Comment on "Deep Regression Learning with Optimal Loss Function"

Zhen Li*

Abstract

OpenReview benefits the peer-review system by promoting transparency, openness, and collaboration. By making reviews, comments, and author responses publicly accessible, the platform encourages constructive feedback, reduces bias, and allows the research community to engage directly in the review process. This level of openness fosters higher-quality reviews, greater accountability, and continuous improvement in scholarly communication. In the statistics community, such a transparent and open review system has not traditionally existed. This lack of transparency has contributed to significant variation in the quality of published papers, even in leading journals, with some containing substantial errors in both proofs and numerical analyses. To illustrate this issue, this note examines several results from Wang et al. (2025) and highlights potential errors in their proofs, some of which are strikingly obvious. This raises a critical question: how important are mathematical proofs in statistical journals, and how should they be rigorously verified? Addressing this question is essential not only for maintaining academic rigor but also for fostering the right attitudes toward scholarship and quality assurance in the field. A plausible approach would be for arXiv to provide an anonymous discussion section, allowing readers–whether anonymous or not–to post comments, while also giving authors the opportunity to respond.

1 Overview

The recent work by Wang et al. (2025) studies the nonparametric estimation of the regression function $g^* : \mathcal{X} \to \mathbb{R}$. The estimation is based on independent and identically distributed (i.i.d.) samples $S = \{(\mathbf{X}_i, Y_i)\}_{i=1}^n$, which satisfy the model

$$Y_i = g^*(\mathbf{X}_i) + \epsilon_i, \quad i = 1, \ldots, n,$$

where $X_i \in \mathcal{X} \subseteq \mathbb{R}^d$, and ϵ_i is an error term independent of $\mathbf{X}_i$ with zero mean. The traditional nonparametric least squares estimator is known to underperform when ϵ_i exhibits heavy tails or follows a multimodal or heterogeneous distribution. Assuming

*Independent Researcher. E-mail: statsbeginner@yahoo.com.

further that the density function of ϵ_i exists, denoted by $f(\cdot)$, the authors propose a quasi-likelihood-type estimator defined as

$$\widehat{g} = \text{argmax}_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \log \widehat{f}(Y_i - g(\mathbf{X}_i)),$$

where $\mathcal{G}$ is a function class with domain$\mathcal{X}$, and $\widehat{f}$ is an estimator of the error density f. Specifically, $\mathcal{G}$ is chosen to be a ReLU-activated feedforward neural network (ReLU-FNN), and $\widehat{f}$ is a Nadaraya-Watson kernel estimator using a kernel function $K(\cdot)$ and a bandwidth $h > 0$.

To study the convergence properties of $\widehat{g}$, the authors first define an oracle estimator that assumes the density function f is known:

$$\widehat{g}_{\text{oracle}} \in \text{argmax}_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \log f(Y_i - g(X_i)).$$

Under the assumption that the error term ϵ is independent of the covariates $\mathbf{X}$, and that $\mathbb{E}[|\log f(\epsilon)|]$ exists, the underlying regression function g^* can be identified as the minimizer of the expected negative log-likelihood:

$$g^* = \text{argmin}_{g:\mathbb{R}^d \to \mathbb{R}} \mathbb{E}\big[- \log(Y - g(\mathbf{X}))\big].$$

The idea of this quasi-likelihood-based method is natural and is expected to perform well in practice. The main theoretical challenge, however, stems from the complexity of the resulting loss function. A similar approach, albeit with a more elaborate implementation, was employed in the recent work of Feng et al. (2024). In comparison, while Feng et al. (2024) develops a sophisticated methodology, the proofs in Wang et al. (2025) appear straightforward at first glance. Yet, my attempt to work through these proofs has raised numerous additional concerns and questions, which I summarize below for open and constructive discussion.

2 Proof of Theorem 1

Before discussing the proof of Theorem 1 in Wang et al. (2025), I would like to make a brief remark: Proposition S.1 (p. 6) and Lemma S.3 (p. 33) of the Supplementary Material (SM) do not appear to be used anywhere in the main paper or in its proofs.

Define

$$g^*_{\mathcal{G}} = \text{argmin}_{g \in \mathcal{G}} \mathcal{R}(g), \tag{2.1}$$

as the constrained population risk minimizer, where the constraint is that g belongs to the given function class of ReLU-FNNs, and the true density function is used. Thus, $g^*_{\mathcal{G}}$ is a deterministic function that serves as a network-based approximation to g^*. Given some small $\delta > 0$, let $g_1, g_2, \ldots, g_{\mathcal{N}_{2n}}$ denote the centers of the balls forming a δ-uniform covering of $\mathcal{G}$, where $\mathcal{N}_{2n} = \mathcal{N}_{2n}(\delta, \|\cdot\|\infty, \mathcal{G}|_{(\mathbf{x}_1,\ldots,\mathbf{x}_n,\mathbf{x}'_1,\ldots,\mathbf{x}'_n)})$ and $\mathcal{G}|_{\mathbf{x}} = \{g(\mathbf{x}_1), g(\mathbf{x}_2), \ldots, g(\mathbf{x}_n) : g \in \mathcal{X}\}$ with $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$.

By definition, there exists some $q^* \in \{1, \dots, \mathcal{N}_{2n}\}$ such that $\|\widehat{g}_{\text{oracle}} - g_{q^*}\|_\infty \le \delta$. Note that q^* is random, as it depends on the location of the oracle estimator. In the middle of page 8 of SM, the authors write

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{D_n}\Big|\log f(\widehat{g}_{\text{oracle}}(\mathbf{X}_i) - Y_i) - \log f(g_{q^*}(\mathbf{X}_i) - Y_i)\Big| \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{D_n}\left|\frac{f'(\epsilon_i)}{f(\epsilon_i)}\{\widehat{g}_{\text{oracle}}(\mathbf{X}_i) - g_{q^*}(\mathbf{X}_i)\}\right|.
\end{aligned}$$

The problem is that g_{q^*} is random, so clearly $Y_i - g_{q^*}(\mathbf{X}_i) \neq \epsilon_i$. This could be properly restated by introducing a mean-value (intermediate) point or by including a remainder term. However, using ϵ_i, the model error, directly in the derivative is not justified.

3 Proof of Theorem 2

Theorem 1 concerns the oracle estimator constructed using the true density function. The primary challenge, however, arises when the Nadaraya-Watson density estimator is employed in defining the loss function.

In the middle of page 14 of the SM, the following equation

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{D_n}\Big|\log f(\widehat{g}_c(\mathbf{X}_i) - Y_i) - \log f(g_{q^*}(\mathbf{X}_i) - Y_i)\Big| \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{D_n}\left|\frac{f'(\epsilon_i)}{f(\epsilon_i)}\{\widehat{g}_c(\mathbf{X}_i) - g_{q^*}(\mathbf{X}_i)\}\right|.
\end{aligned}$$

suffers from the same issue as previously noted. Here, $\widehat{g}_c$ is a corrected version of $\widehat{g}$, and q^* depends on $\widehat{g}_c$ and is therefore random as well.

Moving on to page 18, the authors state: 'We first show that for $\|g_1 - g^*\|_\infty = o_p(1)$ and $h \to 0$, we have (S.13).' Subsequently, g_1 is replaced by the estimator $\widehat{g}_c$, yet no justification is provided for why $\widehat{g}_c$ should be sufficiently close to the true function under the ℓ_∞-norm. Establishing such proximity is substantially more challenging than demonstrating consistency or convergence under the ℓ_2-norm.

On page 21, it is unclear how the authors are able to bound

$$\frac{f^{(2)}(Y_i - \widehat{g}_c(\mathbf{X}_i))}{f(Y_i - \widehat{g}_c(\mathbf{X}_i))} \quad \text{and} \quad \frac{f^{(2)}(Y_i - g_{q^*}(\mathbf{X}_i))}{f(Y_i - g_{q^*}(\mathbf{X}_i))}$$

using t_3, which is defined at the top of page 20. Neither $Y_i - \widehat{g}_c(\mathbf{X}_i)$ nor $Y_i - g_{q^*}(\mathbf{X}_i)$ corresponds directly to ϵ_i.

The most problematic argument appears on page 22. In the second-to-last step of bounding $I_{2,1}$, the authors introduce the term $\|g_{q^*} - g^*\|_\infty^2$. Recall that g_{q^*} belongs to the δ-uniform covering and satisfies $\|g_{q^*} - \widehat{g}_c\|_\infty \le \delta$. However, the proximity of g_{q^*} to g^* is entirely unknown–indeed, establishing this closeness is precisely what needs to be proved. Surprisingly, in the final step on the right-hand side of (S.15), the term $\|g_{q^*} - g^*\|_\infty^2$ disappears and is replaced by $\|g^*_{\mathcal{G}} - g^*\|_\infty^2$. Recall that $g^*_{\mathcal{G}}$ is defined in (2.1). Unlike g_{q^*}, $\|g^*_{\mathcal{G}} - g^*\|_\infty^2$ is deterministic and represents the approximation error, a completely different quantity. Similarly, inequality (S.16) involves the estimation error $\|\widehat{g}_c - g^*\|_\infty$. Yet, in the final bound at the bottom of page 25, only the approximation error term $\|g^*_{\mathcal{G}} - g^*\|_\infty^2$ remains. In fact, the authors' analysis provides no valid basis for this conclusion. The implications of their reasoning are so misguided that they offer little practical value to statisticians working on related problems.

On page 26, although deriving a bound for $\|\widehat{g}_c - \tilde{g}\|_\infty$ is of secondary importance, the proof — and in particular the chain of inequalities that leads to (S.18) — contains errors. It is unclear how the authors obtain the bound

$$\Big| \int \widehat{g}(\mathbf{x}) d(F_{n,\mathbf{x}}(\mathbf{x}) - F_{\mathbf{x}}(\mathbf{x})) \Big| \le \frac{\mathcal{B}}{n^{1/2}}.$$

At the end of the proof, the authors assert that

$$\mathbb{E}(\mathcal{R}(g_1) - \mathcal{R}(g_2)) \simeq \|g_1 - g_2\|_\infty^2.$$

We note that an equivalent assertion is stated in the main paper, immediately following Corollary 2 on page 1309. However, this claim is unfounded. Even after replacing $\|\cdot\|_\infty$ with $\|\cdot\|_2$, such an equivalence typically requires strong additional assumptions on the error density. In particular, without conditions that ensure local quadratic curvature of the risk (e.g., bounded and bounded-away-from-zero conditional density at the relevant residuals and appropriate smoothness), the risk difference need not be comparable to a squared norm. The proposed 'equivalence' to squared uniform error has no basis and is, in general, false.

To summarize, the proof of Theorem 2 is fundamentally flawed, in addition to containing minor issues[1], and it relies on numerous unfounded statements. Given my current knowledge, I do not see an immediate way to correct these fundamentally incorrect arguments.

4 Proof of Proposition 1

Proposition 1 is intended to demonstrate that the proposed estimator achieves the smallest variance among all asymptotically unbiased FNN-based estimators. However, the proof begins with an unfounded claim that

> 'Accoding to Theorem 2, the variance of $\tilde{g}$ equals to the variance of $\widehat{g}_{\text{oracle}}$ (Substituting f with $\hat{f}$ and rescaling does not introduce additional variance).'

[1] At the end of page 13, the authors claim that one can 'handle it with truncation and the classical chaining technique of empirical processes.' However, this purported 'classical chaining technique' does not appear anywhere in the proof.

Not to mention that the proof of Theorem 2 is largely incorrect; even if it were correct, I do not see anywhere that it is rigorously shown that the variance of $\tilde{g}$ equals the variance of $\widehat{g}_{\text{oracle}}$. How can this claim even hold from a common-sense perspective? Replacing a known function with a nonparametric estimator $\hat{f}$ will inevitably introduce additional variance; otherwise, why would there be an extensive literature on double robustness and cross-fitting methods?

The final technical flaw that stands out to me appears in the middle of page 28, where the authors, presumably by invoking 'the chain rule,' derive that

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial \ell(Y_i, \mathbf{X}_i; g(\cdot;\theta))}{\partial g(\mathbf{x};\theta)}\Big|_{\theta=\check{\theta}} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \ell(Y_i, \mathbf{X}_i; g(\cdot;\theta))}{\partial \theta}\frac{\partial \theta}{\partial g(\mathbf{x};\theta)}\Big|_{\theta=\check{\theta}}.$$

This constitutes a misuse of the chain rule. The left-hand side represents the derivative of the loss with respect to the function output $g(\mathbf{x};\theta)$, which itself depends on θ. One cannot simply invert the chain rule to write $\partial\theta/\partial g(\mathbf{x};\theta)$, as this derivative is not well-defined: the mapping $\theta \mapsto g(\mathbf{x};\theta)$ is typically many-to-one and nonlinear (hence non-invertible). In the present case, the mapping is given by a neural network, which is clearly nonlinear. Moreover, the expression suffers from dimensional inconsistency.

5 Summary

With the ever-increasing number of submissions to statistics journals and the growing complexity of the problems and their associated technical analyses, it is becoming increasingly unlikely that Associate Editors and reviewers have sufficient time and resources to thoroughly verify all proofs. Consequently, it is understandable that papers–particularly those containing long and intricate proofs–may occasionally include logical errors. What is surprising, however, is that the proofs of every major result in Wang et al. (2025)–including Theorem 1, Theorem 2, and Proposition 1–contain errors ranging from minor to critical. Nevertheless, it remains the authors' responsibility to rigorously cross-check both their theoretical analyses and numerical implementations. In this case, it seems that a reviewer could have identified at least some of these errors without devoting much time to the supplementary materials.

References

Oliver Y. Feng, Yu-Chun Kao, Min Xu, and Richard J. Samworth. Optimal convex M-estimation via score matching. *arXiv:2403.16688*, 2024.

Xuancheng Wang, Ling Zhou, and Huazhen Lin. Deep regression learning with optimal loss function. *Journal of the American Statistical Association*, 120(550): 1305–1317, 2025.