Risk Bounds on Aleatoric Uncertainty Recovery

Yikai Zhang[*†] **Jiahe Lin**[*] **Fengpei Li**[*] **Yeshaya Adler**
Kashif Rasul **Anderson Schneider** **Yuriy Nevmyvaka**

Machine Learning Research, Morgan Stanley
First Name.Last Name@morganstanley.com

Abstract

Quantifying aleatoric uncertainty is a challenging task in machine learning. It is important for decision making associated with data-dependent uncertainty in model outcomes. Recently, many empirical studies in modeling aleatoric uncertainty under regression settings primarily rely on either a Gaussian likelihood or moment matching. However, the performance of these methods varies for different datasets whereas discussions on their theoretical guarantees are lacking. In this work, we investigate the theoretical aspects of these approaches and establish risk bounds for their estimates. We provide conditions that are sufficient to guarantee the PAC-learnability of the aleatoric uncertainty. The study suggests that the likelihood- and moment matching-based methods enjoy different types of guarantee in their risk bounds, i.e., they calibrate different aspects of the uncertainty and thus exhibit distinct properties in different regimes of the parameter space. Finally, we conduct empirical study which shows promising results and supports our theorems.

1 INTRODUCTION

Quantifying the aleatoric uncertainty of model predictions has long been an active problem in various domains (Cook and Weisberg, 1983; Muller and Stadtmuller, 1987; Long and Ervin, 2000; Der Kiureghian and Ditlevsen, 2009; Neverova et al., 2019). In the machine learning literature, uncertainty typically can be categorized into two types (Der Kiureghian and Ditlevsen, 2009): the epistemic uncertainty such as model uncertainty and approximation error, incurred by a lack of knowledge, and the aleatoric uncertainty, which is due to the inherent randomness in the data-generating process (DGP) (Abdar et al., 2021). As data become increasingly accessible, powerful machine learning models have significantly reduced the uncertainty of the former type (Seitzer et al., 2022; Bousquet et al., 2003); however, the latter remains a challenge in a wide range of applications, such as earthquake prediction, weather forecasting, finance, and aerospace engineering (Walters et al., 2022; Beyer and Sendhoff, 2007; Krinitzsky, 2002; Yao et al., 2011). One reason for such difficulty is that aleatoric uncertainty can often be a function of the input data (Kendall and Gal, 2017; Der Kiureghian and Ditlevsen, 2009); as such, its estimation is directly related to the recovery of the data-dependent uncertainty function in the DGP. However, in the absence of specific assumptions on the uncertainty function, the task remains challenging in general.

Recently, the ML community has made several advancements with the quantification of aleatoric uncertainty which, accompanied by an accurate mean-estimation, has seen success in various tasks including semantic segmentation (Kendall and Gal, 2017; Mukhoti et al., 2021), reinforcement learning (Kahn et al., 2017; Chua et al., 2018; Seitzer et al., 2021) and selective regression (Zaoui et al., 2020). In the absence of an explicitly stated parametric form for the variance function, existing methods ubiquitously resort to modern deep neural network-based methods for the estimation of aleatoric uncertainty, with a majority of them minimizing a Gaussian likelihood-based loss or matching moments of the empirical variance. These neural networks are typically over-parameterized, which is necessary to empower the uncertainty learner with sufficient capacity; the potential over-fitting issues can often be empirically mitigated with a large sample size. Despite the empirical success of using these methods, little investigation has been done on the theoretical aspects of such estimation procedure, in particular, the PAC-learnability of the variance function, the sample complexity of different learners, and the forms of the statistical error.

In this work, we investigate the theoretical aspects of these

[*] Equal Contribution.†Corresponding Author.

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

two widely-used methods in modeling/quantifying aleatoric uncertainty, under general heteroscedastic regression settings. Specifically, we study the following DGP with mean function $\mu(\cdot)$ and data-dependent uncertainty or noise term $\xi(\cdot)$[1],

$$Y = \mu(X) + \xi(X), \;\; X \in \mathbb{R}^d, \; Y \in \mathbb{R}, \tag{1}$$

where $\mu(\boldsymbol{x}) \triangleq \mathbb{E}(Y|X = \boldsymbol{x})$, $\sigma^2(\boldsymbol{x}) \triangleq \text{Var}(Y|X = \boldsymbol{x})$ is the data-dependent variance function. Here

This is a general setting from standard heteroscedastic regressions (Goldberg et al., 1997; Le et al., 2005); In particular, it has been shown (see, e.g. Mohri et al. (2018)) that under regularity conditions, estimators for the *mean* using an empirical risk minimization (ERM) procedure possess PAC-learnable guarantees (Kearns and Vazirani, 1994). In this work, besides the mean function, we want to provide theoretical properties of the variance estimators, in the presence of data-dependent noise. For simplicity, we obtain the estimators from the ERM procedure, based on the Gaussian negative log-likelihood (NLL) loss and the mean-squared-error (MSE) loss, respectively. We find that, although both estimators recover the ground truth variance function with a similar PAC-guarantee (i.e., asymptotic rate), the finite-sample bounds take different forms. In particular, the PAC-bound based on the NLL loss provides guarantees on the learnability of the *precision* function, i.e., $1/\sigma^2(\cdot)$, while that based on the MSE loss is on the variance function $\sigma^2(\cdot)$ directly.

Summary of Contributions. The main contribution of this work is the theoretical analysis on modeling aleatoric uncertainty when the noise variance function is data-dependent. To summarize our contribution,

- We investigate the theoretical properties of empirical risk minimizers that are based on NLL and MSE losses, respectively, and provide, to the best of our knowledge, the first PAC-type finite-sample risk bounds for variance recovery in this setting;
- The theoretical analysis indicates two estimators possess different forms of the PAC-type guarantee, which allows one to identify the potential discrepancy in performance between the two estimators in different regions. In particular, the NLL loss-based estimator tends to have a lower error in the *low variance regime* while the MSE loss-based one tends to have a lower error in the *high variance regime*. This is further corroborated by our experiments.
- A novel technical contribution in Lemma 8 and 9 based on Talagrand contraction lemma. The technical novelty here allows our theorems to hold for a general calibration procedure of μ, irrespective of whether $\widehat{\mu}$ is learned jointly or separately with $\widehat{\sigma^2}$. Standard techniques require either additional assumption (typically the independence between the estimation of $\widehat{\mu}$ and $\widehat{\sigma^2}$, e.g., via split data) or additional treatment to enable such flexibility.

[1]To be specified in Section 3.

2 RELATED WORK

We provide a brief overview of existing approaches to quantify aleatoric uncertainty. These approaches fall into three major categories: (quasi)-maximum likelihood-based methods, moment matching, and Gaussian process regression.

2.1 (Quasi)-Maximum Likelihood

The likelihood function provides a full characterization of the specified density, although often there are only a few likelihood functions that are easy to work with analytically. To that end, people typically choose a working likelihood—with Gaussian being the most popular one, despite that it does not necessarily conform to the underlying true data distribution (misspecification). Such misspecification gives rise to the quasi-maximum likelihood estimation (QMLE) framework, whose estimators' properties have been extensively studied in linear settings (e.g., White, 1982; Bollerslev and Wooldridge, 1992). Nowadays, Gaussian likelihood has become the workhorse for modeling the data distribution, with flexible, non-linear functional forms (e.g., neural networks) considered for the mean and variance specification. In summary, a Gaussian negative log-likelihood (NLL)-based objective is convenient and robust, despite the potential misspecification issue.

Specifically, the NLL for the regression model posited in (1) is given by

$$\sum_{i=1}^{n} \Big[\log \sigma^2(\boldsymbol{x}_i) + \frac{(y_i - \mu(\boldsymbol{x}_i))^2}{\sigma^2(\boldsymbol{x}_i)} \Big] + \text{constant}. \tag{2}$$

In the statistics literature where the mean function is typically assumed to be linear in $\boldsymbol{x}$, that is, $\mu(\boldsymbol{x}) := \boldsymbol{\beta}^\top \boldsymbol{x}$, it reduces to the traditional MLE which goes back to the 1980s (e.g., Jobson and Fuller, 1980; Carroll and Ruppert, 1982b), wherein theoretical properties of the estimators are extensively studied and established (Jobson and Fuller, 1980), and their robustness also investigated (Carroll and Ruppert, 1982a). However, typically the variance function is not assumed to be data-dependent (i.e., constant σ not dependent on $\boldsymbol{x}_i$) until recently (e.g., Daye et al. (2012) considers the high dimensional setting and with the following parametrization: $\sigma^2(\boldsymbol{x}) = \sigma_0^2 \exp(\alpha^\top \boldsymbol{x})$).

The introduction of non-linearity into the specification of the mean and the variance is advanced by kernel-based methods, such as Reproducing Kernel Hilbert Space (RKHS)

regression. In particular, leveraging the representer theorem (Kimeldorf and Wahba, 1970, 1971), Cawley et al. (2004, 2005) propose heteroscedastic kernel ridge regression to model heteroscedastic data, where both the mean and the log standard deviation are modeled through a linear combination of the mean kernels k^{μ} and variance kernels k^{σ}.

In modern machine learning, attempts to capture aleatoric uncertainty typically rely on the power of neural networks (NN) to parameterize the mean and the variance, with NLL as the loss function (e.g., Kendall and Gal, 2017; Vandal et al., 2018; Chua et al., 2018; Skafte et al., 2019). For example, in Lakshminarayanan et al. (2017), the authors consider the use of adversarial training and ensembles to improve the estimation of predictive uncertainty, where the underlying NNs that constitute the ensembles are trained based on NLL loss. To address optimization issues related to estimating $\sigma^2(\cdot)$ using MLE, Skafte et al. (2019) also considers a local NLL framework where mini-batching is location-aware. Moreover, Seitzer et al. (2022) introduces a variance-weighting term to the Gaussian NLL loss, which acts as an adaptive, input-dependent learning rate.

2.2 Moment Matching Methods

Moment matching (MM) methods focus on aligning sample moments to the population ones dictated by the data generating process, without characterization of distribution. Ordinary least square based on MSE loss can be viewed as the simplest case of the MM estimator. For heteroscedastic linear regression, the generalized method of moments (GMM) introduced in the seminal work of Hansen (1982) simultaneously obtains the mean and variance estimates (Wooldridge, 2001); this is achieved by incorporating extra moments and minimizing a quadratic form of moment conditions. However, in practice, one typically adopts a two-step procedure where the mean and the data-dependent variance are estimated sequentially, so the moment matching of variance is based on the squared residual:

$$\min_{\sigma^2(\cdot)} \frac{1}{n} \sum_{i=1}^{n} \left(\sigma^2(\boldsymbol{x}_i) - \left(y - \widehat{\mu}_{\text{estimated mean}}(\boldsymbol{x}_i) \right)^2 \right)^2. \quad (3)$$

In modern ML, MM methods based on Maximum Mean Discrepancy (MMD) objective have also been used in generative models (e.g., Li et al., 2015; Ren et al., 2016; Li et al., 2017). Several recent works on selective regression apply moment matching-type methods to learn the variance function (Zaoui et al., 2020; Shah et al., 2022). More recently, Zaoui et al. (2020) applies a k-nearest neighborhood (kNN) scheme which aggregates local samples around $\boldsymbol{x} \in \mathbb{R}^d$ to estimate the variance using MM. However, the sample complexity provided in Zaoui et al. (2020) has an exponential dependency on dimension d due to the curse of dimensionality in kNN. We show that for universal approximators, such as the family of NNs, the exponential sample complexity can be reduced as a result of the generalization power of NN-based models.

2.3 Gaussian Process Regression

Under the Bayesian nonparametric framework, a Gaussian Process prior with mean $\mathrm{m}(\boldsymbol{x})$ and kernel $\mathrm{k}^{\mu}(\boldsymbol{x}, \boldsymbol{x}'; \theta^{\mu})$ is assumed over the mean vector $\boldsymbol{\mu}$ obtained by evaluating the mean function $\mu(\boldsymbol{x})$ on the samples, which then gives rise to a multivariate Gaussian prior on $\boldsymbol{\mu}$. In the absence of heteroscedasticity, one can solve for the hyperparameters θ^{μ} by maximizing the marginal log-likelihood.

When the noise variance becomes feature-dependent, additional assumptions on $\sigma^2(\boldsymbol{x})$ is often required, for example, Goldberg et al. (1997) assumes $\sigma^2(\boldsymbol{x}) = \exp\{g(\boldsymbol{x})\}$ and a GP prior is placed over $g(\cdot)$. Additionally, the marginal Gaussian likelihood could become analytically intractable and one needs to resort to either MCMC (Goldberg et al., 1997), or to other workarounds such as the EM-approximation in Kersting et al. (2007) or variational lower bound (Lázaro-Gredilla and Titsias, 2011). These methods—without further sparse approximation (e.g., Snelson and Ghahramani, 2007; Titsias and Lawrence, 2010) or stochastic variational inference (SVI) (Hensman et al., 2013)—can be computationally expensive and requires $O(n^3)$, which scales in a similar rate to standard GPs.

Despite the flexibility GP offers, it is subject to the limitation of kernel-based methods and does not scale well for high-dimensional input; in particular, it suffers from the curse of dimensionality and needs exponentially (relative to the input dimension) more data to maintain the same error bound (Bengio et al., 2005). Additionally, as most kernels are isotropic, GP has difficulty capturing correlated input dimensions unless one explicitly considers the covariance structure, which then requires many more hyperparameters. Finally, the cubic-scaled computational cost can often become a bottleneck for large datasets, and the issue can only be alleviated to a limited extent through some of the aforementioned techniques that aim at reducing such cost.

3 PRELIMINARIES

Notation. Vectors are denoted by bold-faced letters and $[\boldsymbol{x}]_i$ denotes the i-th entry of vector $\boldsymbol{x}$. $\boldsymbol{x}^T\boldsymbol{y}$ represents the inner product of vectors. We use $\|\cdot\|_p$ to denote the ℓ_p norm of vectors. Let $[n] = \{1, 2, ..., n\}$ and $\boldsymbol{x}_{1:n} = \{\boldsymbol{x}_i\}_{i=1}^n$. Random variables are denoted by uppercase letters, e.g., X; we write $X \sim \mathcal{D}$ when X follows distribution $\mathcal{D}$ and let $X_{1:n} \sim \mathcal{D}^n$ denote n i.i.d. random samples drawn from $\mathcal{D}$. Let $\mathbb{E}_{X\sim\mathcal{D}}$ denote the expectation taken w.r.t. X under $\mathcal{D}$ or $\mathbb{E}_X$ for short whenever there is no ambiguity. We use $\overset{\mathrm{D}}{=}$ to denote two random variables being equal in distribution. We use $O(\cdot)$, the standard big-O notation for upper bounds, and $\tilde{O}(\cdot)$ for the upper bounds omitting the log factor. Finally,

we use short hand notations $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$.

Problem Setup. Without loss of generality, consider a regression task in the feature space $\mathcal{X}$ (e.g., $\mathbb{R}^d$) with response Y taking values in $\mathcal{Y} \subseteq \mathbb{R}$. The joint distribution of $Z \triangleq (X, Y)$ over $\mathcal{Z} \triangleq \mathcal{X} \times \mathcal{Y}$ is denoted by $\bar{\mathcal{D}}$, with $\boldsymbol{z}_{1:n} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ denoting n i.i.d. samples drawn from $\bar{\mathcal{D}}$. In this paper, we postulate the following joint data-generating process (DGP):

$$X \sim \mathcal{D}; \quad Y|(X = \boldsymbol{x}) \overset{\mathrm{D}}{=} \mu(\boldsymbol{x}) + \xi(\boldsymbol{x}). \tag{4}$$

In other words, the feature $X \sim \mathcal{D}$ follows some underlying distribution on $\mathcal{D}$, and the response Y has feature-dependent mean and noise structure specified by μ and ξ. In particular, $\mu(\boldsymbol{x})$ is the conditional mean function that characterizes $\mathbb{E}[Y|X = \boldsymbol{x}]$ and $\{\xi(\boldsymbol{x})\}_{\boldsymbol{x}\in\mathcal{X}}$ is a family of noise distributions on $\mathbb{R}$ indexed by $\mathcal{X}$ (e.g., Gaussian zero-mean noise with variance $\sigma^2(\boldsymbol{x})$), which specifies $\mathbb{P}(Y \in \mathcal{B}|X = \boldsymbol{x}) = \mathbb{P}(\xi(\boldsymbol{x}) \in \mathcal{B})$ for Borel sets $\mathcal{B} \in \mathscr{B}$. Formally, in this context, both the use of $\mathbb{E}[\cdot|X = \boldsymbol{x}]$ and $\mathbb{P}[\cdot|X = \boldsymbol{x}]$ are to be understood as *regular conditional probability* (RCP) (Durrett, 2019). In general, regular conditional probability exists on *nice* probability spaces (e.g., polish space with Borel sigma algebra such as $\mathbb{R}^d$ with $\mathscr{B}^d$) and is defined almost surely for $X(\omega)$ to provide versions of $\mathbb{P}(Y \in B|\mathcal{F}_X)(\omega)$ where $\mathcal{F}_X$ is the sigma algebra generated from X. We informally write $\mathbb{P}[\cdot|X = \boldsymbol{x}]$, which can also be taken as a continuous version of RCP (Zabell, 1979).

In addition, we assume $-1 \leq \mu \leq 1$, w.o.l.g. Further, we allow the family of noise distribution $\{\xi(\boldsymbol{x})\}_{\boldsymbol{x}\in\mathcal{X}}$ to be general and only specify its first two moments (and assume the existence of its fourth moment, for moment matching with variance function). In particular, let $\mathscr{R}$ be the space of random variable distributed on $\mathbb{R}$ with finite variance, then the feature-dependent noise $\xi(\cdot) : \mathcal{X} \to \mathscr{R}$ satisfies $\mathbb{E}[\xi(\boldsymbol{x})] = 0$ and variance $\sigma^2(\boldsymbol{x}) \triangleq \mathrm{Var}(\xi(\boldsymbol{x}))$. As suggested in the form of (4), we assume an additive noise structure where $\xi(\boldsymbol{x}_i)$ is conditionally independent of $\xi(\boldsymbol{x}_j)$ given $X_i = \boldsymbol{x}_i$, $X_j = \boldsymbol{x}_j$. The DGP model (4) completely characterizes $Z \triangleq (X, Y) \sim \bar{\mathcal{D}}$, and allows for the general form of heteroscedasticity of Y.

Assumption 1. *We assume the noise $\xi(\boldsymbol{x})$ in DGP is bounded and satisfies $|\xi| \leq \sqrt{M}$ for some $M > 0$. Moreover, let $\mathcal{F}_\mu := \{f : \mathcal{X} \mapsto [-1, 1]\}$ and $\mathcal{F}_{\sigma^2} := \{\sigma^2 : \mathcal{X} \mapsto [m, M]\}$ with $0 < m < M$, be two hypothesis classes with finite pseudo-dimension $d_P(\mathcal{F}_\mu) < \infty, d_P(\mathcal{F}_{\sigma^2}) < \infty$ that include the true DGP as dictated in* (4)*:*

$$\mu^\star \in \mathcal{F}_\mu, \text{ and } (\sigma^2)^\star \in \mathcal{F}_{\sigma^2};$$

$\mu^\star(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{x}]$ is the true mean function and $(\sigma^2)^\star(\boldsymbol{x}) = \mathit{Var}(Y|\boldsymbol{x})$ is the true feature-dependent variance.

The remainder of this paper is dedicated to theoretical and empirical analyses on recovering the variance function $(\sigma^2)^\star$ using n i.i.d. samples $\boldsymbol{z}_{1:n} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ from the DGP, respectively based on MSE and NLL loss functions, under Assumption 1. Concretely, the two loss functions are defined as follows.

- NLL loss $\ell_{\mathrm{NLL}} : \mathcal{G} \times \mathcal{F} \times \mathcal{Z} \mapsto \mathbb{R}$ where

$$\ell_{\mathrm{NLL}}(\sigma^2, \mu, z) \triangleq \log(\sigma^2(\boldsymbol{x})) + \frac{(y - \mu(\boldsymbol{x}))^2}{\sigma^2(\boldsymbol{x})}. \tag{5}$$

- MSE loss $\ell_{\mathrm{MSE}} : \mathcal{G} \times \mathcal{F} \times \mathcal{Z} \mapsto \mathbb{R}$ where

$$\ell_{\mathrm{MSE}}(\sigma^2, \mu, z) \triangleq \big(\sigma^2(\boldsymbol{x}) - \big(y - \mu(\boldsymbol{x})\big)^2\big)^2. \tag{6}$$

  Note that MSE loss is a MM-type loss that directly matches the second moment of the residuals.

3.1 Complexity of Function Classes

To derive our main results, we use the *Rademacher complexity* of a hypothesis class (Bartlett and Mendelson, 2002; Bartlett et al., 2005) and VC-class Van der Vaart and Wellner (1996); Vapnik and Chervonenkis (2015). Lemmas in this section are standard results, reference details are specified in the Appendix.

Definition 1. *Given $\{\boldsymbol{x}_{1:n}\}$ where $\boldsymbol{x}_i \in \mathcal{X}$ and a hypothesis class $\mathcal{F} = \{f : \mathcal{X} \mapsto [l, u]\}$, the empirical Rademacher complexity of the hypothesis class $\mathcal{F}$ is defined as*

$$\widehat{\mathcal{R}}_n(\mathcal{F}; \boldsymbol{x}_{1:n}) \triangleq \mathbb{E}_{\sigma_{1:n}\sim\mathcal{D}_\sigma^n}[\sup_{f\in\mathcal{F}} \frac{1}{n} \sum_{i=1}^n \sigma_i f(\boldsymbol{x}_i)], \tag{7}$$

where $\sigma_{1:n}$ are i.i.d. Rademacher random variables with distribution $\mathcal{D}_\sigma$, i.e., $\mathbb{P}(\sigma_i = 1) = \mathbb{P}(\sigma_i = -1) = 1/2$. Furthermore, given $X \sim \mathcal{D}$, the Rademacher average of $\mathcal{F}$ is defined as

$$\mathcal{R}_n(\mathcal{F}) \triangleq \mathbb{E}_{X_{1:n}\sim\mathcal{D}^n}[\widehat{\mathcal{R}}_n(\mathcal{F}; X_{1:n})]. \tag{8}$$

Definition 2 (VC-dimension)**.** *The VC-dimension $d_{VC}(\mathcal{F})$ of a hypothesis class $\mathcal{F} = \{f : \mathcal{X} \mapsto \{1, -1\}\}$ is the largest cardinality of any set $S \subseteq \mathcal{X}$ such that $\forall \bar{S} \subseteq S$, $\exists f \in \mathcal{F}$:*

$$f(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \boldsymbol{x} \in \bar{S} \\ -1 & \text{if } \boldsymbol{x} \in S \setminus \bar{S} \end{cases} \tag{9}$$

Definition 3 (Pseudo-dimension)**.** *The Pseudo-dimension $d_P(\mathcal{F})$ of a real-valued hypothesis class $\mathcal{F} = \{f : \mathcal{X} \mapsto [l, u]\}$ is the VC-dimension of the hypothesis class*

$$\begin{aligned} \mathcal{H} = \{&h : \mathcal{X} \times \mathbb{R} \mapsto \{-1, 1\} \mid \\ &h(\boldsymbol{x}, t) = \mathit{sign}(f(\boldsymbol{x}) - t), f \in \mathcal{F}, t \in \mathbb{R}\}. \end{aligned} \tag{10}$$

Rademacher complexity is extensively used in deriving bounds for the statistical error; such a bound is intimately

related to the complexity of the hypothesis class and the sample size n. To better characterize the relationships among sample size n, the complexity of a hypothesis class, and the statistical error, we use the following lemma to explicitly bound the Rademacher complexity using pseudo-dimension and sample size. The lemma follows from Dudley's chaining bound (Dudley, 1967) and the concept of "covering number" (Pollard, 1982; Rakhlin, 2020) that is built upon Haussler's bound and the combinatorial dimension Haussler (1995); Van der Vaart and Wellner (1996).

Lemma 1. *Let $\mathcal{F} = \{f : \mathcal{X} \mapsto [-M, M]\}$ with $M > 1$ be a hypothesis class with finite pseudo-dimension $d_P(\mathcal{F}) < \infty$. Let $\widehat{\mathcal{R}}_n(\mathcal{F}; \boldsymbol{x}_{1:n})$ be the empirical Rademacher complexity defined in* (7)*, then there exists some universal constant K_1 such that for any $\boldsymbol{x}_{1:n} \subseteq \mathcal{X}$*

$$\widehat{\mathcal{R}}_n(\mathcal{F}; \boldsymbol{x}_{1:n}) \leq K_1 \sqrt{\frac{d_P(\mathcal{F})}{n}}.$$

Consequently, we also have $\mathcal{R}_n(\mathcal{F}) \leq K_1 \sqrt{\frac{d_P(\mathcal{F})}{n}}$ for any distribution $X \sim \mathcal{D}$.

The proof of Lemma 1 can be found in (Bartlett et al., 2005; Rakhlin, 2020).

Note that the complexity of a function class is related to the efficiency in its sample complexity. To be specific, consider the example of learning a parity variance function, one can recover the variance function using ERM by fitting a model within some hypothesis class, or apply local methods by leveraging the neighborhood information. Let $\boldsymbol{x}$ be a d-dimensional vector sampled uniformly from the vertices of a hyper-cube $\mathcal{X} : \{0, 1\}^d$, and let the ground truth variance function be a d-bit parity function $\sigma^2 : \mathcal{X} \mapsto \{0, 1\}$ with $\sigma^2(\boldsymbol{x}) = 1$ if and only if $\mathbf{1}^\top \boldsymbol{x}$ is odd. Given training samples $\boldsymbol{x}_{1:n}$, the task is to obtain estimates for $\sigma^2(\cdot)$. Local methods, such as the kNN predictor (Györfi et al., 2002) and kernel methods that leverage radial basis kernels (Cawley et al., 2004; Bengio et al., 2005), aggregate information around the neighborhood and compute the variance accordingly. Since all neighborhood vertices in a hyper-cube have different parities, to accurately estimate $\sigma^2(\boldsymbol{x})$, these methods effectively restrict themselves to the data that are identical to $\boldsymbol{x}$, rendering an exponential sample requirement $\Omega(2^d)$. This sample complexity is also observed in Zaoui et al. (2020). On the other hand, if we restrict the complexity of the function class, the ERM-based methods can improve the sample complexity. For example, note that the parity function can be represented using a one layer neural network with $O(d^2)$ parameters (Rumelhart et al., 1988), and it has VC-dimension at most $O(d^2 \log(d))$ (Blum et al., 2020). Such a neural network is PAC-learnable with sample complexity polynomial in d and hence is much more sample-efficient than the exponential one. Note the *computational complexity* of learning such a neural network can be intractable in general (Haussler, 1992), but this is beyond the scope of our discussion. Here we focus on the gain in sample efficiency by restricting the estimation within a certain hypothesis class; yet such a class can be flexible enough (e.g., neural networks) to approximate the ground truth sufficiently well. Additionally, learning the parity variance function using training data polynomial in d also implies that the learner is able to *generalize* well on data not seen before. Thus, the complexity of function classes is the entry point of our analysis. In the next section, we rigorously quantify the generalization power of a hypothesis class in estimating aleatoric uncertainty.

4 MAIN RESULTS

We first give an overview of our theoretical results. Both theorems establish that under Assumption 1, the variance function estimators $\widehat{\sigma^2}$ recover the ground truth variance $(\sigma^2)^\star$ with PAC-type guarantee, as long as the estimated mean function $\widehat{\mu}$ is reasonably close to the ground truth mean $\mu^\star$ in $\mathscr{L}_2$, with the latter following from standard results in ML. The sample complexity is polynomial in the pseudo dimension of the hypothesis class of the variance function $d_P(\mathcal{F}_{\sigma^2})$.

A Road Map. Due to space constraints, we only present a proof sketch, while deferring all lemmas and detailed proofs to the supplementary material. Since $\mathcal{F}_\mu$ and $\mathcal{F}_\sigma$ are uniformly bounded hypothesis classes, for any $\sigma^2 \in \mathcal{F}_{\sigma^2}, \mu \in \mathcal{F}_\mu$ and $\boldsymbol{z}$, one can obtain uniform bounds for $\ell_{\text{NLL}}(\sigma^2, \mu, z)$ and $\ell_{\text{MSE}}(\sigma^2, \mu, z)$. In particular, the statistical error of these empirical losses—relative to their population counterpart—can be bounded using McDiarmid inequality, and is characterized using Rademacher complexity in Proposition 1 and Proposition 2, respectively. Theorem 1 and Theorem 2 provide a formal statement of the main results, with the sample size chosen sufficiently large to control the statistical error of the losses; the latter implies the recovery of the ground truth variance function.

Proposition 1 (Statistical Error of ℓ_{NLL})**.** *Under Assumption 1, given n i.i.d. samples $Z_{1:n}$ drawn from the DGP, the following holds uniformly for all $\sigma^2 \in \mathcal{F}_{\sigma^2}, \mu \in \mathcal{F}_\mu$:*

$$\begin{aligned}
&\left| \mathbb{E}_Z \left[\ell_{NLL}(\sigma^2, \mu, Z) \right] - \frac{1}{n} \sum_{i=1}^{n} \ell_{NLL}(\sigma^2, \mu, Z_i) \right| \\
\leq & O\Bigg(\frac{1 + m + M}{m^2} \mathcal{R}_n(\mathcal{F}_{\sigma^2}) + (1 + m^{-1} + M) \mathcal{R}_n(\mathcal{F}_\mu) \\
& + \left(\frac{1 + M}{m} \right) \sqrt{\frac{1}{n} \log\left(\frac{1}{\delta} \right)} \Bigg),
\end{aligned} \tag{11}$$

with probability at least $1 - \delta$.

Proposition 2 (Statistical Error of ℓ_{MSE})**.** *Under Assumption 1, given n i.i.d. samples $Z_{1:n}$ from the DGP, the follow-*

ing holds uniformly for all $\sigma^2 \in \mathcal{F}_{\sigma^2}, \mu \in \mathcal{F}_\mu$:

$$\begin{aligned}&\left|\mathbb{E}_Z\left[\ell_{MSE}(\sigma^2,\mu,Z)\right]-\frac{1}{n}\sum_{i=1}^{n}\ell_{MSE}(\sigma^2,\mu,Z_i)\right|\\ \leq& O\bigg((1+M)\mathcal{R}_n(\mathcal{F}_{\sigma^2})+(1+M^{\frac{3}{2}})\mathcal{R}_n(\mathcal{F}_\mu)\\ &+(1+M^2)\sqrt{\frac{1}{n}\log\frac{1}{\delta}}\bigg),\end{aligned}\tag{12}$$

with probability at least $1-\delta$.

The analysis in Proposition 1 and Proposition 2 holds for any $\mu \in \mathcal{F}_\mu$ and $\sigma^2 \in \mathcal{F}_{\sigma^2}$ and thus permitting μ and σ^2 to be trained using the same training set. This poses a major challenge in the proof as one needs to decouple the Rademacher complexity of class products $\mathcal{R}_n(\mathcal{F}_\mu \cdot \mathcal{F}_{\sigma^2})$ into their own respective terms $\mathcal{R}_n(\mathcal{F}_\mu)$ and $\mathcal{R}_n(\mathcal{F}_{\sigma^2})$ (see proof of Lemmas 5-9 in the supplementary material.

Theorem 1 (Risk Bound of Empirical Risk Minimizer using ℓ_{NLL})**.** *Suppose Assumption 1 holds. For any given $\widehat{\mu} \in \mathcal{F}$ that satisfies*

$$\mathbb{E}_X(\mu^*(X)-\widehat{\mu}(X))^2 \leq \varepsilon,$$

let

$$\widehat{\sigma^2}=\operatorname*{argmin}_{\sigma^2\in\mathcal{F}_{\sigma^2}}\frac{1}{n}\sum_{i=1}^{n}\ell_{NLL}(\sigma^2;\widehat{\mu},Z_i).$$

Then for $\varepsilon < \frac{1}{2}$, the following holds

$$\mathbb{E}_X\left|\frac{1}{\widehat{\sigma^2}(X)}-\frac{1}{(\sigma^2)^\star(X)}\right|^2 \leq \varepsilon\tag{13}$$

with probability at least $1-\delta$, provided that n is sufficiently large

$$\begin{aligned}n=&O\bigg(\frac{1}{\min\{m^6,m^4\}\varepsilon^2}\bigg((\frac{1}{m^2}+\frac{1}{m^6}+\frac{M^2}{m^4})d_P(\mathcal{F}_{\sigma^2})\\ &+(1+\frac{1}{m^2}+M)d_P(\mathcal{F}_\mu)+\frac{1+M^2}{m^2}\log(\frac{2}{\delta})\bigg)\bigg).\end{aligned}$$

Theorem 2 (Risk Bound of Empirical Risk Minimizer using ℓ_{MSE})**.** *Suppose Assumption 1 holds. For any given $\widehat{\mu} \in \mathcal{F}$ that satisfies*

$$\mathbb{E}_X(\mu^*(X)-\widehat{\mu}(X))^2 \leq \varepsilon,$$

let

$$\widehat{\sigma^2}=\operatorname*{argmin}_{\sigma^2\in\mathcal{F}_{\sigma^2}}\frac{1}{n}\sum_{i=1}^{n}\ell_{MSE}(\sigma^2;\widehat{\mu},Z_i).$$

Then for $\varepsilon < \frac{1}{2}$, the following holds with probability at least $1-\delta$

$$\mathbb{E}_X\left(\sigma^{*2}(X)-\widehat{\sigma^2}(X)\right)^2 \leq \varepsilon,$$

provided that n is sufficiently

$$\begin{aligned}n=&O\bigg(\frac{1}{\varepsilon^2}\bigg((M^2+1)d_P(\mathcal{F}_{\sigma^2})+(M^3+1)d_P(\mathcal{F}_\mu)\\ &+(1+M^4)\log(\frac{2}{\delta})\bigg)\bigg).\end{aligned}$$

Remark 1. *The bounds in Theorems 1 and 2 are in their most general forms. In most cases, m is typically a universal constant and M can vary to a large value when compared to the mean. In practice, typically there are two cases:*

- *The mean function $\widehat{\mu}$ is estimated with different data. In this case, we can simply pick $\mathcal{F}_\mu = \{\widehat{\mu}\}$ so that $d_P(\mathcal{F}_\mu) = 0$. As long as $\mathbb{E}_X(\mu^*(X)-\widehat{\mu}(X))^2 \leq \varepsilon$, Theorems 1 and 2 will hold and the dominating term in sample complexity is $O(\frac{M^2 d_P(\mathcal{F}_{\sigma^2})}{\varepsilon^2})$*
- *The mean function $\widehat{\mu}$ is estimated with the same data and jointly with $\widehat{\sigma^2}$. Whether $\widehat{\mu}$ is estimated with a different procedure or jointly with $\widehat{\sigma^2}$ by ERM, note one will always have $\widehat{\sigma^2} = \operatorname{argmin}_{\sigma^2\in\mathcal{F}_{\sigma^2}}\frac{1}{n}\sum_{i=1}^{n}\ell(\sigma^2;\widehat{\mu},Z_i)$, regardless how $\widehat{\mu}$ is obtained. Thus Theorems 1 and 2 will still hold.*

Finally, we briefly comment on the $\mathbb{E}_X(\mu^*(X)-\widehat{\mu}(X))^2 \leq \varepsilon$ condition. This condition can be achieved using standard ERM, with sample size $\widetilde{O}(d_P(\mathcal{F}_\mu)/\varepsilon)$ under mild assumptions (e.g., bounded regression and realizability (Bartlett et al., 2005; Massart and Nédélec, 2006)).

Remark 2. *More importantly, the bound for the ℓ_{NLL}-based estimator has risk bound in the precision, that is, $\mathbb{E}_X\left|\frac{1}{\widehat{\sigma^2}(X)}-\frac{1}{(\sigma^2)^\star(X)}\right|^2 \leq \varepsilon$ which can be equivalently written in some* ratio *form $\mathbb{E}_X\left[\left|\frac{\widehat{\sigma^2}(X)-(\sigma^2)^\star(X)}{\widehat{\sigma^2}(X)(\sigma^2)^\star(X)}\right|^2\right] \leq \varepsilon$; on the other hand, ℓ_{MSE}-based empirical risk minimizer in an additive manner: $\mathbb{E}_X((\sigma^2)^\star(X)-\widehat{\sigma^2}(X))^2 \leq \varepsilon$. The "ratio" error suggests an advantage of ℓ_{NLL} over ℓ_{MSE} when $\widehat{\sigma}^2(\boldsymbol{x})$ and $(\sigma^2)^\star(\boldsymbol{x})$ are small, i.e., in a low variance regime, where an "additive" error suggests an advantage of ℓ_{MSE} over ℓ_{NLL} when $\widehat{\sigma}^2(\boldsymbol{x})$ and $(\sigma^2)^\star(\boldsymbol{x})$ are large, i.e., a high variance regime. Our empirical study corroborates these observations. We shall observe this phenomenon in the experiments section.*

5 EXPERIMENTS

We conduct a series of synthetic data experiments to support our theoretical findings in the previous section. Our experiments consist of two parts: in the first part, we consider settings S1, S2 and S3 (details to be introduced later) with the goal of showing that one can use neural networks to approximate the ground truth variance function by minimize the NLL or MSE loss, when provided with a sufficient

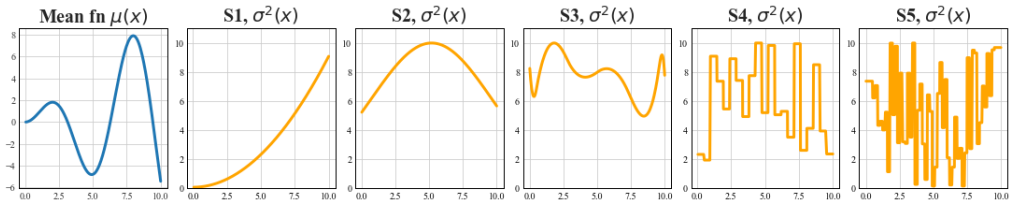

Figure 1: Visualization for the true $\mu^\star(x)$ (blue) and $(\sigma^2)^\star(x)$ (orange) used in synthetic data experiment settings S1: quadratic in x; S2: σ is parametrized through an rbf network; S3: summation of B-spline bases; S4 and S5: summation of indicator function bases with different number of knots.

amount of data; in the second part, we consider settings S4 and S5, aiming to show that the empirical minimizers using two losses have their respective merit in performance, depending on the magnitude of underlying true variance.

Part I. We consider a similar setting to those in Skafte et al. (2019); Seitzer et al. (2022), where the mean function is a sinusoidal curve and the small additive noise is heteroscedastic with the magnitude of the variance depending on the value of $\boldsymbol{x}$. Concretely, the true data-generating process is given as follows for some univariate regressor $x \in [0, 10]$ and response y:

$$y = \mu^\star(x) + \sqrt{(\sigma^2)^\star(x)}u, \quad \mathbb{E}(u) = 0;\ \mathrm{Var}(u) = 1; \tag{14}$$

the ground truth mean function is given by $\mu^\star(x) := x\sin(x)$; for the data-dependent noise scale function $(\sigma^2)^\star(x)$ and u, we consider several different settings where the complexity of the functional class and the distribution of u varies:

(S1) $(\sigma^2)^\star(x) := 0.09(1 + x^2)$, $u \sim \mathcal{N}(0, 1)$;

(S2) $(\sigma^2)^\star(x) := \tilde{\sigma}^2(x)$, where $\tilde{\sigma}(x)$ is parametrized through a radial basis function (rbf) network, that is,

$$\tilde{\sigma}(x) := \sum\nolimits_{k=1}^{K} w_k \cdot \exp\{-\beta_k(x - c_k)^2\}.$$

$K \equiv 50$ is the number of neurons in the hidden layer, c_k is the center and β_k is the scale, with the former randomly generated from $\mathsf{Unif}(0, 10)$ and the latter from $\mathsf{Unif}(0.01, 0.02)$; the weights w_k's are generated in an identical way to Xavier initialization (Glorot and Bengio, 2010); $u \sim t_8$, that is, a centered student t-distribution with degree of freedom being 8;

(S3) $(\sigma^2)^\star(x) := \sum_{k=1}^{K}(\beta_{1,k}B_{k,3}(x) + \beta_{2,k}B_{k,4}(x))$, i.e., the summation of 3rd and 4th order B-spline (De Boor, 1978) bases with the underlying piecewise polynomials having degree 2 and 3; $K \equiv 5$ is the number of knots and $\beta_{1,k}, \beta_{2,k}$ are positive coefficients randomly generated from $\mathsf{Unif}(0, 5)$; $u \sim \mathcal{N}(0, 1)$.

Finally, we ensure that the domain of $(\sigma^2)^\star(x)$ is $[0.1, 10]$, by applying scaling or flooring wherever necessary. Note that settings (S2) and (S3) are challenging as a result of the high complexity of the functional class that $(\sigma^2)^\star(x)$ falls into. A visualization for these settings is given in Figure 1.

We use two feed-forward neural networks (w/ additional residual connections) to fit $\mu(x)$ and $\sigma^2(x)$, respectively. Model training proceeds in two stages: at Stage I, the mean function is fitted based on mean-squared-error loss, that is,

$$\widehat{\mu}(\cdot) = \arg\min_{\mu(\cdot)} \sum_{i=1}^{n_{\text{train}}} \big(y_i - \mu(x_i)\big)^2.$$

Once the mean network is trained, we obtain the residual $\widehat{\xi}_i := y_i - \widehat{\mu}(x_i)$ and proceed to Stage II to fit the variance function, respectively based on the mean-squared-error loss (or equivalently, moment-matching) and the negative log-likelihood loss (NLL):

$$\widehat{\sigma^2}_{\text{MSE}}(\cdot) := \arg\min_{\sigma^2(\cdot)} \sum_{i=1}^{n_{\text{train}}} \big(\widehat{\xi}_i^2 - \sigma^2(x_i)\big)^2,$$

$$\widehat{\sigma^2}_{\text{NLL}}(\cdot) := \arg\min_{\sigma^2(\cdot)} \sum_{i=1}^{n_{\text{train}}} \Big(\frac{\widehat{\xi}_i^2}{\sigma^2(x_i)} + \log\sigma^2(x_i)\Big). \tag{15}$$

For each setting, we fit the model to 10 replicas of the generated data. Specifically, the networks are trained on sets with different training sizes {1e3, 3e3, 5e3, 1e4, 3e4}, whose performance is then evaluated on a test set of fixed size 1e3, by comparing the estimates against the underlying ground truth. Table 1 shows the RMSE of the mean estimate $\widehat{\mu}$ and variance estimates $\widehat{\sigma^2}(x)$ across different settings, with the model trained on samples of different sizes. Figure 2 displays the confidence band from the truth and the estimated variance within 2 standard deviation.

Based on the results in Table 1 and Figure 2, we observe: (1) empirically, both moment matching and NLL-based estimators are capable of learning the underlying true variance function, with their respective performance similar, as manifested by the similar magnitude in RMSE; (2) there is some degradation in performance for the variance estimate when the functional class becomes more complicated, whereas the adverse impact to the mean estimate is somewhat limited.

Part II. To explore what the theoretical bounds in Theorems 1 and 2 imply in empirical settings—in particular,

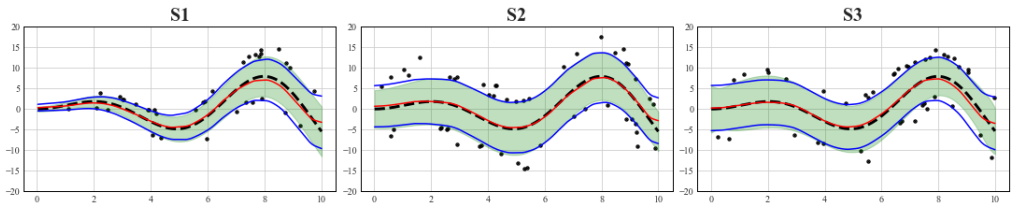

Figure 2: True confidence band and estimated ones. In particular, the black dashed line corresponds to the true mean $\mu^\star(x)$ and the red line corresponds to the estimated one; the $\pm 2\sigma$ area based on the truth is shaded in green and that from the estimated is outlined by the blue line. Black dots corresponds to points that fall outside of the 2σ region.

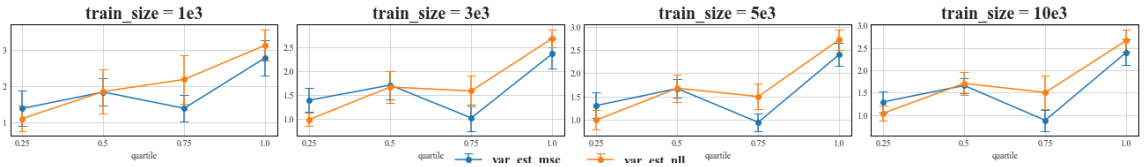

Figure 3: Setting S4; comparison in RMSE between $\widehat{\sigma^2}_{\text{MSE}}(x)$ (in blue) and $\widehat{\sigma^2}_{\text{NLL}}(x)$ (in orange) by quartile

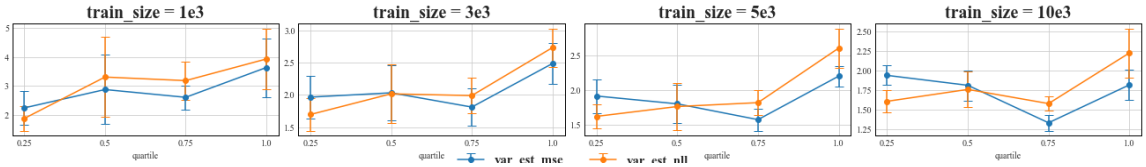

Figure 4: Setting S5; comparison in RMSE between $\widehat{\sigma^2}_{\text{MSE}}(x)$ (in blue) and $\widehat{\sigma^2}_{\text{NLL}}(x)$ (in orange) by quartiles

Table 1: RMSE for $\widehat{\mu}(x)$ and $\widehat{\sigma^2}(x)$ across different settings and train sizes. The reported number corresponds to the average across 10 replicas, with the standard deviation report in parentheses.

		train size				
		1e3	3e3	5e3	1e4	3e4
S1	$\widehat{\mu}$	0.74(.21)	0.58(.16)	0.55(.13)	0.47(.19)	0.48(.10)
	$\widehat{\sigma^2}_{\text{MSE}}$	0.50(.53)	0.34(.31)	0.40(.24)	0.45(.29)	0.31(.10)
	$\widehat{\sigma^2}_{\text{NLL}}$	0.89(.65)	0.49(.32)	0.44(.20)	0.42(.45)	0.45(.16)
S2	$\widehat{\mu}$	0.64(.13)	0.62(.18)	0.56(.12)	0.54(.10)	0.44(.12)
	$\widehat{\sigma^2}_{\text{MSE}}$	1.23(2.29)	0.93(.77)	0.83(.38)	1.00(.32)	0.75(.19)
	$\widehat{\sigma^2}_{\text{NLL}}$	1.59(.49)	1.49(.58)	0.94(.51)	0.70(.46)	0.74(.21)
S3	$\widehat{\mu}$	0.58(.17)	0.58(.15)	0.55(.12)	0.70(.19)	0.56(.21)
	$\widehat{\sigma^2}_{\text{MSE}}$	1.22(.21)	1.21(.25)	1.16(.25)	1.34(.31)	0.98(.45)
	$\widehat{\sigma^2}_{\text{NLL}}$	1.44(.33)	1.02(.30)	0.95(.20)	1.22(.32)	1.18(.35)

in terms of how estimates based on the two losses would compare in different value regions—we consider the following special setup. Data is still generated according to (14), with $\mu^\star(x) := x\sin(x)$. The variance function $(\sigma^\star)^2(x)$ is parameterized through the sum of indicator functions, that is

$$(\sigma^2)^\star(x) = \sum_{k=1}^{K} \beta_k B_{k,1}(x),$$

where $B_{k,1}(x)$'s are B-spline basis function with order 1 that are effectively indicator functions, and β_k's are positive coefficients. The specific values for K and the noise distribution are given below, with the settings marked as S4 and S5; the latter corresponds to a functional class with a higher complexity, as a result of using more knots/bases.

Table 2: RMSE for $\widehat{\sigma^2}(x)$ in different bins based on quartiles of $(\sigma^2)^\star(x)$, across different settings and sample sizes

			bin by quartile			
			0-25%	25-50%	50-75%	75-100%
S4	1e3	$\widehat{\sigma^2}_{\text{MSE}}$	1.39(.49)	**1.84(.38)**	**1.39(.37)**	**2.78(.48)**
		$\widehat{\sigma^2}_{\text{NLL}}$	**1.11(.34)**	1.85(.60)	2.19(.66)	3.12(.43)
	3e3	$\widehat{\sigma^2}_{\text{MSE}}$	1.40(.25)	1.71(.29)	**1.03(.27)**	**2.37(.31)**
		$\widehat{\sigma^2}_{\text{NLL}}$	**0.99(.13)**	**1.67(.32)**	1.59(.32)	2.69(.19)
	5e3	$\widehat{\sigma^2}_{\text{MSE}}$	1.30(.28)	**1.67(.20)**	**0.94(.19)**	**2.40(.25)**
		$\widehat{\sigma^2}_{\text{NLL}}$	**0.99(.21)**	1.68(.29)	1.50(.28)	2.72(.22)
	1e4	$\widehat{\sigma^2}_{\text{MSE}}$	1.29(.24)	**1.66(.17)**	**0.88(.24)**	**2.40(.29)**
		$\widehat{\sigma^2}_{\text{NLL}}$	**1.04(.17)**	1.71(.26)	1.50(.38)	2.67(.25)
S5	1e3	$\widehat{\sigma^2}_{\text{MSE}}$	2.24(.58)	**2.87(1.20)**	**2.60(.42)**	**3.64(1.02)**
		$\widehat{\sigma^2}_{\text{NLL}}$	**1.87(.43)**	3.31(1.38)	3.18(.67)	3.93(1.05)
	3e3	$\widehat{\sigma^2}_{\text{MSE}}$	1.97(.33)	2.03(.43)	**1.81(.29)**	**2.49(.31)**
		$\widehat{\sigma^2}_{\text{NLL}}$	**1.70(.25)**	**2.02(.45)**	1.99(.27)	2.74(.29)
	5e3	$\widehat{\sigma^2}_{\text{MSE}}$	1.92(.25)	1.80(.28)	**1.57(.16)**	**2.20(.15)**
		$\widehat{\sigma^2}_{\text{NLL}}$	**1.62(.17)**	**1.76(.34)**	1.82(.18)	2.61(.28)
	1e4	$\widehat{\sigma^2}_{\text{MSE}}$	1.94(.13)	1.82(.19)	**1.32(.10)**	**1.82(.19)**
		$\widehat{\sigma^2}_{\text{NLL}}$	**1.61(.14)**	**1.76(.23)**	1.58(.09)	2.23(.31)

(S4) $K = 20$, $u \sim \mathcal{N}(0,1)$;

(S5) $K = 50$, $u \sim t_8$, where t_8 is the Student-t distribution with 8 degrees of freedom

To directly compare the performance of variance estimation, we let the ground truth mean function be given and we are only estimating the variance function. Further, instead of parameterizing the variance function through a neural network and optimizing for the weights, we directly provide a set of basis functions whose span include the underlying true variance function, so that we can solve for the weights through deterministic optimizers. This allows one to largely circumvent empirical instability in the training of neural networks and make the comparison more robust. The specific choice of choosing indicator functions as basis function is not only due to its flexibility, but also due to optimization consideration; specifically, we note that the objective function in (15) is convex in $\frac{1}{\sigma^2}$ and having indicator functions as basis enables one to directly parameterize $\frac{1}{\sigma^2(\cdot)}$ and keeping the optimization error small. The model is trained on samples of sizes 1e3, 3e3, 5e3, and 10e3 and applied to a test set of size 1e3.

Table 2 (see also Figures 3 and 4 for visualization) compare the RMSE between $\widehat{\sigma^2}_{\text{MSE}}(x)$ and $\widehat{\sigma^2}_{\text{NLL}}(x)$ in different value regions, as separated based on the quartiles of $(\sigma^2)^\star(x)$, i.e., 25%, 50%, 75%. Across both settings and all sample sizes considered, we again observe that $\widehat{\sigma^2}_{\text{NLL}}(x)$ outperforms $\widehat{\sigma^2}_{\text{MSE}}(x)$ in the low-value region (as manifested by a lower RMSE) , while vice versa in the high-value one. This is consistent with the discussion in Remark 2, that based on the bounds in Theorems 1 and 2, one expects $\widehat{\sigma^2}_{\text{NLL}}(x)$ to underperform in the case where $\sigma^\star(x)$ has a large value, due to the ratio form in the risk bound established.[2]

6 CONCLUSION & FUTURE WORKS

In this paper, we provide theoretical analysis for the estimation of variance based on MSE and NLL losses, respectively, and show that their respective PAC-type risk bounds have distinct behavior under different regimes that are dictated by the magnitude of uncertainty. A pertinent future research direction is how to devise estimation procedures that leverage the aforementioned results, where one could potentially combine two loss functions in an adaptive fashion, to provide robust and accurate uncertainty estimation across the entire domain. It is also worth investigating the connection between our approach and those in the nuisance parameters literature (e.g. Foster and Syrgkanis, 2019), since our analysis circumvents using sample splitting for estimating μ, which is required by the latter line of literature and poses some limitation (in this context μ can be regarded as a nuisance parameter).

[2]See `https://github.com/morganstanley/MSML/tree/main/papers/Risk_Bounds_Aleatoric_Uncertainty` to replicate the settings and runs adopted.

Acknowledgments

We thank anonymous reviewers for their helpful and constructive comments, which improve the quality of this work.

References

Abdar, M., F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, et al. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion 76*, 243–297.

Bartlett, P. L., O. Bousquet, and S. Mendelson (2005). Local Rademacher complexities. *The Annals of Statistics 33*(4), 1497–1537.

Bartlett, P. L. and S. Mendelson (2002). Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research 3*(Nov), 463–482.

Bengio, Y., O. Delalleau, and N. Roux (2005). The curse of highly variable functions for local kernel machines. *Advances in Neural Information Processing Systems 18*.

Beyer, H.-G. and B. Sendhoff (2007). Robust optimization–a comprehensive survey. *Computer methods in applied mechanics and engineering 196*(33-34), 3190–3218.

Blum, A., J. Hopcroft, and R. Kannan (2020). *Foundations of Data Science*. Cambridge University Press.

Bollerslev, T. and J. M. Wooldridge (1992). Quasi-maximum likelihood estimation and inference in dynamic models with time-varying covariances. *Econometric reviews 11*(2), 143–172.

Bousquet, O., S. Boucheron, and G. Lugosi (2003). Introduction to statistical learning theory. In *Summer school on machine learning*, pp. 169–207. Springer.

Carroll, R. J. and D. Ruppert (1982a). A comparison between maximum likelihood and generalized least squares in a heteroscedastic linear model. *Journal of the American Statistical Association 77*(380), 878–882.

Carroll, R. J. and D. Ruppert (1982b). Robust estimation in heteroscedastic linear models. *The Annals of Statistics*, 429–441.

Cawley, G. C., N. L. Talbot, and O. Chapelle (2005). Estimating predictive variances with kernel ridge regression. In *Machine Learning Challenges Workshop*, pp. 56–77. Springer.

Cawley, G. C., N. L. Talbot, R. J. Foxall, S. R. Dorling, and D. P. Mandic (2004). Heteroscedastic kernel ridge regression. *Neurocomputing 57*, 105–124.

Chua, K., R. Calandra, R. McAllister, and S. Levine (2018). Deep reinforcement learning in a handful of trials using probabilistic dynamics models. *Advances in Neural Information Processing Systems 31*.

Cook, R. D. and S. Weisberg (1983). Diagnostics for heteroscedasticity in regression. *Biometrika 70*(1), 1–10.

Daye, Z. J., J. Chen, and H. Li (2012). High-dimensional heteroscedastic regression with an application to eQTL data analysis. *Biometrics 68*(1), 316–326.

De Boor, C. (1978). *A practical guide to splines*, Volume 27. Springer Verlag, New York.

Der Kiureghian, A. and O. Ditlevsen (2009). Aleatory or epistemic? Does it matter? *Structural safety 31*(2), 105–112.

DeSalvo, G., M. Mohri, and U. Syed (2015). Learning with deep cascades. In *International Conference on Algorithmic Learning Theory*, pp. 254–269. Springer.

Dudley, R. M. (1967). The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *Journal of Functional Analysis 1*(3), 290–330.

Durrett, R. (2019). *Probability: theory and examples*, Volume 49. Cambridge university press.

Foster, D. J. and V. Syrgkanis (2019). Orthogonal statistical learning. *arXiv preprint arXiv:1901.09036*.

Glorot, X. and Y. Bengio (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256. JMLR Workshop and Conference Proceedings.

Goldberg, P., C. Williams, and C. Bishop (1997). Regression with input-dependent noise: A Gaussian process treatment. *Advances in Neural Information Processing Systems 10*.

Györfi, L., M. Kohler, A. Krzyzak, H. Walk, et al. (2002). *A distribution-free theory of nonparametric regression*, Volume 1. Springer.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the econometric society*, 1029–1054.

Haussler, D. (1992). Overview of the probably approximately correct (PAC) learning framework. *Information and Computation 100*(1), 78–150.

Haussler, D. (1995). Sphere packing numbers for subsets of the boolean n-cube with bounded Vapnik-Chervonenkis dimension. *Journal of Combinatorial Theory, Series A 69*(2), 217–232.

Hensman, J., N. Fusi, and N. D. Lawrence (2013). Gaussian processes for big data. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pp. 282–290.

Jobson, J. and W. Fuller (1980). Least squares estimation when the covariance matrix and parameter vector are functionally related. *Journal of the American Statistical Association 75*(369), 176–181.

Kääriäinen, M. (2004). Relating the Rademacher and VC bounds. Technical report, Citeseer.

Kahn, G., A. Villaflor, V. Pong, P. Abbeel, and S. Levine (2017). Uncertainty-aware reinforcement learning for collision avoidance. *arXiv preprint arXiv:1702.01182*.

Kearns, M. J. and U. Vazirani (1994). *An introduction to computational learning theory*. MIT press.

Kendall, A. and Y. Gal (2017). What uncertainties do we need in Bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems 30*.

Kersting, K., C. Plagemann, P. Pfaff, and W. Burgard (2007). Most likely heteroscedastic gaussian process regression. In *Proceedings of the 24th International Conference on Machine Learning*, pp. 393–400.

Kimeldorf, G. and G. Wahba (1971). Some results on Tchebycheffian spline functions. *Journal of mathematical analysis and applications 33*(1), 82–95.

Kimeldorf, G. S. and G. Wahba (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics 41*(2), 495–502.

Krinitzsky, E. L. (2002). Epistematic and aleatory uncertainty: a new shtick for probabilistic seismic hazard analysis. *Engineering Geology 66*(1-2), 157–159.

Lakshminarayanan, B., A. Pritzel, and C. Blundell (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems 30*.

Lázaro-Gredilla, M. and M. K. Titsias (2011). Variational heteroscedastic Gaussian process regression. In *ICML*.

Le, Q. V., A. J. Smola, and S. Canu (2005). Heteroscedastic Gaussian process regression. In *Proceedings of the 22nd International Conference on Machine Learning*, pp. 489–496.

Ledoux, M. and M. Talagrand (1991). *Probability in Banach Spaces: isoperimetry and processes*, Volume 23. Springer Science & Business Media.

Li, C.-L., W.-C. Chang, Y. Cheng, Y. Yang, and B. Póczos (2017). Mmd gan: Towards deeper understanding of moment matching network. *Advances in Neural Information Processing Systems 30*.

Li, Y., K. Swersky, and R. Zemel (2015). Generative moment matching networks. In *International Conference on Machine Learning*, pp. 1718–1727. PMLR.

Long, J. S. and L. H. Ervin (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician 54*(3), 217–224.

Massart, P. and É. Nédélec (2006). Risk bounds for statistical learning. *The Annals of Statistics 34*(5), 2326–2366.

Mohri, M. and A. M. Medina (2014). Learning theory and algorithms for revenue optimization in second price auctions with reserve. In *International Conference on Machine Learning*, pp. 262–270. PMLR.

Mohri, M., A. Rostamizadeh, and A. Talwalkar (2018). *Foundations of machine learning*. MIT press.

Mukhoti, J., A. Kirsch, J. van Amersfoort, P. H. Torr, and Y. Gal (2021). Deterministic neural networks with appropriate inductive biases capture epistemic and aleatoric uncertainty. *arXiv preprint arXiv:2102.11582*.

Muller, H.-G. and U. Stadtmuller (1987). Estimation of heteroscedasticity in regression analysis. *The Annals of Statistics*, 610–625.

Neverova, N., D. Novotny, and A. Vedaldi (2019). Correlated uncertainty for learning dense correspondences from noisy labels. *Advances in Neural Information Processing Systems 32*, 920–928.

Pollard, D. (1982). A central limit theorem for empirical processes. *Journal of the Australian Mathematical Society 33*(2), 235–248.

Rakhlin, A. (2020). *Mathematical Statistics: A Non-Asymptotic Approach*. Lecture Notes.

Ren, Y., J. Zhu, J. Li, and Y. Luo (2016). Conditional generative moment-matching networks. *Advances in Neural Information Processing Systems 29*.

Rumelhart, D. E., J. L. McClelland, P. R. Group, et al. (1988). *Parallel distributed processing*, Volume 1. IEEE New York.

Sauer, N. (1972). On the density of families of sets. *Journal of Combinatorial Theory, Series A 13*(1), 145–147.

Seitzer, M., B. Schölkopf, and G. Martius (2021). Causal influence detection for improving efficiency in reinforcement learning. *Advances in Neural Information Processing Systems 34*.

Seitzer, M., A. Tavakoli, D. Antic, and G. Martius (2022). On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks. In *ICLR*.

Shah, A., Y. Bu, J. K. Lee, S. Das, R. Panda, P. Sattigeri, and G. W. Wornell (2022). Selective regression under fairness criteria. In *International Conference on Machine Learning*, pp. 19598–19615. PMLR.

Shalev-Shwartz, S. and S. Ben-David (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Skafte, N., M. Jørgensen, and S. Hauberg (2019). Reliable training and estimation of variance networks. *Advances in Neural Information Processing Systems 32*.

Snelson, E. and Z. Ghahramani (2007). Local and global sparse Gaussian process approximations. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, Volume 2, pp. 524–531. PMLR.

Titsias, M. and N. D. Lawrence (2010). Bayesian Gaussian process latent variable model. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 844–851. JMLR Workshop and Conference Proceedings.

Van der Vaart, A. W. and J. Wellner (1996). *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media.

Vandal, T., E. Kodra, J. Dy, S. Ganguly, R. Nemani, and A. R. Ganguly (2018). Quantifying uncertainty in discrete-continuous and skewed data with Bayesian deep learning. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2377–2386. `https://arxiv.org/pdf/1802.04742.pdf`.

Vapnik, V. N. and A. Y. Chervonenkis (2015). On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of Complexity*, pp. 11–30. Springer.

Walters, D. J., G. Ülkümen, D. Tannenbaum, C. Erner, and C. R. Fox (2022). Investor behavior under epistemic versus aleatory uncertainty. *Available at SSRN 3695316*.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the econometric society*, 1–25.

Wooldridge, J. M. (2001). Applications of generalized method of moments estimation. *Journal of Economic perspectives 15*(4), 87–100.

Yao, W., X. Chen, W. Luo, M. Van Tooren, and J. Guo (2011). Review of uncertainty-based multidisciplinary design optimization methods for aerospace vehicles. *Progress in Aerospace Sciences 47*(6), 450–479.

Zabell, S. (1979). Continous versions of regular conditional distributions. *The Annals of Probability 7*(1), 159–165.

Zaoui, A., C. Denis, and M. Hebiri (2020). Regression with reject option and application to knn. *Advances in Neural Information Processing Systems 33*, 20073–20082.

A PRELIMINARIES

Notation. Vectors are denoted by bold-faced letters and $[\boldsymbol{x}]_i$ denotes the i-th entry of vector $\boldsymbol{x}$. $\boldsymbol{x}^T\boldsymbol{y}$ represents the inner product vectors. We use $\|\cdot\|_p$ to denote the ℓ_p norm of vectors. Let $[n] = \{1, 2, ..., n\}$ and $\boldsymbol{x}_{1:n} = \{\boldsymbol{x}_i\}_{i=1}^n$. Denote uppercase letter X to be random variable. Write $X \sim \mathcal{D}$ when X follows distribution $\mathcal{D}$ and let $X_{1:n} \sim \mathcal{D}^n$ denote n i.i.d. samples from $\mathcal{D}$. Let $\mathbb{E}_{X\sim\mathcal{D}}$ denote the expectation taken w.r.t. X under $\mathcal{D}$ or simply $\mathbb{E}_X$ when $\mathcal{D}$ is clear. We use $\overset{\mathrm{D}}{=}$ to denote two random variables are equal in distribution. Finally, we use short hand $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$.

Problem Setup. Without loss of generality, consider a regression task in feature space $\mathcal{X}$ (typically $\mathbb{R}^d$) with response Y taking value in $\mathcal{Y} \subseteq \mathbb{R}$. The joint underlying distribution of $Z \triangleq (X, Y)$ over $\mathcal{Z} \triangleq \mathcal{X} \times \mathcal{Y}$ is denoted by $\bar{\mathcal{D}}$ (i.e., $\boldsymbol{z}_{1:n} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ denote n i.i.d. samples drawn from $\bar{\mathcal{D}}$). In this paper, we use the following joint data generative process (DGP) to model the heteroscedasticity:

$$\begin{aligned} X \sim& \mathcal{D} \\ Y|_{X=\boldsymbol{x}} \overset{\mathrm{D}}{=}& \mu(\boldsymbol{x}) + \xi(\boldsymbol{x}), \end{aligned} \tag{16}$$

where the feature $X \sim \mathcal{D}$ follows some underlying distribution $\mathcal{D}$, and the response Y has feature-dependent mean and noise structure specified by μ and ξ. Here, $\mu(\boldsymbol{x})$ is the mean function that characterizes $\mathbb{E}[Y|X = \boldsymbol{x}]$, which we henceforth assume $-1 \leq \mu \leq 1$, w.o.l.g. More importantly, $\{\xi(\boldsymbol{x})\}_{\boldsymbol{x}\in\mathcal{X}}$ is a family of noise distribution on $\mathbb{R}$ indexed by $\mathcal{X}$. We allow this family of distribution to be general and we only specify its first two moments (and assume its existence). In particular, let $\mathscr{R}$ be the space of random variable distributed on $\mathbb{R}$ with finite variance, then the feature-dependent noise $\xi(\cdot) : \mathcal{X} \to \mathscr{R}$ satisfies $\mathbb{E}[\xi(\boldsymbol{x})] = 0$ and variance $\sigma^2(\boldsymbol{x}) \triangleq \mathrm{Var}(\xi(\boldsymbol{x}))$. As suggested in the addition sign in (16), we assume an additive noise structure where $\xi(\boldsymbol{x}_i)$ is conditional independent of $\xi(\boldsymbol{x}_j)$ given X_i, X_j. The DGP model (4) completely characterizes $Z \triangleq (X, Y) \sim \bar{\mathcal{D}}$, and allows for general form of heteroscedasticity on Y.

Next, we give definitions for the terms used in the ensuing technical development.

Definition 4. *Given $\boldsymbol{x}_{1:n} \subseteq \mathcal{X}$ and a hypothesis class $\mathcal{F} = \{f : \mathcal{X} \to [l, u]\}$. The empirical Rademacher complexity of the hypothesis class $\mathcal{F}$ is defined to be*

$$\widehat{\mathcal{R}}_n(\mathcal{F}; \boldsymbol{x}_{1:n}) \triangleq \mathbb{E}_{\sigma_{1:n}\sim\mathcal{D}_\sigma^n}[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i f(\boldsymbol{x}_i)], \tag{17}$$

where $\sigma_{1:n}$ follows from i.i.d. Rademacher distribution $\mathcal{D}_\sigma$, i.e., $\mathbb{P}(\sigma_i = 1) = \mathbb{P}(\sigma_i = -1) = \frac{1}{2}$. Furthermore, given $X \sim \mathcal{D}$, the Rademacher average of $\mathcal{F}$ is defined as

$$\mathcal{R}_n(\mathcal{F}) \triangleq \mathbb{E}_{X_{1:n}\sim\mathcal{D}^n}[\widehat{\mathcal{R}}_n(\mathcal{F}; X_{1:n})]. \tag{18}$$

In general, the Rademacher complexity is difficult to calculate (NP-hardness, see (Kääriäinen, 2004)). However, VC-dimension typically can provide a simple bound.

Definition 5 (VC-dimension)**.** *The VC-dimension $d_{VC}(\mathcal{F})$ of a hypothesis class $\mathcal{F} = \{f : \mathcal{X} \to \{1, -1\}\}$ is the largest cardinality of any set $S \subseteq \mathcal{X}$ such that $\forall \bar{S} \subseteq S$, $\exists f \in \mathcal{F}$:*

$$f(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \boldsymbol{x} \in \bar{S} \\ -1 & \text{if } \boldsymbol{x} \in S \setminus \bar{S} \end{cases} \tag{19}$$

Definition 6 (Pseudo-dimension)**.** *The Pseudo-dimension $d_P(\mathcal{F})$ of a real-valued hypothesis class $\mathcal{F} = \{f : \mathcal{X} \to [l, u]\}$ is the VC-dimension of the hypothesis class*

$$\mathcal{H} = \{h : \mathcal{X} \times \mathbb{R} \to \{-1, 1\} \mid h(\boldsymbol{x}, t) = \mathit{sign}(f(\boldsymbol{x}) - t), f \in \mathcal{F}, t \in \mathbb{R}\}.$$

B TECHNICAL LEMMAS

The following lemma follows from Dudley's chaining bound (Dudley, 1967) and concepts of "covering number" Pollard (1982); Rakhlin (2020) built upon Haussler's bound and combinatorial dimension Haussler (1995); Van der Vaart and Wellner

(1996), one can bound the Rademacher complexity with VC-dimension. We note an alternative to bound Rademacher complexity is using Sauer's lemma Sauer (1972) and Massart lemma Shalev-Shwartz and Ben-David (2014). Compared to the first method, these methods can achieve upper bounds of the same order, up to a logarithmic factor.

Lemma 2. *Let $\mathcal{F} = \{f : \mathcal{X} \to [-M, M]\}$ be a bounded (i.e. by $M > 0$) hypothesis class with finite Pseudo-dimension $d_P(\mathcal{F}) < \infty$. Let $\widehat{\mathcal{R}}_n(\mathcal{F}; \boldsymbol{x}_{1:n})$ be the empirical Rademacher complexity defined in* (17)*, then there exists some universal constant K_1 (depending only on M and $\mathcal{X}$) such that*

$$\widehat{\mathcal{R}}_n(\mathcal{F}; \boldsymbol{x}_{1:n}) \leq K_1 \sqrt{\frac{d_P(\mathcal{F})}{n}}$$

for any $\boldsymbol{x}_{1:n} \subseteq \mathcal{X}$. Consequently, we also have $\mathcal{R}_n(\mathcal{F}) \leq K_1 \sqrt{\frac{d_P(\mathcal{F})}{n}}$ for any distribution $X \sim \mathcal{D}$.

Lemma 3. *Let $\mathcal{G} = \{g : \mathcal{X} \to [m, M]\}$ be a hypothesis class of positive functions from $\mathcal{X}$ to $[m, M]$, $M \geq m > 0$. Then, if we define the class*

$$\mathcal{G}^{-1} \triangleq \left\{ h : \mathcal{X} \to [\frac{1}{M}, \frac{1}{m}], h = \frac{1}{g} \text{ for } g \in \mathcal{G} \right\},$$

we have:

$$\mathcal{R}_n(\mathcal{G}^{-1}) \leq \frac{1}{m^2} \mathcal{R}_n(\mathcal{G}) \tag{20}$$

Proof. For $\Phi(t) = \frac{1}{t}$, defined for $t \in [m, M]$, we first show that Φ is $\frac{1}{m^2}$-Lipschitz for $t, s \in [m, M]$:

$$\left| \frac{1}{t} - \frac{1}{s} \right| = \left| \frac{t - s}{ts} \right| \leq \frac{1}{m^2} |t - s| \tag{21}$$

Now if we define, for some $\Phi : \mathbb{R} \to \mathbb{R}$, the function class

$$\Phi \circ \mathcal{G} \triangleq \{h : h = \Phi(g) \text{ for } g \in \mathcal{G}\},$$

we clearly have

$$\mathcal{G}^{-1} = \Phi \circ \mathcal{G}$$

since $g \in [m, M]$. Thus, we can invoke standard Talagrand Contraction Lemma Ledoux and Talagrand (1991) to directly obtain the conclusion. See also Lemma 5.7 in Mohri et al. (2018) or Lemma 8 in Mohri and Medina (2014). □

Lemma 4. *Let $\mathcal{G} = \{g : \mathcal{X} \to [m, M]\}$ be a hypothesis class of positive functions from $\mathcal{X}$ to $[m, M]$, $M > m > 0$. Then we define the class*

$$\log \circ \mathcal{G} \triangleq \left\{ h : \mathcal{X} \to \mathbb{R}, h = \log(g) \text{ for } g \in \mathcal{G} \right\},$$

we can show:

$$\mathcal{R}_n(\log \circ \mathcal{G}) \leq \frac{1}{m} \mathcal{R}_n(\mathcal{G}). \tag{22}$$

Proof. In the domain $t, s \in [m, M]$, we have $\log$ is $\frac{1}{m}$-Lipschitz:

$$\left| \log(t) - \log(s) \right| \leq \frac{1}{m} |t - s|, \tag{23}$$

The rest follows in the same way as lemma 3, a proof could also be found in DeSalvo et al. (2015). □

Lemma 5. *Let $\mathcal{G} = \{g : \mathcal{X} \to [-M, M]\}$ be a hypothesis class of functions for $\mathcal{X}$ to $[-M, M]$. Then we define the class*

$$\mathcal{G}^2 \triangleq \left\{ h : \mathcal{X} \to \mathbb{R}, h = g^2 \text{ for } g \in \mathcal{G} \right\},$$

we can show:

$$\mathcal{R}_n(\mathcal{G}^2) \leq 2M \mathcal{R}_n(\mathcal{G}). \tag{24}$$

Proof. In the domain $t, s \in [-M, M]$, we have t^2 is $2M$-Lipschitz:

$$|t^2 - s^2| = |t - s||t + s| \leq 2M|t - s|.$$

The rest follows in the same way as lemma 3. □

Lemma 6. *(Rademacher Complexity for Class Sum) Let $\mathcal{F}$ and $\mathcal{G}$ be two classes of function mappings from $\mathcal{X}$ to $\mathbb{R}$ and define the classes*

$$\begin{aligned}\mathcal{G} + \mathcal{F} \triangleq &\{h : \mathcal{X} \to \mathbb{R}, h = f + g, f \in \mathcal{F}, g \in \mathcal{G}\}\\ \mathcal{G} - \mathcal{F} \triangleq &\{h : \mathcal{X} \to \mathbb{R}, h = f - g, f \in \mathcal{F}, g \in \mathcal{G}\}.\end{aligned}$$

Then, the Rademacher average defined in (18) *satisfies*

$$\begin{aligned}\mathcal{R}_n(\mathcal{F} + \mathcal{G}) \leq &\mathcal{R}_n(\mathcal{F}) + \mathcal{R}_n(\mathcal{G})\\ \mathcal{R}_n(\mathcal{F} - \mathcal{G}) \leq &\mathcal{R}_n(\mathcal{F}) + \mathcal{R}_n(\mathcal{G}).\end{aligned}$$

Proof. It follows from (18) that

$$\begin{aligned}\mathcal{R}_n(\mathcal{F} + \mathcal{G}) =&\mathbb{E}_{\sigma_{1:n}, X_{1:n}}[\sup_{h \in \mathcal{F}+\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i h(X_i)]\\ =&\mathbb{E}_{\sigma_{1:n}, X_{1:n}}[\sup_{f \in \mathcal{F}, g \in \mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i (f(X_i) + g(X_i))]\\ \leq&\mathbb{E}_{\sigma_{1:n}, X_{1:n}}[\sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(X_i) + \sup_{g \in \mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i g(X_i)]\\ =&\mathcal{R}_n(\mathcal{F}) + \mathcal{R}_n(\mathcal{G})\end{aligned}$$

Similarly,

$$\begin{aligned}\mathcal{R}_n(\mathcal{F} - \mathcal{G}) =&\mathbb{E}_{\sigma_{1:n}, X_{1:n}}[\sup_{f \in \mathcal{F}-\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i h(X_i)]\\ =&\mathbb{E}_{\sigma_{1:n}, X_{1:n}}[\sup_{f \in \mathcal{F}, g \in \mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i (f(X_i) - g(X_i))]\\ \leq&\mathbb{E}_{\sigma_{1:n}, X_{1:n}}[\sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(X_i) + \sup_{g \in \mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} -\sigma_i g(X_i)]\\ =&\mathcal{R}_n(\mathcal{F}) + \mathcal{R}_n(\mathcal{G})\end{aligned}$$

where the last equality follows from the fact the Rademacher distribution is symmetric, i.e., $-\sigma \sim \mathcal{D}_\sigma$ if $\sigma \sim \mathcal{D}_\sigma$. □

Lemma 7 (Rademacher Complexity for class product)**.** *Let $\mathcal{F}$ and $\mathcal{G}$ be two classes of function mappings from $\mathcal{X}$ to $\mathbb{R}$ and define the class*

$$\mathcal{F} \cdot \mathcal{G} \triangleq \{h : \mathcal{X} \to \mathbb{R}, h = f \cdot g, f \in \mathcal{F}, g \in \mathcal{G}\}.$$

Suppose $\mathcal{F}$ and $\mathcal{G}$ are bounded, in the sense that there exists a constant $|f(\cdot)| < B_1, g(\cdot) < B_2$. Then if we let $B = B_1 + B_2$, the Rademacher average defined in (18) *satisfies*

$$\mathcal{R}_n(\mathcal{F} \times \mathcal{G}) \leq 2B \cdot \Big(\mathcal{R}_n(\mathcal{F}) + R_n(\mathcal{G})\Big).$$

Proof. Define $\mathcal{F} + \mathcal{G}$ and $\mathcal{F} - \mathcal{G}$ as in lemma 6 and define $\phi(\cdot) : \mathbb{R} \to \mathbb{R}$ as $\phi(x) = \frac{(|x| \wedge 2B)^2}{4}$. It can be checked that ϕ is B-Lipschitz and

$$xy = \frac{(x + y)^2}{4} - \frac{(x - y)^2}{4} = \phi(x + y) - \phi(x - y)$$

for any $|x| \leq B, |y| \leq B$. Thus, it follows from 6 and Talagrand's contraction inequality that

$$\begin{aligned}\mathcal{R}_n(\mathcal{F}\cdot\mathcal{G}) \leq& \mathcal{R}_n(\phi\circ(\mathcal{F}+\mathcal{G})) + \mathcal{R}_n(\phi\circ(\mathcal{F}-\mathcal{G}))\\ \leq& B\mathcal{R}_n(\mathcal{F}+\mathcal{G}) + B\mathcal{R}_n(\mathcal{F}-\mathcal{G})\\ \leq& 2B\Big(\mathcal{R}_n(\mathcal{F}) + R_n(\mathcal{G})\Big),\end{aligned}$$

which concludes the proof. □

Lemma 8. *Let $\mathcal{F} = \{f : \mathcal{X} \to [-1,1]\}$ be a hypothesis class of functions from $\mathcal{X}$ to $[-1,1]$. Then we define the class*

$$L_S \circ \mathcal{F} \triangleq \Big\{h : \mathcal{Z} \to \mathbb{R}, h(\boldsymbol{z}) = (y - f(\boldsymbol{x}))^2, f \in \mathcal{F}\Big\},$$

Assuming $\mathcal{Y} \subseteq \{y : |y| \leq 1 + \sqrt{M}\}$ and recall $\mathcal{R}_n(L_S \circ \mathcal{F}) \triangleq \mathbb{E}_{Z_{1:n}\sim\bar{\mathcal{D}}^n}[\widehat{\mathcal{R}}_n(L_S \circ \mathcal{F}; Z_{1:n})]$, one can show:

$$\mathcal{R}_n(L_S \circ \mathcal{F}) \leq (4 + 2\sqrt{M})\mathcal{R}_n(\mathcal{F}) \tag{25}$$

Proof. Since $|y - f(\boldsymbol{x})| \leq 2 + \sqrt{M}$ for all $\boldsymbol{z}$, by Lemma 5 we have

$$\begin{aligned}\mathcal{R}_n(L_S \circ \mathcal{F}) \leq& (4 + 2\sqrt{M})\mathbb{E}_{Z_{1:n}}\left[\mathbb{E}_{\sigma_{1:n}}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i(Y_i - f(X_i))\right]\right]\\ =& (4 + 2\sqrt{M})\mathbb{E}_{X_{1:n}}\left[\mathbb{E}_{\sigma_{1:n}}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i f(X_i)\right]\right]\\ =& (4 + 2\sqrt{M})\mathcal{R}_n(\mathcal{F}).\end{aligned}$$

□

Lemma 9 (Rademacher Complexity for NLL loss)**.** *Let $\mathcal{F} = \{f : \mathcal{X} \to [-1,1]\}$ a be hypothesis class of real-valued functions from $\mathcal{X}$ to $[-1,1]$. Let $\mathcal{G} = \{g : \mathcal{X} \to [m, M]\}$ a be hypothesis class of positive functions from $\mathcal{X}$ to $[m, M]$. Define*

$$L_{NLL} \circ (\mathcal{F}\times\mathcal{G}) \triangleq \{\ell : \mathcal{Z} \to \mathbb{R}\,,\ \ell(\boldsymbol{z}) = \log(g(\boldsymbol{x})) + \frac{(y - f(\boldsymbol{x}))^2}{g(\boldsymbol{x})}, f \in \mathcal{F}, g \in \mathcal{G}\}.$$

Assume $\mathcal{Y} \subseteq \{y : |y| \leq 1 + \sqrt{M}\}$ and recall $\mathcal{R}_n(L_{NLL} \circ \mathcal{F}) \triangleq \mathbb{E}_{Z_{1:n}\sim\bar{\mathcal{D}}^n}[\widehat{\mathcal{R}}_n(L_{NLL} \circ \mathcal{F}; Z_{1:n})]$, one can show:

$$\mathcal{R}_n(L_{NLL} \circ (\mathcal{F}\times\mathcal{G})) \leq (\frac{1}{m} + \frac{8 + 8\sqrt{M} + 2M}{m^2} + \frac{6}{m^3})\mathcal{R}_n(\mathcal{G}) + 4(2 + \frac{2}{m} + \sqrt{M})\mathcal{R}_n(\mathcal{F}). \tag{26}$$

Proof. Note we have that

$$\begin{aligned}&\mathcal{R}_n(L_{NLL} \circ (\mathcal{F}\times\mathcal{G}))\\ =&\mathbb{E}_{Z_{1:n}}\left[\mathbb{E}_{\sigma_{1:n}}\left[\sup_{f,g} \frac{1}{n}\sum_{i=1}^n \sigma_i\Big(\log(g(X_i)) + \frac{(Y_i - f(X_i))^2}{g(X_i)}\Big)\right]\right]\\ =&\mathbb{E}_{Z_{1:n}}\left[\mathbb{E}_{\sigma_{1:n}}\left[\sup_{f,g} \frac{1}{n}\sum_{i=1}^n \sigma_i\Big(\log(g(X_i)) + \frac{Y_i^2}{g(X_i)} - \frac{2Y_i f(X_i)}{g(X_i)} + \frac{f^2(X_i)}{g(X_i)}\Big)\right]\right]\\ \leq& \underbrace{\mathbb{E}_{X_{1:n}}\mathbb{E}_{\sigma_{1:n}}\left[\sup_{g} \frac{1}{n}\sum_{i=1}^n \sigma_i \log(g(X_i))\right]}_{\leq \frac{1}{m}\mathcal{R}_n(\mathcal{G}),\text{ by Lemma 4}} + \underbrace{\mathbb{E}_{Z_{1:n}}\mathbb{E}_{\sigma_{1:n}}\left[\sup_{g} \frac{1}{n}\sum_{i=1}^n \sigma_i \frac{Y_i^2}{g(X_i)}\right]}_{\leq \frac{(2+2M+4\sqrt{M})}{m^2}\mathcal{R}_n(\mathcal{G}),\text{ by Lemma 3 and Lemma 7}}\\ &+ \underbrace{\mathbb{E}_{Z_{1:n}}\mathbb{E}_{\sigma_{1:n}}\left[\sup_{f,g} \sum_{i=1}^n \sigma_i \frac{2Y_i f(X_i)}{g(X_i)}\right]}_{\leq 4(1+\sqrt{M}+\frac{1}{m})(\frac{1}{m^2}\mathcal{R}_n(\mathcal{G})+\mathcal{R}_n(\mathcal{F})),\text{ by Lemma 3 and Lemma 7}} + \underbrace{\mathbb{E}_{X_{1:n}}\mathbb{E}_{\sigma_{1:n}}\left[\sup_{f,g} \sum_{i=1}^n \sigma_i \frac{f^2(X_i)}{g(X_i)}\right]}_{2(1+\frac{1}{m})(\frac{1}{m^2}\mathcal{R}_n(\mathcal{G})+2\mathcal{R}_n(\mathcal{F})),\text{ by Lemma 3, Lemma 5 and Lemma 7}}\\ \leq& (\frac{1}{m} + \frac{8 + 8\sqrt{M} + 2M}{m^2} + \frac{6}{m^3})\mathcal{R}_n(\mathcal{G}) + 4(2 + \frac{2}{m} + \sqrt{M})\mathcal{R}_n(\mathcal{F})\end{aligned}$$

□

Lemma 10 (Rademacher Complexity for MSE loss)**.** *Let $\mathcal{F} = \{f : \mathcal{X} \to [-1, 1]\}$ a be hypothesis class of real-valued functions from $\mathcal{X}$ to $[-1, 1]$. Let $\mathcal{G} = \{g : \mathcal{X} \to [m, M]\}$ a be hypothesis class of positive functions from $\mathcal{X}$ to $[m, M]$. Define*

$$L_{MSE} \circ (\mathcal{F} \times \mathcal{G}) \triangleq \{\ell : \mathcal{Z} \to \mathbb{R}\,,\ \ell(\boldsymbol{z}) = \big(g(\boldsymbol{x}) - (y - f(\boldsymbol{x}))^2\big)^2, f \in \mathcal{F}, g \in \mathcal{G}\}.$$

Assume $\mathcal{Y} \subseteq \{y : |y| \leq 1 + \sqrt{M}\}$ and recall $\mathcal{R}_n(L_{MSE} \circ \mathcal{F}) \triangleq \mathbb{E}_{Z_{1:n} \sim \bar{\mathcal{D}}^n}[\widehat{\mathcal{R}}_n(L_{MSE} \circ \mathcal{F}; Z_{1:n})]$, one can show

$$\mathcal{R}_n(\ell_{MSE} \circ (\mathcal{F} \times \mathcal{G})) \leq (16 + 16\sqrt{M} + 6M)\mathcal{R}_n(\mathcal{G}) + 8(8 + 12\sqrt{M} + 6M + M^{\frac{3}{2}})\mathcal{R}_n(\mathcal{F}) \tag{27}$$

Proof. Since $|y| \leq 1 + \sqrt{M}$, we have $|y - f(\boldsymbol{x})|^2 \leq 4 + 4\sqrt{M} + M$. Then we have that

$$\begin{aligned}
&\mathcal{R}_n(\ell_{MSE} \circ (\mathcal{F} \times \mathcal{G})) \\
=&\mathbb{E}_{Z_{1:n}}\mathbb{E}_{\sigma_{1:n}}\left[\sup_{f,g} \frac{1}{n}\sum_{i=1}^n \sigma_i \Big(g(X_i) - \big(Y_i - f(X_i)\big)^2\Big)^2\right] \\
=&\mathbb{E}_{Z_{1:n}}\mathbb{E}_{\sigma_{1:n}}\left[\sup_{f,g} \frac{1}{n}\sum_{i=1}^n \sigma_i \Big(g^2(X_i) - 2g(X_i)\big(Y_i - f(X_i)\big)^2 + \big(Y_i - f(X_i)\big)^4\Big)\right] \\
\leq& \underbrace{\mathbb{E}_{X_{1:n}}\mathbb{E}_{\sigma_{1:n}}\left[\sup_{g} \frac{1}{n}\sum_{i=1}^n \sigma_i g^2(\boldsymbol{x}_i)\right]}_{\leq 2M\mathcal{R}_n(\mathcal{G}), \text{ by Lemma 5}} + \underbrace{\mathbb{E}_{Z_{1:n}}\mathbb{E}_{\sigma_{1:n}}\left[\sup_{f,g} \sum_{i=1}^n \sigma_i 2g(X_i)\big(Y_i - f(X_i)\big)^2\right]}_{\leq 4(4+4\sqrt{M}+M)(\mathcal{R}_n(\mathcal{G})+(2+\sqrt{M})\mathcal{R}_n(\mathcal{F})), \text{ by Lemma 7 and Lemma 8}} \\
&+ \underbrace{\mathbb{E}_{Z_{1:n}}\mathbb{E}_{\sigma_{1:n}}\left[\sup_{f} \sum_{i=1}^n \sigma_i (Y_i - f(X_i))^4\right]}_{4(8+12\sqrt{M}+6M+M^{\frac{3}{2}})\mathcal{R}_n(\mathcal{F}), \text{ by Lemma 5 and Lemma 8}} \\
\leq&(16 + 16\sqrt{M} + 6M)\mathcal{R}_n(\mathcal{G}) + 8(8 + 12\sqrt{M} + 6M + M^{\frac{3}{2}})\mathcal{R}_n(\mathcal{F})
\end{aligned} \tag{28}$$

□

C MISSING PROOF FOR MAIN RESULTS

Assumption 2. *We assume the noise $\xi(\boldsymbol{x})$ in DGP satisfies $|\xi| \leq \sqrt{M}$, almost surely. Moreover, let $\mathcal{F}_\mu = \{f : \mathcal{X} \to [-1, 1]\}$ and $\mathcal{F}_{\sigma^2} = \{\sigma^2 : \mathcal{X} \to [m, M]\}$ for some $0 < m < M$ be two hypothesis classes with finite Pseucdo-dimension $d_P(\mathcal{F}) < \infty$, $d_P(\mathcal{G}) < \infty$. We assume that the DGP further satisfies:*

$$\mu^\star \in \mathcal{F}_\mu, \text{ and } (\sigma^2)^\star \in \mathcal{F}_{\sigma^2}$$

where $\mu^\star(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{x}]$ is the mean function and $(\sigma^2)^\star(\boldsymbol{x}) = Var[Y|\boldsymbol{x}]$ is the feature-dependent variance.

Proposition 3. *Under Assumption 2, given n i.i.d. samples $Z_{1:n}$ from the DGP, we have:*

$$\begin{aligned}
&\sup_{\mu \in \mathcal{F}_\mu, \sigma^2 \in \mathcal{F}_{\sigma^2}} \left|\mathbb{E}_Z\left[\ell_{NLL}(\sigma^2, \mu, Z)\right] - \frac{1}{n}\sum_{i=1}^n \ell_{NLL}(\sigma^2, \mu, Z_i)\right| \\
\leq& 2\Big(\frac{1}{m} + \frac{8 + 8\sqrt{M} + 2M}{m^2} + \frac{6}{m^3}\Big)\mathcal{R}_n(\mathcal{G}) + 8\Big(2 + \frac{2}{m} + \sqrt{M}\Big)\mathcal{R}_n(\mathcal{F}) + \frac{3 + 8M}{m} \cdot \sqrt{\frac{1}{2n}\log\frac{2}{\delta}}
\end{aligned} \tag{29}$$

with probability at least $1 - \delta$.

Proof. First, let $Z_{1:n}^{(-1)}$ be any sample that differs from $Z_{1:n}$ by exactly one point, e.g., $Z_1^{(-1)} \neq Z_1$ but $Z_i = Z_i^{(-1)}$ for

$2 \leq i \leq n$. First, note we have

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n}\ell_{NLL}(\sigma^2,\mu,Z_i)-\frac{1}{n}\sum_{i=1}^{n}\ell_{NLL}(\sigma^2,\mu,Z_i')\right| \qquad (30)\\
\leq&\frac{1}{n}\left|\log\left(\frac{\sigma^2(X_1)}{\sigma^2(X_1^{(-1)})}\right)+\frac{(Y_1-\mu(X_1))^2}{\sigma^2(X_1)}-\frac{(Y_1^{(-1)}-\mu(X_1^{(-1)}))^2}{\sigma^2(X_1^{(-1)})}\right|\\
\leq&\frac{1}{n}\left(\left|\log\left(\frac{\sigma^2(X_1)}{\sigma^2(X_1^{(-1)})}\right)\right|+\left|\frac{(Y_1-\mu(X_1))^2}{\sigma^2(X_1)}-\frac{(Y_1^{(-1)}-\mu(X_1^{(-1)})^2)}{\sigma^2(X_1^{(-1)})}\right|\right)\\
\leq&\frac{1}{n}\left(\left|\log\left(\frac{\sigma^2(X_1)}{\sigma^2(X_1^{(-1)})}\right)\right|+\max_{Z\in\{Z_1,Z_1^{(-1)}\}}\frac{(y-\mu(X))^2}{\sigma^2(X)}\right)\\
\leq&\frac{1}{n}\Big(\log\frac{M}{m}+\frac{8+2M}{m}\Big)\leq\frac{8+3M}{nm}. \qquad (31)
\end{aligned}$$

where the last line follows from the assumption $|\xi| \leq \sqrt{M}$, $m \leq \sigma^2 \leq M$ and elementary inequality $(a+b)^2 \leq 2a^2+2b^2$. On the other hand, define the random variable

$$\Delta^{\mathcal{F}_{\sigma^2},\mathcal{F}_\mu}(Z_{1:n}) \triangleq \sup_{\sigma^2\in\mathcal{F}_{\sigma^2},\mu\in\mathcal{F}_\mu}\left(\mathbb{E}_Z[\ell_{NLL}(\sigma^2,\mu,Z)]-\frac{1}{n}\sum_{i=1}^{N}\ell_{NLL}(\sigma^2,\mu,Z_i)\right), \qquad (32)$$

we must have

$$\mathbb{E}_Z[\ell_{NLL}(\sigma^2,\mu,Z)]-\frac{1}{n}\sum_{i=1}^{N}\ell_{NLL}(\sigma^2,\mu,Z_i)\leq\Delta^{\mathcal{F}_{\sigma^2},\mathcal{F}_\mu}(Z_{1:n}).$$

To analyze the quantity $\Delta^{\mathcal{F}_{\sigma^2},\mathcal{F}_\mu}(Z_{1:n})$, one can check that

$$\begin{aligned}
\Delta^{\mathcal{F}_{\sigma^2},\mathcal{F}_\mu}(Z_{1:n})-\Delta^{\mathcal{F}_{\sigma^2},\mathcal{F}_\mu}(Z_{1:n}^{(-1)})\leq&\sup_{\sigma^2\in\mathcal{F}_{\sigma^2},\mu\in\mathcal{F}_\mu}\left(\frac{1}{n}\sum_{i=1}^{N}\ell_{NLL}(\sigma^2,\mu,Z_i^{(-1)})-\frac{1}{n}\sum_{i=1}^{N}\ell_{NLL}(\sigma^2,\mu,Z_i)\right)\\
=&\frac{1}{n}\sup_{\sigma^2\in\mathcal{F}_{\sigma^2},\mu\in\mathcal{F}_\mu}\left(\ell_{NLL}(\sigma^2,\mu,Z_1^{(-1)})-\ell_{NLL}(\sigma^2,\mu,Z_1))\right)\\
\leq&\frac{3+8M}{nm},
\end{aligned}$$

where the last inequality follows from (30). Similarly, we can show

$$\Delta^{\mathcal{F}_{\sigma^2},\mathcal{F}_\mu}(Z_{1:n}^{(-1)})-\Delta^{\mathcal{F}_{\sigma^2},\mathcal{F}_\mu}(Z_{1:n})\leq\frac{3+8M}{nm}$$

which gives

$$|\Delta^{\mathcal{F}_{\sigma^2},\mathcal{F}_\mu}(Z_{1:n}^{(-1)})-\Delta^{\mathcal{F}_{\sigma^2},\mathcal{F}_\mu}(Z_{1:n})|\leq\frac{3+8M}{nm}. \qquad (33)$$

Thus, using McDiarmid's inequality, we can show that

$$\Delta^{\mathcal{F}_{\sigma^2},\mathcal{F}_\mu}(Z_{1:n})\leq\mathbb{E}_{Z_{1:n}}[\Delta^{\mathcal{F}_{\sigma^2},\mathcal{F}_\mu}(Z_{1:n})]+\frac{3+8M}{m}\cdot\sqrt{\frac{1}{2n}\log\frac{2}{\delta}} \qquad (34)$$

with probability at least $1-\frac{\delta}{2}$. Now, to bound $\mathbb{E}[\Delta^{\mathcal{F}_{\sigma^2},\mathcal{F}_\mu}(Z_{1:n})]$, let $Z_{1:n}$ and $Z'_{1:n}$ be two i.i.d. samples of size n, we have

$$
\begin{aligned}
&\mathbb{E}[\Delta^{\mathcal{F}_{\sigma^2},\mathcal{F}_\mu}(Z_{1:n})]\\
=&\mathbb{E}_{Z_{1:n}}\left[\sup_{\sigma^2\in\mathcal{F}_{\sigma^2},\mu\in\mathcal{F}_\mu}\left(\mathbb{E}_Z[\ell_{NLL}(\sigma^2,\mu,Z)]-\frac{1}{n}\sum_{i=1}^N\ell_{NLL}(\sigma^2,\mu,Z_i)\right)\right]\\
=&\mathbb{E}_{Z_{1:n}}\left[\sup_{\sigma^2\in\mathcal{F}_{\sigma^2},\mu\in\mathcal{F}_\mu}\left(\mathbb{E}_{Z'_{1:n}}\left[\frac{1}{n}\sum_{i=1}^N\ell_{NLL}(\sigma^2,\mu,Z'_i)\right]-\frac{1}{n}\sum_{i=1}^N\ell_{NLL}(\sigma^2,\mu,Z_i)\right)\right]\\
\le&\mathbb{E}_{Z_{1:n},Z'_{1:n}}\left[\sup_{\sigma^2\in\mathcal{F}_{\sigma^2},\mu\in\mathcal{F}_\mu}\frac{1}{n}\sum_{i=1}^n\Big(\ell_{NLL}(\sigma^2,\mu,Z_i)-\ell_{NLL}(\sigma^2,\mu,Z'_i)\Big)\right]\\
=&\mathbb{E}_{Z_{1:n},Z'_{1:n},\sigma_{1:n}}\left[\sup_{\sigma^2\in\mathcal{F}_{\sigma^2},\mu\in\mathcal{F}_\mu}\frac{1}{n}\sum_{i=1}^n\sigma_i\Big(\ell_{NLL}(\sigma^2,\mu,Z_i)-\ell_{NLL}(\sigma^2,\mu,Z'_i)\Big)\right]\\
\le&\mathbb{E}_{Z_{1:n},\sigma_{1:n}}\left[\sup_{\sigma^2\in\mathcal{F}_{\sigma^2},\mu\in\mathcal{F}_\mu}\frac{1}{n}\sum_{i=1}^n\sigma_i\ell_{NLL}(\sigma^2,\mu,Z_i)\right]+\mathbb{E}_{Z'_{1:n},\sigma_{1:n}}\left[\sup_{\sigma^2\in\mathcal{F}_{\sigma^2},\mu\in\mathcal{F}_\mu}\frac{1}{n}\sum_{i=1}^n-\sigma_i\ell_{NLL}(\sigma^2,\mu,Z'_i)\right]\\
=&2\mathbb{E}_{Z_{1:n},\sigma_{1:n}}\left[\sup_{\sigma^2\in\mathcal{F}_{\sigma^2},\mu\in\mathcal{F}_\mu}\frac{1}{n}\sum_{i=1}^n\sigma_i\ell_{NLL}(\sigma^2,\mu,Z_i)\right]\\
\le&2(\frac{1}{m}+\frac{8+8\sqrt{M}+2M}{m^2}+\frac{6}{m^3})\mathcal{R}_n(\mathcal{G})+8(2+\frac{2}{m}+\sqrt{M})\mathcal{R}_n(\mathcal{F}),
\end{aligned}
\tag{35}
$$

where σ is the Radamacher random variable and the last line follows from Lemma 9. The other direction of equation 29 can be proved similarly, which concludes the proof. □

Proposition 4. *Under Assumption 2, given n i.i.d. samples $Z_{1:n}$ from the DGP, we have:*

$$
\begin{aligned}
&\sup_{\mu\in\mathcal{F}_\mu,\sigma^2\in\mathcal{F}_{\sigma^2}}\left|\mathbb{E}_Z\left[\ell_{MSE}(\sigma^2,\mu,Z)\right]-\frac{1}{n}\sum_{i=1}^n\ell_{MSE}(\sigma^2,\mu,Z_i)\right|\\
\le&2(16+16\sqrt{M}+6M)\mathcal{R}_n(\mathcal{G})+16(8+12\sqrt{M}+6M+M^{\frac{3}{2}})\mathcal{R}_n(\mathcal{F})+4(M^2+8M+16)\cdot\sqrt{\frac{1}{2n}\log\frac{2}{\delta}}
\end{aligned}
\tag{36}
$$

with probability at least $1-\delta$.

Proof. The proof is similar to the proof in Proposition 3. First, let $Z_{1:n}^{(-1)}$ be any sample that differs from $Z_{1:n}$ by exactly one point, e.g., $Z_1^{(-1)}\neq Z_1$ but $Z_i=Z_i^{(-1)}$ for $2\le i\le n$. First, note we have

$$
\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^n\ell_{MSE}(\sigma^2,\mu,Z_i)-\frac{1}{n}\sum_{i=1}^n\ell_{MSE}(\sigma^2,\mu,Z_i^{(-1)})\right|\\
=&\frac{1}{n}\left|\left(\sigma^2(X_1)-(Y-\mu(X_1))^2\right)^2-\left(\sigma^2(X_1^{(-1)})-(Y^{(-1)}-\mu(X_1^{(-1)}))^2\right)^2\right|\\
\le&\frac{1}{n}\sup_{Z\in\mathcal{Z}}\left(\sigma^2(X)-(Y-\mu(X))^2\right)^2\le\frac{4(M^2+8M+16)}{n},
\end{aligned}
\tag{37}
$$

where the last line follows from $\sigma^2\le M,(Y-f(X))^2\le 8+2M$. Then, similarly as the proof in Proposition 3, we can define:

$$
\Delta^{\mathcal{F}_{\sigma^2},\mathcal{F}_\mu}(Z_{1:n})\triangleq\sup_{\sigma^2\in\mathcal{F}_{\sigma^2},\mathcal{F}_\mu}\left(\mathbb{E}_Z[\ell_{MSE}(\sigma^2,\mu,Z)]-\frac{1}{n}\sum_{i=1}^n\ell_{MSE}(\sigma^2,\mu,Z_i)\right)
\tag{38}
$$

and use McDiarmid's inequality to show

$$
\Delta^{\mathcal{F}_{\sigma^2},\mathcal{F}_\mu}(Z_{1:n})\le\mathbb{E}_{Z_{1:n}}[\Delta^{\mathcal{F}_{\sigma^2},\mathcal{F}_\mu}(Z_{1:n})]+4(M^2+8M+16)\cdot\sqrt{\frac{1}{2n}\log\frac{2}{\delta}}
\tag{39}
$$

with probability at least $1-\frac{\delta}{2}$. Then, using the argument similar to Equation 35, one can show

$$\begin{aligned}\mathbb{E}_Z\left[\Delta^{\mathcal{F}_{\sigma^2},\mathcal{F}_\mu}(Z_{1:n})\right] \leq & 2\mathbb{E}_{Z_{1:n},\sigma_{1:n}}\left[\sup_{\sigma^2\in\mathcal{F}_{\sigma^2},\mu\in\mathcal{F}_\mu}\frac{1}{n}\sum_{i=1}^n \sigma_i \ell_{MSE}(\sigma^2,\mu,Z_i)\right] \\ \leq & 2(16+16\sqrt{M}+6M)\mathcal{R}_n(\mathcal{G})+16(8+12\sqrt{M}+6M+M^{\frac{3}{2}})\mathcal{R}_n(\mathcal{F})\end{aligned}$$

where the last inequality follows from Lemma 10. The other direction of equation 36 can be proved in a similar way, which concludes the proof.

The other direction of equation 36 can be proved in a similar way. □

Theorem 3. *Suppose Assumption 2 holds, given $\widehat{\mu}\in\mathcal{F}$ satisfied that*

$$\mathbb{E}_X\left[(\mu^*(X)-\widehat{\mu}(X))^2\right]\leq\varepsilon,$$

and let

$$\widehat{\sigma^2}=\operatorname*{argmin}_{\sigma^2\in\mathcal{F}_{\sigma^2}}\frac{1}{n}\sum_{i=1}^n \ell_{NLL}(\sigma^2;\widehat{\mu},Z_i).$$

Then for $\varepsilon<\frac{1}{2}$, as long as n is large enough

$$n=O\bigg(\frac{1}{\min\{m^6,m^4\}\varepsilon^2}\bigg((\frac{1}{m^2}+\frac{1}{m^6}+\frac{M^2}{m^4})d_P(\mathcal{F}_{\sigma^2})+(1+\frac{1}{m^2}+M)d_P(\mathcal{F}_\mu)+\frac{1+M^2}{m^2}\log(\frac{2}{\delta})\bigg)\bigg),$$

we have

$$\mathbb{E}_X\left[\left|\frac{1}{\widehat{\sigma^2}(X)}-\frac{1}{(\sigma^2)^\star(X)}\right|^2\right]\leq\varepsilon \tag{40}$$

with probability at least $1-\delta$.

Proof. By Proposition 3 and Lemma 2, we have with probability at least $1-\delta$, that uniformly for all $\sigma^2\in\mathcal{F}_{\sigma^2},\mu\in\mathcal{F}_\mu$,

$$\begin{aligned}&\left|\mathbb{E}_Z\left[\ell_{NLL}(\sigma^2,\mu,Z)\right]-\frac{1}{n}\sum_{i=1}^n\ell_{NLL}(\sigma^2,\mu,Z_i)\right| \\ \leq & 2K_1(\frac{1}{m}+\frac{8+8\sqrt{M}+2M}{m^2}+\frac{6}{m^3})\sqrt{\frac{d_P(\mathcal{F}_{\sigma^2})}{n}}+8K_1(2+\frac{2}{m}+\sqrt{M})\sqrt{\frac{d_P(\mathcal{F}_\mu)}{n}}+\frac{3+8M}{m}\cdot\sqrt{\frac{1}{2n}\log\frac{2}{\delta}}\end{aligned}$$

Then for

$$n\geq\frac{1}{\varepsilon^2}\max\{(36K_1^2(\frac{1}{m}+\frac{8+8\sqrt{M}+2M}{m^2}+\frac{6}{m^3})^2 d_P(\mathcal{F}_{\sigma^2}),24^2K_1^2(2+\frac{2}{m}+\sqrt{M})^2 d_P(\mathcal{F}_\mu),\frac{9(3+8M)^2}{2m^2}\log(\frac{2}{\delta})\},$$

we have with probability at least $1-\delta$:

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n\ell_{NLL}((\sigma^2)^\star,\widehat{\mu},Z_i)\leq\mathbb{E}_Z\left[\ell_{NLL}((\sigma^2)^\star,\widehat{\mu},Z)\right]+\varepsilon \\ &\mathbb{E}_Z\left[\ell_{NLL}(\widehat{\sigma^2},\widehat{\mu},Z)\right]\leq\frac{1}{n}\sum_{i=1}^n\ell_{NLL}(\widehat{\sigma^2},\widehat{\mu},Z_i)+\varepsilon\end{aligned} \tag{41}$$

Using fact that $\widehat{\sigma^2}(\boldsymbol{x})$ is the empirical risk minimizer, we have:

$$\frac{1}{n}\sum_{i=1}^n\ell_{NLL}(\widehat{\sigma^2};\widehat{\mu},Z_i)\leq\frac{1}{n}\sum_{i=1}^n\ell_{NLL}((\sigma^2)^\star;\widehat{\mu},Z_i) \tag{42}$$

Combining Equation 41 and Equation 42, with probability $1-\delta$:

$$\begin{aligned}
&\mathbb{E}_Z\left[\ell_{NLL}(\widehat{\sigma^2};\widehat{\mu},Z)\right] \le \mathbb{E}_Z\left[\ell_{NLL}((\sigma^2)^\star;\widehat{\mu},Z)\right]+2\varepsilon \\
\implies &\mathbb{E}_{X,Y}\left[\frac{(Y-\widehat{\mu}(X))^2}{\widehat{\sigma^2}(X)}+\log(\widehat{\sigma^2}(X))\right] \le \mathbb{E}_{X,Y}\left[\frac{(Y-\widehat{\mu}(X))^2}{(\sigma^2)^\star(X)}+\log((\sigma^2)^\star(X))\right]+2\varepsilon \\
\implies &\mathbb{E}_X\left[\frac{(\sigma^2)^\star(X)}{\widehat{\sigma^2}(X)}+\frac{(\mu^*(X)-\widehat{\mu}(X))^2}{\widehat{\sigma^2}(X)}+\log(\widehat{\sigma}^2(X))\right] \le \mathbb{E}_X\left[1+\frac{(\mu^*(X)-\widehat{\mu}(X))^2}{(\sigma^2)^\star(X)}+\log((\sigma^2)^\star(X))\right]+2\varepsilon \\
\implies &\mathbb{E}_X\left[\frac{(\sigma^2)^\star(X)}{\widehat{\sigma^2}(X)}+\log(\widehat{\sigma}^2(X))\right] \le \mathbb{E}_X\left[1+\log((\sigma^2)^\star(X))\right]+2\varepsilon+\frac{\varepsilon}{m} \\
\implies &\mathbb{E}_X\left[\frac{(\sigma^2)^\star(X)}{\widehat{\sigma}^2(X)}-1+\log\left(\frac{\widehat{\sigma^2}(X)}{(\sigma^2)^\star(X)}\right)\right] \le 2\varepsilon+\frac{\varepsilon}{m}
\end{aligned} \tag{43}$$

Now, it can be checked that

$$\log(\frac{1}{t}) \ge 1-t+\frac{(\min(t,t^{-1})-1)^2}{2}, \tag{44}$$

for all $t>0$. Then, (44) and (43) together implies that:

$$\begin{aligned}
&\frac{1}{2}\mathbb{E}_X\left[\left(\min\{\frac{\widehat{\sigma^2}(X)}{(\sigma^2)^\star(X)},\frac{(\sigma^2)^\star(X)}{\widehat{\sigma^2}(X)}\}-1\right)^2\right] \le 2\varepsilon+\frac{\varepsilon}{m} \\
\implies &\mathbb{E}_X\left[\mathbf{1}\left\{\frac{(\sigma^2)^\star(X)}{\widehat{\sigma^2}(X)}\ge 1\right\}\left|\frac{\widehat{\sigma^2}(X)}{(\sigma^2)^\star(X)}-1\right|^2\right]+\mathbb{E}_X\left[\mathbf{1}\left\{\frac{(\sigma^2)^\star(X)}{\widehat{\sigma^2}(X)}<1\right\}\left|\frac{(\sigma^2)^\star(X)}{\widehat{\sigma^2}(X)}-1\right|^2\right] \le 4\varepsilon+\frac{2\varepsilon}{m}.
\end{aligned}$$

Finally, dividing m^2 by both side we have

$$\begin{aligned}
&\mathbb{E}_X\left[\mathbf{1}\left\{\frac{(\sigma^2)^\star(X)}{\widehat{\sigma^2}(X)}\ge 1\right\}\left|\frac{1}{(\sigma^2)^\star(X)}-\frac{1}{\widehat{\sigma^2}(X)}\right|^2\frac{(\widehat{\sigma^2})^2(X)}{m^2}\right] \\
+&\mathbb{E}_X\left[\mathbf{1}\left\{\frac{(\sigma^2)^\star(X)}{\widehat{\sigma^2}(X)}<1\right\}\left|\frac{1}{(\sigma^2)^\star(X)}-\frac{1}{\widehat{\sigma^2}(X)}\right|^2\frac{((\sigma^2)^\star(X))^2}{m^2}\right] \\
\le&\frac{4\varepsilon}{m^2}+\frac{2\varepsilon}{m^3}.
\end{aligned}$$

which, since $(\sigma^2)\ge m$, implies

$$\mathbb{E}_X\left[\left|\frac{1}{(\sigma^2)^\star(X)}-\frac{1}{\widehat{\sigma^2}(X)}\right|^2\right] \le \frac{4\varepsilon}{m^2}+\frac{2\varepsilon}{m^3},$$

with probability $1-\delta$. □

Theorem 4. *Suppose Assumption 2 holds, given $\widehat{\mu}\in\mathcal{F}$ satisfied that*

$$\mathbb{E}_X\left[(\mu^*(X)-\widehat{\mu}(X))^2\right] \le \varepsilon,$$

and let

$$\widehat{\sigma^2}=\operatorname*{argmin}_{\sigma^2\in\mathcal{F}_{\sigma^2}}\frac{1}{n}\sum_{i=1}^{n}\ell_{MSE}(\sigma^2;\widehat{\mu},Z_i).$$

Then for $\varepsilon<\frac{1}{2}$, as long as n is large enough

$$n=O\left(\frac{1}{\varepsilon^2}\left((M^2+1)d_P(\mathcal{F}_{\sigma^2})+(M^3+1)d_P(\mathcal{F}_\mu)+(1+M^4)\log(\frac{2}{\delta})\right)\right),$$

we have

$$\mathbb{E}_X\left((\sigma^2)^\star(X) - \widehat{\sigma^2}(X)\right)^2 \leq \varepsilon$$

with probability at least $1-\delta$.

Proof. By Proposition 4 and Lemma 2, we have with probability at least $1-\delta$, that uniformly for all $\sigma^2 \in \mathcal{F}_{\sigma^2}, \mu \in \mathcal{F}_\mu$,

$$\begin{aligned}
&\left|\mathbb{E}_Z\left[\ell_{MSE}(\sigma^2,\mu,Z)\right] - \frac{1}{n}\sum_{i=1}^n \ell_{MSE}(\sigma^2,\mu,Z_i)\right| \\
\leq& 2K_1(16+16\sqrt{M}+6M)\sqrt{\frac{d_P(\mathcal{F}_{\sigma^2})}{n}} + 16K_1(8+12\sqrt{M}+6M+M^{\frac{3}{2}})\sqrt{\frac{d_P(\mathcal{F}_\mu)}{n}} \\
&+4(M^2+8M+16)\cdot\sqrt{\frac{1}{2n}\log\frac{2}{\delta}}
\end{aligned}$$

Then for

$$\begin{aligned}
n \geq \frac{1}{\varepsilon^2}\max\Bigg\{&(36K_1^2(16+16\sqrt{M}+6M)^2 d_P(\mathcal{F}_{\sigma^2}), \\
&48^2K_1^2(8+12\sqrt{M}+6M+M^{\frac{3}{2}})^2 d_P(\mathcal{F}_\mu), \frac{12^2(M^2+8M+16)^2}{2m^2}\log(\frac{2}{\delta})\Bigg\},
\end{aligned}$$

we have with probability at least $1-\delta$:

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \ell_{MSE}(\sigma^{*2},\widehat{\mu},Z_i) \leq \mathbb{E}_Z\left[\ell_{MSE}(\sigma^{*2},\widehat{\mu},Z)\right] + \varepsilon \\
&\mathbb{E}_Z\left[\ell_{MSE}(\widehat{\sigma^2},\widehat{\mu},Z)\right] \leq \frac{1}{n}\sum_{i=1}^n \ell_{MSE}(\widehat{\sigma^2},\widehat{\mu},Z_i) + \varepsilon
\end{aligned} \tag{45}$$

Using fact that $\widehat{\sigma^2}(\boldsymbol{x})$ is the empirical risk minimizer, we have:

$$\frac{1}{n}\sum_{i=1}^n \ell_{MSE}(\widehat{\sigma^2};\widehat{\mu},Z_i) \leq \frac{1}{n}\sum_{i=1}^n \ell_{MSE}((\sigma^2)^\star;\widehat{\mu},Z_i) \tag{46}$$

Combining Equation 45 and Equation 46:

$$
\begin{aligned}
&\mathbb{E}_Z\left[\ell_{MSE}(\widehat{\sigma^2};\widehat{\mu},Z)\right] \le \mathbb{E}_Z\left[\ell_{MSE}((\sigma^2)^\star;\widehat{\mu},Z)\right] + 2\varepsilon \\
\implies &\mathbb{E}_{X,Y}\left[\left(\widehat{\sigma^2}(X) - (Y-\widehat{\mu}(X))^2\right)^2\right] \le \mathbb{E}_{X,Y}\left[\left((\sigma^2)^\star(X) - (Y-\widehat{\mu}(X))^2\right)^2\right] + 2\varepsilon \\
\implies &\mathbb{E}_{X,Y}\Bigg[\left(\widehat{\sigma^2}(X) - (\sigma^2)^\star(X)\right)^2 \\
&\quad + 2\left(\widehat{\sigma^2}(X) - (\sigma^2)^\star(X)\right)\left((\sigma^2)^\star(X) - (Y-\widehat{\mu}(X))^2\right) + \left((\sigma^2)^\star(X) - (Y-\widehat{\mu}(X))^2\right)^2\Bigg] \\
&\le \mathbb{E}_{X,Y}\left[\left((\sigma^2)^\star(X) - (Y-\widehat{\mu}(X))^2\right)^2\right] + 2\varepsilon \\
\implies &\mathbb{E}_{X,Y}\left[\left(\widehat{\sigma^2}(X) - (\sigma^2)^\star(X)\right)^2 + 2\left(\widehat{\sigma^2}(X) - (\sigma^2)^\star(X)\right)\left((\sigma^2)^\star(X) - (Y-\widehat{\mu}(X))^2\right)\right] \le 2\varepsilon \\
\implies &\mathbb{E}_{X,Y}\left[2\left(\widehat{\sigma^2}(X) - (\sigma^2)^\star(X)\right)\left((\sigma^2)^\star(X) - (\mu^*(X)-\widehat{\mu}(X))^2\right)\right] \\
&-\mathbb{E}_{X,Y}\left[2\left(\widehat{\sigma^2}(X) - (\sigma^2)^\star(X)\right)\left((Y-\mu^*(X))^2 + 2(Y-\mu^*(X))(\mu^*(X)-\widehat{\mu}(X))\right)\right] \\
&+\mathbb{E}_X\left[\left(\widehat{\sigma^2}(X) - (\sigma^2)^\star(X)\right)^2\right] \le 2\varepsilon \\
\implies &\mathbb{E}_X\left[\left(\widehat{\sigma^2}(X) - (\sigma^2)^\star(X)\right)^2 - 2\left(\widehat{\sigma^2}(X) - (\sigma^2)^\star(X)\right)\left(\mu^*(X)-\widehat{\mu}(X)\right)^2\right] \le 2\varepsilon \\
\implies &\mathbb{E}_X\left[\left(\widehat{\sigma^2}(X) - (\sigma^2)^\star(X)\right)^2 - 4\left|\widehat{\sigma^2}(X) - (\sigma^2)^\star(X)\right|\left|\mu^*(X)-\widehat{\mu}(X)\right|\right] \le 2\varepsilon \\
\implies &\mathbb{E}_X\left[\left(\widehat{\sigma^2}(X) - (\sigma^2)^\star(X)\right)^2\right] - 4\sqrt{\mathbb{E}_X\left[\left(\widehat{\sigma^2}(X) - (\sigma^2)^\star(X)\right)^2\right]\mathbb{E}_X\left(\mu^*(X)-\widehat{\mu}(X)\right)^2} \le 2\varepsilon.
\end{aligned}
$$

Now, define $a \triangleq \mathbb{E}_X\left[\left(\widehat{\sigma^2}(X) - (\sigma^2)^\star(X)\right)^2\right]$, since $E_X\left(\mu^*(X)-\widehat{\mu}(X)\right)^2 \le \varepsilon$, we have

$$a - 4\sqrt{a\varepsilon} \le 2\varepsilon,$$

which implies $a \le 25\varepsilon$, or equivalently

$$\mathbb{E}_X\left[\left(\widehat{\sigma^2}(X) - (\sigma^2)^\star(X)\right)^2\right] \le 25\varepsilon$$

with probability $1-\delta$. □