Error Bounds for Online Predictions of Linear-Chain Conditional Random Fields. Application to Activity Recognition for Users of Rolling Walkers

Mathieu Sinn, Pascal Poupart
David R. Cheriton School of Computer Science
University of Waterloo
Waterloo, Ontario, Canada
Email: msinn@cs.uwaterloo.ca, ppoupart@cs.uwaterloo.ca

***Abstract*—Linear-Chain Conditional Random Fields (L-CRFs) are a versatile class of models for the distribution of a sequence of hidden states ("labels") conditional on a sequence of observable variables. In general, the exact conditional marginal distributions of the labels can be computed only after the complete sequence of observations has been obtained, which forbids the prediction of labels in an online fashion. This paper considers approximations of the marginal distributions which only take into account past observations and a small number of observations in the future. Based on these approximations, labels can be predicted close to real-time. We establish rigorous bounds for the marginal distributions which can be used to assess the approximation error at runtime. We apply the results to an L-CRF which recognizes the activity of rolling walker users from a stream of sensor data. It turns out that if we allow for a prediction delay of half of a second, the online predictions achieve almost the same accuracy as the offline predictions based on the complete observation sequences.**

I. Introduction

Conditional Random Fields are a widely popular class of models for the distribution of a set of hidden states ("labels") conditional on a set of observable variables [3]. An important subclass are Linear-Chain Conditional Random Fields (L-CRFs) where the labels and observations both form linear sequences. L-CRFs have been successfully applied, e.g., in natural language processing, to the analysis of genome data, and for human activity recognition [8], [1], [4]. A common approach to predict the labels is to choose the arguments maximizing the conditional marginal distributions. In theory, the exact marginal distributions can be computed only after the complete sequence of observations has been obtained. For many real-world applications this is impracticable; for example, in human activity recognition the sequence of observations is often an "endless" stream of sensor data while it is necessary to predict the human activities in an online fashion.

In this paper, we study approximations of the conditional marginal distributions which, at time t, only take into account the observations up to time $t+s$, where $s \geq 0$ corresponds to some (short) prediction delay. Based on these approximations, the label at time t can be predicted after s additional observations have been obtained. Our main theoretical results are error bounds for the approximate marginal distributions. An attractive property of these bounds is that they also depend only on the observations up to time $t+s$, so the potential impact of future observations on the online predictions can be assessed at runtime.

Fig. 1. Rolling walker instrumented with sensors

We apply the results to an L-CRF which recognizes the activity of rolling walker users from sensors measuring the speed, the acceleration, and the load on the four wheels. A prototype of the walker is displayed in Fig. 1. It turns out that if we allow for a prediction delay of half of a second, the accuracy of online and offline predictions is almost the same. Moreover, in 50-70% of the cases the error bounds are tight enough to rule out the possibility that after obtaining more observations in the future a different label would be predicted.

The paper is structured as follows: In Sec. II we review the

definition of L-CRFs. Sec. III introduces the problem of online predictions and establishes the bounds for the approximation errors. In Sec. IV we provide more background on activity recognition for rolling walker users. Empirical results are shown in Sec. V, and Sec. VI concludes the paper.

II. Conditional Random Fields

Linear-Chain Conditional Random Fields (L-CRF) specify the distribution of a sequence of labels $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ conditional on a sequence of observations $\boldsymbol{X} = (X_1, \ldots, X_n)$. In the following, suppose the observations range in $\mathcal{X}$ and the labels in $\mathcal{Y} = \{1, 2, \ldots, \ell\}$. The L-CRFs are parameterized by vectors of feature functions, $\boldsymbol{f}$, and model weights, $\boldsymbol{\lambda}$. For $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathcal{X}^n$ and $\boldsymbol{y} = (y_1, \ldots, y_n) \in \mathcal{Y}^n$, the probability of $\boldsymbol{Y} = \boldsymbol{y}$ conditional on $\boldsymbol{X} = \boldsymbol{x}$ is given by

$$P(\boldsymbol{Y} = \boldsymbol{y} \,|\, \boldsymbol{X} = \boldsymbol{x}) \quad = \quad \frac{1}{Z(\boldsymbol{x})} \exp\left(\boldsymbol{\lambda}^T \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y})\right) \qquad (1)$$

where $Z(\boldsymbol{x})$ is a normalizing constant. It is usually assumed that the inner product $\boldsymbol{\lambda}^T \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y})$ can be written in the following form:

$$\begin{aligned} \boldsymbol{\lambda}^T \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}) \quad &= \quad \sum_{t=1}^{n} \boldsymbol{\mu}^T \boldsymbol{f}^{\text{state}}(x_t, y_t) \\ &+ \quad \sum_{t=2}^{n} \boldsymbol{\nu}^T \boldsymbol{f}^{\text{trans}}(x_t, y_{t-1}, y_t) \end{aligned} \qquad (2)$$

where $\boldsymbol{f}^{\text{state}}$ and $\boldsymbol{f}^{\text{trans}}$ denote vectors of state and transition features. Typically, during the training of L-CRFs, the features are fixed and the model weights are learned by maximizing the conditional log-likelihood of labeled training data.

Given a trained L-CRF and an unlabeled sequence of observations $\boldsymbol{x} = (x_1, \ldots, x_n)$, the goal is to predict the sequence of labels $\boldsymbol{y} = (y_1, \ldots, y_n)$. A common approach is to determine $\boldsymbol{y}$ componentwise by maximizing the conditional marginal distributions of $\boldsymbol{Y}$, i.e.,

$$y_t \quad = \quad \arg\max_{j \in \mathcal{Y}} P(Y_t = j \,|\, \boldsymbol{X} = \boldsymbol{x}) \qquad (3)$$

for $t = 1, \ldots, n$. [1] The conditional marginal distributions of $\boldsymbol{Y}$ can be efficiently computed in the following way: For $x \in \mathcal{X}$, let $\boldsymbol{\alpha}(x)$ be the ℓ-dimensional vector with the ith component $\exp(\boldsymbol{\mu}^T \boldsymbol{f}^{\text{state}}(x, i))$. Furthermore, let $\boldsymbol{A}(x)$ denote the $\ell \times \ell$-matrix with the (h, i)-th component

$$a(x, h, i) \quad = \quad \exp\left(\boldsymbol{\mu}^T \boldsymbol{f}^{\text{state}}(x, i) \,+\, \boldsymbol{\nu}^T \boldsymbol{f}^{\text{trans}}(x, h, i)\right).$$

For any sequence $\boldsymbol{x} \in \mathcal{X}^n$ and $m \leq n$, define

$$\boldsymbol{A}_t^m(\boldsymbol{x}) \quad := \quad \boldsymbol{A}(x_t) \ldots \boldsymbol{A}(x_m)$$

and write $a_t^m(\boldsymbol{x}, h, i)$ for the (h, i)-th component of $\boldsymbol{A}_t^m(\boldsymbol{x})$. Finally, let $\alpha_t(\boldsymbol{x}, i)$ and $\beta_t^m(\boldsymbol{x}, i)$ denote the ith components of the vectors

$$\begin{aligned} \boldsymbol{\alpha}_t(\boldsymbol{x}) \quad &:= \quad \boldsymbol{A}_2^t(\boldsymbol{x})^T \boldsymbol{\alpha}(x_1), \\ \boldsymbol{\beta}_t^m(\boldsymbol{x}) \quad &:= \quad \boldsymbol{A}_t^m(\boldsymbol{x})(1, 1, \ldots, 1)^T. \end{aligned}$$

Then, as a well-known fact, the conditional marginal distributions of $\boldsymbol{Y}$ can be evaluated using the formula

$$P(Y_t = j \,|\, \boldsymbol{X} = \boldsymbol{x}) \quad = \quad \frac{\alpha_t(\boldsymbol{x}, j) \beta_t^n(\boldsymbol{x}, j)}{\boldsymbol{\alpha}_t(\boldsymbol{x})^T \boldsymbol{\beta}_t^n(\boldsymbol{x})} \qquad (4)$$

for $j \in \mathcal{Y}$ (see [8]). This formula will also be a key ingredient to our analysis of error bounds for online predictions.

III. Online Predictions

The main problem of predicting the labels by means of formula (3) is that the exact conditional marginal distributions of $\boldsymbol{Y}$ are available only after the complete sequence of observations has been obtained. For many real-world applications such a delay is prohibitive; typically, the sequences of observations are very long (up to several hours) while it is desirable to predict the labels close to real-time.

Throughout the rest of this section, let $s \geq 0$ and consider time points t for which $t + s \leq n$. By $\boldsymbol{X}_{t+s}$ and $\boldsymbol{x}_{t+s}$ we denote the projection of $\boldsymbol{X}$ and $\boldsymbol{x}$ onto their first $t + s$ components. In order to allow for online predictions, we consider marginal distributions of $\boldsymbol{Y}$ conditional on $\boldsymbol{X}_{t+s} = \boldsymbol{x}_{t+s}$. The parameter s can be regarded as the delay of the predictions. Below we will show that, as s increases, the difference between $P(Y_t = j \,|\, \boldsymbol{X}_{t+s} = \boldsymbol{x}_{t+s})$ and $P(Y_t = j \,|\, \boldsymbol{X} = \boldsymbol{x})$ tends to 0. How large s can be in practice depends on the particular application. For the activity recognition of rolling walker users, we find that a prediction delay of up to $s = 25$ time points is acceptable (corresponding to half of a second).

A. Error Bounds

The following result can be used to bound the difference between $P(Y_t = j \,|\, \boldsymbol{X}_{t+s} = \boldsymbol{x}_{t+s})$ and $P(Y_t = j \,|\, \boldsymbol{X} = \boldsymbol{x})$. Similarly as in [7], a key in the mathematical analysis are results from the theory of weak ergodicity.

Theorem 1: Suppose that $s > 0$. Then, for any sequence of observations $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathcal{X}^n$ and any label $j \in \mathcal{Y}$, both $P(Y_t = j \,|\, \boldsymbol{X}_{t+s} = \boldsymbol{x}_{t+s})$ and $P(Y_t = j \,|\, \boldsymbol{X} = \boldsymbol{x})$ are bounded from below by

$$\left(\sum_{i=1}^{\ell} \frac{\alpha_t(\boldsymbol{x}, i)}{\alpha_t(\boldsymbol{x}, j)} \left[\max_{k \in \mathcal{Y}} \frac{a_t^{t+s}(\boldsymbol{x}, i, k)}{a_t^{t+s}(\boldsymbol{x}, j, k)} \right] \right)^{-1}. \qquad (5)$$

An upper bound is given by

$$\left(\sum_{i=1}^{\ell} \frac{\alpha_t(\boldsymbol{x}, i)}{\alpha_t(\boldsymbol{x}, j)} \left[\min_{k \in \mathcal{Y}} \frac{a_t^{t+s}(\boldsymbol{x}, i, k)}{a_t^{t+s}(\boldsymbol{x}, j, k)} \right] \right)^{-1}. \qquad (6)$$

Clearly, the difference between $P(Y_t = j \,|\, \boldsymbol{X}_{t+s} = \boldsymbol{x}_{t+s})$ and $P(Y_t = j \,|\, \boldsymbol{X} = \boldsymbol{x})$ is bounded by the difference between the upper and the lower bound.

[1] Alternatively, $\boldsymbol{y}$ can be determined by maximizing the probability of the complete label sequence. In human activity recognition, this approach often leads to poor results due to a tendency of over-smoothing the label sequences and missing many of the short-lasting activities (see, e.g., [2]).

Proof: According to formula (4), we have

$$P(Y_t = j \,|\, \boldsymbol{X}_{t+s} = \boldsymbol{x}_{t+s}) \;=\; \left(\sum_{i=1}^{\ell} \frac{\alpha_t(\boldsymbol{x}, i)}{\alpha_t(\boldsymbol{x}, j)} \frac{\beta_t^{t+s}(\boldsymbol{x}, i)}{\beta_t^{t+s}(\boldsymbol{x}, j)} \right)^{-1} .$$

By definition, $\beta_t^{t+s}(\boldsymbol{x}, i)$ and $\beta_t^{t+s}(\boldsymbol{x}, j)$ are the sums of the elements in the ith and jth row of the matrix $\boldsymbol{A}_t^{t+s}(\boldsymbol{x})$, respectively. Therefore,

$$\frac{\beta_t^{t+s}(\boldsymbol{x}, i)}{\beta_t^{t+s}(\boldsymbol{x}, j)} \;\leq\; \max_{k \in \mathcal{Y}} \frac{a_t^{t+s}(\boldsymbol{x}, i, k)}{a_t^{t+s}(\boldsymbol{x}, j, k)} \tag{7}$$

which proves the lower bound for $P(Y_t = j \,|\, \boldsymbol{X}_{t+s} = \boldsymbol{x}_{t+s})$. Now, by a fundamental result on products of positive matrices, the quantity on the right hand side of (7) is non-increasing with s (see, e.g., Lemma 3.4 in [5]). Consequently,

$$\frac{\beta_t^{n}(\boldsymbol{x}, i)}{\beta_t^{n}(\boldsymbol{x}, j)} \;\leq\; \max_{k \in \mathcal{Y}} \frac{a_t^{t+s}(\boldsymbol{x}, i, k)}{a_t^{t+s}(\boldsymbol{x}, j, k)}$$

which proves the lower bound for $P(Y_t = j \,|\, \boldsymbol{X} = \boldsymbol{x})$. The upper bounds are established by similar arguments. ∎

Let us make a few remarks:

- It is easy to see that both the lower and upper bound for $P(Y_t = j \,|\, \boldsymbol{X} = \boldsymbol{x})$ lie within the interval $[0, 1]$. Moreover, as the proof of Theorem 1 shows, the lower bound is non-decreasing with s, while the upper bound is non-increasing. Using results from the theory of weak ergodicity one can show that, under mild regularity conditions on the feature functions, the difference between the bounds tends to 0 with a geometric rate. Hence, the larger the prediction delay s, the more precisely $P(Y_t = j \,|\, \boldsymbol{X} = \boldsymbol{x})$ can be localized.
- In order to establish bounds for $s = 0$, suppose that the matrices $\boldsymbol{A}(x)$ are bounded away from zero and infinity. Then, with a_{inf} and a_{sup} denoting the infimum and supremum over all $a(x, h, i)$ with $x \in \mathcal{X}$ and $h, i \in \mathcal{Y}$, lower and upper bounds are obtained by replacing the max and min in Theorem 1 with $\frac{a_{\text{sup}}}{a_{\text{inf}}}$ and $\frac{a_{\text{inf}}}{a_{\text{sup}}}$, respectively.
- An important property is that the error bounds themselves only depend on the observations up to time $t + s$. Therefore, at the same time where $P(Y_t = j \,|\, \boldsymbol{X}_{t+s} = \boldsymbol{x}_{t+s})$ is available, it can already be assessed how far this approximation is from the exact conditional marginal distribution $P(Y_t = j \,|\, \boldsymbol{X} = \boldsymbol{x})$. In some cases, it is even possible to determine at time $t + s$ the $j_0 \in \mathcal{Y}$ which maximizes $P(Y_t = j \,|\, \boldsymbol{X} = \boldsymbol{x})$ (although most of the sequence $\boldsymbol{x}$ might not have been observed yet), namely, if the lower bound of $P(Y_t = j_0 \,|\, \boldsymbol{X}_{t+s} = \boldsymbol{x}_{t+s})$ is greater than the upper bounds of $P(Y_t = j \,|\, \boldsymbol{X}_{t+s} = \boldsymbol{x}_{t+s})$ for any other $j \in \mathcal{Y}$.

We conclude this section by noting that (4) allows for an efficient iterative computation of $P(Y_t = j \,|\, \boldsymbol{X}_{t+s} = \boldsymbol{x}_{t+s})$. In particular, $\boldsymbol{\alpha}_{t+1}(\boldsymbol{x})$ can be obtained from $\boldsymbol{\alpha}_t(\boldsymbol{x})$ by means of one vector-matrix multiplication; computing the updated matrix $\boldsymbol{A}_{t+1}^{t+s+1}(\boldsymbol{x})$ from $\boldsymbol{A}_t^{t+s}(\boldsymbol{x})$ requires one matrix inversion and two matrix multiplications; finally, $\boldsymbol{\beta}_{t+1}^{t+s+1}(\boldsymbol{x})$ can be obtained by one extra vector-matrix multiplication. Note that, given $\boldsymbol{\alpha}_{t+1}(\boldsymbol{x})$ and $\boldsymbol{A}_{t+1}^{t+s+1}(\boldsymbol{x})$, the error bounds provided by Theorem 1 can be computed at very low additional cost.

B. Example

Let us illustrate the application of the error bounds for a simple toy example. Here we assume the observations are real-valued, $\mathcal{X} = \mathbb{R}$, and there are only two different labels, $\mathcal{Y} = \{1, 2\}$. Consider the following set of weighted features:

$$\boldsymbol{\mu}^T \boldsymbol{f}^{\text{state}}(x_t, y_t) \;=\; \begin{cases} x_t/10 & \text{if } y_t = 1 \\ -x_t/10 & \text{if } y_t = 2 \end{cases}$$

$$\boldsymbol{\nu}^T \boldsymbol{f}^{\text{trans}}(x_t, y_{t-1}, y_t) \;=\; \begin{cases} 2 & \text{if } y_{t-1} = y_t \\ 0 & \text{if } y_{t-1} \neq y_t \end{cases}$$

We use an autoregressive model for the sequence of observations, $X_t = 0.75 X_{t-1} + \epsilon_t$ where the ϵ_t are iid standard normal. Note that, by means of the state features, $X_t > 0$ increases the probability that $Y_t = 1$, and $X_t < 0$ the probability that $Y_t = 2$. According to the transition features, it is more likely that $Y_t = Y_{t-1}$ rather than $Y_t \neq Y_{t-1}$.

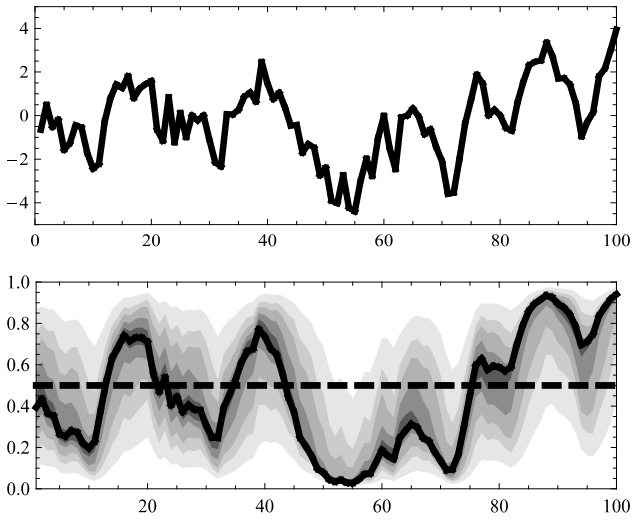

Fig. 2. Observations (upper plot) and online predictions (lower plot)

An example is shown in Fig. 2. The upper plot displays 100 time points of a realization of $\boldsymbol{X}$. The lower plot shows the resulting conditional probabilities $P(Y_t = 1 \,|\, \boldsymbol{X} = \boldsymbol{x})$ for $t = 1, 2, \ldots, 100$ (black solid line), and the error bounds for $s = 1, 2, 3, 5, 8$ (gray shaded regions). As can be seen, the difference between the upper and lower bounds decreases rather quickly as s increases. Note that, in the case of two labels, the lower bound for $Y_t = 1$ and the upper bound for $Y_t = 2$ (and, vice versa, the upper bound for $Y_t = 1$ and the lower bound for $Y_t = 2$) sum up to 1. Hence, if the lower bound for $Y_t = 1$ exceeds 0.5 (depicted by the dashed line in Fig. 2), then it is certain that the argument maximizing $P(Y_t = j \,|\, \boldsymbol{X} = \boldsymbol{x})$ is equal to 1. Similarly, if the upper bound for $Y_t = 1$ is below 0.5, then the argument maximizing $P(Y_t = j \,|\, \boldsymbol{X} = \boldsymbol{x})$ is equal to 2.

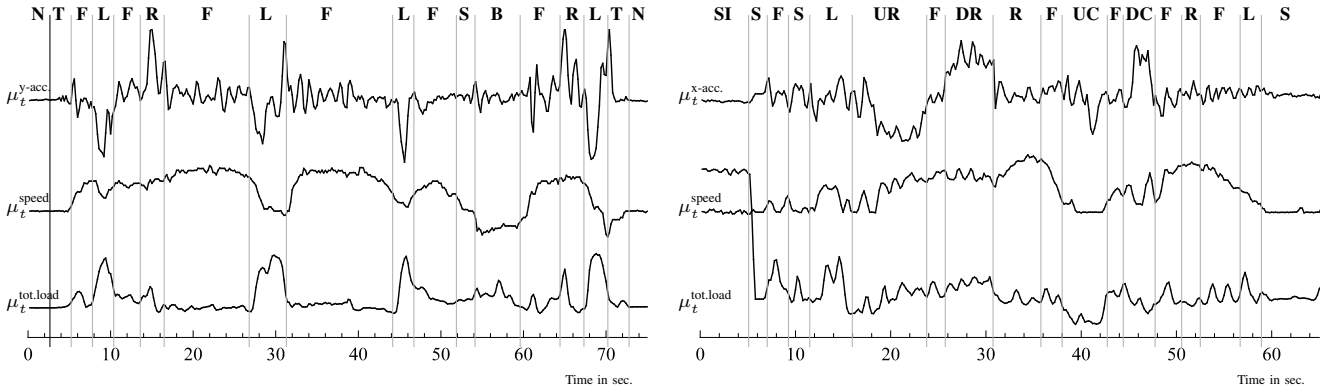

Fig. 3. Two segments of the walker data, with measurements averaged over time windows of size $w = 25$. The users perform the following activities: not touching the walker (**N**), stopping (**S**), walking forward (**F**), walking backwards (**B**), turning left (**L**), turning right (**R**), transferring between the walker and a chair (**T**), sitting on the walker (**SI**), going up a ramp (**UR**), going down a ramp (**DR**), going up a curb (**UC**), going down a curb (**DC**).

IV. Application: Activity Recognition for Users of Rolling Walkers

The prototype of the instrumented walker shown in Fig. 1 has been developed at the Toronto Rehabilitation Institute [9]. The goal of this project is to build a smart walker companion that assists users, caregivers and clinicians, e.g., by monitoring the user's stability, assessing her motoric abilities, and supervising the execution of daily excercises. The key step in implementing these functionalities is the recognition of the user activity from the sensor data.

The raw sensor data consists of 8 measurements at every time point:

- $x_t^{\text{dist.}}$, the walked distance measured by an encoder mounted at the rear right wheel;
- $x_t^{\text{fr.le.}}$, $x_t^{\text{fr.ri.}}$, $x_t^{\text{re.le.}}$, $x_t^{\text{re.ri.}}$, the load on the four wheels (front/rear, left/right) measured by load sensors;
- $x_t^{\text{x-acc.}}$, $x_t^{\text{y-acc.}}$, $x_t^{\text{z-acc.}}$, the acceleration in x-, y-, z-direction measured by a 3D-accelerometer.

The measurements are digitized with a sampling rate of 50 Hz and 16-bit resolution. In particular, $x_t^{\text{dist.}}$ is the walked distance modulo 2^{16}, where walking backwards results in a decrease of $x_t^{\text{dist.}}$. Besides the sensors, the walker is equipped with two video cameras recording the environment and the position of the lower limbs. At present, the video recordings are only used to manually label the data collected in the experiments.

¿From the raw sensor data, we compute the following measures:

- the speed: $x_t^{\text{speed}} = x_t^{\text{dist.}} - x_{t-1}^{\text{dist.}}$ where we add (substract) 2^{16} if an overflow (underflow) of $x_t^{\text{dist.}}$ is detected;
- the total load: $x_t^{\text{tot. load}} = x_t^{\text{fr.le.}} + x_t^{\text{fr.ri.}} + x_t^{\text{re.le.}} + x_t^{\text{re.ri.}}$;
- the frontal plane center of pressure (FCOP), measuring the relative difference between the load on the left and the right wheels:

$$x_t^{\text{FCOP}} \quad = \quad \frac{d_f(x_t^{\text{fr.le.}} - x_t^{\text{fr.ri.}}) + d_r(x_t^{\text{re.le.}} - x_t^{\text{re.ri.}})}{x_t^{\text{fr.le.}} + x_t^{\text{fr.ri.}} + x_t^{\text{re.le.}} + x_t^{\text{re.ri.}}}$$

  where the constants $d_f = 22.25$ and $d_r = 26.6$ are the distances of the front/rear load cells to the midline of the walker (in centimeters);

- the sagittal plane center of pressure (SCOP), measuring the relative difference between the load on the rear and the front wheels:

$$x_t^{\text{SCOP}} \quad = \quad \frac{(x_t^{\text{fr.le.}} - x_t^{\text{re.le.}}) + (x_t^{\text{fr.ri.}} - x_t^{\text{re.ri.}})}{x_t^{\text{fr.le.}} + x_t^{\text{fr.ri.}} + x_t^{\text{re.le.}} + x_t^{\text{re.ri.}}}.$$

The advantage of using FCOP and SCOP instead of the raw load sensor measurements is that they measure the *relative* difference between the load on the left/right and rear/front wheels, so they do not depend on the user's individual body weight. In order to include information on the past, we also compute the mean and the variance of the previous w measurements, for instance,

$$\mu_t^{\text{speed}}(w) \quad = \quad \frac{1}{w}\sum_{k=0}^{w-1} x_{t-k}^{\text{speed}}.$$

Of course, the variances for $w = 1$ are equal to 0.

Fig. 3 shows two segments of the walker data. Note that the measurements have been averaged over time windows of size $w = 25$; the original recordings are much more noisy. The gray vertical lines display the parts during which the user performed a particular activity. As can be seen, certain activities can be well distinguished by the sensor data: for example, the speed tells us whether the user is standing, walking forward or backwards. The total load is a good indicator whether the user is sitting on the walker or performing turns. The x-acceleration allows us to distinguish between going up and down a ramp/curb, while the y-acceleration can be used to discriminate between left and right turns.

A. Architecture of the L-CRF

In order to recognize the user activities, we use 35-dimensional observational vectors, consisting of the means and the variances of x_t^{speed}, $x_t^{\text{x-acc.}}$, $x_t^{\text{y-acc.}}$, $x_t^{\text{z-acc.}}$ $x_t^{\text{tot.load}}$, x_t^{FCOP}, x_t^{SCOP} for the time horizons $w = 1, 5, 25$ (excluding the zero variances for $w = 1$). For the state features of the L-CRF we use simple linear functions of the following form:

$$\boldsymbol{\mu}^T \boldsymbol{f}^{\text{state}}(x_t, i) \quad = \quad \alpha_i + \boldsymbol{\beta}_i^T x_t,$$

that is, for each label $i \in \mathcal{Y}$ the L-CRF learns an intercept α_i and a 35-dimensional vector of linear coefficients $\boldsymbol{\beta}_i$. Note that this model is able to learn correspondences like "the higher the speed, the more likely it is that the user is walking forward" or "the higher the acceleration in y-direction, the more likely it is that she is turning left". In practice, it is advantageous to normalize the sensor measurements (i.e., to substract the means and divide the standard deviations) to avoid that large values of α_i and $\boldsymbol{\beta}_i$ are penalized during the training of the model.

The transition features of our L-CRF simply reflect whether or not an activity persists:

$$\boldsymbol{\nu}^T \boldsymbol{f}^{\text{trans}}(x_t, y_{t-1}, y_t) \quad = \quad \begin{cases} \nu & \text{if } y_{t-1} = y_t \\ 0 & \text{if } y_{t-1} \neq y_t \end{cases}$$

where ν is a scalar weight. The reason for using this simplistic model is that we want to avoid a bias towards certain transitions due to the design of the experimental courses.

TABLE I
CONFUSION MATRIX FOR EXPERIMENT 1

	N	S	F	L	R	B	T
N	**93.9**	1.1	0.2	0	0	0	5.2
S	4.0	**72.7**	6.2	6.5	0.3	6.3	2.9
F	0.1	0.2	**95.8**	2.4	1.7	0	0.0
L	0	4.4	15.0	**75.8**	0.5	2.8	1.3
R	0.6	0	27.2	5.4	**64.4**	0	2.4
B	0	2.7	0.4	4.2	0.2	**91.3**	1.2
T	23.5	7.9	3.3	6.6	0.1	2.3	**55.9**

TABLE II
CONFUSION MATRIX FOR EXPERIMENT 2

	S	F	L	R	SI	UR	DR	UC	DC
S	**89.9**	5.3	3.0	0.3	0.7	0.1	0.1	0.6	0.0
F	4.7	**85.8**	4.0	3.5	0	0.7	0.2	0.7	0.4
L	12.2	25.4	**60.1**	1.9	0	0.3	0.1	0	0
R	5.8	41.1	0.9	**49.2**	0	0	1.8	0.2	1.1
SI	1.1	0.0	0	0	**98.7**	0.1	0.1	0.0	0
UR	1.9	17.3	0.1	0	0	**76.4**	0	4.4	0
DR	1.7	17.6	0.1	9.8	0	0	**68.0**	0	2.8
UC	12.7	10.0	0	5.5	0	6.4	0	**62.6**	2.7
DC	1.6	27.5	0	3.1	0	0.4	10.7	6.1	**50.6**

B. Experiments

We evaluate the proposed methods on real user data collected in two different experiments. In the first experiment, 12 healthy young subjects (19-53 years old) were asked to walk twice through a predefined course. The activities exhibited during this course are the ones shown in the left plot of Fig. 3. In total, the data set consists of 98,259 time points. The participants of the second experiment were 15 older adults (80-97 years old), 8 of which were residents of a long term health care facility and regularly using a walker. The participants were asked to walk through two different courses; the resulting activities are shown in the right plot of Figure 3. While they were performing the courses, we asked the participants to execute real-life tasks, e.g., picking up objects from the ground, turning in a confined space, or walking at different speeds. We also recorded some spontaneous activity in between the two courses. In total, the second data set consists of 130,195 time points.

We use leave-one-out cross validation to evaluate the methods. For the maximization of the conditional log-likelihood of the training set, we apply the L-BFGS algorithm (see, e.g., [6]). In order to avoid overfitting, we use an L_2-regularizer the strength of which is determined by cross validation.

V. RESULTS

Table I and II show the confusion matrices for the two experiments. The (h, i)-th entry corresponds to the percentage of time points at which the true label is h, while the algorithm predicts i. The abbreviations for the different activities are the same as in Fig. 3. In total, the accuracy of the predictions is 88.5% for Experiment 1 (standard error 3.2%), and 83.0% for Experiment 2 (standard error 4.0%).

As can be seen from Table I, in Experiment 1 the algorithm has problems to identify transfers (T), which is not surprising since this is an intermediate activity between not touching the walker (N) and stopping (S). Turning left (L) and right (R) is sometimes confused with walking forward (F), however, we found it very difficult to define when a turn exactly starts or ends, so the ground truth is not always unambiguous. Overall, the results for Experiment 1 are better than those for Experiment 2. A possible explanation is that the participants in Experiment 1 performed the course twice, so the training data set always includes one recording of the person for which the activity is currently predicted. Furthermore, some of the activities in Experiment 2 are highly individual; for example, the participants used very different strategies to move the walker up and down the curb. Even for simple activities there is a higher variability in Experiment 2, because the participants were instructed to walk at different speeds or execute various real-life tasks during the course.

Fig. 4 and 5 show online prediction results for Experiment 1 and 2. The plots on the left hand side display the total accuracy for various values of s. As can be seen, the accuracy increases with s. For $s = 25$, which corresponds to a prediction delay of half of a second, the accuracy almost reaches the offline levels of 88.5% and 83.0%.

The plots on the right hand side of Fig. 4 and 5 display the *certainty*, by which we mean the percentage of time points at which the error bounds in Theorem 5 assert that the given online predictions coincide with the predictions based on the complete sequence of observations (as mentioned before, this is the case if there exists an activity whose lower bound is larger than the upper bound of any other activity). As can be seen, also the certainty increases with s and reaches 50-70% for $s = 25$. Hence, if we allow for a prediction delay of half of a second, the vast majority of online predictions is equal to the predictions based on the complete observational sequence, and in 50-70% of the cases we already know this for certain at runtime. Note the particular shape of the certainty curve for Experiment 2: we found that the increase of certainty for $s \leq 5$ is mainly due to accurate predictions of whether a

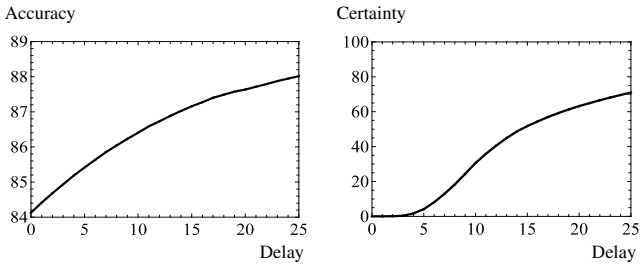

Fig. 4. Online predictions for Experiment 1

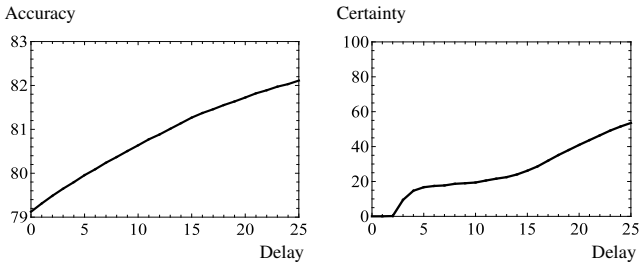

Fig. 5. Online predictions for Experiment 2

person is sitting on a walker or walking forward. In order to predict more complex activities, a larger prediction delay is necessary, leading to the plateau of the curve between $s = 5$ and $s = 15$.

An detailed case study is shown in Fig. 6: the black curve displays the probabilities of the offline predictions, and the gray regions the error bounds for $s = 1, 5, 10, 20, 25$. The black points on the x-axis indicate time points at which the user activity is actually mispredicted. As can be seen, the bounds for $s = 25$ are fairly tight. Interestingly, the certainty varies a lot for different activities. While it is typically high for predictions of walking forward or backwards, it is much lower for turns or intermediate activities like stopping and transfers.

VI. Conclusion

In this paper, we have studied error bounds for online predictions of Linear-Chain Conditional Random Fields. The motivation for this work is that, in many real-world applications, the sequences of observations tend to be very long (up to several hours) while it is desirable to predict the labels close to real-time. Our main result is Theorem 5 which establishes error bounds for online predictions based only on past observations and a small number of observations in the future. An appealing property of these error bounds is that they can be computed at runtime. Hence, at the same time where the online predictions are available, it can be assessed how volatile these predictions are with respect to future observations.

We have demonstrated the practical applicability of our results for a system which recognizes the activity of rolling walker users from a stream of sensor data. We found that, if we allow for a prediction delay of half of a second, the online predictions achieve almost the same accuracy as the offline predictions which take into account the complete sequence of observations. In 50-70% of the cases, the error bounds allow us to determine at runtime whether the online and offline predictions coincide.

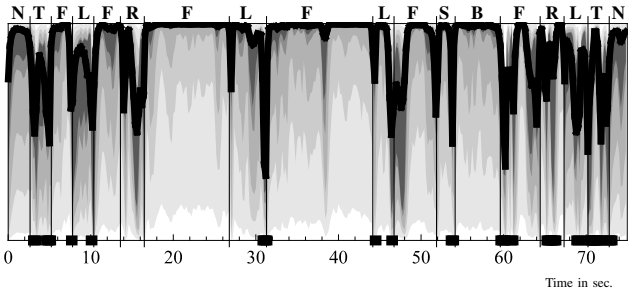

Fig. 6. Error bounds for the left data set of Fig. 3.

References

[1] A. Culotta, D. Kulp and A. McCallum, Gene Prediction with Conditional Random Fields, *Technical Report UM-CS-2005-028*, University of Massachusetts, Amherst, 2005.

[2] T. L. M. van Kasteren, G. Englebienne and B. J. A. Kröse, An Activity Monitoring System for Elderly Care Using Generative and Discriminative Models, *Pers. Ubiquit. Comput.* 14, 489-498, 2010.

[3] J. Lafferty, A. McCallum and F. C. N. Pereira, Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, *Proceedings of the 18th International Conference on Machine Learning*, 2001.

[4] F. Omar, M. Sinn, J. Truszkowski, P. Poupart, J. Tung and A. Caine, Comparative Analysis of Probabilistic Models for Activity Recognition with an Instrumented Walker, *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence*, 2010.

[5] E. Seneta, *Non-negative Matrices and Markov Chains. Revised Edition*, New York, NY: Springer, 2006.

[6] F. Sha and F. Pereira, Shallow parsing with conditional random fields, *Proceedings of the HLT-NAACL*, 2003.

[7] M. Sinn and P. Poupart, Asymptotic Theory for Linear-Chain Conditional Random Fields, *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, 2011.

[8] C. Sutton and A. McCallum, An Introduction to Conditional Random Fields for Relational Learning, *Introduction to Statistical Relational Learning*, L. Getoor and B. Taskar (eds.), Cambridge, Massachusetts: MIT Press, 2006.

[9] J. Tung, W. H. Gage, K. Zabjek, D. Brooks, B. E. Maki, A. Mihailidis, G. Fernie and W. E. McIlroy, iWalker: a 'Real-World' Mobility Assessment Tool. *Proceedings of the 30th Canadian Medical & Biological Engineering Society*, 2007.