Heteroscedastic Sequences: Beyond Gaussianity

Oren Anava OANAVA@TX.TECHNION.AC.IL
Technion, Haifa, Israel

Shie Mannor SHIE@EE.TECHNION.AC.IL
Technion, Haifa, Israel

Abstract

We address the problem of sequential prediction in the heteroscedastic setting, when both the signal and its variance are assumed to depend on explanatory variables. By applying regret minimization techniques, we devise an efficient online learning algorithm for the problem, *without* assuming that the error terms comply with a specific distribution. We show that our algorithm can be adjusted to provide confidence bounds for its predictions, and provide an application to ARCH models. The theoretical results are corroborated by an empirical study.

1. Introduction

Heteroscedasticity refers to the case in which the variability of the dependent variable (also called *signal*) is unequal across the range of values of the explanatory variables (also called *features*). In statistical modeling, the variability is usually characterized through the conditional variance, which is a key parameter in many statistical applications such as volatility estimation in finance, disease phenotypes prediction in medicine, and more.

Much work has been done on parameter estimation and signal prediction using heteroscedastic models, mostly relying on statistical assumptions on the error terms such as Gaussianity or other symmetrical distributions. These assumptions allow the use of Maximum Likelihood (ML) techniques to recover consistent estimators for the signal and its conditional variance. However, if these assumptions are not met in practice, the resulting estimators are no longer consistent and the following statement from (Whittaker & Robinson, 1967) is sometimes quoted:

> *"Everybody believes in the exponential law of errors: the experimenters, because they think it can be proved by mathematics; and the mathematicians, because they believe it has been established by observation."*

In this paper we argue that traditional modeling assumptions on the signal generation can be substantially relaxed while still maintaining the ability to solve the problem. Moreover, we offer a novel *online learning approach* that allows the signal to be partially adversarial and partially stochastic. We show that our approach is more general than a ML-based approach, and is capable of coping with rather complex scenarios and models.

An important aspect of our work is bridging the gap between a statistical approach and a "pure" online learning approach for sequential prediction problems. We claim that while the statistical approach fails to model real-world data due to strict distributional assumptions, the online learning approach fails to do so due to lack of such assumptions, and the actual "truth" lies somewhere in between them.

1.1. Main contribution

In this work we propose a new approach to handle heteroscedasticity —*an online learning approach*— that does not require the error terms to be Gaussian, nor to comply with a specific distribution as is common in the statistical approach to the problem. The main contributions of this work are as follows:

Casting the problem of heteroscedastic signal prediction as an online learning problem, in which two terms of regret are minimized in parallel: one captures the prediction accuracy, and the other measures the quality of the conditional variance estimation. This casting is the key idea that enables handling non-Gaussian error terms in this setting.

Design of an online learning algorithm that is suitable to work with biased gradients. The necessity in our case arises since the conditional variance is not observed, and can only be estimated with bias.

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

Derivation of worst case confidence bounds for the prediction in each round, using regret analysis. Here, worst case refers to the distribution of the error terms, which is assumed to be unknown and might even vary from round to round.

Application to ARCH Prediction in which the robustness of our approach to non-Gaussian error terms is demonstrated on synthetic data.

1.2. Related Work

Heteroscedastic models have been extensively studied, mainly in the context of time series and regression. In regression, perhaps the earliest heteroscedastic model that was considered (see (Lee, 1973) for example) assumes the following generation of the signal:

$$y_t = u^\top x_t + \epsilon_t \quad \forall t, \tag{1}$$

where x_t is a known feature vector and $\epsilon_t \sim \mathcal{N}(0, \sigma_t^2)$ for $\sigma_t^2 = v^\top x_t$. In this setting, u and v are unknown to us and have to be regressed by the algorithm. Though many such regression algorithms exist, they are all based on the following principle: apply a known technique (e.g., least squares or maximum likelihood) to recover $\tilde{u}$, and then regress over $\tilde{\epsilon}_t^2 = (y_t - \tilde{u}^\top x_t)^2$ to recover $\tilde{v}$. A comparison of several such algorithms appears in (Amemiya, 1977).

Clearly, the basic model (Equation (1)) suffers from many shortcomings, such as lack of generality and the need for strong assumptions on the error distribution. In the following years, many works addressed these issues and proposed more general frameworks to model heteroscedasticity. Initially, various parametric models of the conditional variance were considered:

$$\left.\begin{array}{l}\sigma_t = \sigma(1 + |v^\top x_t|)^\lambda \\ \sigma_t = \sigma |v^\top x_t|^\lambda\end{array}\right\} \quad \text{(Box \& Hill, 1974);}$$

$$\sigma_t = \sigma e^{\lambda v^\top x_t} \quad \text{(Bickel, 1978);}$$

$$\sigma_t = \sqrt{1 + (v^\top x_t)^2} \quad \text{(Jobson \& Fuller, 1980).}$$

Later works showed that qualitatively similar results can be obtained for any smooth (and known in advance) function $\sigma_t = \sigma(x_t)$ (Davidian & Carroll, 1987; Muller & Stadtmuller, 1987). In a parallel line of work, (Fuller & Rao, 1978) circumvented the need to come up with a specific parametric form of the conditional variance by assuming that the signal is divided into finite number of groups, where the variance within each group is equal. Their results rely as well on Gaussian distributional assumptions of the error terms.

The problem of conditional variance estimation was also studied in the context of nonparametric models. (Fan & Yao, 1998) considered two-dimensional strictly stationary processes $\{(y_t, x_t)\}$ of the form

$$y_t = m(x_t) + \sigma(x_t)\epsilon_t \quad \forall t,$$

where $\mathbb{E}[\epsilon_t \mid x_t] = 0$ and $\mathbb{E}[\epsilon_t^2 \mid x_t] = 1$, and offered residual based estimators which are locally linear. The idea of nonparametric estimators was further studied in the works of (Yu & Jones, 2004) who proposed likelihood-based locally linear estimators; (Brown et al., 2007) who applied the difference sequence idea to estimate the conditional variance; and (Mishra et al., 2010) who incorporated parametric and nonparametric estimators in a multiplicative way.

In the field of time series analysis and prediction, it was the seminal work of (Engle, 1982), in which the autoregressive conditional heteroscedastic (ARCH) model was introduced, that led the development of a plethora of heteroscedastic models. Among the many extensions to the ARCH model, one can find the GARCH model (Bollerslev, 1986), the EGARCH model (Nelson, 1991), and many other models that were shown to be highly effective in practice.

In the learning literature, the study of heteroscedasticity is rather limited. We note the work of (Zhu et al., 2013) that is aimed at coping with high-dimensional heteroscedastic data. Perhaps the closest works to ours, at least in spirit, are the works of (Anava et al., 2013; 2015) who considered the sequential prediction problem using the AR and ARMA models in a partially (or fully) adversarial setting. We also note the works of (Even-Dar et al., 2009; Yu et al., 2009), who considered a hybrid setting in the context of Markov decision processes.

2. Preliminaries and Model

Before defining our setting, we provide some useful background about kernel methods and the framework of Online Convex Optimization (OCO).

2.1. Kernel Methods

A *kernel* is a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ (for some $\mathcal{X} \subset \mathbb{R}^d$), which is usually assumed to be continuous. A kernel is a *Mercer kernel* if for any finite set of points $\{x_1, \ldots, x_n\}$ the $n \times n$ matrix K, where K_{ij} is defined to be $k(x_i, x_j)$, is positive semi-definite. For such kernels, it is well-known that there exists a Hilbert space $\mathcal{H}$ and a mapping $\phi : \mathcal{X} \to \mathcal{H}$ such that $k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$, where $\langle \cdot, \cdot \rangle$ is the inner product in $\mathcal{H}$. Two well-known examples of kernels are the polynomial kernel $k(x_i, x_j) = \left(1 + x_i^\top x_j\right)^p$ and the Gaussian kernel $k(x_i, x_j) = \exp\left(-\frac{1}{2\sigma^2}\|x_i - x_j\|^2\right)$, which are particularly useful for modeling nonnegative entities. See (Shawe-Taylor & Cristianini, 2004) for more information on Kernel methods.

Algorithm 1 LAZY OGD (on the ℓ_2 unit ball)

Input: learning rate η_t.
Set $a_1 = 0$.
for $t = 1$ to T **do**
  Play a_t and observe loss $\ell_t(a_t)$
  Set $a_{t+1} = -\frac{\eta_t \sum_{i=1}^{t} \nabla \ell_i(a_i)}{\max\{1, \eta_t \| \sum_{i=1}^{t} \nabla \ell_i(a_i) \|\}}$
end for

In our context, we will use kernel functions to characterize the first and second moments of the signal. The motivation is to gain a modeling power without paying big computational costs. This important property does not come without drawbacks; perhaps the most prominent is the fact that not all online algorithms are compatible with kernels. We turn to present an online algorithm that is suited for kernels: LAZY ONLINE GRADIENT DESCENT (OGD), which is also known to be a special instance of FOLLOW THE REGULARIZED LEADER (FTRL) algorithm.

2.2. Online Convex Optimization and LAZY OGD

One of the most well-studied frameworks of online learning is *Online Convex Optimization* (OCO). In this framework, an online player iteratively chooses an action $a_t \in \mathcal{K}$, and then suffers loss that is equal to $\ell_t(a_t)$. The action set $\mathcal{K}$ is assumed to be a closed and bounded convex subset of $\mathbb{R}^d$, and the loss functions $\{\ell_t\}_{t=1}^T$ are assumed to be convex functions from $\mathcal{K}$ to $[0, 1]$. The performance of the player is measured using the *regret* criterion, defined as follows:

$$\mathcal{R}_T(\ell_1, \ldots, \ell_T) = \sum_{t=1}^{T} \ell_t(a_t) - \min_{a \in \mathcal{K}} \sum_{t=1}^{T} \ell_t(a),$$

where T is a predefined integer that denotes the total number of rounds played. The goal in this framework is to design efficient algorithms, whose regret grows sublinearly in T, corresponding to an average per-round regret going to zero as T increases.

One of the popular algorithms for OCO is the LAZY OGD algorithm (Algorithm 1), for which the following regret bound is known[1]:

$$\begin{aligned}\mathcal{R}_T^{\text{Lazy}}(\ell_1, \ldots, \ell_T) &= \sum_{t=1}^{T} \ell_t(a_t) - \min_{a \in \mathcal{K}} \sum_{t=1}^{T} \ell_t(a) \\ &\leq \frac{\|a^*\|^2}{\eta_T} + \sum_{t=1}^{T} \eta_t \|\nabla \ell_t(a_t)\|^2. \quad (2)\end{aligned}$$

where $a^* = \arg\min_{a \in \mathcal{K}} \sum_{t=1}^{T} \ell_t(a)$. If the action set $\mathcal{K}$ is assumed to be the ℓ_2 unit ball, then we have $\mathcal{R}_T^{\text{Lazy}}(\ell_1, \ldots, \ell_T) \leq 2G\sqrt{T}$ for a properly chosen η_t and $G = \max_{a,t} \{\|\nabla \ell_t(a)\|\}$. A complete analysis can be found in (Hazan, 2011; Shalev-Shwartz, 2012).

[1] We use $\|\cdot\|$ throughout the paper to denote the ℓ_2-norm.

2.3. Online Prediction of Heteroscedastic Signals

Assume the following iterative game between a player and nature (which might be adversarial): at round t, nature chooses $x_t \in \mathcal{X} \subset \mathbb{R}^d$ and generates $y_t \in \mathbb{R}$ such that $\mathbb{E}[\, y_t \mid x_t \,] = u_0^\top \phi(x_t)$ and $\text{Var}[\, y_t \mid x_t \,] = v_0^\top \psi(x_t)$, where ϕ and ψ are induced by a Mercer kernel, and $\|u_0\|, \|v_0\| \leq 1$ are set beforehand by nature. A special case, for instance, is the standard linear model

$$y_t = u_0^\top x_t + \sqrt{v_0^\top x_t} \cdot \epsilon_t,$$

where the error terms ϵ_t are assumed to be distributed $\mathcal{N}(0, 1)$. Note that the adversarial behavior of nature is expressed in (1) the selection of x_t; (2) the choice of the parameters u_0, v_0; and (3) the distribution of y_t given x_t.

The player receives x_t and has to provide a prediction $\tilde{y}_t = u_t^\top \phi(x_t)$ for y_t and an estimation $\tilde{\sigma}_t^2 = v_t^\top \psi(x_t)$ for its conditional variance. The player, of course, is not aware of nature's selection of u_0 and v_0, but is aware of the functions ϕ and ψ (in the sense of knowing to compute the inner products $\langle \phi(x_i), \phi(x_j) \rangle$ and $\langle \psi(x_i), \psi(x_j) \rangle$ for any $x_i, x_j \in \mathcal{X}$). After committing to $\tilde{y}_t$ and $\tilde{\sigma}_t^2$, the player incurs two *losses*, one for the prediction error and the other for the conditional variance inaccuracy:

$$\begin{aligned}\ell_t^{\text{Sig}}(u_t) &= \left(y_t - u_t^\top \phi(x_t)\right)^2 \\ \ell_t^{\text{Var}}(v_t) &= \left(v_t^\top \psi(x_t) - v_0^\top \psi(x_t)\right)^2.\end{aligned}$$

Naturally, the goal of the player is to (separately) minimize the sum of losses, over a predefined number of rounds T. Here also, we choose the regret to measure the performance of the online player. Thus, for the signal prediction task we have

$$\mathcal{R}_T(\ell_1^{\text{Sig}}, \ldots, \ell_T^{\text{Sig}}) = \sum_{t=1}^{T} \ell_t^{\text{Sig}}(u_t) - \min_{\|u\| \leq 1} \sum_{t=1}^{T} \ell_t^{\text{Sig}}(u),$$

and for the conditional variance estimation we have

$$\mathcal{R}_T(\ell_1^{\text{Var}}, \ldots, \ell_T^{\text{Var}}) = \sum_{t=1}^{T} \ell_t^{\text{Var}}(v_t) - \min_{\|v\| \leq 1} \sum_{t=1}^{T} \ell_t^{\text{Var}}(v). \quad (3)$$

Notice that $\ell_t^{\text{Var}}(v)$ cannot be directly computed since $v_0^\top \psi(x_t)$ is unknown at any stage, and thus a reasonable goal would be to minimize $\mathbb{E}\left[\mathcal{R}_T(\ell_1^{\text{Var}}, \ldots, \ell_T^{\text{Var}})\right]$.

Readers, especially those familiar with online learning, might wonder if our setting cannot be strengthened in certain ways. First of all, in many online learning applications, nothing is assumed about the data generation, and we might

envision a scenario in which the signal y_t is generated arbitrarily. However, this setting is ill-defined in the context of conditional variance estimation, as the notion of variance does not exist for non-stochastic signal. This fact causes the regret term in Equation (3), along with learning in this setting, to be meaningless.

2.4. Our Assumptions

Throughout this work we assume the following:

(1) For any t, it holds that $\mathbb{E}[\, y_t \mid x_t \,] = u_0^\top \phi(x_t)$ and $\mathrm{Var}[\, y_t \mid x_t \,] = v_0^\top \psi(x_t)$, where ϕ and ψ are functions induced by a Mercer kernel, and $\|u_0\|, \|v_0\| \le 1$.

(2) It holds that $\|\phi(x)\|, \|\psi(x)\| \le 1$ for any $x \in \mathcal{X}$.

(3) It holds that $y_t \in [-1, 1]$ for any t.

We note that these assumptions can be relaxed, and are merely here to simplify the exposition and calculations. In assumption (1), it would be sufficient to require $\mathbb{E}[\, y_t \mid x_t \,] \approx u_0^\top \phi(x_t)$ and $\mathrm{Var}[\, y_t \mid x_t \,] \approx v_0^\top \psi(x_t)$, for a small enough bias (that would be added to our regret bound). Assumption (2) can be replaced by the assumption that $\|\phi(x)\|, \|\psi(x)\| \le C$ for some $C < \infty$ (which need not be known in advance, as the algorithm can be easily adjusted to handle this case by using a standard doubling trick). Assumption (3) can be relaxed to light tail assumption. That is, we can assume that y_t lies in a finite interval with high probability (which holds in particular for Gaussian errors), without increasing the complexity of the problem at hand.

3. Our Approach

The main challenge in our setting is the fact that the conditional variance is not revealed to us at any stage, and yet we wish to compete against the best conditional variance estimator in hindsight. To circumvent this issue, we use *biased* estimators of the conditional variance instead of the actual ones. The bias follows from the fact that the expected value of signal is also unknown, and can only be approximated by our online algorithm.

Next, we take a step back from our setting, and present a general working scheme for OCO with biased gradient estimators. This scheme was considered before (e.g., in the work of (Huh & Rusmevichientong, 2013)), but the existing algorithms are not suitable to our setting due to lack of generality (and in particular, inability of coping with kernels). Throughout this section, we denote by $\mathcal{F}_{t-1}$ the sigma-algebra that is generated by all the actions played up to round t, and all losses occurred up to round $t-1$.

3.1. OCO with Biased Gradient Estimators

Consider the OCO framework described in Section 2.2, with the following change: after committing to an action a_t, the online player only receives a feedback in the form of a biased gradient estimator at a_t. For this framework, we can prove the following:

Proposition 3.1. *Let $\{\ell_t\}_{t=1}^T$ be a sequence of loss functions, and $\{\tilde{\ell}_t\}_{t=1}^T$ a sequence of corresponding approximation functions for which it holds that*

$$\mathbb{E}[\nabla \tilde{\ell}_t(a) \mid \mathcal{F}_{t-1}] = \nabla \ell_t(a) + b_t(a),$$

for any t and $a \in \mathcal{K}$. Denote by $\{a_t\}_{t=1}^T$ the sequence of actions that a first-order algorithm $\mathcal{A}$ outputs for $\{h_t\}_{t=1}^T$, where $h_t(a) = \ell_t(a) + a^\top(\nabla \tilde{\ell}_t(a_t) - \nabla \ell_t(a_t))$. Then,

$$\mathbb{E}[\mathcal{R}_T(\ell_1, \ldots, \ell_T)] = \sum_{t=1}^T \mathbb{E}\left[\ell_t(a_t)\right] - \sum_{t=1}^T \ell_t(a^*)$$

$$\le \mathbb{E}\left[\mathcal{B}_T^{\mathcal{A}}(h_1, \ldots, h_T)\right] - \sum_{t=1}^T \mathbb{E}\left[(a_t - a^*)^\top b_t(a_t)\right],$$

where $a^* = \arg\min_{a \in \mathcal{K}} \sum_{t=1}^T \ell_t(a)$, and $\mathcal{B}_T^{\mathcal{A}}(h_1, \ldots, h_T)$ is a regret bound of algorithm $\mathcal{A}$ applied to $\{h_t\}_{t=1}^T$.

Basically, the proposition states that one can provide biased gradient estimators as an input to any first-order online algorithm, and incur a corresponding additional term in the regret. The proof of this Proposition appears in the supplementary material.

3.2. Algorithm and Analysis

We turn to present our algorithm (Algorithm 2) along with its analysis. We start by defining

$$\tilde{\ell}_t^{\mathrm{Var}}(v_t) = \left(v_t^\top \psi(x_t) - (y_t - u_t^\top \phi(x_t))^2\right)^2, \qquad (4)$$

which is an approximation to $\ell_t^{\mathrm{Var}}(v_t)$. This definition plays an important role in our analysis, since $\ell_t^{\mathrm{Var}}(v_t)$ is unobserved at any stage.

Note that despite the inefficient representation of Algorithm 2, in practice the predictions and the variance estimations are generated efficiently using a simple kernel trick. This form is easier to analyze and is thus stated here. For Algorithm 2 we can prove the following:

Theorem 3.2. *Algorithm 2 generates online sequences $\{u_t\}_{t=1}^T$ and $\{v_t\}_{t=1}^T$, for which it holds that*

$$\mathcal{R}_T(\ell_1^{\mathit{Sig}}, \ldots, \ell_T^{\mathit{Sig}}) = \sum_{t=1}^T \ell_t^{\mathit{Sig}}(u_t) - \min_{\|u\| \le 1} \sum_{t=1}^T \ell_t^{\mathit{Sig}}(u) \le 8T^{1/2},$$

Algorithm 2

Input: learning rates $\eta_{\text{Sig}}, \eta_{\text{Var}}$.
Set $u_1 = 0$ and $v_1 = 0$.
for $t = 1$ to T **do**
  Play u_t and observe loss $\ell_t^{\text{Sig}}(u_t)$
  Play v_t and suffer loss $\ell_t^{\text{Var}}(v_t)$
  Set $u_{t+1} = -\frac{\eta_{\text{Sig}} \sum_{i=1}^{t} \nabla \ell_i^{\text{Sig}}(u_i)}{\max\{1, \eta_{\text{Sig}} \| \sum_{i=1}^{t} \nabla \ell_i^{\text{Sig}}(u_i) \|\}}$
  Set $v_{t+1} = -\frac{\eta_{\text{Var}} \sum_{i=1}^{t} \nabla \tilde{\ell}_i^{\text{Var}}(v_i)}{\max\{1, \eta_{\text{Var}} \| \sum_{i=1}^{t} \nabla \tilde{\ell}_i^{\text{Var}}(v_i) \|\}}$
end for

and also

$$\mathbb{E}\left[\mathcal{R}_T(\ell_1^{Var}, \ldots, \ell_T^{Var})\right] = \sum_{t=1}^{T} \mathbb{E}\left[\ell_t^{Var}(v_t)\right] - \min_{\|v\| \le 1} \sum_{t=1}^{T} \ell_t^{Var}(v) \le 64 T^{1/2},$$

if we choose $\eta_{Sig} = \eta_{Var} = \frac{1}{2\sqrt{T}}$.

Proof. Notice that applying Algorithm 2 to $\{\ell_t^{\text{Sig}}\}_{t=1}^T$ is equivalent to applying LAZY OGD, and thus it trivially holds that $\mathcal{R}_T(\ell_1^{\text{Sig}}, \ldots, \ell_T^{\text{Sig}}) \le 8T^{1/2}$. For $\{\ell_t^{\text{Var}}\}_{t=1}^T$ we have to work somewhat harder. Our proof relies on the fact that Algorithm 2 generates an online sequence $\{u_t\}_{t=1}^T$ for which it holds that

$$\sum_{t=1}^{T} \mathbb{E}\left[\left(u_t^\top \phi(x_t) - u_0^\top \phi(x_t)\right)^2\right] - \min_{\|u\| \le 1} \sum_{t=1}^{T} \left(u^\top \phi(x_t) - u_0^\top \phi(x_t)\right)^2 \le 8T^{1/2}.$$

This observation is proved in the supplementary material. This immediately implies that

$$\sum_{t=1}^{T} \mathbb{E}\left[\left(u_t^\top \phi(x_t) - u_0^\top \phi(x_t)\right)^2\right] \le 8T^{1/2},$$

since $\min_{\|u\| \le 1} \sum_{t=1}^{T} \left(u^\top \phi(x_t) - u_0^\top \phi(x_t)\right)^2 = 0$ for the case where $\|u_0\| \le 1$, which holds by assumption (1).

Next, we use the definition of $\tilde{\ell}_t^{\text{Var}}$ in Equation 4 to derive

$$\mathbb{E}\left[\nabla \tilde{\ell}_t^{\text{Var}}(v_t) \mid \mathcal{F}_{t-1}\right] = \nabla \ell_t^{\text{Var}}(v_t) + b_t(v_t),$$

where $b_t(v_t) = -2\psi(x_t)(u_0^\top \phi(x_t) - u_t^\top \phi(x_t))^2$. Notice that $b_t(v_t)$ does not depend on v_t in this case. Defining $h_t(v) = \ell_t^{\text{Var}}(v) + v^\top(\nabla \tilde{\ell}_t^{\text{Var}}(v_t) - \nabla \ell_t^{\text{Var}}(v_t))$ and applying LAZY OGD to $\{h_t\}_{t=1}^T$ gives (by Proposition 3.1):

$$\mathbb{E}[\mathcal{R}_T(\ell_1^{\text{Var}}, \ldots, \ell_T^{\text{Var}})] = \sum_{t=1}^{T} \mathbb{E}\left[\ell_t^{\text{Var}}(v_t)\right] - \sum_{t=1}^{T} \ell_t^{\text{Var}}(v^*) \le \mathbb{E}\left[\mathcal{B}_T(h_1, \ldots, h_T)\right] - \sum_{t=1}^{T} \mathbb{E}\left[(v_t - v^*)^\top b_t(v_t)\right],$$

where $v^* = \arg\min_v \sum_{t=1}^{T} \ell_t(v)$, and $\mathcal{B}_T(h_1, \ldots, h_T)$ is the regret bound of LAZY OGD for $\{h_t\}_{t=1}^T$. It can be easily shown that $\mathbb{E}\left[\mathcal{B}_T(h_1, \ldots, h_T)\right] \le 32T^{1/2}$.

Finally, we can bound

$$\sum_{t=1}^{T} \mathbb{E}\left[(v_t - v^*)^\top b_t(v_t)\right] = -2 \sum_{t=1}^{T} \mathbb{E}\left[(v_t - v^*)^\top \psi(x_t)(u_0^\top \phi(x_t) - u_t^\top \phi(x_t))^2\right] \le 4 \sum_{t=1}^{T} \mathbb{E}\left[(u_0^\top \phi(x_t) - u_t^\top \phi(x_t))^2\right] \le 32T^{1/2},$$

which completes the proof. □

3.3. Worst-Case Confidence Bounds

We are now interested in using the results from the previous section to generate confidence bounds for our prediction. More formally, given a probability $\alpha \in (0, 1)$ our task is to provide a sequence $\{c_t\}_{t=1}^T$ for which it holds that:

$$\frac{1}{T} \sum_{t=1}^{T} P\left(|u_t^\top \phi(x_t) - y_t| \ge c_t\right) \le \alpha. \tag{5}$$

In words, the expected proportion of the predictions for which the distance to the actual signal exceeds the corresponding c_t is at most α. The above trivially holds if we choose large enough constants $\{c_t\}_{t=1}^T$, and thus we are interested not only in finding such constants, but also in showing that they are tight in a sense.

To derive Equation (5), we somewhat abuse notations and define $\ell_t^{\text{Sig}}(u) = \frac{1}{c_t^2}\left(y_t - u^\top \phi(x_t)\right)^2$ and $\ell_t^{\text{Var}}(v) = \frac{1}{c_t^4}\left(v^\top \psi(x_t) - v_0^\top \psi(x_t)\right)^2$. Notice that here also we need an estimated loss for ℓ_t^{Var} as it is not revealed to us. Thus, we define

$$\tilde{\ell}_t^{\text{Var}}(v) = \frac{1}{c_t^4}\left(v^\top \psi(x_t) - \left(y_t - u_t^\top \phi(x_t)\right)^2\right)^2.$$

Note that c_t might be random, yet it must hold that $\mathbb{E}\left[c_t \mid \mathcal{F}_{t-1}\right] = c_t$. That is, c_t is known given the actions played up to round t, and the losses occurred up to time $t-1$. Otherwise, the losses are not well defined. Now, we can prove the following:

Proposition 3.3. *Let ℓ_t^{Sig}, ℓ_t^{Var} and $\tilde{\ell}_t^{Var}$ be as defined above and let $\alpha \in (0, 1)$. Then, Algorithm 2 generates online sequences $\{u_t\}_{t=1}^T$ and $\{v_t\}_{t=1}^T$, for which it holds that:*

$$\frac{1}{T} \sum_{t=1}^{T} P\left(|u_t^\top \phi(x_t) - y_t| \ge c_t\right) \le \alpha,$$

for $c_t = \sqrt{\frac{2 \max\{\beta, v_t^\top \psi(x_t)\}}{\alpha}}$ and $\beta = 16T^{-1/4}$.

Remark: The constant $\sqrt{2}$ can be further improved to 1 (infinitesimally) by adjusting the value assigned to β. For simplicity, we prove only the result stated in the claim.

As mentioned before, finding a sequence $\{c_t\}_{t=1}^T$ for which $\frac{1}{T}\sum_{t=1}^T P\left(|u_t^\top \phi(x_t) - y_t| \geq c_t\right) \leq \alpha$ is meaningless, unless this sequence is tight in some sense. In our context, we need to show that there exist an error distribution for which the above holds in the other direction. Thus, we set some $k \geq 1$, and define the following error distribution:

$$\epsilon_t = \left(y_t - u_0^\top \phi(x_t)\right) = \begin{cases} -k\sqrt{v_0^\top \psi(x_t)}, & \text{w.p. } \frac{1}{2k^2} \\ 0, & \text{w.p. } 1 - \frac{1}{k^2} \\ k\sqrt{v_0^\top \psi(x_t)}, & \text{w.p. } \frac{1}{2k^2} \end{cases}$$

One can easily verify that for this distribution $\mathbb{E}\left[y_t \mid x_t\right] = u_0^\top \phi(x_t)$ and $\mathrm{Var}\left[y_t \mid x_t\right] = v_0^\top \psi(x_t)$. Now, note that for the choice $c_t = k\sqrt{v_0^\top \psi(x_t)}$ it holds that:

$$\begin{aligned} P\left(|y_t - u_t^\top \phi(x_t)| \geq c_t\right) &\geq P\left(|\epsilon_t| \geq c_t\right) \\ &= P\left(|\epsilon_t| \geq k\sqrt{v_0^\top \psi(x_t)}\right) = \frac{1}{k^2}. \end{aligned}$$

Setting $k = \sqrt{\frac{1}{\alpha}}$ gives the result.

4. Extensions and Applications

Here we extend the result of the previous section to several interesting cases. The first is the multivariate case, in which $y_t \in \mathbb{R}^n$ and the conditional variance then takes the form of a matrix. The second is an extension of our approach to higher moments, which enable the derivation of tighter confidence bounds for the prediction.

4.1. The Multivariate Case

In the multivariate case, the considered game is described as follows. At round t, nature chooses $x_t \in \mathbb{R}^d$ and generates $y_t \in \mathbb{R}^n$ such that:

(1) $\mathbb{E}\left[\, y_t \mid x_t \,\right] = U_0\phi(x_t)$, where $\phi : \mathbb{R}^d \to \mathbb{R}^{\hat{d}}$ and U_0 is an $n \times \hat{d}$ matrix with $\|U_0\|_F \leq 1$. Here and on, $\|\cdot\|_F$ refers to the Frobenius norm.

(2) $\mathrm{Var}\left[\, y_t \mid x_t \,\right] = V_0\psi(x_t)$, where $\psi : \mathbb{R}^d \to \mathbb{R}^{\bar{d}}$ and V_0 is an $n \times n \times \bar{d}$ tensor such that $\|V_0\|_F \leq 1$ and $V_{ijk} = V_{jik}$ for any i, j and k.

Here also, the player receives x_t and has to provide a prediction $\tilde{y}_t = U_t\phi(x_t)$, and an estimation $\tilde{\Sigma}_t^2 = V_t\psi(x_t)$ for its covariance matrix. After committing to $\tilde{y}_t$ and $\tilde{\Sigma}_t^2$, the player suffers two *losses*:

$$\begin{aligned} \ell_t^{\text{Sig}}(U_t) &= \left\|y_t - U_t\phi(x_t)\right\|^2, \\ \ell_t^{\text{Var}}(V_t) &= \left\|V_t\psi(x_t) - V_0\psi(x_t)\right\|_F^2. \end{aligned}$$

The regret is defined accordingly. Notice that in this setting as well an estimation for ℓ_t^{Var} is required, and thus we define an approximation $\tilde{\ell}_t^{\text{Var}}$ as follows:

$$\left\|V\psi(x_t) - (y_t - U_t\phi(x_t))(y_t - U_t\phi(x_t))^\top\right\|_F^2.$$

Applying Algorithm 2 to the extended setting yields the following result:

Corollary 4.1. *Algorithm 2 generates online sequences $\{U_t\}_{t=1}^T$ and $\{V_t\}_{t=1}^T$, for which it holds that*

$$\sum_{t=1}^T \ell_t^{Sig}(U_t) - \min_{\|U\|_F \leq 1} \sum_{t=1}^T \ell_t^{Sig}(U) \leq 8\sqrt{nT},$$

and also

$$\sum_{t=1}^T \mathbb{E}\left[\ell_t^{Var}(V_t)\right] - \min_{\|V\|_F \leq 1} \sum_{t=1}^T \ell_t^{Var}(V) = 64n\sqrt{T}.$$

The proof resembles the proof of Theorem 3.2, and is thus omitted here.

4.2. Higher Moments

The motivation in this section is to refine the result from Section 3.3 to higher moments. That is, to derive confidence bounds for the prediction which account for higher moments (other than the first and the second). We present here only a high level description of our approach and defer the technical parts to future work. We start by providing some useful background.

Recall that for a random variable X that has a moment generating function M_X that is finite in some open interval $\mathcal{I}$ about 0, it holds that: (1) X has moments of all orders; and (2) we can represent $M_X(s) = \sum_{n=0}^\infty \frac{\mathbb{E}[X^n]s^n}{n!}$ for $s \in \mathcal{I}$. By Chernoff bounds, we know we can upper bound the tail events of X as follows:

$$P(X \geq x) \leq e^{-sx}M_X(s) \text{ for } s > 0,$$

and

$$P(X \leq x) \leq e^{-sx}M_X(s) \text{ for } s < 0.$$

The above can further be optimized over s to derive tight bounds. In our context, these facts can be used to generate confidence bounds of the form we are interested in (as in Equation (5)).

Thus, assume that for each y_t and its corresponding feature vector x_t it holds that:

$$\mathbb{E}\left[(y_t - \mathbb{E}\left[\, y_t \,\right])^n\right] = u_n^\top \phi_n(x_t),$$

for $n \in \{1, \ldots, k\}$. In words, the n-th moment of the error term is given by the inner product between some

vector u_n and a function of x_t that is known in advance. This assumption generalizes the assumptions presented in Section 2 to higher moments. In addition, assume that $\sum_{n=0}^{\infty} \frac{u_n^\top \phi_n(x_t) s^n}{n!} \approx \sum_{n=1}^{k} \frac{u_n^\top \phi_n(x_t) s^n}{n!}$ for some k, which is independent of T. This assumption holds, for instance, when $(y_t - \mathbb{E}[\, y_t \mid x_t \,]) \in [-1, 1]$. Then, for a given y_t we can derive the following bound:

$$P\left(y_t - \mathbb{E}[\, y_t \,] \geq c\right) \leq e^{-sc} \sum_{n=0}^{k} \frac{u_n^\top \phi_n(x_t) s^n}{n!},$$

for $s > 0$, and the symmetric inequality for $s < 0$. This, again, can be optimized over s to derive the optimal bound. Notice, however, that $u_1, \ldots, u_k$ are unknown to us, and thus we need the regret analysis from which we can derive vectors $u_{1,t}, \ldots, u_{k,t}$ with the following guarantee:

$$\mathbb{E}\left[\mathcal{R}_T^n\right] = \sum_{t=1}^{T} \mathbb{E}\left[\ell_t^n(u_{n,t})\right] - \min_{\|u\| \leq 1} \sum_{t=1}^{T} \ell_t^n(u) = o(T),$$

for $n \in \{1, \ldots, k\}$. Here, as before, we use the definition $\ell_t^n(u) = (u^\top \phi_n(x_t) - u_n^\top \phi_n(x_t))^2$ and its corresponding estimate $\tilde{\ell}_t^n(u) = (u^\top \phi_n(x_t) - y_t^n)^2$ to derive the regret bounds.

5. Application to ARCH Models

We turn to present an application of our result to ARCH models. We first provide some background and then proceed to formally define the adaptation to our framework and our main result.

5.1. Background

Let $\{y_t\}_{t=1}^T$ be a time series (that is, a series of signal observations). The traditional $\mathrm{ARCH}(p)$ (short for autoregressive conditional heteroskedasticity) model of (Engle, 1982) is parameterized by lag p and coefficient vectors $u_0, v_0 \in \mathbb{R}^{p+1}$. The model assumes that y_t is a noisy linear combination of the previous p observations. That is,

$$y_t = u_0(0) + \sum_{k=1}^{p} u_0(k) y_{t-k} + \epsilon_t. \tag{6}$$

The error term ϵ_t in this model is assumed to be split into a stochastic piece z_t and a time-dependent standard deviation σ_t, characterizing the typical size of the error terms so that $\epsilon_t = \sigma_t z_t$. The random variable z_t is usually assumed to be a white noise process. The term σ_t^2 complies with the following model:

$$\sigma_t^2 = v_0(0) + \sum_{k=1}^{p} v_0(k) \epsilon_{t-k}^2, \tag{7}$$

where $v_0(0) > 0$, and $v_0(k) \geq 0$ for all $k > 0$. Notice that the $\mathrm{AR}(p)$ model is a special case of the $\mathrm{ARCH}(p)$ model, where the $v_0(k)$ coefficients are all zero for $k > 0$.

5.2. Adaptation to Our Setting

In our context, we will describe the setting as follows: First, some coefficient vectors (u_0, v_0) are fixed by the adversary. At each round t, the adversary generates ϵ_t from some zero-mean distribution with variance σ_t^2, and then use it to determine y_t via Equation (6). We emphasize that (u_0, v_0) and the noise terms (along with their distribution) are not revealed to us at any point.

At round t, we need to make a prediction $\tilde{y}_t$ for the signal and another prediction $\tilde{\sigma}_t^2$ for its conditional variance. After that, we incur two *losses*:

$$\ell_t^{\text{Sig}}(u_t) = \left(y_t - \tilde{y}_t(u_t)\right)^2 \text{ and } \ell_t^{\text{Var}}(v_t) = \left(\sigma_t^2 - \tilde{\sigma}_t^2(v_t)\right)^2.$$

Naturally, $\tilde{y}_t(u_t)$ should be of the form $\tilde{y}_t(u_t) = u_t(0) - \sum_{k=1}^{p} u_t(k) y_{t-k}$, but the main question is what should be the form of $\tilde{\sigma}_t^2(v_t)$? Recall that $\epsilon_{t-p}, \ldots, \epsilon_{t-1}$ are unobserved, and thus we cannot straightforwardly apply our approach (which requires knowing the feature vector).

5.3. Main Result

Our main result relies on the fact that Proposition 3.1 does not require the feature vector to be explicitly given, but only that the gradient of the approximated loss are close enough to the gradient of the original loss. Thus, if we let $\tilde{\epsilon}_t^2(u_t) = (y_t - \tilde{y}_t(u_t))^2$ and consequently define

$$\tilde{\ell}_t^{\text{Var}}(v) = \left(\tilde{\epsilon}_t^2(u_t) - v(0) - \sum_{k=1}^{p} v(k) \tilde{\epsilon}_{t-k}^2(u_{t-k})\right)^2,$$

then we can prove the following result:

Corollary 5.1. *Let ℓ_t^{Sig}, ℓ_t^{Var}, and $\tilde{\ell}_t^{Var}$ be as defined above. Then, Algorithm 2 generates online sequences $\{u_t\}_{t=1}^T$ and $\{v_t\}_{t=1}^T$, for which it holds that*

$$\mathcal{R}_T(\ell_1^{Sig}, \ldots, \ell_T^{Sig}) \leq \mathcal{O}(T^{1/2}),$$

and also that

$$\mathbb{E}\left[\mathcal{R}_T(\ell_1^{Var}, \ldots, \ell_T^{Var})\right] \leq \mathcal{O}(T^{1/2}),$$

if we choose $\eta_{Sig} = \eta_{Var} = \frac{1}{2\sqrt{T}}$.

The corollary relies on the fact that attaining a regret bound for $\ell_1^{\text{Sig}}, \ldots, \ell_T^{\text{Sig}}$ is a standard task in OCO. This regret bound implies that $\tilde{\epsilon}_t^2(u_t)$ is close in average to ϵ_t^2, which in turn implies that $\nabla \tilde{\ell}_t^{\text{Var}}$ is close to $\nabla \ell_t^{\text{Var}}$.

5.4. Experimental Results

Most of the works on time series prediction consider what we call the *offline setting*: given a time series, compute the model parameters (in our case, the ARCH coefficients) and

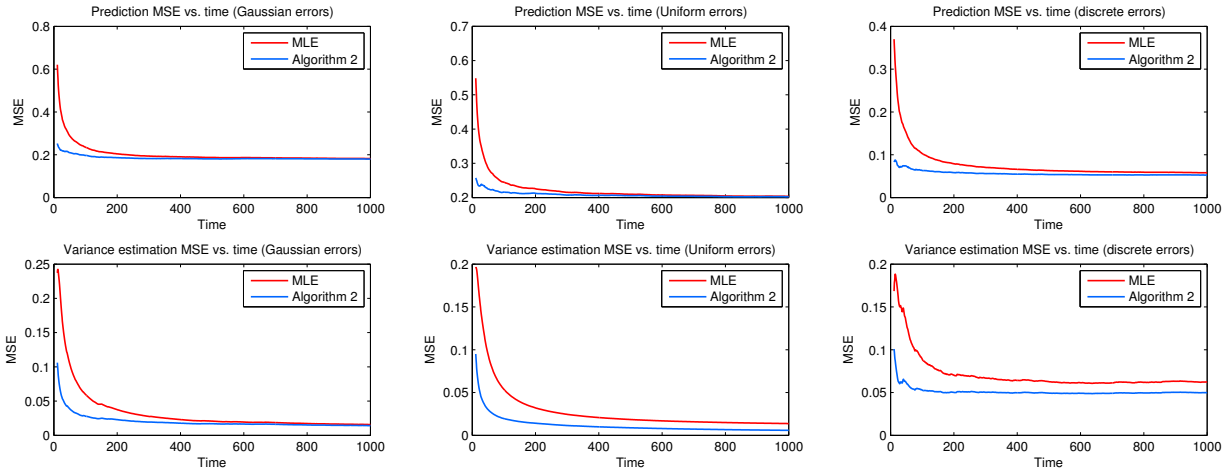

Figure 1. MSE of the prediction and the conditional variance estimation as a function of time.

	MSE@1000 for the signal prediction task			MSE@1000 for the variance estimation task		
	Gaussian errors	**Uniform errors**	**Discrete errors**	**Gaussian errors**	**Uniform errors**	**Discrete errors**
MLE	0.1823 (0.0159)	0.2037 (0.0137)	0.0580 (0.0119)	0.0157 (0.0139)	0.0138 (0.0121)	0.0625 (0.0207)
Algorithm 2	**0.1799 (0.0134)**	**0.2025 (0.0065)**	**0.0526 (0.0108)**	**0.0141 (0.0021)**	**0.0058* (0.0087)**	**0.0499* (0.0092)**

Table 1. MSE@1000 of the prediction and the conditional variance estimation. Bold font marks the best results, and asterisk indicates significance level of 0.05. Standard deviations of the results are presented in brackets.

measure the prediction error. Our *online setting* can be seen as a sequential offline setting, in which at round t we are given the time series values up to round $t-1$ and our task is to predict the signal and its variance at round t. In light of this, we adapt the state-of-the-art offline baseline (MLE) to the online setting. Note that this adaptation does not weaken the offline baseline in any way, and is used only for comparison purposes. In the plots we present the average accumulated loss up to round t for $t = 1, \ldots, 1000$.

To test the robustness of our approach to different error distributions, we generate three time series using the ARCH model (Equations (6) and (7)) with $u_0 = (0, 0.55, 0.11)$ and $v_0 = (0.1, 0.25, 0.25)$, each differs only in its error distribution. We consider the following distributions of the stochastic piece of the error terms: $z_t \sim \mathcal{N}(0,1)$; $z_t \sim \text{Uni}(-\sqrt{3}, \sqrt{3})$; and a discrete distribution

$$z_t = \begin{cases} k, & \text{w.p. } \frac{1}{2k^2} \\ 0, & \text{w.p. } 1 - \frac{1}{k^2} \\ -k, & \text{w.p. } \frac{1}{2k^2} \end{cases}$$

where $k = \sqrt{20}$. In all cases, one can easily verify that $\mathbb{E}[z_t] = 0$ and $\mathbb{E}[z_t^2] = 1$. This implies that the first and second moments of y_t conditioned on the history are equal.

We apply our algorithm and the MLE baseline (Engle, 1982) to the three time series. Figure 1 presents the MSE of the prediction and the MSE of the estimated conditional variance with respect to the true conditional variance (the latter is known as the data is synthetically generated). Table 1 presents the MSE at the end of both tasks, where bold font marks the best results, and asterisk indicates significance level of 0.05. To ensure stability, we average the results over 50 runs.

As evidenced by Figure 1 and Table 1, both algorithms perform roughly the same in the signal prediction task regardless of the error distribution. However, the online algorithm significantly surpasses the standard MLE in the variance estimation task, mainly when the time series exhibits some complicated error distribution (uniform or discrete). These empirical findings support the theoretical results and validate the generality of the online approach in practice.

6. Conclusion and Discussion

In this paper we presented an approach for the problem of sequential signal prediction in heteroscedastic environment. The main novelty of our approach is the fact that we allow the signal to be partially adversarial, in contrast to traditional methods that require it to be fully stochastic. To date, we are not aware of many works pursuing this direction (even outside the scope of sequential prediction of heteroscedastic sequences). We hope to extend this approach more broadly to bridge the gap between the statistical approach and online learning.

Acknowledgments

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 306638 (SUPREL), and the European Unions Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 336078 – ERC-SUBLRN.

References

Amemiya, Takeshi. A note on a heteroscedastic model. *Journal of Econometrics*, 6(3):365–370, 1977.

Anava, Oren, Hazan, Elad, Mannor, Shie, and Shamir, Ohad. Online learning for time series prediction. *arXiv preprint arXiv:1302.6927*, 2013.

Anava, Oren, Hazan, Elad, and Zeevi, Assaf. Online time series prediction with missing data. In *ICML*, 2015.

Bickel, Peter J. Using residuals robustly i: Tests for heteroscedasticity, nonlinearity. *The Annals of Statistics*, pp. 266–291, 1978.

Bollerslev, Tim. Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, 31(3):307–327, 1986.

Box, George E. P. and Hill, William J. Correcting inhomogeneity of variance with power transformation weighting. *Technometrics*, 16(3):385–389, 1974. doi: 10.1080/00401706.1974.10489207.

Brown, Lawrence D, Levine, M, et al. Variance estimation in nonparametric regression via the difference sequence method. *The Annals of Statistics*, 35(5):2219–2232, 2007.

Davidian, Marie and Carroll, Raymond J. Variance function estimation. *Journal of the American Statistical Association*, 82(400):1079–1091, 1987.

Engle, R.F. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50:987–1007, 1982.

Even-Dar, Eyal, Kakade, Sham M, and Mansour, Yishay. Online markov decision processes. *Mathematics of Operations Research*, 34(3):726–736, 2009.

Fan, Jianqing and Yao, Qiwei. Efficient estimation of conditional variance functions in stochastic regression. *Biometrika*, 85(3):645–660, 1998.

Fuller, Wayne A. and Rao, J. N. K. Estimation for a linear regression model with unknown diagonal covariance matrix. *The Annals of Statistics*, 6(5):1149–1158, 09 1978.

Hazan, Elad. The convex optimization approach to regret minimization. *Optimization for machine learning*, pp. 287, 2011.

Huh, Woonghee Tim and Rusmevichientong, Paat. Online sequential optimization with biased gradients: theory and applications to censored demand. *INFORMS Journal on Computing*, 26(1):150–159, 2013.

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

Lee, T. C. Nonlinear methods in econometrics : S.M. Goldfeld and R.E. Quandt, (North-Holland Publ. Co., Amsterdam and London, 1972). *Journal of Econometrics*, 1 (4):399–401, December 1973.

Mishra, Santosh, Su, Liangjun, and Ullah, Aman. Semiparametric estimator of time series conditional variance. *Journal of Business & Economic Statistics*, 28(2):256–274, 2010.

Muller, Hans-Georg and Stadtmuller, Ulrich. Estimation of heteroscedasticity in regression analysis. *Ann. Statist.*, 15(2):610–625, 06 1987. doi: 10.1214/aos/1176350364.

Nelson, Daniel B. Conditional heteroskedasticity in asset returns: A new approach. *Econometrica: Journal of the Econometric Society*, pp. 347–370, 1991.

Shalev-Shwartz, Shai. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2012.

Shawe-Taylor, John and Cristianini, Nello. *Kernel methods for pattern analysis*. Cambridge university press, 2004.

Whittaker, E. T. and Robinson, G. *The calculus of observations: An introduction to numerical analysis*, chapter 11, pp. 285–316. Dover Publications, 1967.

Yu, Jia Yuan, Mannor, Shie, and Shimkin, Nahum. Markov decision processes with arbitrary reward processes. *Mathematics of Operations Research*, 34(3): 737–757, 2009.

Yu, K and Jones, MC. Likelihood-based local linear estimation of the conditional variance function. *Journal of the American Statistical Association*, 99(465):139–144, 2004.

Zhu, Liping, Dong, Yuexiao, and Li, Runze. Semiparametric estimation of conditional heteroscedasticity via single-index modeling. *Statistica Sinica*, 23(3):1215, 2013.