BRIEF: Bayesian Regression of Infinite Expert Forecasters for Single and Multiple Time Series Prediction

Ming Jin and Costas J. Spanos

***Abstract*— Bayesian Regression of Infinite Expert Forecasters (BRIEF) as proposed in the study is a prediction algorithm for time-varying systems. The method is based on regret minimization by tracking the performance of an inifinite pool of experts for single and multiple time series. The inverse correlation weighted error (ICWE) employed in BRIEF takes into account the dependency structure among multiple time series, which can also be adapted to multi-step ahead predictions. Theoretical bounds show that the cumulative regret grows at rate $O(\log T)$ with respect to the oracle that can select the best strategy in retrospect. As the per round regret vanishes, BRIEF is indistinguishable to the oracle when the horizon increases. Also since the bound applies to any choice of input subject to the euclidean norm constraint, the method can be applied to adversarial settings. Experimental results verify that BRIEF excels in single and multiple steps ahead prediction of ARMAX simulated data and building energy consumptions.**

I. Introduction

Consider the indexed class of models, $\mathcal{M} = \{Q_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$, where Θ is a compact convex set of a finite dimensional linear space, and assume for a time-varying system $\exists \boldsymbol{\theta}_t^* \in \Theta$ such that Q_t is reasonably well modeled by the linear model $Q_{\boldsymbol{\theta}_t^*}$, the task is to *predict* output y_t given input $\boldsymbol{x}_t$, i.e. $\hat{y}_t = \boldsymbol{x}_t^\top \boldsymbol{\theta}_t$, where $\boldsymbol{x}_t$ can be past values of y_t or exogenous variables $\{u_t\}$. Inasmuch as we are not limited in the selection of $\boldsymbol{x}_t \in \mathbb{R}^d$, we can deal with nonlinearity through the process of "lifting" by augmenting the set of predictors with nonlinear transformations of the features.

Due to the familiarity of the problem to adaptive control [1], algorithms such as stochastic gradient descent and pseudo linear regression can be applied, where convergence properties for time-varying systems are proven using Lyapunov functions [2]. The assumption on the time-varying parameter, nevertheless, is not suitable in the adverserial setting where the parameter is chosen by the opponent.

Online convex optimization (OCO) has been developed for sequential decision-making in the presence of time-varying uncertainty [3]. The loss of $\boldsymbol{\theta}$, also known as an "expert", up to time t is given by

$$L_t(\boldsymbol{\theta}) = \tfrac{1}{2} \sum_{i=1}^{t} \left(\boldsymbol{x}_i^\top \boldsymbol{\theta} - y_i \right)^2 . \tag{1}$$

The empirical loss incurred by the forecaster is $\hat{L}_t = \frac{1}{2} \sum_{i=1}^{t} \left(\hat{y}_i - y_i \right)^2$. To assess the performance we define the notion of **regret** with respect to the best possible expert $\boldsymbol{\theta}^* = \arg_{\boldsymbol{\theta}} \min L_t(\boldsymbol{\theta})$ in retrospect [3]:

$$R_t = \hat{L}_t - \min_{\boldsymbol{\theta}} L_t(\boldsymbol{\theta}) \tag{2}$$

Note that $\boldsymbol{\theta}^*$ is regarded as an "oracle" as it is not possible to know beforehand which expert performs the best until we have already seen all the trajectory. The *objective of time series prediction is thus to minimize the regret to be sublinear*, i.e. $o(t)$, so the per round regret is vanishing as t grows.

Previous works based on iterative minimization of a convex functional includes proximal point [4], mirror descent algorithm [5], and adptive online gradient descent [6]–[8]. Raginsky et. al. have shown the application to a generalization of classical adaptive control schemes and supervisory controller switching policy [9]. Mixture forecasters have been introduced by Vovk [10] in the optimization literature and Clarke and Barron [11] in the bayesian literature, and later developed by Kakade and Ng [12].

Bayesian Regression of Inifinite Expert Forecasters (BRIEF) belongs to the category of prediction methods with log loss using infinite families of probability distributions [13]. In addition to the closed form prediction which brings computational advantage, we have also proven that the regret grows at rate $O(\log T)$ as performance guarantee.

The rest of the paper is organized as follow. Section II-A and II-B introduce the BRIEF algorithm for single and multiple time series respectively, where the bounds on regret is proven in Section II-C. Section III reports experimental results on simulation and building energy prediction. We conclude in Section IV with future works.

II. BRIEF Methodology

Generally, BRIEF is a method of tracking the performance of "experts" in an *infinite pool* in order to make forecasts in a *systematic* way. In the following we first derive the algorithm for single and multiple time series (Section II-A and II-B), then we provide bounds on regret as a guarantee of performance (Section II-C).

This research is funded by the Republic of Singapore's National Research Foundation through a grant to the Berkeley Education Alliance for Research in Singapore (BEARS) for the Singapore-Berkeley Building Efficiency and Sustainability in the Tropics (SinBerBEST) Program. BEARS has been established by the University of California, Berkeley as a center for intellectual excellence in research and education in Singapore.

M. Jin and C. J. Spanos are with the Department of Electrical Engineering and Computer Sciences at the University of California Berkeley, USA. Emails: {jinming, spanos}@berkeley.edu

The authors would like to thank Peter Bartlett and Lin Zhang for the meaningful discussions and constructive comments for the paper.

A. BRIEF with Single Time Series

Compared to the mixture of experts model [14] whose forecast is based on a weighted sum of opinions from a finite set of experts, the approach of BRIEF bears similarities with Bayesian regression, such as Ridge regression, where the statistical analysis is undertaken by assuming prior distributions on the model's parameters.

Assume a prior on the distribution of experts $\boldsymbol{\theta} \in \mathbb{R}^d$, $\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, which is the initial weights imposed on the experts. The **transfer function** $\sigma(y, \boldsymbol{\theta} \cdot \boldsymbol{x}) = e^{-(\boldsymbol{\theta}\cdot\boldsymbol{x}-y)^2/2}$ relates the loss incurred at round t for expert $\boldsymbol{\theta}$ to its weight updates, as can be seen in $L_n(\boldsymbol{\theta}) = -\ln \prod_{t=1}^n \sigma(y_t, \boldsymbol{\theta}\cdot\boldsymbol{x}_t)$. The posterior density of experts at round t is thus

$$q_t(\boldsymbol{\theta}) = \frac{q_0(\boldsymbol{\theta})e^{-L_{t-1}(\boldsymbol{\theta})}}{\int q_0(\boldsymbol{v})e^{-L_{t-1}(\boldsymbol{v})}d\boldsymbol{v}} = \frac{q_0(\boldsymbol{\theta})\prod_{s=1}^{t-1}\sigma(y_s, \boldsymbol{\theta}\cdot\boldsymbol{x}_s)}{\int q_0(\boldsymbol{v})\prod_{s=1}^{t-1}\sigma(y_s, \boldsymbol{v}\cdot\boldsymbol{x}_s)d\boldsymbol{v}}, \tag{3}$$

where the cumulative loss is defined in (1). Intuitively, experts with poor performances (larger loss) are downweighed through the exponential updates. The following result provides the core of BRIEF as the expected output $\hat{y}_t$ given the input $\boldsymbol{x}_t$.

Theorem 1: For any prior distribution $\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, if the updates law follows (3), then given $\boldsymbol{x}_t$, the expectation of Y is given by

$$\mathbb{E}_{\hat{p}_t}[Y_t|\boldsymbol{x}_t] = \frac{\boldsymbol{x}_t^\top \boldsymbol{A}_t^{-1}\boldsymbol{a}_{t-1}}{1 - \boldsymbol{x}_t^\top \boldsymbol{A}_t^{-1}\boldsymbol{x}_t} \tag{4}$$

where $\boldsymbol{a}_t = \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \sum_{s=1}^t y_s\boldsymbol{x}_s$, $\boldsymbol{A}_t = \boldsymbol{\Sigma}_0^{-1} + \sum_{s=1}^t \boldsymbol{x}_s\boldsymbol{x}_s^\top$, and $\hat{p}_t(y, \boldsymbol{x}_t) = \int \sigma(y, \boldsymbol{\theta}\cdot\boldsymbol{x}_t)q_t(\boldsymbol{\theta})d\boldsymbol{\theta}$ is the projected distribution given $\boldsymbol{x}_t$.

Proof: Since the prior and transfer function form a conjugate pair, the posterior distribution is still Gaussian whose expectation is contained in the exponent:

$$\begin{aligned}
\hat{p}_t(y|\boldsymbol{x}_t) &= \int \sigma(y, u\cdot\boldsymbol{x}_t)q_t(\boldsymbol{\theta})d\boldsymbol{\theta} \\
&\propto \int e^{-\frac{1}{2}\left((\boldsymbol{\theta}\cdot\boldsymbol{x}_t - y)^2 + (\boldsymbol{\theta}-\boldsymbol{\mu}_0)^\top\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu}_0) + \sum_{s=1}^{t-1}(\boldsymbol{\theta}\cdot\boldsymbol{x}_s - y_s)^2\right)}d\boldsymbol{\theta} \\
&\propto e^{-\frac{1}{2}\left(y^2 + \sum_{s=1}^{t-1}y_s^2 - (y\boldsymbol{x}_t + \boldsymbol{a}_{t-1})^\top\boldsymbol{A}_t^{-1}(y\boldsymbol{x}_t + \boldsymbol{a}_{t-1})\right)} \\
&\propto e^{-\frac{1-\boldsymbol{x}_t^\top\boldsymbol{A}_t^{-1}\boldsymbol{x}_t}{2}\left(y - \frac{\boldsymbol{x}_t^\top\boldsymbol{A}_t^{-1}\boldsymbol{a}_{t-1}}{1-\boldsymbol{x}_t^\top\boldsymbol{A}_t^{-1}\boldsymbol{x}_t}\right)^2},
\end{aligned}$$

from which we have

$$Y_t \sim \mathcal{N}\left(\frac{\boldsymbol{x}_t^\top\boldsymbol{A}_t^{-1}\boldsymbol{a}_{t-1}}{1-\boldsymbol{x}_t^\top\boldsymbol{A}_t^{-1}\boldsymbol{x}_t}, (1-\boldsymbol{x}_t^\top\boldsymbol{A}_t^{-1}\boldsymbol{x}_t)^{-1}\right). \tag{5}$$

■

Remarks: It is interesting to point out the close connection of BRIEF with the well-known *Vovk-Azoury-Warmuth forecaster* [10], which predicts at time t with $\hat{\boldsymbol{w}}_t^\top\boldsymbol{x}_t$, where

$$\hat{\boldsymbol{w}}_t = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d}\left[\|\boldsymbol{\theta}\|^2 + \sum_{s=1}^{t-1}(\boldsymbol{\theta}^\top\boldsymbol{x}_s - y_s)^2 + (\boldsymbol{\theta}^\top\boldsymbol{x}_t)^2\right]. \tag{6}$$

Without the term $(\boldsymbol{\theta}^\top\boldsymbol{x}_t)^2$, the above formulation becomes the *ridge regression forecaster* [15]. The additional term can be viewed as $(\boldsymbol{\theta}^\top\boldsymbol{x}_t - 0)^2$ that shrinks the prediction towards 0. It is easy to derive the solution to (6) as $\boldsymbol{x}_t^\top\hat{\boldsymbol{w}}_t = \boldsymbol{x}_t^\top\boldsymbol{A}_t^{-1}\boldsymbol{a}_{t-1}$, which differs from BRIEF by the scaling factor. Instead of arbitrarily shrinking towards 0, BRIEF keeps track of the posterior distribution and shrinks towards the mean, which makes BRIEF a more sensible approach to adopt.

B. BRIEF with Multiple Time Series

Consider the prediction problem with multiple time series $\{\boldsymbol{y}_t\}$, $\boldsymbol{y}_t \in \mathbb{R}^M$. One might straightforwardly apply the forecaster in the previous section to each series independently. For highly correlated time series, this amounts to using the cumulative loss as the independent sum of individual losses, $\sum_{i=1}^t(\boldsymbol{X}_i^\top\boldsymbol{\theta} - \boldsymbol{y}_i)^\top I(\boldsymbol{X}_i^\top\boldsymbol{\theta} - \boldsymbol{y}_i)$, where

$$\boldsymbol{X}_i^\top = \begin{bmatrix} \boldsymbol{x}_{1,i}^\top & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{x}_{2,i}^\top & \ddots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{x}_{M,i}^\top \end{bmatrix} \tag{7}$$

is the input matrix augmented with input vector $\boldsymbol{x}_{j,i}$ for series j at time i, $\boldsymbol{\theta} = \begin{bmatrix}\boldsymbol{\theta}_1^\top & \cdots & \boldsymbol{\theta}_M^\top\end{bmatrix}$ is the augmented expert vector, and I is the identity matrix.

Motivated by the maximum likelihood estimation of multivariate Gaussian, the following loss function is designed,

$$L_t(\boldsymbol{\theta}) = \tfrac{1}{2}\sum_{i=1}^t(\boldsymbol{X}_i^\top\boldsymbol{\theta} - \boldsymbol{y}_i)^\top\boldsymbol{\Gamma}^{-1}(\boldsymbol{X}_i^\top\boldsymbol{\theta} - \boldsymbol{y}_i), \tag{8}$$

where the prediction error for each time series is weighted by the inverse correlation matrix $\boldsymbol{\Gamma}^{-1}$ to capture the correlated structure.[1] The regret defined in (2) is the discrepancy of loss incurred by the forecaster with a group of oracle experts.

The following theorem provides with the BRIEF forecaster for multiple time series, where the transfer function

$$\sigma(\boldsymbol{y}, \boldsymbol{X}^\top\boldsymbol{\theta}) = e^{-\frac{1}{2}(\boldsymbol{X}^\top\boldsymbol{\theta}-\boldsymbol{y})^\top\boldsymbol{\Gamma}^{-1}(\boldsymbol{X}^\top\boldsymbol{\theta}-\boldsymbol{y})} \tag{9}$$

is used in the posterior weight updates of (3).

Theorem 2: For the prior distribution $\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ on the augmented expert $\boldsymbol{\theta} = \begin{bmatrix}\boldsymbol{\theta}_1^\top & \cdots & \boldsymbol{\theta}_M\end{bmatrix}$, and correlation matrix $\boldsymbol{\Gamma}$, if the updates law follows (3), then given $\boldsymbol{X}_t$ as (7), the expectation of $\boldsymbol{Y} \in \mathbb{R}^M$ is given by

$$\mathbb{E}_{\hat{p}_t}[\boldsymbol{Y}_t|\boldsymbol{X}_t] = \boldsymbol{B}_t^{-1}\boldsymbol{\Gamma}^{-1}\boldsymbol{X}_t^\top\boldsymbol{A}_t^{-1}\boldsymbol{a}_{t-1} \tag{10}$$

[1] The correlation matrix $\boldsymbol{\Gamma}$ has 1 on the diagonal and $\boldsymbol{\Gamma}_{ij} = \frac{Cov(Y_i, Y_j)}{\sqrt{Var(Y_i)Var(Y_j)}} \in [-1, 1]$ as the correlation between time series $\{Y_i\}$ and $\{Y_j\}$ as the off-diagonal entry. In practice $\boldsymbol{\Gamma}$ can be learned from training data or designed by experts to account for the correlation structures.

Algorithm 1: Pseudo-code of BRIEF

Init : Set $\boldsymbol{a}_0 = \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0$, $\boldsymbol{A}_0 = \boldsymbol{\Sigma}_0^{-1}$, $\boldsymbol{B}_0 = \boldsymbol{\Gamma}^{-1}$, where $\boldsymbol{\mu}_0$, $\boldsymbol{\Sigma}_0$, and $\boldsymbol{\Gamma}$ are hyperpriors provided by the users.

for $t = 1, \cdots, T$ **do** // round of predictions

- Observe $\boldsymbol{y}_{t-1} \in \mathbb{R}^M$
- $\boldsymbol{X}_t$ is revealed s.t. $\lambda_{max}(\boldsymbol{X}_t)$ is bounded.
- Perform updates for $\boldsymbol{a}_{t-1}$, $\boldsymbol{A}_t$, $\boldsymbol{B}_t$ as in Theorem 2.
- Predict $\hat{y}_t = \boldsymbol{B}_t^{-1}\boldsymbol{\Gamma}^{-1}\boldsymbol{X}_t^\top \boldsymbol{A}_t^{-1}\boldsymbol{a}_{t-1}$

where

$$\begin{aligned}\boldsymbol{a}_t &= \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \sum_{s=1}^{t} \boldsymbol{X}_s\boldsymbol{\Gamma}^{-1}\boldsymbol{y}_s \\ \boldsymbol{A}_t &= \boldsymbol{\Sigma}_0^{-1} + \sum_{s=1}^{t} \boldsymbol{X}_s\boldsymbol{\Gamma}^{-1}\boldsymbol{X}_s^\top \\ \boldsymbol{B}_t &= \boldsymbol{\Gamma}^{-1} - \boldsymbol{\Gamma}^{-1}\boldsymbol{X}_t^\top \boldsymbol{A}_t^{-1}\boldsymbol{X}_t\boldsymbol{\Gamma}^{-1},\end{aligned}$$

and $\hat{p}_t(\boldsymbol{y}|\boldsymbol{x}_t) = \int \sigma(\boldsymbol{y}, \boldsymbol{X}^\top\boldsymbol{\theta}) q_t(\boldsymbol{\theta}) d\boldsymbol{\theta}$ is the projected distribution given $\boldsymbol{X}_t$.

Proof: The proof is similar to that of the single time series, using the fact that affine transformation of Gaussian random variable is still Gaussian. For presentation let $\|\boldsymbol{x}\|_{\boldsymbol{A}}^2 = \boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x}$, the key steps involves:

$$\begin{aligned}\hat{p}_t(\boldsymbol{y}|\boldsymbol{x}_t) &= \int \sigma(\boldsymbol{y}, \boldsymbol{X}^\top\boldsymbol{\theta}) q_t(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &\propto \int e^{-\frac{1}{2}\left(\|\boldsymbol{X}_t^\top\boldsymbol{\theta} - \boldsymbol{y}\|_{\Gamma^{-1}}^2 + \|\boldsymbol{\theta} - \boldsymbol{\mu}_0\|_{\Sigma_0^{-1}}^2 + \sum_{s=1}^{t-1} \|\boldsymbol{X}_s^\top\boldsymbol{\theta} - \boldsymbol{y}_s\|_{\Gamma^{-1}}^2 \right)} d\boldsymbol{\theta} \\ &\propto e^{-\frac{1}{2}\left(\|\boldsymbol{y}\|_{\boldsymbol{\Gamma}^{-1}}^2 + \sum_{s=1}^{t-1} \|\boldsymbol{y}_s\|_{\boldsymbol{\Gamma}^{-1}}^2 - \|\boldsymbol{X}_t\boldsymbol{\Gamma}^{-1}\boldsymbol{y} + \boldsymbol{a}_{t-1}\|_{\boldsymbol{A}_t^{-1}}^2 \right)} \\ &\propto e^{-\frac{1}{2}\left\| \boldsymbol{B}_t^{1/2}\boldsymbol{y} - \boldsymbol{B}_t^{-1/2}\boldsymbol{\Gamma}^{-1}\boldsymbol{X}_t^\top\boldsymbol{A}_t^{-1}\boldsymbol{a}_{t-1} \right\|^2},\end{aligned}$$

from which we have

$$\boldsymbol{Y}_t \sim \mathcal{N}(\boldsymbol{B}_t^{-1}\boldsymbol{\Gamma}^{-1}\boldsymbol{X}_t^\top\boldsymbol{A}_t^{-1}\boldsymbol{a}_{t-1}, \boldsymbol{B}_t^{-1}). \tag{11}$$

■

The implementation of BRIEF with single and multiple time series (BRIEF-S, BRIEF-M) is shown in Algorithm 1, where the choice of $\boldsymbol{\Gamma} = I$ (identity matrix) for BRIEF-M is equivalent to BRIEF-S when applied independently to individual signals.

Remarks: The inverse correlation weighted error (ICWE) metric proposed in the study is an extension of the mean squared error (MSE) in multiple time series to correlated structures [16]. Potential applications include energy prediction of buildings in the neighborhood that are subject to the same set of exogenous inputs not captured in the model, or prediction with extended horizon where each step is highly correlated with each other. See Section III for the experimental results.

C. Performance Bounds on Regret

The regret as in (2) compares the forecaster with the "oracle forecaster", which chooses its strategy, $\boldsymbol{\theta}^*$, in retrospect once all the $\{y_t\}$ are seen. In the task of prediction, it is impossible to see the values in the future. We provide bounds on the regret as performance guarantees for BRIEF with single and multiple time series as follow.

Proposition 1 (KL divergence for Gaussians [17]): The KL divergence, $D(p_0 \parallel p_1)$, between two Gaussian distributions p_0, p_1 parameterized by mean $\boldsymbol{\mu}_0, \boldsymbol{\mu}_1$ and covariance matrices $\boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_1$ is given by:

$$D(p_0 \parallel p_1) = \frac{1}{2}\left(\ln\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_0|} - d + \mathrm{tr}(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_0) + \|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\|_{\boldsymbol{\Sigma}_1^{-1}}^2 \right) \tag{12}$$

where we use the shorthand notation

$$\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\|_{\boldsymbol{\Sigma}_1^{-1}}^2 = (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_1^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1) \tag{13}$$

for the rest of the paper.

Lemma 1 (Maximum Differential Entropy [17]): For any multivariate density q on $\mathbb{R}^d$ with zero mean $\int \boldsymbol{\theta} q(\boldsymbol{\theta}) d\boldsymbol{\theta} = 0$ and covariance matrix $\boldsymbol{\Sigma}$, the differential entropy $h(q) = -\int q(\boldsymbol{\theta}) \ln q(\boldsymbol{\theta}) d\boldsymbol{\theta}$ is maximized at $\frac{d}{2}\ln(2\pi e) + \frac{1}{2}\ln\det(\boldsymbol{\Sigma})$, which is achieved by the multivariate normal density with covariance matrix $\boldsymbol{\Sigma}$.

Lemma 2 (Descent Lemma [18]): The following holds for $f \in C_L^1$ with Lipschitz continuous gradient, i.e. $\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2 \le L\|\boldsymbol{x} - \boldsymbol{y}\|_2$:

$$f(\boldsymbol{x}) \le f(\boldsymbol{y}) + \langle \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle + \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \tag{14}$$

The following theorems bounds the regret for BRIEF.

Theorem 3 (BRIEF with single time series): The regret of BRIEF, under the same conditions as Theorem 1, with respect to any expert $\boldsymbol{\theta}$ is bounded by:

$$\hat{L}_n - L_n(\boldsymbol{\theta}) \le c_1 + \frac{d}{2}\ln\left(\frac{\mathrm{tr}(\boldsymbol{\Sigma}_0^{-1})}{d} + \frac{nc}{d} \right) \tag{15}$$

where $c_1 = \frac{1}{2}\|\boldsymbol{\mu}_0 - \boldsymbol{\theta}\|_{\boldsymbol{\Sigma}_0^{-1}}^2 + \frac{1}{2}\ln|\boldsymbol{\Sigma}_0|$, $|\boldsymbol{\Sigma}_0|$ is the determinant of $\boldsymbol{\Sigma}_0$.

Proof: The proof is similar to that of Theorem 4. ■

Theorem 4 (BRIEF with multiple time series): For BRIEF with multiple time series forecasting as detailed in Theorem 2, the regret with respect to any expert $\boldsymbol{\theta}$ is bounded by

$$\hat{L}_n - L_n(\boldsymbol{\theta}) \le c_1 + \frac{d}{2}\ln\left(\frac{\mathrm{tr}(\boldsymbol{\Sigma}_0^{-1})}{d} + \frac{ncM\lambda_{max}(\boldsymbol{\Sigma}_{\boldsymbol{x}})}{d} \right) \tag{16}$$

where $c_1 = \frac{1}{2}\|\boldsymbol{\mu}_0 - \boldsymbol{\theta}\|_{\boldsymbol{\Sigma}_0^{-1}}^2 + \frac{1}{2}\ln|\boldsymbol{\Sigma}_0|$, λ_{max} is the upper bound on the eigenvalue of the feature matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \boldsymbol{X}_t\boldsymbol{X}_t^\top$, which is equal to the largest euclidean norm on the input vector $\max_j \|\boldsymbol{x}_{j,t}\|^2$ as in (7).

Proof: Using the argument of Lemma 2.1 in [12], which is essentially the same as Theorem 1 in [19], we introduce the auxiliary density $q_{\boldsymbol{\theta}}^\epsilon$ with mean $\boldsymbol{\theta}$ and covariance matrix $\epsilon^2 I$ which has the cumulative loss $L_n(q_{\boldsymbol{\theta}}^\epsilon)$ given by:

$$L_n(q_{\boldsymbol{\theta}}^\epsilon) = \int L_n(\boldsymbol{v}) q_{\boldsymbol{\theta}}^\epsilon(\boldsymbol{v}) d\boldsymbol{v} \tag{17}$$

Step 1: Relate $L_n(q_{\boldsymbol{\theta}}^\epsilon)$ to the loss incurred by the corresponding expert $\boldsymbol{\theta}$. Let $H_{\boldsymbol{y}}(\boldsymbol{z}) = -\ln\sigma(\boldsymbol{y}, \boldsymbol{z})$ where $\sigma(\boldsymbol{y}, \boldsymbol{z})$

is the transfer function given in (9), then by the descent lemma (2), $\forall \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{z}_0 \in \mathbb{R}^M$, we have

$$H_{\boldsymbol{y}}(\boldsymbol{z}) \le H_{\boldsymbol{y}}(\boldsymbol{z}_0) + \langle \nabla H_{\boldsymbol{y}}(\boldsymbol{z}_0), \boldsymbol{z} - \boldsymbol{z}_0 \rangle + \tfrac{c}{2}\|\boldsymbol{z} - \boldsymbol{z}_0\|_2^2. \quad (18)$$

Denote $\boldsymbol{V}$ to be a random variable with density $q_{\boldsymbol{\theta}}^{\epsilon}$, and let $\boldsymbol{z} = \boldsymbol{X}_t^\top \boldsymbol{V}$, $\boldsymbol{z}_0 = \boldsymbol{X}_t^\top \mathbb{E}[\boldsymbol{V}] = \boldsymbol{X}_t^\top \boldsymbol{\theta}$, by taking the expectation on both sides of (18),

$$\mathbb{E} H_{\boldsymbol{y}}(\boldsymbol{X}_t^\top \boldsymbol{V}) \le H_{\boldsymbol{y}}(\boldsymbol{X}_t^\top \boldsymbol{\theta}) + \tfrac{cM}{2}\epsilon^2 \lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{x}}), \quad (19)$$

where we used the fact that $\mathbb{E}((\boldsymbol{V}-\boldsymbol{\theta})^\top \boldsymbol{X}_t \boldsymbol{X}_t^\top (\boldsymbol{V}-\boldsymbol{\theta})) \le M\epsilon^2 \lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{x}})$ with $\Sigma_{\boldsymbol{x}} = \boldsymbol{X}_t \boldsymbol{X}_t^\top$.

Since $\sum_{t=1}^n H_{\boldsymbol{y}_t}(\boldsymbol{x}_t^\top \boldsymbol{\theta}) = L_n(\boldsymbol{\theta})$ by definition and $\sum_{t=1}^n \mathbb{E} H_{\boldsymbol{y}_t}(\boldsymbol{x}_t^\top \mathbf{V}) = L_n(q_{\boldsymbol{\theta}}^{\epsilon})$, we obtain

$$L_n(q_{\boldsymbol{\theta}}^{\epsilon}) \le L_n(\boldsymbol{\theta}) + \tfrac{ncM}{2}\epsilon^2 \lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{x}}) \quad (20)$$

Step 2: Relate L_n to the loss of expert $\boldsymbol{\theta}$. We proceed by bounding the difference between L_n and $L_n(q_{\boldsymbol{\theta}}^{\epsilon})$:

$$\begin{aligned}
\hat{L}_n - L_n(q_{\boldsymbol{\theta}}^{\epsilon}) &\overset{①}{=} -\ln \prod_{t=1}^n \hat{p}_t(\boldsymbol{y}_t|\boldsymbol{X}_t) + \int q_{\boldsymbol{\theta}}^{\epsilon}(\boldsymbol{v}) L_n(\boldsymbol{v}) d\boldsymbol{v} \\
&= \int q_{\boldsymbol{\theta}}^{\epsilon}(\boldsymbol{v}) \ln \frac{\prod_{t=1}^n \sigma(\boldsymbol{y}_t, \boldsymbol{X}_t^\top \boldsymbol{v})}{\prod_{t=1}^n \hat{p}_t(\boldsymbol{y}_t|\boldsymbol{X}_t)} d\boldsymbol{v} \\
&\overset{②}{=} \int q_{\boldsymbol{\theta}}^{\epsilon}(\boldsymbol{v}) \ln \frac{q_n(\boldsymbol{v})}{q_0(\boldsymbol{v})} d\boldsymbol{v} \\
&= \underbrace{\int q_{\boldsymbol{\theta}}^{\epsilon}(\boldsymbol{v}) \ln \frac{q_{\boldsymbol{\theta}}^{\epsilon}(\boldsymbol{v})}{q_0(\boldsymbol{v})} d\boldsymbol{v}}_{D(q_{\boldsymbol{\theta}}^{\epsilon} \| q_0)} - \underbrace{\int q_{\boldsymbol{\theta}}^{\epsilon}(\boldsymbol{v}) \ln \frac{q_{\boldsymbol{\theta}}^{\epsilon}(\boldsymbol{v})}{q_n(\boldsymbol{v})} d\boldsymbol{v}}_{D(q_{\boldsymbol{\theta}}^{\epsilon} \| q_n)} \\
&\overset{③}{\le} D(q_{\boldsymbol{\theta}}^{\epsilon} \parallel q_0),
\end{aligned}$$

where $\hat{p}_t(\boldsymbol{y}|\boldsymbol{X}_t) = \int \sigma(\boldsymbol{y}, \boldsymbol{\theta} \cdot \boldsymbol{X}_t) q_t(\boldsymbol{\theta}) d\boldsymbol{\theta}$ in ①, and ② follows from the definition of $q_n(\boldsymbol{\theta})$ in (3), and ③ follows from nonnegativity of KL divergence. Together with (20),

$$\hat{L}_n - L_n(\boldsymbol{\theta}) \le D(q_{\boldsymbol{\theta}}^{\epsilon} \parallel q_0) + \tfrac{ncM}{2}\epsilon^2 \lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{x}}) \quad (21)$$

Step 3: Optimizing the bound. By Prop. 1 and Lemma 1, we have the RHS of (21) minimized by the choice of $q_{\boldsymbol{\theta}}^{\epsilon}$ as Normal with mean $\boldsymbol{\theta}$ and covariance $\epsilon^2 I$. Optimizing over ϵ obtains the bound. ∎

Remarks: The bound on regret with respect to a single or a group of best experts for BRIEF involves a constant term and a term that grows logarithmically with n, i.e. $O(\log n)$. Also the role of $\boldsymbol{X}_t$ is very flexible, which can be chosen adversarially subject to the euclidean norm constraint that $\lambda_{max}(\boldsymbol{\Sigma}_{\boldsymbol{x}}) = \max_i \|\boldsymbol{x}_{i,t}\|^2$ is bounded, where $\boldsymbol{x}_{i,t}$ is the input vector for the j-th time series as in (7). It is not required to know the bound a priori for the method to work.

III. Experimental Evaluation

A. Simulation by ARMAX

We first evaluate BRIEF on data artificially generated according to the stochastic autoregressive moving-average model with auxiliary input (ARMAX) [16]:

$$A(\rho^{-1})y(t) = B(\rho^{-1})u(t) + C(\rho^{-1})\omega(t) \quad (22)$$

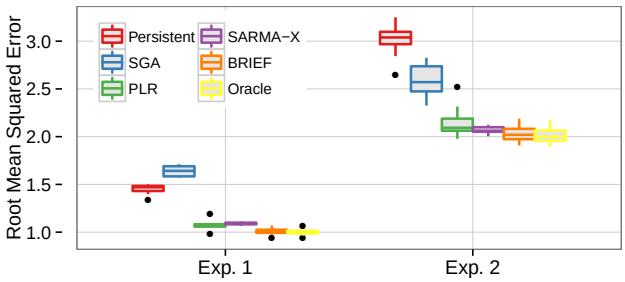

Fig. 1. Distribution of root mean squared error (RMSE) in one-step ahead prediction, evaluated for the persistent model (PS, the prediction is the previous observation), stochastic gradient algorithm (SGA), pseudo linear regression (PLR), seasonal ARMA with auxiliary input (SARMA-X), BRIEF, and oracle (least square regression in retrospect). The two experimental conditions are with noise $\mathcal{N}(0,1)$ and $\mathcal{N}(0,2)$ respectively. The input vector $\boldsymbol{x}_t$ is chosen as the previous 2 points together with the input variable at time t.

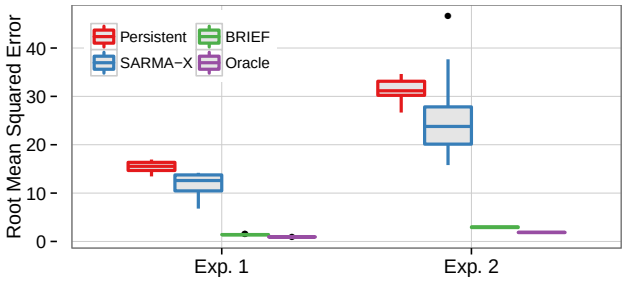

Fig. 2. Distribution of RMSE in multiple step ahead prediction with horizon 100 the same as the period of the input $u(t)$. The experiments with noise level 1 and 2 are performed for the persistent model (taking the previous period as prediction), SARMA-X, BRIEF, and oracle, which have straightforward generalization to multi-step ahead predictions. The input vector $\boldsymbol{x}_t$ is chosen as the previous 2 points, and 1 point at seasonal lag (100), together with the input variable at time t.

where $\{y(t)\}$ and $\{u(t)\}$ denote the output and inputs, and $\{\omega(t)\}$ is the innovations sequence with $\omega(t) \sim \mathcal{N}(0, \sigma_\omega^2)$. $A(\rho^{-1}) = 1 + \sum_{i=1}^n a_i \rho^{-i}$, $B(\rho^{-1}) = \sum_{i=0}^m b_i \rho^{-i}$, $b_0 \neq 0$, $C(\rho^{-1}) = 1 + \sum_{i=1}^l c_i \rho^{-i}$, where ρ is the lag operator. The simulation adopts coefficients with autoregressive order $n = 2$, moving-average order $l = 1$, and auxiliary input order $m = 0$. The input $\{u(t)\}$ is periodic square wave.

BRIEF is evaluated against Seasonal ARMAX (SARMA-X), stochastic gradient algorithm (SGA), and pseudo linear regression algorithm (PLR) in the task of one-step and multi-step ahead predictions.[2] Results are illustrated in Fig. 1 and 2, where the root mean squared error (RMSE) is given by $\sqrt{\frac{1}{n}\sum_{i=1}^n (\hat{y}_i - y_i)^2}$. As can be seen, BRIEF is comparable to the oracle in both tasks, and outperforms the persistent model and seasonal ARMAX (SARMA-X) model in the multi-step ahead prediction.

As is shown in Fig. 3, which illustrates the per round regret

[2]See Chpater 7 and 9 in [1] for implementation details of SGA and PLR. The parameters of the SARMA-X models are estimated by maximum likelihood, which is implemented by the standard package in R.

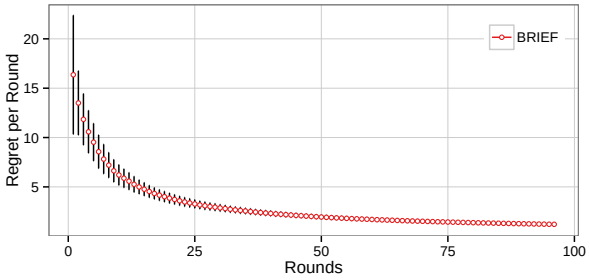

Fig. 3. Per round regret R_t/t as a function of rounds t for BRIEF-M in the multi-step ahead prediction task. The mean (red dot) and 1 standard deviation (black bar) are shown for 10 independent trials.

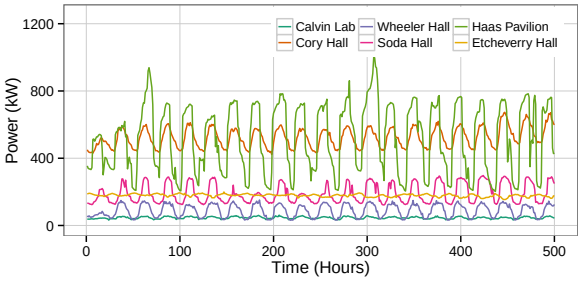

Fig. 4. Typical energy consumptions of the selected set of buildings on the UC Berkeley campus with the sampling resolution of 1 hour.

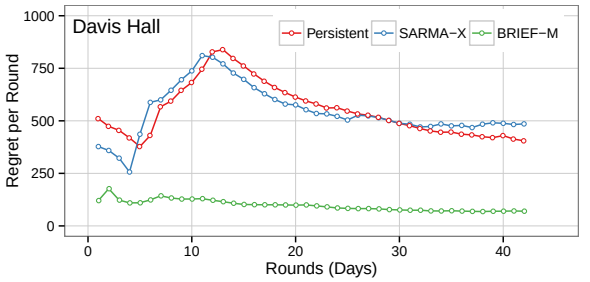

Fig. 5. Per round regret R_t/t in one-day ahead energy prediction of Davis Hall. Each point shows the mean regret in the 5-fold cross-validation.

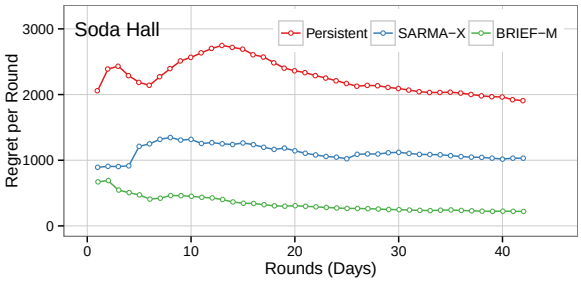

Fig. 6. Per round regret R_t/t in one-day ahead energy prediction of Soda Hall. The performance of the persistent model is worse than SARMA-X, while BRIEF remains the superior.

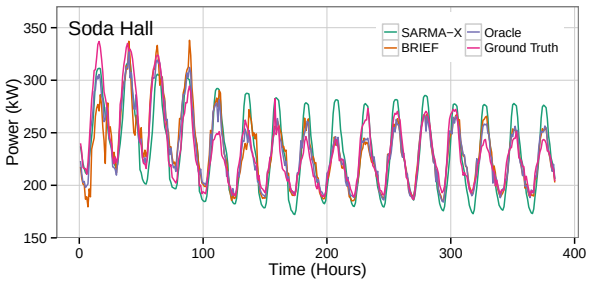

Fig. 7. Selected period of energy consumption ground truth and predicted time series by SARMA-X, BRIEF, and oracle for Soda Hall, where the consumption remains regular throughout the period.

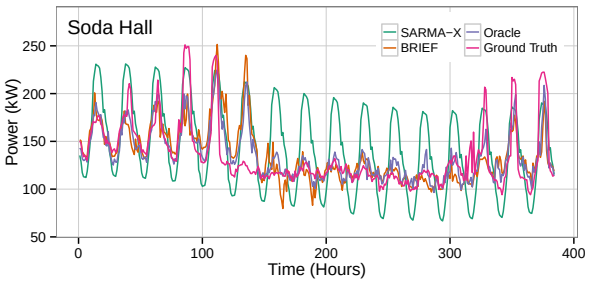

Fig. 8. Selected period of energy consumption ground truth and predicted time series by SARMA-X, BRIEF, and oracle for Soda Hall, where there is a sudden drop in consumption during the period.

R_t/t where R_t is given in (2), the performance of BRIEF converges to the oracle.

B. Use Case: Building Energy Prediction

Prediction of building energy consumption, which accounts for approximately 40% of all energy usage in the U.S., facilitates demand-response for energy savings [20], [21]. We collected energy consumption of 10 buildings on UC Berkeley campus at a resolution of 15 minutes and 1 hour, together with local weather, from 2014 Feb. 10 to 2015 Feb. 10, as shown in Fig. 4.

The results of the one-day ahead prediction by BRIEF (for single and multiple time series) and two popular models among utility companies, i.e. ARIMA, and persistent model (taking the previous period as prediction), are shown in Table I.[3] As can be seen, the performance of BRIEF-M is closer to the oracle than BRIEF-S due to the employment of the inverse-correlation matrix weighted cost (8), and both methods excel the SARMA-X and PS models in most cases.

The per round regrets R_t/t for the prediction of two buildings (Davis and Soda Hall) are shown in Fig. 5 and 6. While the persistent and SARMA-X models have large per round regret, BRIEF exhibits convergence to the oracle as t increases.

Close inspection of the predicted time series reveal that while SARMA-X behaves reasonably well for periodic sig-

[3]Since building consumption exhibits strong seasonality, we extract the seasonal trend learned from training data before predictions. The correlation matrix $\mathbf{\Gamma}$ is calculated for the seasonally adjusted data. For the choice of $\boldsymbol{x}_t$ we choose the power consumption in the prevoius 2 hours together with the consumption one day before at the same time, and the temperature forecast for the next hour.

TABLE I. Root mean squared error (kW) in one-day ahead energy prediction for 10 buildings on the UC Berkeley campus by the persistent model, Seasonal seasonal ARMAX (SARMA-X), BRIEF with single and multiple time series (BRIEF-S, BRIEF-M), and the oracle.

	Persistent	**SARMA-X**	**BRIEF-S**	**BRIEF-M**	**Oracle**
Davis Hall	24.28	22.72	18.67	**18.49**	16.37
Calvin Lab	7.08	3.48	2.55	**2.53**	2.22
Cory Hall	69.61	31.23	25.00	**24.65**	21.82
Stanley Hall	80.00	45.39	43.63	**42.69**	35.08
Doe Library	**14.01**	14.77	18.06	16.00	13.08
Lawrence Lab	21.60	20.38	19.65	**19.22**	16.59
Wheeler Hall	20.14	17.09	14.94	**14.81**	12.77
Soda Hall	47.98	31.16	26.60	**26.25**	23.87
Haas Pavilion	128.21	116.73	102.60	**97.13**	80.48
Etcheverry Hall	9.48	6.00	5.79	**5.74**	5.22
Mean error:	42.27	30.89	27.75	**26.75**	22.75

nals (see Fig. 7) , it fails to capture the change in consumption as in Fig. 8. On the contrary BRIEF quickly responds to the change, as in Fig. 8, similar to the oracle.

IV. Conclusion

Bayesian Regression of Inifinite Expert Forecaster (BRIEF) approaches time series prediction by consulting an infinite pool of experts, whose performances are tracked to provide the weights of their opinions. The method, based on regret minimization, is free of any strong assumptions on the parameters of the time-varying system.

Based on the posterior updates of the experts' weights by accounting for their past performances, simple update rules have been derived which offer computational advantage. For multiple time series prediction, the inverse correlation weighted error (ICWE) is employed as a guiding criterion motivated by maximum likelihood estimation of multivariate Gaussian distributions. Performance bounds show that the cumulative regret increases at rate $O(\log T)$, which guarantees that BRIEF behaves indistinguishably as the oracle with the per round regret R_t/t decreases at rate $O(\frac{\log T}{T})$. As the proofs stand, we can allow adversarial choice of the $\boldsymbol{X}_t$ subject to the euclidean norm constraint. Additionally for the algorithm to work it is not required to have prior knowledge on the constraints.

Through simulation and real data on building energy prediction, BRIEF is demonstrated to outperform other models, such as the persistent model, stochastic gradient algorithm, pseudo linear regression, and SARMA-X. The per round regret is also shown to be converging as the horizon increases.

As we can select any forms of transfer functions subject to the regularization conditions required by the proof of regret bounds, the logistic loss $\sigma(y, \boldsymbol{u}^\top \boldsymbol{x}) = \frac{1}{1+e^{-y\boldsymbol{u}^\top \boldsymbol{x}}}$ is a viable candidate for binary valued data as an extension of the current method on real valued signal. We also want to investigate the application of BRIEF to adaptive and robust controls to benefit from its performance guarantee.

References

[1] G. C. Goodwin and K. S. Sin, Adaptive filtering prediction and control. Courier Corporation, 2014.

[2] S. Meyn and L. Guo, "Stability, convergence, and performance of an adaptive control algorithm applied to a randomly varying system," in Decision and Control, 1989., Proceedings of the 28th IEEE Conference on. IEEE, 1989, pp. 2108–2113.

[3] N. Cesa-Bianchi and G. Lugosi, Prediction, learning, and games. Cambridge University Press, 2006.

[4] R. T. Rockafellar, "Monotone operators and the proximal point algorithm," SIAM journal on control and optimization, vol. 14, no. 5, 1976, pp. 877–898.

[5] A. Beck and M. Teboulle, "Mirror descent and nonlinear projected subgradient methods for convex optimization," Operations Research Letters, vol. 31, no. 3, 2003, pp. 167–175.

[6] E. Hazan, A. Rakhlin, and P. L. Bartlett, "Adaptive online gradient descent," in Advances in Neural Information Processing Systems, 2007, pp. 65–72.

[7] M. Zinkevich, "Online convex programming and generalized infinitesimal gradient ascent," 2003.

[8] E. Hazan, A. Agarwal, and S. Kale, "Logarithmic regret algorithms for online convex optimization," Machine Learning, vol. 69, no. 2-3, 2007, pp. 169–192.

[9] M. Raginsky, A. Rakhlin, and S. Yuksel, "Online convex programming and regularization in adaptive control," in Decision and Control (CDC), 2010 49th IEEE Conference on. IEEE, 2010, pp. 1957–1962.

[10] V. Vovk, "Competitive on-line statistics," International Statistical Review, vol. 69, no. 2, 2001, pp. 213–248.

[11] B. S. Clarke and A. R. Barron, "Information-theoretic asymptotics of bayes methods," Information Theory, IEEE Transactions on, vol. 36, no. 3, 1990, pp. 453–471.

[12] S. M. Kakade and A. Y. Ng, "Online bounds for bayesian algorithms," in Advances in neural information processing systems, 2004, pp. 641–648.

[13] P. D. Grünwald, The minimum description length principle. MIT press, 2007.

[14] J. Kivinen and M. K. Warmuth, "Exponentiated gradient versus gradient descent for linear predictors," Information and Computation, vol. 132, no. 1, 1997, pp. 1–63.

[15] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," Technometrics, vol. 12, no. 1, 1970, pp. 55–67.

[16] H. Lütkepohl, New introduction to multiple time series analysis. Springer Science & Business Media, 2007.

[17] T. M. Cover and J. A. Thomas, Elements of information theory. John Wiley & Sons, 2012.

[18] A. Ben-Tal and A. Nemirovski, Lectures on modern convex optimization: analysis, algorithms, and engineering applications. Siam, 2001, vol. 2.

[19] Y. Freund, R. E. Schapire, Y. Singer, and M. K. Warmuth, "Using and combining predictors that specialize," in Proceedings of the twenty-ninth annual ACM symposium on Theory of computing. ACM, 1997, pp. 334–343.

[20] J. Zhang, J. Han, R. Wang, and G. Hou, "Day-ahead electricity price forecasting based on rolling time series and least square-support vector machine model," in Control and Decision Conference (CCDC), 2011 Chinese. IEEE, 2011, pp. 1065–1070.

[21] F. Oldewurtel, A. Ulbig, A. Parisio, G. Andersson, and M. Morari, "Reducing peak electricity demand in building climate control using real-time pricing and model predictive control," in Decision and Control (CDC), 2010 49th IEEE Conference on. IEEE, 2010, pp. 1927–1932.