1+s)(k_0-2)\arccos(\langle\boldsymbol{\theta}^*,\boldsymbol{\theta}\rangle)+\tfrac{1}{\pi}t\arccos(\langle\boldsymbol{\theta}^*,\boldsymbol{\theta}\rangle)} \\
&\qquad\qquad\Big| L \le (1+s)\tfrac{1}{\pi}n\arccos(\langle\boldsymbol{\theta}^*,\boldsymbol{\theta}\rangle)\Big) \\
&\qquad + P\left(L > (1+s)\frac{1}{\pi}n\arccos(\langle\boldsymbol{\theta}^*,\boldsymbol{\theta}\rangle)\right) \\
&\quad\le e^{-\frac{1}{2\pi(1+s)}nt^2\arccos(\langle\boldsymbol{\theta}^*,\boldsymbol{\theta}\rangle)}+e^{-\frac{1}{3\pi}ns^2\arccos(\langle\boldsymbol{\theta}^*,\boldsymbol{\theta}\rangle)}.
\end{aligned}$$

▶ by Equation (203) and Lemma 33

Recalling the equivalences in distribution described earlier in the proofs, it directly follows that

$$\begin{aligned}
&P\left(\left\|\frac{\bar{h}_J(\boldsymbol{\theta}^*,\boldsymbol{\theta})}{\sqrt{2\pi}}-\mathbb{E}\left[\frac{\bar{h}_J(\boldsymbol{\theta}^*,\boldsymbol{\theta})}{\sqrt{2\pi}}\right]\right\|_2 > \sqrt{\tfrac{1}{\pi n}(1+s)(k_0-2)\arccos(\langle\boldsymbol{\theta}^*,\boldsymbol{\theta}\rangle)+\tfrac{1}{\pi}t\arccos(\langle\boldsymbol{\theta}^*,\boldsymbol{\theta}\rangle)}\right) \\
&\qquad\le e^{-\frac{1}{2\pi(1+s)}nt^2\arccos(\langle\boldsymbol{\theta}^*,\boldsymbol{\theta}\rangle)}+e^{-\frac{1}{3\pi}ns^2\arccos(\langle\boldsymbol{\theta}^*,\boldsymbol{\theta}\rangle)}
\end{aligned}$$

for any single choice of $J \in \mathcal{J}$ and $\boldsymbol{\theta} \in \tilde{\mathcal{C}}$. Then, union bounds over $\mathcal{J}$ and $\tilde{\mathcal{C}}$ yields the concentration inequality in Equation (142):

$$\begin{aligned}
&P\left(\exists J\in\mathcal{J},\boldsymbol{\theta}\in\tilde{\mathcal{C}}\quad \left\|\frac{\bar{h}_J(\boldsymbol{\theta}^*,\boldsymbol{\theta})}{\sqrt{2\pi}}-\mathbb{E}\left[\frac{\bar{h}_J(\boldsymbol{\theta}^*,\boldsymbol{\theta})}{\sqrt{2\pi}}\right]\right\|_2 > \sqrt{\frac{(1+s)(k_0-2)\arccos(\langle\boldsymbol{\theta}^*,\boldsymbol{\theta}\rangle)}{\pi n}+\frac{t\arccos(\langle\boldsymbol{\theta}^*,\boldsymbol{\theta}\rangle)}{\pi}}\right) \\
&\qquad\le |\mathcal{J}||\tilde{\mathcal{C}}|e^{-\frac{1}{2\pi(1+s)}nt^2\arccos(\langle\boldsymbol{\theta}^*,\boldsymbol{\theta}\rangle)}+|\tilde{\mathcal{C}}|e^{-\frac{1}{3\pi}ns^2\arccos(\langle\boldsymbol{\theta}^*,\boldsymbol{\theta}\rangle)},
\end{aligned}$$

as claimed. ■

D.4. Proof of Lemma 31

Proof Lemma 31 The proof of the lemma is split across a few subsections within this section, Appendix D.4: Appendix D.4.1 is devoted to Equations (148) and (146), while Appendix D.4.2 derives Equations (149) and (147). Lastly, Appendix D.4.3 proves an intermediate result.

D.4.1. PROOF OF EQUATIONS (146) AND (148)

Fix $\boldsymbol{\theta}^* \in \Theta$ and $J'' \in \mathcal{J}''$ arbitrarily. Write $\tilde{\mathbf{x}}_i \triangleq T_{\mathrm{supp}(\boldsymbol{\theta}^*)\cup J''}(\mathbf{x}_i)$, $i \in [n]$. As similarly seen earlier, $\frac{1}{\sqrt{2\pi}}h_{f;J''}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)$ can be written as follows:

$$\frac{1}{\sqrt{2\pi}}h_{f;J''}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*) = T_{\mathrm{supp}(\boldsymbol{\theta}^*)\cup J''}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\,\frac{1}{2}\left(f(\langle\mathbf{x},\boldsymbol{\theta}^*\rangle)-\mathrm{sign}\left(\langle\mathbf{x},\boldsymbol{\theta}^*\rangle\right)\right)\right) \tag{204}$$

▶ by the definition of $h_{f;J}$ in Equation (21) (205)

$$= \frac{1}{n}\sum_{i=1}^{n}T_{\mathrm{supp}(\boldsymbol{\theta}^*)\cup J''}\left(\mathbf{x}_i\right)\frac{1}{2}\left(f(\langle\mathbf{x},\boldsymbol{\theta}^*\rangle)-\mathrm{sign}\left(\langle\mathbf{x},\boldsymbol{\theta}^*\rangle\right)\right) \tag{206}$$

$$\blacktriangleright \text{ by the linearity of the subset thresholding operation (see, Section 3)} \tag{207}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \tilde{\mathbf{x}}_i \, \frac{1}{2} \left(f(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle) - \operatorname{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle \right) \right) \tag{208}$$

$$\blacktriangleright \text{ by the definition of } \tilde{\mathbf{x}}_i,\ i \in [n] \tag{209}$$

$$= -\frac{1}{n} \sum_{i=1}^{n} \tilde{\mathbf{x}}_i \operatorname{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle \right) \mathbb{I}\big(f(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle) \neq \operatorname{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle \right) \big), \tag{210}$$

and thus,

$$\begin{aligned} \left\langle \frac{1}{\sqrt{2\pi}} h_{f;J''}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*), \boldsymbol{\theta}^* \right\rangle &= -\frac{1}{n} \sum_{i=1}^{n} \langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle \operatorname{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle \right) \mathbb{I}\big(f(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle) \neq \operatorname{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle \right) \big) \\ &= -\frac{1}{n} \sum_{i=1}^{n} \left| \langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle \right| \mathbb{I}\big(f(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle) \neq \operatorname{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle \right) \big). \end{aligned} \tag{211}$$

Note that justifications for some of the steps taken above can be obtained by extending those appearing in the proof of Lemma 30.

The first step towards deriving Equations (146) and (148) is characterizing the distribution of each i^{th} summand, $i \in [n]$, in Equation (211). Let $Z_i \sim \mathcal{N}(0,1)$ and $R_i \triangleq \mathbb{I}(f(Z_i) \neq \operatorname{sign}(Z_i))$, $i \in [n]$. Then, each i^{th} summand, $i \in [n]$, follows the same distribution as

$$\left| \langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle \right| \mathbb{I}\big(f(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle) \neq \operatorname{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle \right) \big) \sim W_i \triangleq |Z_i| \, \mathbb{I}\big(f(Z_i) \neq \operatorname{sign}(Z_i) \big) = |Z_i| R_i.$$

The density and mass functions of $|Z_i|$ and R_i, respectively, are given by

$$f_{|Z_i|}(z) = \begin{cases} \sqrt{\frac{2}{\pi}} e^{-\frac{1}{2}z^2}, & \text{if } z \geq 0, \\ 0, & \text{if } z = 0, \end{cases} \tag{212}$$

$$f_{R_i}(r) = \begin{cases} 1 - \alpha, & \text{if } r = 0, \\ \alpha, & \text{if } r = 1. \end{cases} \tag{213}$$

Additionally, the mass function of the conditioned random variable $R_i = 1 \mid Z_i$ is given by

$$f_{R_i \mid Z_i}(1 \mid z) = \begin{cases} p(z), & \text{if } z < 0, \\ 1 - p(z), & \text{if } z \geq 0, \end{cases} \tag{214}$$

and the mass function of the conditioned random variable $R_i = 1 \mid |Z_i|$ is given by

$$f_{R_i \mid |Z_i|}(1 \mid z) = \begin{cases} 0, & \text{if } z < 0, \\ \frac{1}{2}(1 - p(z) + p(-z)), & \text{if } z \geq 0, \end{cases} \tag{215}$$

where the latter case—when $z \geq 0$—is obtained as follows:

$$f_{R_i||Z_i|}(1 \mid z) = f_{R_i||Z_i|,Z_i}(1 \mid z, z) f_{Z_i||Z_i|}(z|z) + f_{R_i||Z_i|,Z_i}(1 \mid z, -z) f_{Z_i||Z_i|}(z| - z)$$

$$
\begin{aligned}
&\quad \blacktriangleright \text{ by the law of total probability and the definition of conditional probabilities} \\
&\quad \text{(and the observation that } f_{Z_i||Z_i|}(z'|z) = 0 \text{ whenever } |z'| \neq z,\, z' \in \mathbb{R}, z \geq 0) \\
&= \frac{1}{2} f_{R_i||Z_i|,Z_i}(1 \,|\, z, z) + \frac{1}{2} f_{R_i||Z_i|,Z_i}(1 \,|\, z, -z) \\
&\quad \blacktriangleright \text{ by symmetry} \\
&= \frac{1}{2} f_{R_i|Z_i}(1 \,|\, z) + \frac{1}{2} f_{R_i|Z_i}(1 \,|\, -z) \\
&\quad \blacktriangleright \text{ because } Z_i \text{ completely determines } |Z_i|, \text{ which implies } (R_i \,|\, |Z_i|, Z_i) \sim (R_i \,|\, Z_i) \\
&= \frac{1}{2}(1 - p(z)) + \frac{1}{2} p(-z) \\
&\quad \blacktriangleright \text{ by Equation (214)} \\
&= \frac{1}{2}(1 - p(z) + p(-z)).
\end{aligned}
$$

Note that $(W_i \,|\, R_i = 1) \sim (|Z_i| R_i \,|\, R_i = 1) \sim (|Z_i| \,|\, R_i = 1)$. Thus, via Bayes' theorem, for $z \in \mathbb{R}$,

$$
\begin{aligned}
f_{W_i|R_i}(z \,|\, 1) &= f_{|Z_i||R_i}(z \,|\, 1) \\
&\quad \blacktriangleright \text{ by the above remark} \\
&= \frac{f_{|Z_i|}(z) f_{R_i||Z_i|}(1 \,|\, z)}{f_{R_i}(1)} \\
&\quad \blacktriangleright \text{ by Bayes' theorem} \\
&= \frac{\sqrt{\frac{2}{\pi}} e^{-\frac{1}{2}z^2} \frac{1}{2}(1 - p(z) + p(-z))}{\alpha} \\
&\quad \blacktriangleright \text{ by Equations (212), (213), and (215)} \\
&= \frac{1}{\sqrt{2\pi}\alpha} e^{-\frac{1}{2}z^2} (1 - p(z) + p(-z)),
\end{aligned}
$$

and therefore, taking together the above work, the density of the conditioned random variable $W_i \,|\, R_i$ is given for $r \in \{0, 1\}$ and $z \in \mathbb{R}$ by

$$
f_{W_i|R_i}(z \,|\, r) = \begin{cases} 0, & \text{if } r = 0, z \neq 0, \\ 1, & \text{if } r = 0, z = 0, \\ 0, & \text{if } r = 1, z < 0, \\ \frac{1}{\sqrt{2\pi}\alpha} e^{-\frac{1}{2}z^2} (1 - p(z) + p(-z)), & \text{if } r = 1, z \geq 0. \end{cases} \tag{216}
$$

In expectation, when conditioning on $R_i = 0$,

$$
\begin{aligned}
\mathbb{E}[W_i \,|\, R_i = 0] &= \int_{z=-\infty}^{z=\infty} z f_{W_i|R_i}(z \,|\, 0) dz \\
&= 0 f_{W_i|R_i}(0 \,|\, 0) \\
&\quad \blacktriangleright \text{ due to Equation (216)} \\
&= 0,
\end{aligned} \tag{217}
$$

and when conditioning on $R_i = 1$,

$$\begin{aligned}
\mathbb{E}[W_i \,|\, R_i = 1] &= \int_{z=-\infty}^{z=\infty} z f_{W_i|R_i}(z \,|\, 1) dz \\
&= \int_{z=0}^{z=\infty} \frac{1}{\sqrt{2\pi}\alpha} z e^{-\frac{1}{2}z^2} (1 - p(z) + p(-z)) dz \\
&\qquad \blacktriangleright \text{ by Equation (216)} \\
&= \int_{z=0}^{z=\infty} \frac{1}{\sqrt{2\pi}\alpha} z e^{-\frac{1}{2}z^2} dz - \int_{z=0}^{z=\infty} \frac{1}{\sqrt{2\pi}\alpha} z e^{-\frac{1}{2}z^2} (p(z) - p(-z)) dz \\
&= \frac{1}{\sqrt{2\pi}\alpha} - \frac{\gamma}{2\alpha} \\
&\qquad \blacktriangleright \text{ by the definition of } \gamma \\
&= \frac{\sqrt{2/\pi} - \gamma}{2\alpha}.
\end{aligned} \tag{218}$$

With this preliminary work completed, we now proceed to the derivations of Equations (146) and (148), starting with the former.

Verification of Equation (146) Having obtained the density function and expectations for the conditioned random variable $W_i \,|\, R_i$, the expectation of the random variable $|\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle| \, \mathbb{I}(f(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle) \neq \text{sign}(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle))$ is now calculated as follows:

$$\begin{aligned}
\mathbb{E}[|\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle| \, \mathbb{I}\big(f(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle) \neq \text{sign}\,(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle)\big)] &= \mathbb{E}[W_i] \\
&= f_{R_i}(0)\, \mathbb{E}[W_i \,|\, R_i = 0] + f_{R_i}(1)\, \mathbb{E}[W_i \,|\, R_i = 1] \\
&\qquad \blacktriangleright \text{ by the law of total expectation} \\
&= (1-\alpha)0 + \alpha \frac{\sqrt{2/\pi} - \gamma}{2\alpha} \\
&\qquad \blacktriangleright \text{ by Equations (213), (217), and (218)} \\
&= \frac{\sqrt{2/\pi} - \gamma}{2}.
\end{aligned}$$

By the linearity of expectation, it follows that

$$\begin{aligned}
\mathbb{E}\left[\langle h_{f;J''}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*), \boldsymbol{\theta}^* \rangle\right] &= \mathbb{E}\left[-\frac{\sqrt{2\pi}}{n} \sum_{i=1}^{n} |\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle| \, \mathbb{I}\big(f(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle) \neq \text{sign}\,(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle)\big)\right] \\
&= -\frac{\sqrt{2\pi}}{n} \sum_{i=1}^{n} \mathbb{E}\left[|\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle| \, \mathbb{I}\big(f(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle) \neq \text{sign}\,(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle)\big)\right] \\
&= -\frac{\sqrt{2\pi}}{n} \sum_{i=1}^{n} \frac{\sqrt{2/\pi} - \gamma}{2} \\
&= -\left(1 - \sqrt{\frac{\pi}{2}}\gamma\right),
\end{aligned}$$

as claimed. This completes the derivation of Equation (146).

Verification of Equation (148) Next, Equation (148) is derived. This derivation is based on the mgfs of the centered conditioned random variables $(W_i \,|\, R_i) - \mathbb{E}[W_i \,|\, R_i]$ and $(-W_i \,|\, R_i) - \mathbb{E}[-W_i \,|\, R_i]$, which are denoted by $\psi_{(W_i|R_i)-\mathbb{E}[W_i|R_i]}$ and $\psi_{(-W_i|R_i)-\mathbb{E}[-W_i|R_i]}$, respectively. Write $\mu_0 \triangleq \mathbb{E}[W_i \,|\, R_i = 0] = 0$ and $\mu_1 \triangleq \mathbb{E}[W_i \,|\, R_i = 1] = \frac{\sqrt{2/\pi}-\gamma}{2\alpha}$, where these expectations were calculated previously in Equations (217) and (218). Conditioned on $R_i = 1$, the mgfs are given at $s \in [0, \infty)$ by

$$\begin{aligned}\psi_{(W_i|R_i=1)-\mathbb{E}[W_i|R_i=1]}(s) &= \mathbb{E}\left[e^{s(W_i-\mathbb{E}[W_i])}\middle| R_i = 1\right]\\ &= \int_{z=0}^{z=\infty} \frac{1}{\sqrt{2\pi}\alpha} e^{s(z-\mu_1)} e^{-\frac{1}{2}z^2}(1-p(z)+p(-z))dz\\ &= \frac{1}{\alpha}e^{\frac{1}{2}s^2}\frac{1}{\sqrt{2\pi}}e^{-s\mu_1}\int_{z=0}^{z=\infty} e^{-\frac{1}{2}(z-s)^2}(1-p(z)+p(-z))dz\\ &= \frac{1}{\alpha}e^{\frac{1}{2}s^2} f_1(s),\end{aligned}$$

and

$$\begin{aligned}\psi_{(-W_i|R_i=1)-\mathbb{E}[-W_i|R_i=1]}(s) &= \mathbb{E}\left[e^{s(-W_i-\mathbb{E}[-W_i])}\middle| R_i = 1\right]\\ &= \int_{z=0}^{z=\infty} \frac{1}{\sqrt{2\pi}\alpha} e^{s(z+\mu_1)} e^{-\frac{1}{2}z^2}(1-p(z)+p(-z))dz\\ &= \frac{1}{\alpha}e^{\frac{1}{2}s^2}\frac{1}{\sqrt{2\pi}}e^{s\mu_1}\int_{z=0}^{z=\infty} e^{-\frac{1}{2}(z+s)^2}(1-p(z)+p(-z))dz\\ &= \frac{1}{\alpha}e^{\frac{1}{2}s^2} f_2(s),\end{aligned}$$

where

$$f_1(s) \triangleq \frac{1}{\sqrt{2\pi}}e^{-s\mu_1}\int_{z=0}^{z=\infty} e^{-\frac{1}{2}(z-s)^2}(1-p(z)+p(-z))dz, \tag{219}$$

$$f_2(s) \triangleq \frac{1}{\sqrt{2\pi}}e^{s\mu_1}\int_{z=0}^{z=\infty} e^{-\frac{1}{2}(z+s)^2}(1-p(z)+p(-z))dz. \tag{220}$$

Before proceeding, the following lemma is introduced to facilitate upper bounds on the mgfs, $\psi_{(W_i|R_i=1)-\mathbb{E}[W_i|R_i=1]}$ and $\psi_{(-W_i|R_i=1)-\mathbb{E}[-W_i|R_i=1]}$. Its proof is deferred to Appendix D.4.3.

Lemma 35 *Let $f_1, f_2 : \mathbb{R} \to \mathbb{R}$ be the functions defined in Equations* (219) *and* (220)*. Then,*

$$\sup_{s\geq 0} f_1(s) = f_1(0), \tag{221}$$

$$\sup_{s\geq 0} f_2(s) = f_2(0), \tag{222}$$

where

$$f_1(0) = \alpha, \tag{223}$$

$$f_2(0) = \alpha. \tag{224}$$

Due to Equations (221)–(224) in Lemma 35, the mgfs of $(W_i \mid R_i = 1) - \mathbb{E}[W_i \mid R_i = 1]$ and $(-W_i \mid R_i = 1) - \mathbb{E}[-W_i \mid R_i = 1]$ are now upper bounded by

$$\psi_{(W_i|R_i=1)-\mathbb{E}[W_i|R_i=1]}(s) = \frac{1}{\alpha} e^{\frac{1}{2}s^2} f_1(s) \leq e^{\frac{1}{2}s^2}, \tag{225}$$

$$\psi_{(-W_i|R_i=1)-\mathbb{E}[-W_i|R_i=1]}(s) = \frac{1}{\alpha} e^{\frac{1}{2}s^2} f_2(s) \leq e^{\frac{1}{2}s^2}. \tag{226}$$

On the other hand, when conditioned on $R_i = 0$, the mgfs of these random variables $(W_i \mid R_i = 0) - \mathbb{E}[W_i \mid R_i = 0]$ and $(-W_i \mid R_i = 0) - \mathbb{E}[-W_i \mid R_i = 0]$, written as $\psi_{(W_i|R_i=0)-\mathbb{E}[W_i|R_i=0]}$ and $\psi_{(-W_i|R_i=0)-\mathbb{E}[-W_i|R_i=0]}$, are obtained as follows for $s \in [0, \infty)$:

$$\begin{aligned}
\psi_{(W_i|R_i=0)-\mathbb{E}[W_i|R_i=0]}(s) &= \mathbb{E}\left[e^{s(W_i - \mu_0)} \middle| R_i = 0 \right] \\
&= \mathbb{E}\left[e^{sW_i} \middle| R_i = 0 \right] \\
&= e^{s \cdot 0} f_{W_i|R_i}(0 \mid 0) \\
&= 1,
\end{aligned} \tag{227}$$

and likewise,

$$\psi_{(-W_i|R_i=0)-\mathbb{E}[-W_i|R_i=0]}(s) = 1. \tag{228}$$

Now, consider the sum of the random variables W_i, $i \in [n]$. Write $W \triangleq \sum_{i=1}^{n} W_i$, and let $\mathbf{R} \triangleq (R_1, \ldots, R_n)$. Fixing $\mathbf{r} \in \{0,1\}^n$, the mgf of $(W \mid \mathbf{R} = \mathbf{r}) - \mathbb{E}[W \mid \mathbf{R} = \mathbf{r}]$, written $\psi_{(W|\mathbf{R}=\mathbf{r})-\mathbb{E}[W|\mathbf{R}=\mathbf{r}]}(s)$, is given and upper bounded at $s \in [0, \infty)$ by

$$\begin{aligned}
\psi_{(W|\mathbf{R}=\mathbf{r})-\mathbb{E}[W|\mathbf{R}=\mathbf{r}]}(s) &= \mathbb{E}\left[e^{s \sum_{i=1}^{n} W_i - \mathbb{E}[W_i]} \middle| \mathbf{R} = \mathbf{r} \right] \\
&\qquad \blacktriangleright \text{ by the definition of mgfs and by the definition of } W \\
&= \prod_{i=1}^{n} \mathbb{E}\left[e^{s(W_i - \mathbb{E}[W_i])} \middle| R_i = r_i \right] \\
&\qquad \blacktriangleright \text{ each } W_i,\ i \in [n], \text{ is independent of } \{R_{i'}\}_{i' \neq i}; \text{ and} \\
&\qquad \blacktriangleright\ \{(W_1 \mid R_1), \ldots, (W_n \mid R_n)\} \text{ are mutually independent} \\
&= \prod_{\substack{i=1: \\ r_i=1}}^{n} \psi_{(W_i|R_i=1)-\mathbb{E}[W_i|R_i=1]}(s) \prod_{\substack{i=1: \\ r_i=0}}^{n} \psi_{(W_i|R_i=0)-\mathbb{E}[W_i|R_i=0]}(s) \\
&\qquad \blacktriangleright \text{ by partitioning the values of the index of multiplication} \\
&\qquad\quad \text{and by the definition of } \psi_{(W_i|R_i)-\mathbb{E}[W_i|R_i]},\ i \in [n] \\
&\leq \prod_{\substack{i=1: \\ r_i=1}}^{n} e^{\frac{1}{2}s^2} \prod_{\substack{i=1: \\ r_i=0}}^{n} 1 \\
&\qquad \blacktriangleright \text{ by Equations (225) and (227)} \\
&= e^{\frac{1}{2}\|\mathbf{r}\|_0 s^2}.
\end{aligned} \tag{229}$$

It follows that

$$P\left(\frac{W}{n} - \mathbb{E}\left[\frac{W}{n}\right] > \alpha t \middle| \mathbf{R} = \mathbf{r}\right) \leq \inf_{s \geq 0} e^{-\alpha n s t} \psi_{(W|\mathbf{R}=\mathbf{r}) - \mathbb{E}[W|\mathbf{R}=\mathbf{r}]}(s)$$

▶ due to Bernstein (see, e.g., Vershynin (2018))

$$\leq \inf_{s \geq 0} e^{-\alpha n s t} e^{\frac{1}{2}\|\mathbf{r}\|_0 s^2}. \tag{230}$$

▶ by Equation (229)

Additionally, note that

$$f_{\mathbf{R}}(\mathbf{r}) = \alpha^{\|\mathbf{r}\|_0}(1-\alpha)^{n-\|\mathbf{r}\|_0}. \tag{231}$$

Then,

$$P\left(\frac{W}{n} - \mathbb{E}\left[\frac{W}{n}\right] > \alpha t\right) = \sum_{\mathbf{r} \in \{0,1\}^n} f_{\mathbf{R}}(\mathbf{r})\, P\left(\frac{W}{n} - \mathbb{E}\left[\frac{W}{n}\right] > \alpha t \middle| \mathbf{R} = \mathbf{r}\right)$$

▶ by the law of total expectation

$$\leq \sum_{\mathbf{r} \in \{0,1\}^n} \alpha^{\|\mathbf{r}\|_0}(1-\alpha)^{n-\|\mathbf{r}\|_0} \inf_{s \geq 0} e^{-\alpha n s t} e^{\frac{1}{2}\|\mathbf{r}\|_0 s^2}$$

▶ by Equations (230) and (231)

$$\leq \inf_{s \geq 0} e^{-\alpha n s t} \sum_{\mathbf{r} \in \{0,1\}^n} (\alpha e^{\frac{1}{2}s^2})^{\|\mathbf{r}\|_0}(1-\alpha)^{n-\|\mathbf{r}\|_0}$$

$$= \inf_{s \geq 0} e^{-\alpha n s t} \sum_{\ell=0}^{n} \binom{n}{\ell} (\alpha e^{\frac{1}{2}s^2})^{\ell}(1-\alpha)^{n-\ell}$$

▶ by partitioning the values of the index of summation according to $\ell = \|\mathbf{r}\|_0$

$$= \inf_{s \geq 0} e^{-\alpha n s t} \left(1 + \alpha(e^{\frac{1}{2}s^2} - 1)\right)^n$$

▶ by the binomial theorem

$$\leq \inf_{s \geq 0} e^{-\alpha n (st - e^{\frac{1}{2}s^2} + 1)}.$$

▶ by a well-known inequality, $\log(1+u) \leq u$ for $u > -1$

As discussed earlier in the proof of Lemma 30, $st - e^{\frac{1}{2}s^2} + 1$ is maximized with respect to s for $s, t \in (0,1)$ when roughly $s \approx t$ because

$$\frac{\partial}{\partial s} st - e^{\frac{1}{2}s^2} + 1\bigg|_{s=t} = t - se^{\frac{1}{2}s^2}\bigg|_{s=t} \approx 0$$

for small $s, t > 0$, and because for all $s \in \mathbb{R}$,

$$\frac{\partial^2}{\partial s^2} st - e^{\frac{1}{2}s^2} + 1 = -(1+s^2)e^{\frac{1}{2}s^2} < 0.$$

Hence, we will take $s = t$. In addition, recall that

$$st - e^{\frac{1}{2}s^2} + 1\Big|_{s=t} \geq \frac{t^2}{3}.$$

It follows that

$$P\left(W - \mathbb{E}[W] > \alpha t\right) \leq e^{-\frac{1}{3}\alpha n t^2},$$

and therefore, due to the design of W,

$$P\left(\left\langle \frac{h_{f;J''}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)}{\sqrt{2\pi}}, \boldsymbol{\theta}^* \right\rangle - \mathbb{E}\left[\left\langle \frac{h_{f;J''}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)}{\sqrt{2\pi}}, \boldsymbol{\theta}^* \right\rangle\right] > \alpha t\right) \leq e^{-\frac{1}{3}\alpha n t^2}.$$

Moreover, by a nearly identical argument (omitted here), the other side of the bound is obtained:

$$P\left(W - \mathbb{E}[W] < -\alpha t\right) = P\left(-W - \mathbb{E}[-W] > \alpha t\right) \leq e^{-\frac{1}{3}\alpha n t^2},$$

and thus,

$$P\left(\left\langle \frac{h_{f;J''}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)}{\sqrt{2\pi}}, \boldsymbol{\theta}^* \right\rangle - \mathbb{E}\left[\left\langle \frac{h_{f;J''}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)}{\sqrt{2\pi}}, \boldsymbol{\theta}^* \right\rangle\right] < -\alpha t\right) \leq e^{-\frac{1}{3}\alpha n t^2}.$$

Combining the above inequalities into a two-sided bound via a union bound yields:

$$P\left(\left|\left\langle \frac{h_{f;J''}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)}{\sqrt{2\pi}}, \boldsymbol{\theta}^* \right\rangle - \mathbb{E}\left[\left\langle \frac{h_{f;J''}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)}{\sqrt{2\pi}}, \boldsymbol{\theta}^* \right\rangle\right]\right| > \alpha t\right) \leq 2e^{-\frac{1}{3}\alpha n t^2}.$$

Lastly, union bounding over all $J'' \in \mathcal{J}''$, the desired uniform concentration inequality follows:

$$P\left(\exists J'' \in \mathcal{J}'' \quad \left|\left\langle \frac{h_{f;J''}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)}{\sqrt{2\pi}}, \boldsymbol{\theta}^* \right\rangle - \mathbb{E}\left[\left\langle \frac{h_{f;J''}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)}{\sqrt{2\pi}}, \boldsymbol{\theta}^* \right\rangle\right]\right| > \alpha t\right) \leq 2|\mathcal{J}''|e^{-\frac{1}{3}\alpha n t^2}.$$

D.4.2. PROOF OF THE EQUATIONS (147) AND (149)

We begin with some preliminary analysis to characterize a few random variables of interest. As in the derivations of Equations (146) and (148) in Appendix D.4.1, consider an arbitrary coordinate subset $J'' \in \mathcal{J}''$, recall Equation (210) and (211):

$$\frac{1}{\sqrt{2\pi}} h_{f;J''}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) = -\frac{1}{n}\sum_{i=1}^{n} \tilde{\mathbf{x}}_i \operatorname{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle\right) \mathbb{I}\big(f(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle) \neq \operatorname{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle\right)\big),$$

$$\left\langle \frac{1}{\sqrt{2\pi}} h_{f;J''}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*), \boldsymbol{\theta}^* \right\rangle = -\frac{1}{n}\sum_{i=1}^{n} \langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle \operatorname{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle\right) \mathbb{I}\big(f(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle) \neq \operatorname{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle\right)\big),$$

where $\tilde{\mathbf{x}}_i \triangleq T_{\operatorname{supp}(\boldsymbol{\theta}^*) \cup J''}(\mathbf{x}_i)$. Thus,

$$\begin{aligned}
&\frac{1}{\sqrt{2\pi}} h_{f;J''}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) - \left\langle \frac{1}{\sqrt{2\pi}} h_{f;J''}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*), \boldsymbol{\theta}^* \right\rangle \boldsymbol{\theta}^* \\
&\quad = -\frac{1}{n}\sum_{i=1}^{n} \left(\tilde{\mathbf{x}}_i - \langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle \boldsymbol{\theta}^*\right) \operatorname{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle\right) \mathbb{I}\big(f(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle) \neq \operatorname{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle\right)\big).
\end{aligned}$$

Let $k' \triangleq |\text{supp}(\boldsymbol{\theta}^*) \cup J''|$, and denote the k'-dimensional subspace of vectors whose support is a (possibly improper) subset of $\text{supp}(\boldsymbol{\theta}^*) \cup J''$ by $\mathcal{V} \triangleq \{\mathbf{v} \in \mathbb{R}^d : \text{supp}(\mathbf{v}) \subseteq \text{supp}(\boldsymbol{\theta}^*) \cup J''\}$. Let $\{\mathbf{v}_1, \dots, \mathbf{v}_{k'}\} \subset \mathcal{V}$ be an orthonormal basis of $\mathcal{V}$, where $\mathbf{v}_{k'} = \boldsymbol{\theta}^*$. Then, for each $i \in [n]$, since the vector $\tilde{\mathbf{x}}_i$ is contained in $\mathcal{V}$, it is orthogonally decomposed with this bases as:

$$\tilde{\mathbf{x}}_i = \sum_{j=1}^{k'} \langle \tilde{\mathbf{x}}_i, \mathbf{v}_j \rangle \mathbf{v}_j,$$

while $\tilde{\mathbf{x}}_i - \langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle \boldsymbol{\theta}^*$—which is likewise an element in the vector subspace $\mathcal{V}$—is orthogonally decomposed as:

$$\begin{aligned}
\tilde{\mathbf{x}}_i - \langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle \boldsymbol{\theta}^* &= \sum_{j=1}^{k'} \big(\langle \tilde{\mathbf{x}}_i, \mathbf{v}_j \rangle - \langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle \langle \boldsymbol{\theta}^*, \mathbf{v}_j \rangle \big) \mathbf{v}_j \\
&= \big(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle - \langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle \langle \boldsymbol{\theta}^*, \boldsymbol{\theta}^* \rangle \big) \boldsymbol{\theta}^* + \sum_{j=1}^{k'-1} \big(\langle \tilde{\mathbf{x}}_i, \mathbf{v}_j \rangle - \langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle \langle \boldsymbol{\theta}^*, \mathbf{v}_j \rangle \big) \mathbf{v}_j \\
&\qquad \blacktriangleright \text{ by separating out the } k'^{\text{th}} \text{ term from the summation,} \\
&\qquad\quad \text{and since } \mathbf{v}_{k'} = \boldsymbol{\theta}^* \text{ by design} \\
&= \sum_{j=1}^{k'-1} \big(\langle \tilde{\mathbf{x}}_i, \mathbf{v}_j \rangle - \langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle \langle \boldsymbol{\theta}^*, \mathbf{v}_j \rangle \big) \mathbf{v}_j \\
&\qquad \blacktriangleright \text{ recalling that } \langle \boldsymbol{\theta}^*, \boldsymbol{\theta}^* \rangle = \|\boldsymbol{\theta}^*\|_2^2 = 1 \\
&= \sum_{j=1}^{k'-1} \langle \tilde{\mathbf{x}}_i, \mathbf{v}_j \rangle \mathbf{v}_j. \\
&\qquad \blacktriangleright \text{ for } j \neq k',\ \langle \boldsymbol{\theta}^*, \mathbf{v}_j \rangle = \langle \mathbf{v}_{k'}, \mathbf{v}_j \rangle = 0 \text{ since } \mathbf{v}_j \perp \mathbf{v}_{k'} \text{ by design}
\end{aligned}$$

Using this orthogonal decomposition,

$$\begin{aligned}
&\frac{1}{\sqrt{2\pi}} h_{f;J''}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*) - \left\langle \frac{1}{\sqrt{2\pi}} h_{f;J''}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*), \boldsymbol{\theta}^* \right\rangle \boldsymbol{\theta}^* \\
&\quad = -\frac{1}{n} \sum_{i=1}^{n} \left(\tilde{\mathbf{x}}_i - \langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle \boldsymbol{\theta}^* \right) \ \text{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle \right) \mathbb{I}\big(f(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle) \neq \text{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle \right) \big) \\
&\quad = -\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k'-1} \langle \tilde{\mathbf{x}}_i, \mathbf{v}_j \rangle \mathbf{v}_j \ \text{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle \right) \mathbb{I}\big(f(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle) \neq \text{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle \right) \big).
\end{aligned} \tag{232}$$

Note that by a well-known property of Gaussian vectors, $\langle \tilde{\mathbf{x}}_i, \mathbf{v}_j \rangle \sim \mathcal{N}(0, 1)$ for each $i \in [n]$, $j \in [k']$, and moreover, due to the orthogonality of $\mathbf{v}_1, \dots, \mathbf{v}_{k'}$, the random variables $\langle \tilde{\mathbf{x}}_i, \mathbf{v}_1 \rangle, \dots, \langle \tilde{\mathbf{x}}_i, \mathbf{v}_{k'} \rangle$ are mutually independent. In particular, the random variable $\langle \tilde{\mathbf{x}}_i, \mathbf{v}_{k'} \rangle = \langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle$ is independent of every $\langle \tilde{\mathbf{x}}_i, \mathbf{v}_j \rangle$, $j \in [k'-1]$. Therefore, for every $i \in [n]$, each j^{th} summand, $j \in [k'-1]$, in (232) follows the same distribution as

$$\langle \tilde{\mathbf{x}}_i, \mathbf{v}_j \rangle \mathbf{v}_j \ \text{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle \right) \mathbb{I}\big(f(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle) \neq \text{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle \right) \big) \sim Y_i Z_{i,j} \mathbf{v}_j = U_{i,j} \mathbf{v}_j, \tag{233}$$

where $Z_{i,1}, \ldots, Z_{i,k'} \sim \mathcal{N}(0,1)$ are i.i.d. Gaussian random variables, and where

$$R_i \triangleq \mathbb{I}\big(f(Z_{i,k'}) \neq \text{sign}\left(Z_{i,k'}\right)\big),$$
$$Y_i \triangleq \text{sign}\left(Z_{i,k'}\right) R_i,$$
$$U_{i,j} \triangleq Y_i Z_{i,j}.$$

Conditioned on $R_i = 1$, the random variable is distributed as $(Y_i \,|\, R_i \neq 0) \sim \{-1, 1\}$, uniformly, and is independently of all $Z_{i,j}$, $j \in [k'-1]$. Hence, conditioned on R_i, the random variable $U_{i,j} \,|\, R_i$ has a density function given for $z \in \mathbb{R}$ and $r \in \{0,1\}$ by

$$\begin{aligned} f_{U_{i,j}|R_i}(z \,|\, r) &= \begin{cases} 0, & \text{if } z \neq 0, r = 0, \\ 1, & \text{if } z = 0, r = 0, \\ f_{Z_{i,j}}(z) f_{Y_i|R_i}(1 \,|\, 1) + f_{-Z_{i,j}}(z) f_{Y_i|R_i}(-1 \,|\, 1), & \text{if } r = 1, \end{cases} \\ &= \begin{cases} 0, & \text{if } z \neq 0, r = 0, \\ 1, & \text{if } z = 0, r = 0, \\ \frac{1}{2} f_{Z_{i,j}}(z) + \frac{1}{2} f_{-Z_{i,j}}(z), & \text{if } r = 1, \end{cases} \\ &= \begin{cases} 0, & \text{if } z \neq 0, r = 0, \\ 1, & \text{if } z = 0, r = 0, \\ \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, & \text{if } r = 1, \end{cases} \end{aligned} \tag{234}$$

where the third case on the right-hand-side of the first equality is due to the law of total probability, the definition of conditional probabilities, the independence of $(Y_i \,|\, R_i = 1)$ and $Z_{i,j}$, $j \in [k'-1]$, and remarks made in the proof of Lemma 30. Equation (234) implies $(U_{i,j} \,|\, R_i = 1) \sim Z_{i,j}$. Additionally, due to an earlier discussion in Appendix D.4.1, the mass function of the random variable R_i is given for $r \in \{0,1\}$ by

$$f_{R_i}(r) = \begin{cases} 1 - \alpha, & \text{if } r = 0, \\ \alpha, & \text{if } r = 1. \end{cases}$$

This concludes the preliminary work. We now proceed to the derivations of Equations (147) and (149), beginning with the latter.

Verification of Equation (149) It is now possible to verify Equation (149). Combining the above arguments with the law of total expectation and the definition of conditional expectations, the expectation of $U_{i,j}$ is calculated as follows:

$$\begin{aligned} \mathbb{E}[U_{i,j}] &= f_{R_i}(0)\, \mathbb{E}[U_{i,j} \,|\, R_i = 0] + f_{R_i}(1)\, \mathbb{E}[U_{i,j} \,|\, R_i = 1] \\ &= (1-\alpha)0 + \alpha \int_{z=-\infty}^{z=\infty} \frac{1}{\sqrt{2\pi}} z e^{-\frac{1}{2}z^2} dz \\ &= 0. \end{aligned} \tag{235}$$

Equation (149) now follows:

$$\mathbb{E}[\bar{h}_{f;J''}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)] = -\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k'-1} \mathbb{E}\left[\langle \tilde{\mathbf{x}}_i, \mathbf{v}_j \rangle \,\text{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle\right) \mathbb{I}\big(f(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle) \neq \text{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle\right)\big)\right] \mathbf{v}_j$$

$$
\begin{aligned}
&\quad \blacktriangleright \text{ by the definition of } \bar{h}_{f;J''} \text{ in Equation (49) and by Equation (232)} \\
&\quad\;\; \text{and the linearity of expectation} \\
&= -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k'-1}\mathbb{E}[U_{i,j}]\mathbf{v}_j \\
&\quad \blacktriangleright \text{ by Equation (233)} \\
&= \mathbf{0}, \\
&\quad \blacktriangleright \text{ by Equation (235)}
\end{aligned}
$$

as desired.

Verification of Equation (147) The verification of Equation (147) will largely rely on the work already accomplished above, as well as Lemma 34, which is restated below for convenience as Lemma 36.

Lemma 36 *Fix $t'', \sigma > 0$ and $0 < m \le d$. Let $J'' \subseteq [d]$, $|J''| = m$. Let $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \sum_{j \in J''} \mathbf{e}_j \mathbf{e}_j^T)$. Then,*

$$
P\left(\|\mathbf{X} - \mathbb{E}[\mathbf{X}]\|_2 > \sqrt{m}\sigma + t''\right) \le P\left(\|\mathbf{X} - \mathbb{E}[\mathbf{X}]\|_2 > \mathbb{E}[\|\mathbf{X}\|_2] + t''\right) \le e^{-\frac{1}{2\sigma^2}t''^2}. \tag{236}
$$

Recall that for each $j \in [k']$,

$$
\langle \tilde{\mathbf{x}}_i, \mathbf{v}_j \rangle \mathbf{v}_j \operatorname{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle\right) \mathbb{I}\big(f(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle) \neq \operatorname{sign}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\theta}^* \rangle\right)\big) \sim Y_i Z_{i,j} \mathbf{v}_j = U_{i,j} \mathbf{v}_j
$$

per Equation (233). Due to the rotational invariance of Gaussians, we will assume going forward, without loss of generality, that the basis vectors, $\mathbf{v}_1, \dots, \mathbf{v}_{k'}$, are simply the first k' standard basis vectors of $\mathbb{R}^d$, i.e., $\mathbf{v}_j = \mathbf{e}_j$, $j \in [k']$, where $\mathbf{e}_j \in \mathbb{R}^d$ is the vector in which the j^{th} entry is 1 and all other entries are 0. Note that under this assumption, $\boldsymbol{\theta}^* = \mathbf{v}_{k'} = \mathbf{e}_{k'}$, but this, again, does not lose any generality. For $j \in [k'-1]$, let

$$
U_j \triangleq \frac{1}{n}\sum_{i=1}^{n} U_{i,j},
$$

and let

$$
\mathbf{U} \triangleq \sum_{j=1}^{k'-1} U_j \mathbf{e}_j.
$$

Writing the random vector $\mathbf{R} \triangleq (R_1, \dots, R_n)$, and fixing $\mathbf{r} \in \{0,1\}^n$, notice that

$$
\begin{aligned}
(U_j \,|\, \mathbf{R} = \mathbf{r}) &= \frac{1}{n}\sum_{i \in \operatorname{supp}(\mathbf{r})} (U_{i,j} \,|\, \mathbf{R} = \mathbf{r}) \\
&= \frac{1}{n}\sum_{i \in \operatorname{supp}(\mathbf{r})} (U_{i,j} \,|\, R_i = 1) \\
&\quad \blacktriangleright \text{ each } U_{i,j},\, i \in [n], \text{ is independent of } \{R_{i'}\}_{i' \neq i},
\end{aligned}
$$

and since for each $i \in \operatorname{supp}(\mathbf{r})$, $r_i = 1$

$$\sim \frac{1}{n} \sum_{i \in \operatorname{supp}(\mathbf{r})} Z_{i,j},$$

▶ by the density of $U_{i,j} \mid R_i = 1$ in Equation (234)

and therefore, $(U_j \mid \mathbf{R} = \mathbf{r}) \sim \mathcal{N}(0, \frac{\|\mathbf{r}\|_0}{n^2})$. Moreover, letting $L \triangleq \|\mathbf{R}\|_0$ and $\ell \in \{0, \ldots, n\}$, it also happens that $(U_j \mid L = \ell) \sim \mathcal{N}(0, \frac{\ell}{n^2})$ since the random variables, $Z_{i,j}$, $i \in [n]$, are i.i.d. (This can be more formally argued using the density function of the conditioned random variable $U_j \mid \mathbf{R}$, combined with the law of probability.) Writing the $(k'-1)$-sparse random vector

$$\mathbf{U} \triangleq \sum_{j=1}^{k'-1} U_j \mathbf{e}_j,$$

it follows that the conditioned random vector $\mathbf{U} \mid L = \ell$ follows a zero-mean Gaussian distribution such that

$$(\mathbf{U} \mid L = \ell) = \left(\sum_{j=1}^{k'-1} U_j \mathbf{e}_j \middle| L = \ell \right) \sim \mathcal{N}\left(\mathbf{0}, \frac{\ell}{n^2} \sum_{j=1}^{k'-1} \mathbf{e}_j \mathbf{e}_j^T \right),$$

and hence, $\mathbf{U} \mid L \leq \ell$ is at most $\frac{\sqrt{\ell}}{n}$-subgaussian with mean $\mathbb{E}[\mathbf{U} \mid L \leq \ell] = \mathbf{0}$ and support of cardinality $\|\mathbf{U}\|_0 = k' - 1$. Therefore, by Lemma 36 and standard properties of Gaussians,

$$P\left(\|\mathbf{U} - \mathbb{E}[\mathbf{U}]\|_2 > \frac{\sqrt{(k'-1)\ell}}{n} + \alpha_0 t \middle| L \leq \ell \right) \tag{237}$$

$$\leq P\left(\|\mathbf{U} - \mathbb{E}[\mathbf{U}]\|_2 > \frac{\sqrt{(k'-1)\ell}}{n} + \alpha_0 t \middle| L = \ell \right) \tag{238}$$

$$\leq e^{-\frac{n^2 \alpha_0^2 t^2}{2\ell}}. \tag{239}$$

Additionally, it is folklore that $L = \sum_{i=1}^n R_i \sim \operatorname{Binomial}(n, \alpha)$ with $\mathbb{E}[L] = \mathbb{E}[\sum_{i=1}^n R_i] = \alpha n$ (see, e.g., Charikar (2002)). Thus, letting $L_0 \sim \operatorname{Binomial}(n, \alpha_0)$ be a binomial random variable with mean $\mathbb{E}[L_0] = \alpha_0 n$, by recalling that $\alpha_0 \triangleq \max\{\alpha, \delta\} \geq \alpha$, and by a standard concentration inequality for binomial random variables, for $s' \in (0, 1)$,

$$P(L > (1+s')\alpha_0 n) \leq P(L_0 > (1+s')\alpha_0 n) \leq e^{-\frac{1}{3}\alpha_0 n s'^2}. \tag{240}$$

Applying the law of total probability gives way to

$$P\left(\|\mathbf{U} - \mathbb{E}[\mathbf{U}]\|_2 > \sqrt{\frac{\alpha_0(1+s')(k'-1)}{n}} + \alpha_0 t \right)$$

$$= P\left(\|\mathbf{U} - \mathbb{E}[\mathbf{U}]\|_2 > \sqrt{\frac{\alpha_0(1+s')(k'-1)}{n}} + \alpha_0 t \middle| L \leq (1+s')\alpha_0 n \right) P(L \leq (1+s')\alpha_0 n)$$

$$+ P\left(\|\mathbf{U}-\mathbb{E}[\mathbf{U}]\|_2 > \sqrt{\frac{\alpha_0(1+s')(k'-1)}{n}} + \alpha_0 t \middle| L > (1+s')\alpha_0 n\right) P(L > (1+s')\alpha_0 n)$$

▶ by the law of total probability and the definition of conditional probabilities

$$\leq P\left(\|\mathbf{U}-\mathbb{E}[\mathbf{U}]\|_2 > \sqrt{\frac{\alpha_0(1+s')(k'-1)}{n}} + \alpha_0 t \middle| L \leq (1+s')\alpha_0 n\right) + P(L > (1+s')\alpha_0 n)$$

$$\leq e^{-\frac{1}{2(1+s')}\alpha_0 n t^2} + e^{-\frac{1}{3}\alpha_0 n s'^2}. \tag{241}$$

▶ by Equations (237) and (240) and because $\frac{\alpha_0^2 n^2 t^2}{2\alpha_0 n(1+s')} = \frac{\alpha_0 n t^2}{2(1+s')}$

For an arbitrary fixing of $J'' \in \mathcal{J}''$, the variables k' and k_0'' satisfy

$$k' = |\mathrm{supp}(\boldsymbol{\theta}^*) \cup J''| \leq \min\{|\mathrm{supp}(\boldsymbol{\theta}^*)| + |J''|, d\} \leq \min\{k + \max_{J''' \in \mathcal{J}''} |J'''|, d\} = k_0''.$$

It follows from this and the above bound in Equation (241) that

$$P\left(\left\|\frac{\bar{h}_{f;J''}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)}{\sqrt{2\pi}} - \mathbb{E}\left[\frac{\bar{h}_{f;J''}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)}{\sqrt{2\pi}}\right]\right\|_2 > \sqrt{\frac{\alpha_0(1+s')(k_0''-1)}{n}} + \alpha_0 t\right)$$

$$\leq P\left(\left\|\frac{\bar{h}_{f;J''}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)}{\sqrt{2\pi}} - \mathbb{E}\left[\frac{\bar{h}_{f;J''}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)}{\sqrt{2\pi}}\right]\right\|_2 > \sqrt{\frac{\alpha_0(1+s')(k'-1)}{n}} + \alpha_0 t\right)$$

▶ due to the above remark that $k' \leq k_0''$

$$\leq e^{-\frac{1}{2(1+s')}\alpha_0 n t^2} + e^{-\frac{1}{3}\alpha_0 n s'^2}. \tag{242}$$

▶ by Equation (241) and the definition of $\mathbf{U}$

To obtain a uniform result over all $J'' \in \mathcal{J}''$, a union bound over $\mathcal{J}''$ can be applied to the probability corresponding to the first term on the right-hand-side of the above inequality, (242), yielding Equation (147):

$$P\left(\exists J'' \in \mathcal{J}'' \quad \left\|\frac{\bar{h}_{f;J''}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)}{\sqrt{2\pi}} - \mathbb{E}\left[\frac{\bar{h}_{f;J''}(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)}{\sqrt{2\pi}}\right]\right\|_2 > \sqrt{\frac{\alpha_0(1+s')(k_0''-1)}{n}} + \alpha_0 t\right)$$
$$\leq |\mathcal{J}''| e^{-\frac{1}{2(1+s')}\alpha_0 n t^2} + e^{-\frac{1}{3}\alpha_0 n s'^2},$$

as desired. ■

D.4.3. PROOF OF LEMMA 35

This section establishes the auxiliary result, Lemma 35, stated and used in the proof of Lemma 31.
Proof Lemma 35 Recall the definitions of the functions $f_1, f_2 : \mathbb{R} \to \mathbb{R}$ from Equations (219) and (220), respectively:

$$f_1(s) \triangleq \frac{1}{\sqrt{2\pi}} e^{-s\mu_1} \int_{z=0}^{z=\infty} e^{-\frac{1}{2}(z-s)^2}(1-p(z)+p(-z))dz = \frac{1}{\sqrt{2\pi}} e^{-s\mu_1} \int_{z=0}^{z=\infty} e^{-\frac{1}{2}(z-s)^2} \nu(z) dz,$$
$$f_2(s) \triangleq \frac{1}{\sqrt{2\pi}} e^{s\mu_1} \int_{z=0}^{z=\infty} e^{-\frac{1}{2}(z+s)^2}(1-p(z)+p(-z))dz = \frac{1}{\sqrt{2\pi}} e^{s\mu_1} \int_{z=0}^{z=\infty} e^{-\frac{1}{2}(z+s)^2} \nu(z) dz,$$

where $\mu_1 = \frac{\sqrt{2/\pi}-\gamma}{2\alpha}$ and $\nu(z) = 1-p(z)+p(-z)$, $z \in \mathbb{R}$. Due to Condition (i) of Assumption 2, the function p is nondecreasing over the real line, which implies that ν is nonincreasing. Additionally, by Condition (ii) of Assumption 2, ν satisfies

$$\frac{\nu(z+w)}{\nu(z)} \geq \frac{\nu(z'+w)}{\nu(z')}$$

for $z \leq z' \in [0, \infty)$ and $w > 0$.

Next, we will establish the lemma's result for $f_1(0)$ and $f_2(0)$.

Verification of Equations (223) and (224) Equations (223) and (224) in the lemma are simple to verify:

$$\begin{aligned} f_1(0) &= \frac{1}{\sqrt{2\pi}} e^{-0\mu_1} \int_{z=0}^{z=\infty} e^{-\frac{1}{2}(z-0)^2} (1-p(z)+p(-z)) dz \\ &= \frac{1}{\sqrt{2\pi}} \int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2} (1-p(z)+p(-z)) dz \\ &= \alpha, \end{aligned}$$

and likewise,

$$\begin{aligned} f_2(0) &= \frac{1}{\sqrt{2\pi}} e^{0\mu_1} \int_{z=0}^{z=\infty} e^{-\frac{1}{2}(z+0)^2} (1-p(z)+p(-z)) dz \\ &= \frac{1}{\sqrt{2\pi}} \int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2} (1-p(z)+p(-z)) dz \\ &= \alpha, \end{aligned}$$

where the last equality in each derivation follows directly from the definition of α in Equation (8).

Verification of Equation (221) Moving on to the upper bound on f_1 in Equation (221), it suffices to show that $\frac{d}{ds} f_1(s) \leq 0$ for all $s \geq 0$ since this implies, by basic calculus, that $f_1(s) \leq f_1(0)$ over the interval $s \in [0, \infty)$. Observe:

$$\frac{d}{ds} f_1(s) = e^{-s\mu_1} \frac{1}{\sqrt{2\pi}} \int_{z=0}^{z=\infty} (z-s-\mu_1) e^{-\frac{1}{2}(z-s)^2} \nu(z) dz.$$

When $s = 0$, the desired inequality, $\frac{d}{ds} f_1(s) \leq 0$, is true:

$$\begin{aligned} \frac{d}{ds} f_1(0) &= \frac{1}{\sqrt{2\pi}} \int_{z=0}^{z=\infty} (z-\mu_1) e^{-\frac{1}{2}z^2} \nu(z) dz \\ &= \frac{1}{\sqrt{2\pi}} \int_{z=0}^{z=\infty} z e^{-\frac{1}{2}z^2} \nu(z) dz - \mu_1 \frac{1}{\sqrt{2\pi}} \int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2} \nu(z) dz \\ &= \frac{\sqrt{2/\pi}-\gamma}{2} - \frac{\sqrt{2/\pi}-\gamma}{2\alpha} \alpha \end{aligned}$$

▶ by the definitions of α, γ in Equations (8) and (9), respectively, and an earlier remark in Equation (218) that $\mu_1 = \frac{\sqrt{2/\pi}-\gamma}{2\alpha}$

$= 0.$

On the other hand, the case when $s > 0$ will require more work. Notice that

$$\frac{d}{ds}f_1(s) = e^{-s\mu_1}\frac{1}{\sqrt{2\pi}}\int_{z=0}^{z=\infty}(z-s-\mu_1)e^{-\frac{1}{2}(z-s)^2}\nu(z)dz < 0 \tag{243}$$

if and only if

$$\int_{z=0}^{z=\infty}(z-s-\mu_1)e^{-\frac{1}{2}(z-s)^2}\nu(z)dz < 0, \tag{244}$$

and similarly,

$$\frac{d}{ds}f_1(0) = \frac{1}{\sqrt{2\pi}}\int_{z=0}^{z=\infty}(z-\mu_1)e^{-\frac{1}{2}z^2}\nu(z)dz = 0 \tag{245}$$

if and only if

$$\int_{z=0}^{z=\infty}(z-\mu_1)e^{-\frac{1}{2}z^2}\nu(z)dz = 0. \tag{246}$$

We already have that $\frac{d}{ds}f_1(0) = 0$, which implies by the above observation that Equation (246) also holds.

The next argument focuses in on the former biconditional statement—in particular, the establishment of Equation (244). To derive Equation (244), the interval of integration on its left-hand-side is partitioned into three intervals:

$$\begin{aligned}
&\int_{z=0}^{z=\infty}(z-s-\mu_1)e^{-\frac{1}{2}(z-s)^2}\nu(z)dz \\
&\quad= \int_{z=0}^{z=s}(z-s-\mu_1)e^{-\frac{1}{2}(z-s)^2}\nu(z)dz + \int_{z=s}^{z=s+\mu_1}(z-s-\mu_1)e^{-\frac{1}{2}(z-s)^2}\nu(z)dz \\
&\qquad+ \int_{z=s+\mu_1}^{z=\infty}(z-s-\mu_1)e^{-\frac{1}{2}(z-s)^2}\nu(z)dz \\
&\quad= \int_{z=-s}^{z=0}(z-\mu_1)e^{-\frac{1}{2}z^2}\nu(z+s)dz + \int_{z=0}^{z=\mu_1}(z-\mu_1)e^{-\frac{1}{2}z^2}\nu(z+s)dz \\
&\qquad+ \int_{z=\mu_1}^{z=\infty}(z-\mu_1)e^{-\frac{1}{2}z^2}\nu(z+s)dz,
\end{aligned} \tag{247}$$

where the second equality applies a change of variables. Clearly, the first of the three integrals in the last expression in (247) is negative when $s > 0$:

$$\int_{z=-s}^{z=0}(z-\mu_1)e^{-\frac{1}{2}z^2}\nu(z+s)dz < 0, \tag{248}$$

and thus, if the second and third integrals in the last expression in (247) sum to a nonpositive value, then Equation (244)—and hence also Equation (243)—will hold. We will now show that this nonpositivity indeed occurs. Note the following property of an expression related to the integrand:

$$(z-\mu_1)e^{-\frac{1}{2}z^2}\nu(z) \leq 0, \quad z \in [0, \mu_1],$$

$$(z-\mu_1)e^{-\frac{1}{2}z^2}\nu(z) \geq 0, \quad z \in [\mu_1, \infty),$$

which implies that

$$-(z-\mu_1)e^{-\frac{1}{2}z^2}\nu(z) = |(z-\mu_1)e^{-\frac{1}{2}z^2}\nu(z)|, \quad z \in [0, \mu_1], \tag{249}$$

$$(z-\mu_1)e^{-\frac{1}{2}z^2}\nu(z) = |(z-\mu_1)e^{-\frac{1}{2}z^2}\nu(z)|, \quad z \in [\mu_1, \infty). \tag{250}$$

Then, for the second integral in (247), observe:

$$\begin{aligned}
\int_{z=0}^{z=\mu_1} (z-\mu_1)e^{-\frac{1}{2}z^2}\nu(z+s)dz &= \int_{z=0}^{z=\mu_1} |(z-\mu_1)e^{-\frac{1}{2}z^2}\nu(z)| \left(-\frac{\nu(z+s)}{\nu(z)}\right) dz \\
&\quad \blacktriangleright \text{ by Equation (249)} \\
&\leq \int_{z=0}^{z=\mu_1} |(z-\mu_1))e^{-\frac{1}{2}z^2}\nu(z)| \left(-\frac{\nu(\mu_1+s)}{\nu(\mu_1)}\right) dz \\
&\quad \blacktriangleright \text{ by Condition (ii) of Assumption 2,} \\
&\quad\ \ \text{and because } z \leq \mu_1 \text{ for all } z \in [0, \mu_1] \\
&= \frac{\nu(\mu_1+s)}{\nu(\mu_1)} \int_{z=0}^{z=\mu_1} (z-\mu_1)e^{-\frac{1}{2}z^2}\nu(z)dz. \qquad (251) \\
&\quad \blacktriangleright \text{ by Equation (249)}
\end{aligned}$$

Similarly, for the third integral from (247), observe:

$$\begin{aligned}
\int_{z=\mu_1}^{z=\infty} (z-\mu_1)e^{-\frac{1}{2}z^2}\nu(z+s)dz &= \int_{z=\mu_1}^{z=\infty} |(z-\mu_1)e^{-\frac{1}{2}z^2}\nu(z)| \frac{\nu(z+s)}{\nu(z)} dz \\
&\quad \blacktriangleright \text{ by Equation (250)} \\
&\leq \int_{z=\mu_1}^{z=\infty} |(z-\mu_1)e^{-\frac{1}{2}z^2}\nu(z)| \frac{\nu(\mu_1+s)}{\nu(\mu_1)} dz \\
&\quad \blacktriangleright \text{ by Condition (ii) of Assumption 2,} \\
&\quad\ \ \text{and because } z \geq \mu_1 \text{ for all } z \in [\mu_1, \infty) \\
&= \frac{\nu(\mu_1+s)}{\nu(\mu_1)} \int_{z=\mu_1}^{z=\infty} (z-\mu_1)e^{-\frac{1}{2}z^2}\nu(z)dz. \qquad (252) \\
&\quad \blacktriangleright \text{ by Equation (250)}
\end{aligned}$$

Then, the sum of the two integrals is bounded from above as follows:

$$\begin{aligned}
&\int_{z=0}^{z=\mu_1} (z-\mu_1)e^{-\frac{1}{2}z^2}\nu(z+s)dz + \int_{z=\mu_1}^{z=\infty} (z-\mu_1)e^{-\frac{1}{2}z^2}\nu(z+s)dz \\
&\leq \frac{\nu(\mu_1+s)}{\nu(\mu_1)} \int_{z=0}^{z=\mu_1} (z-\mu_1)e^{-\frac{1}{2}z^2}\nu(z)dz + \frac{\nu(\mu_1+s)}{\nu(\mu_1)} \int_{z=\mu_1}^{z=\infty} (z-\mu_1)e^{-\frac{1}{2}z^2}\nu(z)dz \\
&\quad \blacktriangleright \text{ by Equations (251) and (252)} \\
&= \frac{\nu(\mu_1+s)}{\nu(\mu_1)} \left(\int_{z=0}^{z=\infty} ze^{-\frac{1}{2}z^2}\nu(z)dz - \mu_1 \int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2}\nu(z)dz \right)
\end{aligned}$$

$$= \frac{\nu(\mu_1 + s)}{\nu(\mu_1)} \left(\left(1 - \sqrt{\frac{\pi}{2}}\gamma\right) - \frac{\sqrt{2/\pi} - \gamma}{2\alpha}\sqrt{2\pi}\alpha \right)$$

▶ by the definitions of α, γ in Equations (8) and (9), respectively,

and because $\mu_1 = \frac{\sqrt{2/\pi}-\gamma}{2\alpha}$ as in Equation (218)

$$= 0. \tag{253}$$

Substituting Equations (248) and (253) into Equation (247), it follows that for $s > 0$,

$$\begin{aligned}
&\int_{z=0}^{z=\infty} (z - s - \mu_1) e^{-\frac{1}{2}(z-s)^2} \nu(z) dz \\
&= \int_{z=-s}^{z=0} (z - \mu_1) e^{-\frac{1}{2}z^2} \nu(z + s) dz + \int_{z=0}^{z=\mu_1} (z - \mu_1) e^{-\frac{1}{2}z^2} \nu(z + s) dz \\
&\quad + \int_{z=\mu_1}^{z=\infty} (z - \mu_1) e^{-\frac{1}{2}z^2} \nu(z + s) dz
\end{aligned}$$

▶ by Equation (247)

$$< \int_{z=0}^{z=\mu_1} (z - \mu_1) e^{-\frac{1}{2}z^2} \nu(z + s) dz + \int_{z=\mu_1}^{z=\infty} (z - \mu_1) e^{-\frac{1}{2}z^2} \nu(z + s)$$

▶ by Equation (248)

$$\leq 0.$$

▶ by Equation (253)

In short, the above work has established that

$$\int_{z=0}^{z=\infty} (z - s - \mu_1) e^{-\frac{1}{2}(z-s)^2} \nu(z) dz < 0$$

when $s > 0$, and that $\frac{d}{ds} f_1(0) = 0$. Therefore, by Equations (243) and (244), as well as the earlier discussion, it happens that $\frac{d}{ds} f_1(s) \leq 0$ for all $s \geq 0$. By basic calculus, this implies that

$$\sup_{s \geq 0} f_1(s) = f_1(0),$$

verifying Equation (221).

Verification of Equation (222) Equation (222) can be derived through an analogous approach. As such, most of the analysis to upper bound f_2 falls onto showing that $\frac{d}{ds} f_2(s) \leq 0$ for all $s \geq 0$, from which it will directly follow that $f_2(s) \leq f_2(0)$ for $s \geq 0$. The derivative of f_2 with respect to s is given by

$$\frac{d}{ds} f_2(s) = e^{s\mu_1} \frac{1}{\sqrt{2\pi}} \int_{z=0}^{z=\infty} (\mu_1 - z - s) e^{-\frac{1}{2}(z+s)^2} \nu(z) dz.$$

At $s = 0$, this evaluates to

$$\frac{d}{ds} f_2(0) = \frac{1}{\sqrt{2\pi}} \int_{z=0}^{z=\infty} (\mu_1 - z) e^{-\frac{1}{2}z^2} \nu(z) dz$$

$$
\begin{aligned}
&= \mu_1 \frac{1}{\sqrt{2\pi}} \int_{z=0}^{z=\infty} e^{-\frac{1}{2}z^2}\nu(z)dz - \frac{1}{\sqrt{2\pi}} \int_{z=0}^{z=\infty} ze^{-\frac{1}{2}z^2}\nu(z)dz \\
&= \frac{\sqrt{2/\pi}-\gamma}{2\alpha}\alpha - \frac{\sqrt{2/\pi}-\gamma}{2} \\
&\qquad \blacktriangleright \text{ by the definitions of } \alpha, \gamma \text{ in Equations (8) and (9), respectively,} \\
&\qquad\quad \text{and an earlier remark in Equation (218) that } \mu_1 = \tfrac{\sqrt{2/\pi}-\gamma}{2\alpha} \\
&= 0,
\end{aligned}
\tag{254}
$$

which verifies the desired nonpositivity of $\frac{d}{ds}f_2$ in the case when $s = 0$, and which further implies that

$$
\int_{z=0}^{z=\infty} (\mu_1 - z)e^{-\frac{1}{2}z^2}\nu(z)dz = 0. \tag{255}
$$

On the other hand, towards the case whee $s > 0$, note the following biconditional statement for $s > 0$:

$$
\frac{d}{ds}f_2(s) = e^{s\mu_1}\frac{1}{\sqrt{2\pi}} \int_{z=0}^{z=\infty} (\mu_1 - z - s)e^{-\frac{1}{2}(z+s)^2}\nu(z)dz < 0 \tag{256}
$$

if and only if

$$
\int_{z=0}^{z=\infty} (\mu_1 - z - s)e^{-\frac{1}{2}(z+s)^2}\nu(z)dz < 0. \tag{257}
$$

The next step is establishing the inequality in (257) for $s > 0$. Throughout the upcoming analysis, take $s > 0$ arbitrarily. The interval of integration appearing on the left-hand-side of Equation (257) can be partitioned according to where the integrand takes positive verses nonpositive values: $z \in [0, \mu_1 - s)$ and $z \in [\mu_1 - s, \infty)$, respectively. Hence, the integral in (257) can be rewritten as:

$$
\begin{aligned}
&\int_{z=0}^{z=\infty} (\mu_1 - z - s)e^{-\frac{1}{2}(z+s)^2}\nu(z)dz \\
&= \int_{z=0}^{z=\mu_1-s} (\mu_1 - z - s)e^{-\frac{1}{2}(z+s)^2}\nu(z)dz + \int_{z=\mu_1-s}^{z=\infty} (\mu_1 - z - s)e^{-\frac{1}{2}(z+s)^2}\nu(z)dz.
\end{aligned}
\tag{258}
$$

The first of the two terms on the right-hand-side of Equation (258) is bounded from above by

$$
\begin{aligned}
&\int_{z=0}^{z=\mu_1-s} (\mu_1 - z - s)e^{-\frac{1}{2}(z+s)^2}\nu(z)dz \\
&\quad < \int_{z=0}^{z=\mu_1-s} (\mu_1 - z)e^{-\frac{1}{2}(z+s)^2}\nu(z)dz \\
&\quad < \int_{z=0}^{z=\mu_1} (\mu_1 - z)e^{-\frac{1}{2}(z+s)^2}\nu(z)dz \\
&\qquad \blacktriangleright \text{ the integrand is nonnegative on the interval } z \in [0, \mu_1] \\
&\quad < \int_{z=0}^{z=\mu_1} (\mu_1 - z)e^{-\frac{1}{2}z^2}\nu(z)dz,
\end{aligned}
\tag{259}
$$

while the second term on the right-hand-side of (258) is upper bounded by

$$\begin{aligned}
&\int_{z=\mu_1-s}^{z=\infty}(\mu_1-z-s)e^{-\frac{1}{2}(z+s)^2}\nu(z)dz\\
&\quad=\int_{z=\mu_1}^{z=\infty}(\mu_1-z)e^{-\frac{1}{2}z^2}\nu(z-s)dz\\
&\quad=\int_{z=\mu_1}^{z=\infty}|(\mu_1-z)e^{-\frac{1}{2}z^2}|(-\nu(z-s))dz\\
&\qquad\qquad \blacktriangleright \text{ since } -(\mu_1-z)e^{-\frac{1}{2}z^2}=|(\mu_1-z)e^{-\frac{1}{2}z^2}| \text{ for } z\geq\mu_1\\
&\quad\leq\int_{z=\mu_1}^{z=\infty}|(\mu_1-z))e^{-\frac{1}{2}z^2}|(-\nu(z))dz\\
&\qquad\qquad \blacktriangleright \text{ since } \nu \text{ is nonincreasing}\\
&\quad=\int_{z=\mu_1}^{z=\infty}(\mu_1-z)e^{-\frac{1}{2}z^2}\nu(z)dz.
\end{aligned}\tag{260}$$

Combining Equations (259) and (260) into Equation (258) yields:

$$\begin{aligned}
&\int_{z=0}^{z=\infty}(\mu_1-z-s)e^{-\frac{1}{2}(z+s)^2}\nu(z)dz\\
&\quad<\int_{z=0}^{z=\mu_1}(\mu_1-z)e^{-\frac{1}{2}z^2}\nu(z)dz+\int_{z=\mu_1}^{z=\infty}(\mu_1-z)e^{-\frac{1}{2}z^2}\nu(z)dz\\
&\qquad\qquad \blacktriangleright \text{ by Equations (259) and (260)}\\
&\quad=\int_{z=0}^{z=\infty}(\mu_1-z)e^{-\frac{1}{2}z^2}\nu(z)dz\\
&\quad=0.\\
&\qquad\qquad \blacktriangleright \text{ by Equation (255)}
\end{aligned}$$

Thus, for every $s>0$, Equation (257) holds, implying that Equation (256) is also true due to the biconditional relationship between this pair of equations (which was stated earlier in the proof)—that is, it indeed happens that $\frac{d}{ds}f_2(s)<0$ for $s>0$. Moreover, the derivation in (254) showed that $\frac{d}{ds}f_2(0)=0$. It follows that $\frac{d}{ds}f_2(s)\leq 0$ for all $s\geq 0$, and therefore, due to standard facts about derivatives, Equation (222) holds:

$$\sup_{s\geq 0} f_2(s)=f_2(0),$$

concluding the proof of Lemma 35. ∎