

# Discriminative State Space Models

Minyoung Kim<sup>1</sup> and Vladimir Pavlovic<sup>1</sup>

<sup>1</sup>Department of Computer Science  
Rutgers University  
Piscataway, NJ 08854  
{mikim,vladimir}@cs.rutgers.edu

We consider the problem of tracking or state estimation of time-series motion sequences. The problem can be formulated as estimating a continuous multivariate state sequence,  $\mathbf{x} = \mathbf{x}_1 \cdots \mathbf{x}_T$ , from the measurement sequence,  $\mathbf{y} = \mathbf{y}_1 \cdots \mathbf{y}_T$ , where  $\mathbf{x}_t \in \mathbb{R}^d$  and  $\mathbf{y}_t \in \mathbb{R}^k$ . Its applications include 3D tracking of the human motion and pose estimation for moving objects from sequences of monocular or multi-camera images.

A problem resembling the state estimation in tracking, when  $\mathbf{x}_t$  is a *discrete label* instead of continuous multivariate, is known as sequence tagging or segmentation. The most popular generative model in this realm is the Hidden Markov Model (HMM). Traditional Maximum Likelihood (ML) learning of generative models such as HMMs is not directly compatible with the ultimate goal of label prediction (namely,  $\mathbf{x}$  given  $\mathbf{y}$ ), as it optimizes the fit of the models to data jointly,  $\mathbf{x}$  and  $\mathbf{y}$ . Recently, discriminative models such as Conditional Random Fields (CRFs) and Maximum Entropy Markov Models (MEMMs) were introduced to address the label prediction problem directly, resulting in superior performance to the generative models [3,4].

Despite a broad success of discriminative models in the discrete state domain, the use of discriminative dynamic models for continuous multivariate state estimation is not widespread. One reason for this is that a natural reparameterization-based transformation of generative state space models (or dynamical systems) to conditional models may violate density integrability constraints and can often produce unstable systems. For example, an extension of Linear Dynamical System (LDS) to CRF imposes irregular constraints on the CRF parameters to ensure finiteness of the log-partition function, making convex or general gradient-based optimization complex and prone to numerical failure.

As an alternative to CRF-based models in continuous state sequence domains we propose to learn generative state space models discriminatively. This approach has been well studied in machine learning and automatic speech recognition communities for classification settings: Learning generative models such as Tree-Augmented Naive Bayes (TAN) or HMMs discriminatively via maximizing conditional likelihoods yields better prediction performance than the traditional maximum likelihood estimator [1,5,6]. Our main contribution is to extend this approach to dynamic models and the motion tracking problem. Namely, we learn dynamic models that directly optimize the accuracy of pose predictions rather than jointly increasing the data likelihood.

We introduce two discriminative learning algorithms for generative state space models. The first algorithm, named CML, maximizes the conditional log-likelihood of the entire state sequence  $\mathbf{x}$  given the measurement sequence  $\mathbf{y}$ , that is,  $\arg \max \log P(\mathbf{x}|\mathbf{y})$ . The other, called SCML, rather focuses on the estimation error at each time slice individually, namely,  $\arg \max (1/T) \sum_{t=1}^T \log P(\mathbf{x}_t|\mathbf{y})$ . The SCML objective, directly related to minimization of the Hamming distance (in the discrete case), was previously proposed as an alternative objective for CRF in discrete state domain [2]. Both objectives are in general non-convex with respect to the parameters of the generative model, however, the gradient-based optimization yields superior prediction performance to that of the standard ML learning. In addition, we devise computationally efficient methods for gradient evaluation as a part of the proposed framework.

For several human motions (from CMU 3D motion capture data), we compare the prediction performance of the competing models including nonlinear and latent variable dynamic models. The proposed discriminative learning algorithms on LDS can provide significantly lower prediction error than the standard maximum likelihood estimator, often comparable to estimates of computationally more expensive and parameter sensitive nonlinear or latent variable models (See Table 1 and Figure 1). Thus the discriminative state space models offer a highly desired combination of high estimation accuracy and low computational complexity.

**Topic:** estimation and prediction in sequence modeling, learning algorithms, graphical models, control  
**Preference:** oral/poster

Motions	Err.	ML	CML	SCML	NDS	LVN	Motions	Err.	ML	CML	SCML	NDS	LVN
Walk	SJA	19.20	18.31	17.19	18.91	18.01	Walk	S3P	15.28	14.79	13.53	14.62	13.99
	FJA	22.57	22.73	20.78	20.84	19.05		F3P	20.02	20.28	17.07	16.59	14.96
Pick-up	SJA	35.03	33.15	30.56	33.50	32.23	Pick-up	S3P	22.60	21.27	19.33	21.14	20.49
	FJA	42.28	38.89	36.99	41.25	32.10		F3P	25.20	24.36	23.83	25.35	20.40
Run	SJA	23.35	22.11	19.39	21.26	19.08	Run	S3P	21.52	19.85	16.96	18.41	16.97
	FJA	21.87	22.09	20.92	21.86	19.76		F3P	20.40	20.43	18.43	18.42	17.65

Table 1: Average test L2 errors: Error types abbreviated by 3 letters: **S**moothed ( $= (1/T) \sum_{t=1}^T \|\bar{\mathbf{x}}_t - E[\mathbf{x}_t|\mathbf{y}]\|$ ) or **F**iltered ( $= (1/T) \sum_{t=1}^T \|\bar{\mathbf{x}}_t - E[\mathbf{x}_t|\mathbf{y}_1, \dots, \mathbf{y}_t]\|$ ), where  $\bar{\mathbf{x}} =$  (true) states, followed by either joint angle (JA) or 3D point space (3P) in which the error is measured, (e.g., SJA = smoothed error in the joint angle space). Competing models: LDS learned via ML and proposed methods (CML and SCML). NDS = nonlinear dynamical system. LVN = latent-variable nonlinear model, where we place dynamics on newly introduced latent variables obtained from the low-dim embedding of the joint angle space.

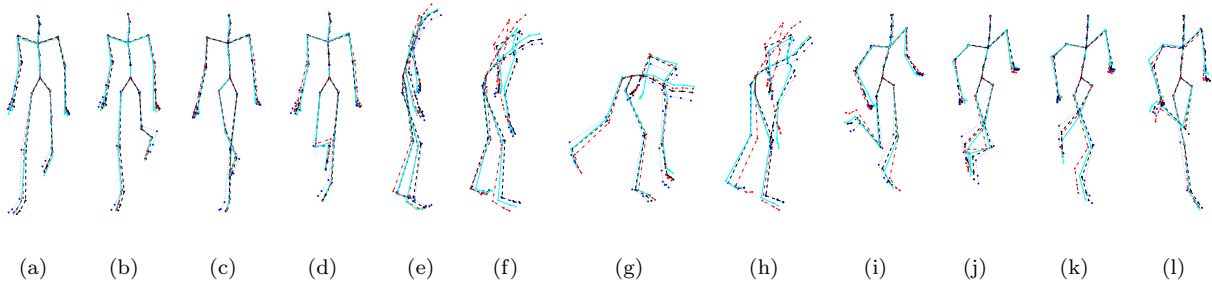


Figure 1: Selected frames of estimated poses for walking (a–d), picking-up a ball (e–h), and running (i–l): The ground-truth is depicted by solid (cyan) lines, ML by dotted (blue), SCML by dashed (black), and LVN by dotted-dashed (red).

## References

- [1] R. Greiner and W. Zhou, “Structural Extension to Logistic Regression: Discriminative Parameter Learning of Belief Net Classifiers,” *Annual National Conference on Artificial Intelligence*, 2002.
- [2] S. Kakade, Y. W. Teh, and S. Roweis, “An Alternative Objective Function for Markovian Fields”, *International Conference on Machine Learning*, 2002.
- [3] J. Lafferty, A. McCallum, and F. Pereira, “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data”, *International Conference on Machine Learning (ICML)*, 2001.
- [4] A. McCallum, D. Freitag, and F. Pereira, “Maximum Entropy Markov Models for Information Extraction and Segmentation”, *International Conference on Machine Learning (ICML)*, 2000.
- [5] A. Y. Ng and M. Jordan, “On Discriminative vs. Generative Classifiers: A Comparison of Logistic Regression and Naive Bayes”, *In Advances in Neural Information Processing Systems (NIPS)*, 2002.
- [6] P. C. Woodland and D. Povey, “Large Scale Discriminative Training For Speech Recognition”, *Proceedings of the Workshop on Automatic Speech Recognition*, 2000.