2021 American Control Conference (ACC)
New Orleans, USA, May 25-28, 2021

Discounted online Newton method for time-varying time series prediction

Dongsheng Ding, Jianjun Yuan, and Mihailo R. Jovanović

Abstract— **We develop an online convex optimization method for predicting time series based on streaming observations. We first approximate the evolution of time-varying autoregressive integrated moving average (ARIMA) processes and then propose a discounted online Newton method for estimating time-varying ARIMA time series. Under practical assumptions, we establish dynamic regret bounds that quantify the tracking performance of our algorithm. To verify the effectiveness and robustness of our method, we conduct experiments on prediction problems based on both artificial data and real-world COVID-19 data. To the best of our knowledge, we are the first to report a COVID-19 prediction that utilizes online learning.**

I. Introduction

Time series prediction studies how to use a model to predict the future based on previously collected observations [1]. Typical modeling schemes: Moving average (MA), Autoregressive (AR), and Integrated (I), have been used in parameter identification and signal prediction, e.g., stock price prediction [2], [3] and pandemic forecasting [4], [5]. Despite its broad applicability, most studies assume fixed underlying models and fit these models with pre-collected data under strong assumptions on noise and loss functions. It is natural to ask whether we can allow time-varying models and more practical assumptions.

In practice, models of time series data often appear to be time-varying. An example is given by COVID-19 time series data, where the effective production number R_0 [6] indicates stages of different pandemic periods. The virus spread changes with public health interventions, e.g., quarantine. For this case, it is crucial to develop prediction methods for time-varying models [6]–[8]. Similar instances include financial time series data [2], [9] and psychological phenomena [10]. Another practical concern is the data usage. In situations with streaming observations, e.g., stock market or COVID-19 pandemic, fitting a model to the entire (or batch) dataset collected in advance is no longer feasible. Instead, a prediction method using online streaming data is more appealing, and more suitable for large-scale datasets.

Our contribution: In this paper, we propose an online convex optimization algorithm to predict streaming time series using time-varying models. Specifically, we propose a discounted online Newton method for estimating time-varying autoregressive integrated moving average (ARIMA) models. Under mild assumptions on noise and loss functions, we establish dynamic regret bounds for tracking performance of our algorithm along the evolution of time-varying ARIMA processes. Our results encompass static regret bound for time-independent stationary models as a special case. Finally, we conduct experiments for prediction of COVID-19 cases to demonstrate the effectiveness and robustness. To the best of our knowledge, our work is the first to utilize online learning for COVID-19 prediction.

Related work: Prediction using time-varying models has been studied in references [2], [11]–[15]. By assuming Gaussian noise, standard estimation methods such as maximum likelihood [11], [14], [15] and least-squares [12] are useful tools for estimating time series models directly. However, these methods require collecting large datasets for offline training and are not applicable in the online setting. More closely related studies for online time series prediction are summarized in references [16], [17]. In reference [16], the authors proposed an online Newton step for ARMA time series prediction and provide a regret bound under weak assumptions on noise and loss functions. In reference [17], the result of reference [16] was extended to ARIMA time series prediction. However, in these studies, time-independent time series models were assumed and only static regret bounds were provided. Since static regret is designed to compare with the best-fixed model in hindsight, it is unsuitable when underlying model is changing over time. We utilize a dynamic regret method to address this challenge.

Paper outline: In Section II, we formulate the online ARIMA prediction problem. In Section III, we propose a discounted online Newton method for ARIMA prediction and provide performance guarantees. We prove our main theorem in Section IV and show computational results in Section V. Finally, we conclude the paper in Section VI.

II. Online ARIMA Prediction

We describe the ARIMA model in Section II-A and formulate the online prediction problem in Section II-B.

A. ARIMA Model

Let X_t be an observation of the time series at time t. The dth order difference of X_t is $\nabla^d X_t$ for $d \geq 1$, e.g., the 1st order difference $\nabla X_t = X_t - X_{t-1}$ and the 2nd order difference $\nabla^2 X_t = \nabla X_t - \nabla X_{t-1}$. In particular, $\nabla^0 X_t = X_t$ and X_t can written as $X_t = \nabla^d X_t + \sum_{k=0}^{d-1} \nabla^k X_{t-1}$.

An ARIMA(p, d, q) depicts the evolution of X_t through a linear combination of past dth order differences and noises,

$$\nabla^d X_t \;=\; \sum_{i=1}^{p} \alpha_t^i \, \nabla^d X_{t-i} \;+\; \sum_{j=1}^{q} \beta_t^j \, \epsilon_{t-j} \;+\; \epsilon_t \qquad (1)$$

Financial support from the National Science Foundation under Awards ECCS-1708906 and ECCS-1809833 is gratefully acknowledged.

D. Ding and M. R. Jovanović are with the Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA 90089. J. Yuan is with Expedia Group. E-mails: dongshed@usc.edu, yuanx270@umn.edu, mihailo@usc.edu

where $\alpha_t := (\alpha_t^1, \ldots, \alpha_t^p)$ and $\beta_t := (\beta_t^1, \ldots, \beta_t^q)$ are the unknown model parameters, (p, d, q) is the given model order, and ϵ_t is the zero-mean noise. At time t, the goal of time series prediction is to forecast X_t based on past observation. When (α_t, β_t) are known at time t, we predict the dth oder difference $\nabla^d X_t$ and X_t as

$$\nabla^d \hat{X}_t \;=\; \sum_{i=1}^{p} \alpha_t^i \, \nabla^d X_{t-i} \;+\; \sum_{j=1}^{q} \beta_t^j \, \epsilon_{t-j} \tag{2a}$$

$$\hat{X}_t(\alpha, \beta) \;=\; \nabla^d \hat{X}_t \;+\; \sum_{k=0}^{d-1} \nabla^k X_{t-1}. \tag{2b}$$

However, the model parameters (α_t, β_t) are often unknown and time-varying. We focus on this setting in this paper.

B. Online ARIMA Prediction

Let $(\hat{\alpha}_t, \hat{\beta}_t)$ be an estimate of (α_t, β_t). We view the time series prediction as an online learning game between a player and an adversary or environment. Before the game starts, the adversary fixes the model parameters (α_t, β_t) and noise ϵ_t, and then generates X_t following the ARIMA model (1). At time t, the player first predicts X_t as $\hat{X}_t(\hat{\alpha}_t, \hat{\beta}_t)$, next observes the true X_t, then suffers loss $\ell_t(\hat{\alpha}_t, \hat{\beta}_t)$,

$$\ell_t(\hat{\alpha}_t, \hat{\beta}_t) \;:=\; \ell_t\big(X_t, \hat{X}_t(\hat{\alpha}_t, \hat{\beta}_t)\big) \tag{3}$$

where $\nabla^d \hat{X}_t = \sum_{i=1}^{p} \hat{\alpha}_t^i \, \nabla^d X_{t-i} + \sum_{j=1}^{q} \hat{\beta}_t^j \, \epsilon_{t-j}$. The prediction loss $\ell_t(\hat{\alpha}_t, \hat{\beta}_t)$ depends on both the current prediction $(\hat{\alpha}_t, \hat{\beta}_t)$ and the evolution of X_t. To measure the prediction performance over T rounds, we utilize the dynamic regret $\mathcal{R}_T^{\epsilon}$ that compares the cumulative loss against the minimum,

$$\mathcal{R}_T^{\epsilon} \;=\; \sum_{t=1}^{T} \Big(\ell_t(\hat{\alpha}_t, \hat{\beta}_t) \;-\; \min_{\alpha,\, \beta} \, \ell_t(\alpha, \beta) \Big)$$

where minimization over α and β yields the optimal model that changes over time t. However, since the noise ϵ_t is not observable, the loss (3) is not computable and classical online learning algorithms cannot be used to minimize $\mathcal{R}_T^{\epsilon}$.

Instead, we use the improper learning [16] to approximate ARIMA(p, d, q) with ARIMA$(p + m, d, 0)$ by adding extra m model parameters and removing the noise, where $m \geq 0$ is a design parameter. Let $\theta_t := (\theta_t^1, \cdots, \theta_t^{p+m}) \in \mathbb{R}^{p+m}$ be the model parameter of ARIMA$(p + m, d, 0)$, and let $\hat{\theta}_t$ be the prediction of θ_t at time t. We predict X_t via

$$\hat{X}_t(\hat{\theta}_t) \;=\; \sum_{i=1}^{p+m} \hat{\theta}_t^i \, \nabla^d X_{t-i} \;+\; \sum_{k=0}^{d-1} \nabla^k X_{t-1} \tag{4}$$

where m is to be determined in analysis. The loss function for prediction $\hat{X}_t(\hat{\theta}_t)$ at time t becomes

$$\ell_t(\hat{\theta}_t) \;:=\; \ell_t\big(X_t, \hat{X}_t(\hat{\theta}_t)\big) \tag{5}$$

which is computable given past observations and predictions. We define a more practical dynamic regret $\mathcal{R}_T$,

$$\mathcal{R}_T \;=\; \sum_{t=1}^{T} \Big(\ell_t(\hat{\theta}_t) \;-\; \min_{\alpha,\, \beta} \, \ell_t(\alpha, \beta) \Big) \tag{6}$$

which effectively tracks prediction performance of the time-varying ARIMA model as we present in Section III.

III. Main Results

In Section III-A, we propose a discounted online Newton method for ARIMA prediction and, in Section III-B, we establish regret guarantees.

A. Discounted Online Newton Method

Let $\mathcal{S}$ be the domain of model parameter $\theta \in \mathbb{R}^{p+m}$. We present our ARIMA prediction method in Algorithm 1. At time t, the player predicts $\hat{X}_t(\hat{\theta}_t)$ using (4) based on history and current model parameter $\hat{\theta}_t$, and then suffers loss $\ell_t(\hat{\theta}_t)$ after X_t is revealed. We next compute gradient $\nabla \ell_t(\hat{\theta}_t)$, approximate the Hessian $\nabla^2 \ell_t(\hat{\theta}_t)$ by matrix P_t, and update $\hat{\theta}_{t+1}$ via a Newton-type step and a P_t-induced projection to the domain $\mathcal{S}$. The projection under the norm induced by $P \succ 0$ is $\Pi_{\mathcal{S}}^{P}(x) := \operatorname{argmin}_{y \in \mathcal{S}} \|y - x\|_P^2$.

Algorithm 1 Discounted Online Newton Step (D-ONS)

1: **Input**: (p, d, q), G, D, ρ, m, ϵ, η, $\gamma \in (0, 1)$
2: **Initialization**: $\hat{\theta}_1 \in \mathbb{R}^{p+m}$, and $P_0 = \epsilon I_{(p+m)\times(p+m)}$
3: **for** time $t = 1, \ldots, T$ **do**
4: Predict $\hat{X}_t(\hat{\theta}_t)$ as

$$\hat{X}_t(\hat{\theta}_t) \;=\; \sum_{i=1}^{p+m} \hat{\theta}_t^i \, \nabla^d X_{t-i} \;+\; \sum_{k=0}^{d-1} \nabla^k X_{t-1}.$$

5: Observe X_t and suffer loss $\ell_t(\hat{\theta}_t) = \ell_t\big(X_t, \hat{X}_t(\hat{\theta}_t)\big)$.
6: Compute $\nabla_t := \nabla \ell_t(\hat{\theta}_t)$ and update P_t via

$$P_t \;=\; (1 - \gamma) \, P_0 \;+\; \gamma \, P_{t-1} \;+\; \nabla_t \nabla_t^{\top} \tag{7}$$

7: Update $\hat{\theta}_{t+1}$ via

$$\hat{\theta}_{t+1} \;=\; \Pi_{\mathcal{S}}^{P_t}\Big(\hat{\theta}_t \;-\; \frac{1}{\eta} \, P_t^{-1} \, \nabla_t \Big). \tag{8}$$

8: **end for**

Let $\nabla_t := \nabla \ell_t(\hat{\theta}_t)$. We approximate the Hessian of $\ell_t(\hat{\theta}_t)$ by (7) using a sum of current estimate $\nabla_t \nabla_t^{\top}$ and a convex combination of P_0 and P_{t-1}. Equivalently, we express (7) as a combination of initial $P_0 \succ 0$ and discounted history,

$$P_t \;=\; P_0 + \sum_{s=1}^{t} \gamma^{t-s} \nabla_s^{\top} \nabla_s.$$

This discounting scheme ensures invertability of P_t in (8) at any time and makes the algorithm numerically more stable than the discounting method [18]. The inverse P_t^{-1} can be efficiently updated using the Sherman-Morrison formula.

B. Dynamic Regret Bound

Assumption 1 (The ARIMA(p, d, q) model): (i) The zero-mean noises ϵ_t are generated independently, $\mathbb{E}\left[|\epsilon_t|\right] \leq M < \infty$ and $\ell_t(X_t, X_t - \epsilon_t) < \infty$; (ii) The coefficients $\{\alpha_t^i\}_{i=1}^{p}$ satisfy $|\alpha_t^i| \leq 1$ for any i and any time t; (iii) The coefficients $\{\beta_t^j\}_{j=1}^{q}$ satisfy $\sum_{j=1}^{q} |\beta_t^j| \leq 1 - \xi$ for $\xi \in (0, 1)$.

Assumption 2 (The loss function): (i) The domain is $\mathcal{S} = \{\theta \in \mathbb{R}^{p+m} \mid |\theta_i| \leq 1, i = 1, \cdots, p+m\}$ with diameter $D := \sup_{\theta,\theta' \in \mathcal{S}} \|\theta - \theta'\|_2 \leq 2\sqrt{p+m}$; (ii) The loss $\ell_t(\cdot)$ is Lipschitz continuous with parameter $L > 0$ and ρ-exp-concave with $\rho > 0$; (iii) The gradient of $\ell_t(\cdot)$ satisfies $\|\nabla \ell_t(\theta)\| \leq G$ for all $\theta \in \mathcal{S}$; (iv) There exists $V \geq 0$ such that $\sum_{t=2}^{T} \|\phi_t - \phi_{t-1}\| \leq V$, where $\phi_t = \operatorname{argmin}_{\theta \in \mathcal{S}} \ell_t(\theta)$.

Assumption 1 is mild since we still allow the noise to be adversarial and the ARIMA(p,d,q) models to be time-varying. Assumption 2 (i) follows Assumption 1 (ii,iii). A ρ-exp concave $\ell_t(\theta)$ in Assumption 2 (ii) makes sure that $\exp\{-\rho \ell_t(\theta)\}$ is concave in $\theta \in \mathcal{S}$, e.g., the quadratic loss satisfies Assumption 2 (ii, iii). In Assumption 2 (iv), boundedness is required on prediction variations in hindsight, which is standard in dynamic regret analysis [19].

Theorem 1 (Dynamic Regret Bound): Let Assumptions 1 and 2 hold. We set $\eta \leq (1/2)\min(1/(4GD), \rho)$, $\epsilon > 0$, and $m = q\log_{1-\xi}(1/(TML))$ in Algorithm 1. Then,

$$\mathcal{R}_T \;\leq\; -b_1 T \log\gamma \;-\; b_1 \log(1-\gamma) \;+\; \frac{b_2}{1-\gamma} V \;+\; b_3 \quad (9)$$

where $b_1 = (p+m)/(2\eta)$, $b_2 = 2\eta D(\epsilon + G^2)$, and $b_3 = b_1 \log(1 + G^2/\epsilon) + \eta D^2 \epsilon/2 + C$, C is an absolute constant.

Theorem 1 shows that the dynamic regret for predicting ARIMA(p,d,q) via ARIMA$(p+m,d,0)$ is upper bounded by an instance-dependent quantity. The constant V bounds the path length of the comparison sequences $\sum_{t=2}^{T} \|\phi_t - \phi_{t-1}\| \leq V$ where $\phi_t = \operatorname{argmin}_{\theta \in \mathcal{S}} \ell_t(\theta)$. Setting different discounting factors γ yields interesting special cases of the regret bound (9). We elaborate them as follows.

Remark 1: When the path length V is unknown, we can take $\gamma = 1 - T^{-s}$ for $s \in (0,1)$. Thus, we have $-T\log\gamma = -T\log(1-T^{-s}) \leq T^{1-s}/(1-T^{-s}) = O(T^{1-s})$, where the inequality $-\log(1-x) \leq x/(1-x)$ for $0 \leq x < 1$ is used. Hence, (9) reduces to

$$\mathcal{R}_T \;\leq\; O\big(T^{1-s} \;+\; s\log T \;+\; T^s V\big)$$

which scales as $O(T^{1-s} + T^s V)$. When $V = 0$, $\mathcal{R}_T$ defines a static regret $O(T^{1-s})$. Thus, for $s \in (0,1)$, both static and dynamic regrets are sublinear if $V < O(T)$.

Remark 2: When the path length V is known, we can take

$$\gamma \;=\; 1 \;-\; \frac{1}{2}\sqrt{\frac{\max(V, (\log^2 T)/T)}{2DT}}.$$

Similarly, (9) simplifies to

$$\mathcal{R}_T \;\leq\; \max\big(O(\log T),\, O(\sqrt{TV})\big)$$

which scales as $O(\sqrt{T(1+V)})$. By setting $V = 0$, the logarithmic static regret in [17] is obtained as a special case. Even if V is unknown, it is straightforward to employ the meta-optimization method [20] to achieve the same bound.

IV. Proof of Theorem 1

In Algorithm 1, we present an online Newton step in minimizing loss $\ell_t(\theta)$ for ARIMA$(p+m,d,0)$ prediction. We can apply the online optimization method [18] to show the following regret bound; see Appendix A for a proof.

Lemma 2: Let Assumption 2 hold. We set $\eta \leq (1/2)\min(1/(4GD), \rho)$ and $\epsilon > 0$ in Algorithm 1. For the regret $\mathcal{R}_T^0 := \sum_{t=1}^{T} \ell_t(\hat{\theta}_t) - \sum_{t=1}^{T} \min_{\theta \in \mathcal{S}} \ell_t(\theta)$, we have

$$\mathcal{R}_T^0 \;\leq\; -a_1 T \log\gamma - a_1 \log(1-\gamma) + \frac{a_2}{1-\gamma} V + a_3 \quad (10)$$

where $a_1 = (p+m)/(2\eta)$, $a_2 = 2\eta D(\epsilon + G^2)$, and $a_3 = ((p+m)/(2\eta))\log\big(1 + G^2/\epsilon\big) + \eta D^2 \epsilon/2$ are constants.

However, there is a discrepancy between the regret $\mathcal{R}_T^0$ and the desired regret $\mathcal{R}_T$ in (6). To fill in the gap, the rest is to study the difference between loss functions $\ell_t(\alpha,\beta)$for ARIMA(p,d,q) and $\ell_t(\theta)$ for ARIMA$(p+m,d,0)$. Towards this objective, we first introduce two auxiliary sequences using different amount of history information. As defined by (1), $\nabla^d X_t$ evolves as an ARMA(p,q). We can define a time series $\nabla^d X_t^\infty$ with parameters (α_t, β_t),

$$\nabla^d X_t^\infty \;=\; \sum_{i=1}^{p} \alpha_t^i \nabla^d X_{t-i} + \sum_{j=1}^{q} \beta_t^j \big(\nabla^d X_{t-j} - \nabla^d X_{t-j}^\infty\big) \quad (11a)$$

$$X_t^\infty(\alpha_t,\beta_t) \;=\; \nabla^d X_t^\infty \;+\; \sum_{k=1}^{d-1} \nabla^k X_{t-1} \quad (11b)$$

where $\nabla^d X_1^\infty = \nabla^d X_1$. For the prediction X_t^∞ at time t, the suffered loss is given by

$$\ell_t^\infty(\alpha_t,\beta_t) \;=\; \ell_t(X_t, X_t^\infty(\alpha_t,\beta_t)). \quad (12)$$

With appropriate coefficients $c_t^i(\alpha_t,\beta_t)$, we can express $X_t^\infty(\alpha_t,\beta_t)$ as a linear combination of all past history,

$$\nabla^d X_t^\infty(\alpha_t,\beta_t) = \sum_{i=1}^{t-1} c_t^i(\alpha_t,\beta_t)\, \nabla^d X_{t-i}.$$

Instead of using whole history, it is efficient to predict using only the most recent $p+m$ observations. Fix $m \in \mathbb{N}$, we define another time series $\nabla^d X_t^m$ with parameters (α_t,β_t),

$$\nabla^d X_t^m \;=\; \sum_{i=1}^{p} \alpha_t^i \nabla^d X_{t-i} + \sum_{j=1}^{q} \beta_t^j \big(\nabla^d X_{t-j} - \nabla^d X_{t-j}^{m-j}\big) \quad (13a)$$

$$X_t^m(\alpha_t,\beta_t) \;=\; \nabla^d X_t^m \;+\; \sum_{k=1}^{d-1} \nabla^k X_{t-1} \quad (13b)$$

where $\nabla^d X_t^s = \nabla^d X_t$ for all t and $s \leq 0$. For the prediction X_t^m at time t, the suffered loss is given by

$$\ell_t^m(\alpha_t,\beta_t) \;=\; \ell_t(X_t, X_t^m(\alpha_t,\beta_t)) \quad (14)$$

Let $(\alpha_t^\star, \beta_t^\star) := \operatorname{argmin}_{\alpha,\beta} \mathbb{E}[\ell_t(\alpha,\beta)]$. By (5) and (14), we have Lemma 3.

Lemma 3: Let Assumption 1 hold. Then,

$$\sum_{t=1}^{T} \min_{\theta \in \mathcal{S}} \ell_t(\theta) \;\leq\; \sum_{t=1}^{T} \ell_t^m(\alpha_t^\star, \beta_t^\star).$$

Proof: At time t, we can set $\hat{\theta}_t^i = c_t^i(\alpha_t^\star, \beta_t^\star)$ for the

loss (5). By (14), we have

$$\ell_t(\hat{\theta}_t) \;=\; \ell_t^m(\alpha_t^\star, \beta_t^\star) \text{ for any } t = 1, \cdots, T.$$

Clearly, $\min_{\theta \in \mathcal{S}} \ell_t(\theta) \leq \ell_t(\hat{\theta}_t)$. Summing up from $t = 1$ to $t = T$ completes the proof. ■

Next, we connect the loss $\ell_t^m(\cdot, \cdot)$ to $\ell_t(\cdot, \cdot)$ via $\ell_t^\infty(\cdot, \cdot)$ in Lemmas 4 and 5. Their proofs are provided in Appendix B.

Lemma 4: Let Assumptions 1 and 2 hold. Then,

$$\left| \sum_{t=1}^{T} \mathbb{E}\left[\ell_t^\infty(\alpha_t^\star, \beta_t^\star)\right] - \sum_{t=1}^{T} \mathbb{E}\left[\ell_t(\alpha_t^\star, \beta_t^\star)\right] \right| \;=\; O(1).$$

Lemma 5: Let Assumptions 1 and 2 hold. Fix $m = q \log_{1-\xi}((TML)^{-1})$. Then,

$$\left| \sum_{t=1}^{T} \mathbb{E}\left[\ell_t^\infty(\alpha_t^\star, \beta_t^\star)\right] - \sum_{t=1}^{T} \mathbb{E}\left[\ell_t^m(\alpha_t^\star, \beta_t^\star)\right] \right| \;=\; O(1).$$

We now combine above lemmas to establish a regret bound for $\mathcal{R}_T$. By Lemma 3, we first obtain a lower bound for $\mathcal{R}_T^0$,

$$\sum_{t=1}^{T} \ell_t(\hat{\theta}_t) - \sum_{t=1}^{T} \ell_t^m(\alpha_t^\star, \beta_t^\star) \;\leq\; \mathcal{R}_T^0. \tag{15}$$

Combining Lemmas 4 and 5 leads to

$$\sum_{t=1}^{T} \mathbb{E}\left[\ell_t^m(\alpha_t^\star, \beta_t^\star)\right] \;=\; \sum_{t=1}^{T} \mathbb{E}\left[\ell_t(\alpha_t^\star, \beta_t^\star)\right] + O(1). \tag{16}$$

Note that $\mathbb{E}[\ell_t^m(\alpha_t^\star, \beta_t^\star)] = \ell_t^m(\alpha_t^\star, \beta_t^\star)$. Finally, substituting (16) and the upper bound (10) into (15) leads to (9).

V. Computational Experiments

We use different datasets to examine the effectiveness and robustness of Algorithm 1 (or D-ONS).

A. Synthetic Data

We generate observations using a time-varying ARIMA model with $d = 1$, and $\alpha_t = (0.6, -0.5, 0.4, 0.4, 0.3), \beta_t = (0.3, 0.2)$ for $t \leq 1000$, and $\alpha_t = (-0.4, -0.5, 0.4, 0.4, 0.1)$, $\beta_t = (-0.3, 0.2)$ for $t > 1000$. The noise terms $\epsilon_t \sim$ Unif$[-0.1, 0.1]$. We run D-ONS with $m = 10$, $T = 2000$, where G, D, and ρ are computed according to Assumption 2. We use the quadratic loss function, display cumulative losses of our D-ONS with different discount factor γ in Fig. 1, and compare them with OGD [17]. The lines show averaged cumulative losses resulting from 12 experiments for each algorithm and the same ARIMA process. We observe that cumulative losses of our D-ONS grow much slower than OGD. By tuning γ, our algorithm with $\gamma = 0.5$ performs better than ARIMA-ONS [17] (D-ONS with $\gamma \approx 1$).

B. Real-world Data of US COVID-19 Cases

We test our D-ONS for predicting COVID-19 cases in the US. We use a dataset from COVID-19 Daily Cases, Deaths, and Hospitalizations [21]. In Fig. 2, we display our prediction results for total number of cases and total deaths in 300 days (from 1/23/2020 to 11/15/2020) for NYC. For the quadratic loss function, we run D-ONS with $m = 10$, $d = 1$, $T = 300$, $\gamma = 0.5$, where G, D, ρ are computed according to Assumption 2.

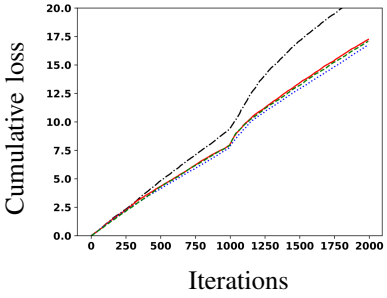

Fig. 1: Performance comparison: OGD [17] (–), D-ONS with: $\gamma = 0.98$ (– –), $\gamma = 0.5$ (····), and $\gamma = 0.1$ (– –). Slow loss growth better performance. We have simulated a time-varying ARIMA model with a jump change.

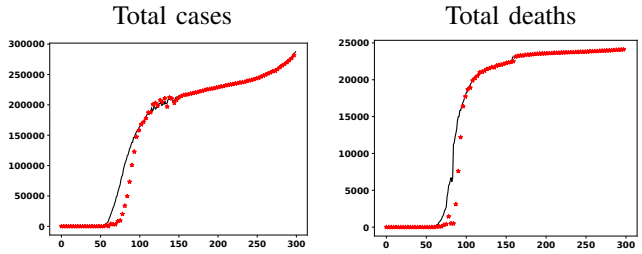

Fig. 2: Prediction of COVID-19 total cases and total deaths for NYC from 1/23/2020 to 11/15/2020: real observation (– –) and D-ONS's prediction (★). Our D-ONS's prediction displays for every 3 days.

As we see in Fig. 2, predictions of our algorithm successfully track the spreading trend of coronavirus, signaling fast outbreak in May for NYC. Before the outbreak (roughly at 100 in Fig. 2), our predictions have delays to track the spreading behavior since cases in January to March were not disclosed. During/after the outbreak, our predictions match real observations, effectively and consistently for different states. We also see that our predictions are robust in quickly responding to any fluctuations. In contrast to studies in references [22], [23], our approach works in an online way, needs very few modeling parameters, and does not require any stationarity assumptions/verifications of the underlying infection process. Our approach is flexible to predict trends for different regions without extra designs [24]. Our observations made for NYC are also visible for other US states.

VI. Concluding Remarks

In this paper, we have developed a new online prediction method – discounted online Newton step (D-ONS) – for predicting time-varying ARIMA time series. For the first time, we provide dynamic regret analysis for non-stationary time series prediction. Our dynamic regret bound $\max(O(\log T), O(\sqrt{TV}))$ captures both the static regret $O(\log T)$ and the path length V of time-varying comparison sequences. We also empirically verify the effectiveness

and robustness of our method on both artificial and real-world datasets. Our paper appears to be the first to report COVID-19 predictions using online learning. Future directions include multi-step ahead prediction, optimal tuning of algorithmic parameters, and other COVID-19 datasets.

APPENDIX

A. Proof of Lemma 2

We begin with the Newton-type step (8). By the non-expansiveness of projection $\Pi_{\mathcal{S}}^{P_t}(\cdot)$, for any $\phi_t \in \mathcal{S}$, we have

$$\begin{aligned}\|\hat{\theta}_{t+1} - \phi_t\|_{P_t}^2 \;&\le\; \|\hat{\theta}_t - \tfrac{1}{\eta} P_t^{-1}\nabla_t - \phi_t\|_{P_t}^2 \\ &=\; \|\hat{\theta}_t - \phi_t\|_{P_t}^2 \;-\; \tfrac{2}{\eta}\nabla_t^\top\big(\hat{\theta}_t - \phi_t\big) \;+\; \tfrac{1}{\eta^2}\nabla_t^\top P_t^{-1}\nabla_t\end{aligned}$$

which implies that

$$\begin{aligned}&\nabla_t^\top(\hat{\theta}_t - \phi_t) \\ &\le\; \tfrac{1}{2\eta}\nabla_t^\top P_t^{-1}\nabla_t \;+\; \tfrac{\eta}{2}\Big(\|\hat{\theta}_t - \phi_t\|_{P_t}^2 - \|\hat{\theta}_{t+1} - \phi_t\|_{P_t}^2\Big).\end{aligned} \tag{17}$$

On the other hand, by induction, we show that $\|P_t\| \le \epsilon + \frac{G^2}{1-\gamma} := c_1$. Clearly, it is true for P_0. Assume that it is true for P_{t-1}. By (7) and $P_0 = \epsilon I$, we have

$$\|P_t\| \;\le\; (1-\gamma)\epsilon \;+\; \gamma\big(\epsilon + \tfrac{G^2}{1-\gamma}\big) \;+\; G^2 \;=\; \epsilon \;+\; \tfrac{G^2}{1-\gamma}.$$

Thus, a lower bound $\|\hat{\theta}_{t+1} - \phi_t\|_{P_t}^2$ is given by

$$\begin{aligned}\|\hat{\theta}_{t+1} - \phi_t\|_{P_t}^2 \;&=\; \|\hat{\theta}_{t+1} - \phi_{t+1}\|_{P_t}^2 \;+\; \|\phi_{t+1} - \phi_t\|_{P_t}^2 \\ &\quad + 2(\hat{\theta}_{t+1} - \phi_{t+1})^\top P_t(\phi_{t+1} - \phi_t) \\ &\ge\; \|\hat{\theta}_{t+1} - \phi_{t+1}\|_{P_t}^2 \;-\; 4Dc_1\|\phi_{t+1} - \phi_t\|.\end{aligned}$$

Substituting the above inequality into (17) yields

$$\begin{aligned}\nabla_t^\top(\hat{\theta}_t - \phi_t) \;&\le\; \tfrac{1}{2\eta}\nabla_t^\top P_t^{-1}\nabla_t \;+\; 2\eta D c_1\|\phi_{t+1} - \phi_t\| \\ &\quad + \tfrac{\eta}{2}\big(\|\hat{\theta}_t - \phi_t\|_{P_t}^2 - \|\hat{\theta}_{t+1} - \phi_{t+1}\|_{P_t}^2\big).\end{aligned}$$

Summing both sides of the above inequality from $t = 1$ to $t = T$ leads to

$$\begin{aligned}&\textstyle\sum_{t=1}^T \nabla_t^\top(\hat{\theta}_t - \phi_t) \\ &\le\; \tfrac{1}{2\eta}\textstyle\sum_{t=1}^T \nabla_t^\top P_t^{-1}\nabla_t \;+\; 2\eta D c_1 V + \tfrac{\eta}{2}\epsilon\|\hat{\theta}_1 - \phi_1\|^2 \\ &\quad + \tfrac{\eta}{2}\textstyle\sum_{t=1}^T \big(\hat{\theta}_t - \phi_t\big)^\top (P_t - P_{t-1})\big(\hat{\theta}_t - \phi_t\big)\end{aligned} \tag{18}$$

where we set set $\phi_{T+1} = \phi_T$ and apply $\sum_{t=2}^T \|\phi_t - \phi_{t-1}\| \le V$; we omit $-\|\hat{\theta}_{T+1} - \phi_{T+1}\|_{P_T}^2$. By (7) and $\gamma \in (0,1)$, $\eta\,(P_t - P_{t-1}) \le \eta\nabla_t\nabla_t^\top$. Hence, (18) becomes

$$\begin{aligned}&\textstyle\sum_{t=1}^T\big(\ell_t(\hat{\theta}_t) - \ell_t(\phi_t)\big) \\ &\le\; \tfrac{1}{2\eta}\textstyle\sum_{t=1}^T \nabla_t^\top P_t^{-1}\nabla_t \;+\; 2\eta D c_1 V \;+\; \tfrac{\eta}{2}D^2\epsilon.\end{aligned} \tag{19}$$

where we use the exp-concave property [19, Lemma 4.2].

The rest is to bound the right-hand side of (19). We note that $\nabla_t^\top P_t^{-1}\nabla_t = \langle P_t^{-1}, \nabla_t\nabla_t^\top\rangle$, $\nabla_t\nabla_t^\top \preceq P_t - \gamma P_{t-1}$, and

$$\begin{aligned}\nabla_t^\top P_t^{-1}\nabla_t \;&\le\; \langle P_t^{-1}, P_t - \gamma P_{t-1}\rangle \;\le\; \log\tfrac{|P_t|}{|\gamma P_{t-1}|} \\ &=\; \log\tfrac{|P_t|}{|P_{t-1}|} - (p+m)\log\gamma.\end{aligned}$$

where $p + m$ is the matrix dimension, and the second inequality is due to: $\langle A^{-1}, A - B\rangle \le \log\frac{|A|}{|B|}$ for any matrices $A \succeq B \succ 0$ (see [19, Lemma 4.5]). Hence,

$$\begin{aligned}&\textstyle\sum_{t=1}^T \nabla_t^\top P_t^{-1}\nabla_t \\ &\le\; \log|P_T| - (p+m)\log\epsilon - (p+m)T\log\gamma \\ &\le\; (p+m)\log\Big(1 + \tfrac{G^2}{\epsilon(1-\gamma)}\Big) - (p+m)T\log\gamma\end{aligned}$$

where we use $\|P_t\| \le c_1$ in the second inequality. Applying the above inequality to the right hand side of (19) yields

$$\begin{aligned}&\textstyle\sum_{t=1}^T\big(\ell_t(\hat{\theta}_t) - \ell_t(\phi_t)\big) \\ &\le\; -\tfrac{(p+m)T}{2\eta}\log\gamma \;+\; \tfrac{p+m}{2\eta}\log\big(1 + \tfrac{G^2}{\epsilon(1-\gamma)}\big) \\ &\qquad + 2\eta D\big(\epsilon + \tfrac{G^2}{1-\gamma}\big)V \;+\; \tfrac{\eta}{2}D^2\epsilon.\end{aligned}$$

We note $1 \le \frac{1}{1-\gamma}$, take appropriate constants a_1, a_2, and a_3, and set $\phi_t = \operatorname{argmin}_{\theta\in\mathcal{S}}\ell_t(\theta)$ to get the desired bound.

B. Proofs of Lemma 4 and Lemma 5

Proof: [Proof of Lemma 4] By Assumption 1, ϵ_t is independent of $\epsilon_1, \cdots, \epsilon_{t-1}$. Thus, the best prediction available at time t has loss at least $\ell_t(X_t, X_t - \epsilon_t)$ in expectation. The ideal ARIMA model in hindsight, i.e., the one that generated signals, has the same loss $\ell_t(X_t, X_t - \epsilon_t)$. By the Lipschitz continuity of $\ell_t(\cdot)$,

$$\begin{aligned}&\big|\ell_t^\infty(\alpha_t^\star, \beta_t^\star) - \ell_t(\alpha_t^\star, \beta_t^\star)\big| \\ &=\; \big|\ell_t(X_t, X_t^\infty(\alpha_t^\star, \beta_t^\star)) - \ell_t(X_t, X_t - \epsilon_t)\big| \\ &\le\; L\big|X_t^\infty(\alpha_t^\star, \beta_t^\star) - X_t + \epsilon_t\big| \\ &=\; L\big|\nabla^d X_t^\infty - \nabla^d X_t + \epsilon_t\big|\end{aligned}$$

where the second equality follows (11b).

We next show by induction that $\mathbb{E}[|\nabla^d X_t^\infty - \nabla^d X_t - \epsilon_t|]$ decays exponentially in t. Let U be a positive constant such that $\mathbb{E}[|\nabla^d X_t^\infty - \nabla^d X_t - \epsilon_t|] \le U$ for $1 \le t \le q$. Assume that $\mathbb{E}[|\nabla^d X_\tau^\infty - \nabla^d X_\tau - \epsilon_\tau|] \le U(1-\xi)^{\tau/q}$ for $q < \tau < t$, as the inductive basis. Next we show that $\mathbb{E}[|\nabla^d X_t^\infty - \nabla^d X_t - \epsilon_t|] \le U(1-\xi)^{t/q}$. Using $\nabla^d X_t^\infty$ in (11a) and $\nabla^d X_t$, we have

$$\begin{aligned}&\mathbb{E}[|\nabla^d X_t - \nabla^d X_t^\infty - \epsilon_t|] \\ &=\; \mathbb{E}\big[\big|\textstyle\sum_{j=1}^q \beta_t^{j,\star}\big(\nabla^d X_{t-j}^\infty - \nabla^d X_{t-j} - \epsilon_{t-j}\big)\big|\big] \\ &\le\; \textstyle\sum_{j=1}^q |\beta_t^{j,\star}|\mathbb{E}\big[\big|\nabla^d X_{t-j}^\infty - \nabla^d X_{t-j} - \epsilon_{t-j}\big|\big] \\ &\le\; \textstyle\sum_{j=1}^q |\beta_t^{j,\star}| U(1-\xi)^{(t-j)/q} \\ &\le\; \textstyle\sum_{j=1}^q |\beta_t^{j,\star}| U(1-\xi)^{(t-q)/q} \\ &\le\; U(1-\xi)^{t/q}\end{aligned} \tag{20}$$

where the second inequality follows the induction basis and the last inequality is due to Assumption 1 (iii).

Therefore,

$$\begin{aligned}&\big|\mathbb{E}\big[\textstyle\sum_{t=1}^T \ell_t^\infty(\alpha_t^\star, \beta_t^\star) \;-\; \sum_{t=1}^T \ell_t(\alpha_t^\star, \beta_t^\star)\big]\big| \\ &\le\; \big|\textstyle\sum_{t=1}^T L\mathbb{E}[|\nabla^d X_t^\infty - \nabla^d X_t + \epsilon_t|]\big| \;\le\; O(1)\end{aligned}$$

which concludes our proof. ∎

Proof: [Proof of Lemma 5] By the Lipschitz continuity,

$$\begin{aligned}
&\left|\ell_t^{\infty}(\alpha_t^{\star},\beta_t^{\star}) - \ell_t^{m}(\alpha_t^{\star},\beta_t^{\star})\right| \\
&= \left|\ell_t(X_t, X_t^{\infty}(\alpha_t^{\star},\beta_t^{\star})) - \ell_t(X_t, X_t^{m}(\alpha_t^{\star},\beta_t^{\star}))\right| \\
&\leq L\left|X_t^{\infty}(\alpha_t^{\star},\beta_t^{\star}) - X_t^{m}(\alpha_t^{\star},\beta_t^{\star})\right| \\
&= L|\nabla^d X_t^{\infty} - \nabla^d X_t^{m}|
\end{aligned}$$

where the inequality follows (11b) and (13b). For $m \in \{0,-1,\cdots,-(1-q)\}$, by (13a), $\nabla^d X_t^m = \nabla^d X_t$ and

$$\begin{aligned}
\left|\nabla^d X_t^m - \nabla^d X_t^{\infty}\right| &\leq \left|\nabla^d X_t - \nabla^d X_t^{\infty} - \epsilon_t\right| + |\epsilon_t| \\
&\leq 2M
\end{aligned} \tag{21}$$

where we employ $|\epsilon_t| \leq M$ from Assumption 1 (i) and the exponential decaying of $|\nabla^d X_t - \nabla^d X_t^{\infty} - \epsilon_t|$ from (20).

We next show by induction that $|\nabla^d X_t^m - \nabla^d X_t^{\infty}| \leq 2M(1-\xi)^{m/q}$. For the inductive basis, it is trivial for $m = 0$ from (21); for $m = 1,\ldots,q-1$, it can be verified as follows,

$$\begin{aligned}
&|\nabla^d X_t^m - \nabla^d X_t^{\infty}| \\
&= \left|\textstyle\sum_{j=1}^{m} \beta_t^{j,\star}\left(\nabla^d X_{t-j}^{\infty} - \nabla^d X_{t-j}^{m-j}\right)\right| \\
&\quad + \left|\textstyle\sum_{j=m+1}^{q} \beta_t^{j,\star}\left(\nabla^d X_{t-j}^{\infty} - \nabla^d X_{t-j}^{m-j}\right)\right| \\
&\leq \textstyle\sum_{j=1}^{m} |\beta_t^{j,\star}|\left|\nabla^d X_{t-j}^{\infty} - \nabla^d X_{t-j}^{m-j}\right| \\
&\quad + \textstyle\sum_{j=m+1}^{q} |\beta_t^{j,\star}|\left|\nabla^d X_{t-j}^{\infty} - \nabla^d X_{t-j}\right| \\
&\leq 2M\left(\textstyle\sum_{j=1}^{m} |\beta_t^{j,\star}|(1-\xi)^{(m-j)/q} + \sum_{j=m+1}^{q} |\beta_t^{j,\star}|\right) \\
&\leq 2M \textstyle\sum_{j=1}^{m} |\beta_t^{j,\star}|(1-\xi)^{(m-q)/q} \\
&\leq 2M(1-\xi)^{m/q}
\end{aligned}$$

where the first inequality follows the triangle inequality and the definition of $\nabla^d X_t^m$ for $m \leq 0$, and the last inequality is due to $1 \leq (1-\xi)^{(m-q)/q}$ for $1 \leq m \leq q-1$.

We now show the inductive step by assuming that $|\nabla^d X_\tau^{m'} - \nabla^d X_\tau^{\infty}| \leq 2M(1-\xi)^{m'/q}$ for $q \leq m' \leq m$ and $\tau < t$ and proving that $|\nabla^d X_t^m - \nabla^d X_t^{\infty}| \leq 2M(1-\xi)^{m/q}$.

$$\begin{aligned}
&|\nabla^d X_t^m - \nabla^d X_t^{\infty}| \\
&= \Big|\textstyle\sum_{j=1}^{q} \beta_t^{j,\star}\left(\nabla^d X_{t-j} - \nabla^d X_{t-j}^{m-j}\right) \\
&\qquad - \textstyle\sum_{j=1}^{q} \beta_t^{j,\star}\left(\nabla^d X_{t-j} - \nabla^d X_{t-j}^{\infty}\right)\Big| \\
&= \left|\textstyle\sum_{j=1}^{q} \beta_t^{j,\star}\left(\nabla^d X_{t-j}^{\infty} - \nabla^d X_{t-j}^{m-j}\right)\right| \\
&\leq 2M \textstyle\sum_{j=1}^{q} |\beta_t^{j,\star}|(1-\xi)^{(m-j)/q} \\
&\leq 2M \textstyle\sum_{j=1}^{q} |\beta_t^{j,\star}|(1-\xi)^{(m-q)/q} \\
&\leq 2M(1-\xi)^{m/q}
\end{aligned}$$

where the last inequality follows Assumption 1 (iii).

Therefore,

$$\begin{aligned}
&\left|\mathbb{E}\left[\textstyle\sum_{t=1}^{T} \ell_t^{\infty}(\hat{\alpha}_t,\hat{\beta}_t) - \sum_{t=1}^{T} \ell_t^{m}(\hat{\alpha}_t,\hat{\beta}_t)\right]\right| \\
&\leq \left|\textstyle\sum_{t=1}^{T} L\mathbb{E}[|\nabla^d X_t^{\infty} - \nabla^d X_t^{m}|]\right| \\
&\leq 2MTL(1-\xi)^{m/q}
\end{aligned}$$

which finishes proof by taking $m = q\log_{1-\xi}((TML)^{-1})$. ∎

REFERENCES

[1] P. J. Brockwell, R. A. Davis, and M. V. Calder, *Introduction to time series and forecasting*. Springer, 2002, vol. 2.

[2] Z. Cai, "Trending time-varying coefficient time series models with serially correlated errors," *J. Econom.*, vol. 136, no. 1, pp. 163–188, 2007.

[3] A. A. Ariyo, A. O. Adewumi, and C. K. Ayo, "Stock price prediction using the ARIMA model," in *UKSim-AMSS 16th International Conference on Computer Modelling and Simulation*, 2014, pp. 106–112.

[4] S. Roy, G. S. Bhunia, and P. K. Shit, "Spatial prediction of COVID-19 epidemic using ARIMA techniques in India," *Model. Earth Syst. Environ.*, pp. 1–7, 2020.

[5] R. K. Singh, M. Rani, A. S. Bhagavathula, R. Sah, A. J. Rodriguez-Morales, H. Kalita, C. Nanda, S. Sharma, Y. D. Sharma, A. A. Rabaan *et al.*, "Prediction of the COVID-19 pandemic for the top 15 affected countries: Advanced autoregressive integrated moving average (ARIMA) model," *JMIR Public Health Surveill.*, vol. 6, no. 2, p. e19115, 2020.

[6] A. L. Bertozzi, E. Franco, G. Mohler, M. B. Short, and D. Sledge, "The challenges of modeling and forecasting the spread of COVID-19," *Proc. Natl. Acad. Sci.*, vol. 117, no. 29, pp. 16 732–16 738, 2020.

[7] M. Kiamari, G. Ramachandran, Q. Nguyen, E. Pereira, J. Holm, and B. Krishnamachari, "COVID-19 risk estimation using a time-varying SIR-model," *arXiv preprint arXiv:2008.08140*, 2020.

[8] H. G. Hong and Y. Li, "Estimation of time-varying reproduction numbers underlying epidemiological processes: A new statistical tool for the COVID-19 pandemic," *PloS one*, vol. 15, no. 7, p. e0236464, 2020.

[9] D. Creala, S. J. Koopmanb, and A. Lucasc, "The estimation of time-varying parameters in multivariate linear time series models," 2011.

[10] J. M. Haslbeck, L. F. Bringmann, and L. J. Waldorp, "A tutorial on estimating time-varying vector autoregressive models," *Multivariate Behav. Res.*, pp. 1–30, 2020.

[11] G. Kitagawa and W. Gersch, "A smoothness priors time-varying AR coefficient modeling of nonstationary covariance time series," *IEEE Trans. Autom. Control*, vol. 30, no. 1, pp. 48–56, 1985.

[12] C. Grillenzoni, "Time-varying parameters prediction," *Ann. Inst. Stat. Math*, vol. 52, no. 1, pp. 108–122, 2000.

[13] R. Prado and G. Huerta, "Time-varying autoregressions with model order uncertainty," *J. Time Ser. Anal.*, vol. 23, no. 5, pp. 599–618, 2002.

[14] M. Ito, A. Noda, and T. Wada, "An alternative estimation method of a time-varying parameter model," *arXiv preprint arXiv:1707.06837*, 2017.

[15] W. Fei and L. Bai, "Time-varying moving average model for autocovariance nonstationary time series," *J. Fiber Bioeng. Inf.*, vol. 7, no. 1, pp. 53–65, 2014.

[16] O. Anava, E. Hazan, S. Mannor, and O. Shamir, "Online learning for time series prediction," in *Proceedings of the 26th Annual Conference on Learning Theory*, 2013, pp. 172–184.

[17] C. Liu, S. C. Hoi, P. Zhao, and J. Sun, "Online ARIMA algorithms for time series prediction," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 2016, pp. 1867–1873.

[18] J. Yuan and A. G. Lamperski, "Trading-off static and dynamic regret in online least-squares and beyond." in *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020, pp. 6712–6719.

[19] E. Hazan, "Introduction to online convex optimization," *Foundations and Trends in Optimization*, vol. 2, no. 3-4, pp. 157–325, 2016.

[20] L. Zhang, S. Lu, and Z.-H. Zhou, "Adaptive online learning in dynamic environments," in *Advances in neural information processing systems*, 2018, pp. 1323–1333.

[21] "Covid-19 daily cases, deaths, and hospitalizations," http://healthdata.gov/dataset/covid-19-daily-cases-deaths-and-hospitalizations, accessed: 2020-09-20.

[22] A. K. Sahai, N. Rath, V. Sood, and M. P. Singh, "ARIMA modelling & forecasting of COVID-19 in top five affected countries," *Diabetes Metab Syndr*, vol. 14, no. 5, pp. 1419–1427, 2020.

[23] Q. Yang, J. Wang, H. Ma, and X. Wang, "Research on COVID-19 based on ARIMA model δtaking hubei, china as an example to see the epidemic in Italy," *J. Infect. Public Health*, 2020.

[24] A. Hernandez-Matamoros, H. Fujita, T. Hayashi, and H. Perez-Meana, "Forecasting of COVID19 per regions using ARIMA models and polynomial functions," *Appl. Soft Comput.*, vol. 96, p. 106610, 2020.