Density Propagation for Continuous Temporal Chains Generative and Discriminative Models

Cristian Sminchisescu and Allan Jepson
Department of Computer Science, Artificial Intelligence Lab, University of Toronto
{crismin,jepson}@cs.toronto.edu, http://www.cs.toronto.edu/~{crismin,jepson}

Abstract

We analyze non-linear, non-Gaussian temporal chain models (dynamical systems) having continuous hidden states and non-linear, non-Gaussian dynamics and observation models. In this setting we study both *discriminative* and *generative* models, describe their underlying independence assumptions, and give the propagation rules for filtering and smoothing. Despite different graphical model structure and independences, the motivation is similar for using either of these models: infer a dynamically varying hidden state, based on sequences of observations. The setting is common in the solution of many inverse problems in artificial intelligence (*e.g.* computer vision, speech) or control theory. See our companion papers for demonstrations of discriminative [10] and generative [9] models in 3D human motion reconstruction from monocular video applications.

Keywords: *generative models, discriminative models, non-linear systems, variational approximation, mixture models.*

1 Introduction

Consider a non-linear, non-Gaussian, continous chain model (dynamical system) having temporal state $\mathbf{x}_t$, $t = 1 \dots T$, prior $p(\mathbf{x}_1)$, observation model $p(\mathbf{r}_t|\mathbf{x}_t)$ with observations $\mathbf{r}_t$, and dynamics $p(\mathbf{x}_t|\mathbf{x}_{t-1})$. We also define the conditional state distribution $p(\mathbf{x}_t|\mathbf{r}_t)$ and a previous state/current observation-based density $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{r}_t)$ for reasons that will become clear shortly. Let $\mathbf{X}_t = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_t)$ be the model joint state estimated over a time series $1 \dots t, t \leq T$, based on observations encoded as $\mathbf{R}_t = (\mathbf{r}_1, \dots, \mathbf{r}_t)$ or $\mathbf{R}_t^k = (\mathbf{r}_{t-k}, \dots, \mathbf{r}_t)$.

We wish to compute quantities like one-slice, *filtered* conditionals $p(\mathbf{x}_t|\mathbf{R}_t)$, as well as *smoothed* conditionals $p(\mathbf{X}|\mathbf{R}) = p(\mathbf{X}_T|\mathbf{R}_T)$.

The filtered density can then be written as:

$$p(\mathbf{x}_t|\mathbf{R}_t) = \int_{\mathbf{X}_{t-1}} p(\mathbf{x}_t, \mathbf{X}_{t-1}|\mathbf{R}_t) \tag{1}$$

For reasons of tractability, we assume a first-order Markov property: $p(\mathbf{x}_t|\mathbf{X}_{t-1}) = p(\mathbf{x}_t|\mathbf{x}_{t-1})$. Then, (1) simplifies to:

$$p(\mathbf{x}_t|\mathbf{R}_t) = \int_{\mathbf{x}_{t-1}} p(\mathbf{x}_t, \mathbf{x}_{t-1}|\mathbf{r}_t, \mathbf{R}_{t-1}) \tag{2}$$

Similarly, the smoothed distribution can be computed in terms of the joint:

$$p(\mathbf{X}|\mathbf{R}) = \frac{p(\mathbf{X}, \mathbf{R})}{p(\mathbf{R})} \tag{3}$$

where the joint is:

$$p(\mathbf{X}_T, \mathbf{R}_T) = p(\mathbf{X}_{T-1}, \mathbf{R}_{T-1}) p(\mathbf{r}_T | \mathbf{X}_{T-1}, \mathbf{R}_{T-1}) p(\mathbf{x}_T | \mathbf{X}_{T-1}, \mathbf{R}_T) \tag{4}$$

The filtered distribution (2) and the joint (4) will be derived for different choices of density types for both generative and discriminative models.

2 Generative Models

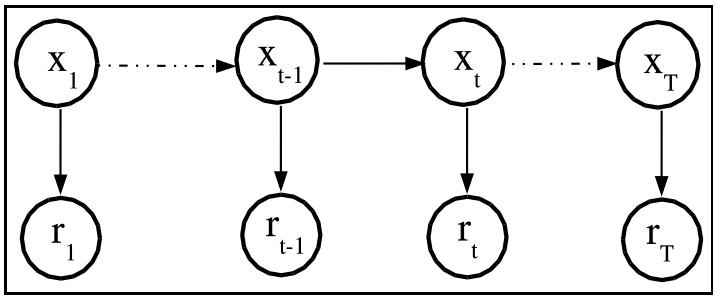

Figure 1: Generative continuous chain model (non-linear dynamical system).

A generative continuous chain model (non-linear dynamical system) is described by the graphical model in fig. 1. In this model the observations are *conditional independent* given the states:

$$\mathbf{r}_t \perp\!\!\!\perp \mathbf{R}_{t-1} | \mathbf{x}_t, \mathbf{x}_{t-1} \tag{5}$$

Also:

$$\mathbf{x}_t \perp\!\!\!\perp \mathbf{R}_{t-1} | \mathbf{x}_{t-1} \tag{6}$$

Different derivations for density propagation in generative models appeared in [2, 6, 4].

2.1 Filtering

The density filtering rule is:

$$p(\mathbf{x}_t | \mathbf{R}_t) = \frac{1}{p(\mathbf{r}_t | \mathbf{R}_{t-1})} p(\mathbf{r}_t | \mathbf{x}_t) \int_{\mathbf{x}_{t-1}} p(\mathbf{x}_t | \mathbf{x}_{t-1}) \, p(\mathbf{x}_{t-1} | \mathbf{R}_{t-1}) \tag{7}$$

Proof of (7)

$$p(\mathbf{x}_t, \mathbf{x}_{t-1} | \mathbf{r}_t, \mathbf{R}_{t-1}) = \frac{p(\mathbf{r}_t, \mathbf{R}_{t-1} | \mathbf{x}_t, \mathbf{x}_{t-1}) p(\mathbf{x}_t, \mathbf{x}_{t-1})}{p(\mathbf{r}_t, \mathbf{R}_{t-1})} = \tag{8}$$

$$= \frac{p(\mathbf{r}_t | \mathbf{x}_t, \mathbf{x}_{t-1}) p(\mathbf{R}_{t-1} | \mathbf{x}_t, \mathbf{x}_{t-1}) p(\mathbf{x}_t, \mathbf{x}_{t-1})}{p(\mathbf{R}_{t-1}) p(\mathbf{r}_t | \mathbf{R}_{t-1})} \tag{9}$$

From(5) and (6), the above transforms to:

$$p(\mathbf{x}_t, \mathbf{x}_{t-1}|\mathbf{r}_t, \mathbf{R}_{t-1}) = \frac{p(\mathbf{r}_t|\mathbf{x}_t)p(\mathbf{R}_{t-1}|\mathbf{x}_{t-1})p(\mathbf{x}_t, \mathbf{x}_{t-1})}{p(\mathbf{R}_{t-1})p(\mathbf{r}_t|\mathbf{R}_{t-1})} = \tag{10}$$

$$= \frac{p(\mathbf{r}_t|\mathbf{x}_t)p(\mathbf{R}_{t-1}|\mathbf{x}_{t-1})p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1})}{p(\mathbf{R}_{t-1})p(\mathbf{r}_t|\mathbf{R}_{t-1})} = \tag{11}$$

$$= \frac{p(\mathbf{r}_t|\mathbf{x}_t)}{p(\mathbf{r}_t|\mathbf{R}_{t-1})}p(\mathbf{x}_t|\mathbf{x}_{t-1})\frac{p(\mathbf{R}_{t-1}|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1})}{p(\mathbf{R}_{t-1})} = \tag{12}$$

$$= \frac{p(\mathbf{r}_t|\mathbf{x}_t)}{p(\mathbf{r}_t|\mathbf{R}_{t-1})}p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{R}_{t-1}) \tag{13}$$

2.1.1 Generative density propagation using conditionals $p(\mathbf{x}_t|\mathbf{r}_t)$ and $p(\mathbf{x}_t|\mathbf{x}_{t-1})$

$$p(\mathbf{x}_t|\mathbf{R}_t) = \frac{p(\mathbf{r}_t)}{p(\mathbf{r}_t|\mathbf{R}_{t-1})}\frac{p(\mathbf{x}_t|\mathbf{r}_t)}{p(\mathbf{x}_t)}\int_{\mathbf{x}_{t-1}} p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{R}_{t-1}) = \tag{14}$$

$$= \frac{p(\mathbf{r}_t)}{p(\mathbf{r}_t|\mathbf{R}_{t-1})}\frac{p(\mathbf{x}_t|\mathbf{r}_t)}{\int_{\mathbf{x}_{t-1}} p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1})}\int_{\mathbf{x}_{t-1}} p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{R}_{t-1}) \tag{15}$$

Proof follows directly from (7) and uses Bayes rule to invert $p(\mathbf{r}_t|\mathbf{x}_t)$.

Given $p(\mathbf{x}_1)$ and $p(\mathbf{x}_t|\mathbf{x}_{t-1})$, we have to recursively propagate both $p(\mathbf{x}_t|\mathbf{R}_t)$ *and* $p(\mathbf{x}_t)$. Over long time series, the latter approaches the state equilibrium distribution, under conditional dynamics $p(\mathbf{x}_t|\mathbf{x}_{t-1})$, and may be approximately precomputed. However, the computation of the filtered posterior involves division of two distributions which may complicate matters, as discussed in §4.3.

2.1.2 A note on forcing a discriminative style distribution $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{r}_t)$

First note that working with $p(\mathbf{r}_t|\mathbf{x}_t, \mathbf{x}_{t-1})$ wouldn't model any additional dependency, since the structure of the graphical model implies $p(\mathbf{r}_t|\mathbf{x}_t, \mathbf{x}_{t-1}) = p(\mathbf{r}_t|\mathbf{x}_t)$. Contrary to the discriminative model §3, in this case $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{r}_t)$ does not have any obvious meaning, *e.g.* as a conditional predicting a node given its parents. However, it is interesting to check whether the propagation rules can be expressed in terms of this distribution. The filtered density can be derived as:

$$p(\mathbf{x}_t|\mathbf{R}_t) = \frac{1}{p(\mathbf{r}_t|\mathbf{R}_{t-1})}\int_{\mathbf{x}_{t-1}} p(\mathbf{r}_t|\mathbf{x}_{t-1})p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{r}_t)p(\mathbf{x}_{t-1}|\mathbf{R}_{t-1}) \tag{16}$$

Proof using (7)

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{r}_t) = \frac{p(\mathbf{x}_{t-1}, \mathbf{r}_t|\mathbf{x}_t)p(\mathbf{x}_t)}{p(\mathbf{x}_{t-1}, \mathbf{r}_t)} = \tag{17}$$

$$= \frac{p(\mathbf{x}_{t-1}|\mathbf{x}_t)p(\mathbf{r}_t|\mathbf{x}_t)p(\mathbf{x}_t)}{p(\mathbf{x}_{t-1})p(\mathbf{r}_t|\mathbf{x}_{t-1})} = \tag{18}$$

$$= \frac{p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{r}_t|\mathbf{x}_t)}{p(\mathbf{r}_t|\mathbf{x}_{t-1})} \tag{19}$$

Despite using $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{r}_t)$, we still can't avoid modeling the current observation based on the previous state, *i.e.* a generative observation conditional $p(\mathbf{r}_t|\mathbf{x}_{t-1})$.

2.2 Smoothing

The joint distribution factorizes as:

$$p(\mathbf{X}_T, \mathbf{R}_T) = p(\mathbf{x}_1) \prod_{t=2}^{T} p(\mathbf{x}_t|\mathbf{x}_{t-1}) \prod_{t=1}^{T} p(\mathbf{r}_t|\mathbf{x}_t) \tag{20}$$

Proof
The following two properties are based on the structure of the graphical model:

$$\mathbf{r}_t \perp\!\!\!\perp \mathbf{X}_{t-2}, \mathbf{R}_{t-1}|\mathbf{x}_{t-1} \tag{21}$$

$$\mathbf{x}_t \perp\!\!\!\perp \mathbf{X}_{t-2}, \mathbf{R}_{t-1}|\mathbf{x}_{t-1} \tag{22}$$

From (21):

$$p(\mathbf{r}_T|\mathbf{X}_{T-1}, \mathbf{R}_{T-1}) = p(\mathbf{r}_T|\mathbf{x}_{T-1}) \tag{23}$$

From (22):

$$p(\mathbf{x}_T|\mathbf{X}_{T-1}, \mathbf{R}_T) = p(\mathbf{x}_T|\mathbf{x}_{T-1}, \mathbf{r}_T) \tag{24}$$

By simplifying (4) using (21) and (22), we obtain:

$$p(\mathbf{X}_T, \mathbf{R}_T) = p(\mathbf{X}_{T-1}, \mathbf{R}_{T-1}) p(\mathbf{r}_T|\mathbf{x}_{T-1}) p(\mathbf{x}_T|\mathbf{x}_{T-1}, \mathbf{r}_T) \tag{25}$$

$$= p(\mathbf{X}_{T-1}, \mathbf{R}_{T-1}) p(\mathbf{r}_T|\mathbf{x}_{T-1}) \frac{p(\mathbf{x}_T, \mathbf{r}_T|\mathbf{x}_{T-1})}{p(\mathbf{r}_T|\mathbf{x}_{t-1})} = \tag{26}$$

$$= p(\mathbf{X}_{T-1}, \mathbf{R}_{T-1}) p(\mathbf{r}_T|\mathbf{x}_T, \mathbf{x}_{T-1}) p(\mathbf{x}_T|\mathbf{x}_{T-1}) \tag{27}$$

Given that:

$$\mathbf{r}_T \perp\!\!\!\perp \mathbf{x}_{T-1}|\mathbf{x}_T \tag{28}$$

We have:

$$p(\mathbf{X}_T, \mathbf{R}_T) = p(\mathbf{X}_{T-1}, \mathbf{R}_{T-1}) p(\mathbf{r}_T|\mathbf{x}_T) p(\mathbf{x}_T|\mathbf{x}_{T-1}) \tag{29}$$

2.2.1 Smoothing using $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ and $p(\mathbf{x}_t|\mathbf{r}_t)$

$$p(\mathbf{X}_T, \mathbf{R}_T) = \prod_{t=1}^{T} p(\mathbf{r}_t) \frac{\prod_{t=2}^{T} p(\mathbf{x}_t|\mathbf{x}_{t-1}) \prod_{t=1}^{T} p(\mathbf{x}_t|\mathbf{r}_t)}{\prod_{t=2}^{T} p(\mathbf{x}_t)} = \tag{30}$$

$$= \prod_{t=1}^{T} p(\mathbf{r}_t) \frac{\prod_{t=2}^{T} p(\mathbf{x}_t|\mathbf{x}_{t-1}) \prod_{t=1}^{T} p(\mathbf{x}_t|\mathbf{r}_t)}{\prod_{t=2}^{T} \int_{\mathbf{x}_{t-1}} p(\mathbf{x}_t|\mathbf{x}_{t-1}) p(\mathbf{x}_{t-1})} \tag{31}$$

Proof follows directly from (20) and (14).

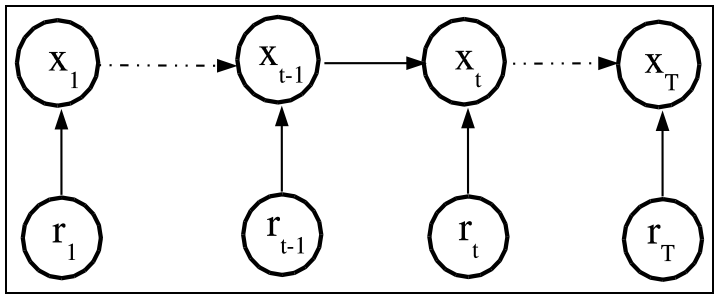

Figure 2: Discriminative continuous chain model.

3 Discriminative Models

A discriminative continuous chain model has the graphical structure shown in fig. 2.

The following properties can be easily verified visually using fig. 2, based on a Bayes ball algorithm:

$$\mathbf{x}_t \perp\!\!\!\perp \mathbf{X}_{t-2} | \mathbf{x}_{t-1} \tag{32}$$

$$\mathbf{r}_t \perp\!\!\!\perp \mathbf{R}_{t-1} \tag{33}$$

$$\mathbf{x}_t \perp\!\!\!\perp \mathbf{R}_{t-1} | \mathbf{x}_{t-1}, \mathbf{r}_t \tag{34}$$

$$\mathbf{X}_{t-1} \perp\!\!\!\perp \mathbf{r}_t \tag{35}$$

In this model the observations are *marginally independent*. Notice how this is different from a generative chain model where the observations are *conditionally independent*, given the states.

3.1 Filtering using $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{r}_t)$

The density propagation rule is:

$$p(\mathbf{x}_t|\mathbf{R}_t) = \int_{\mathbf{x}_{t-1}} p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{r}_t) p(\mathbf{x}_{t-1}|\mathbf{R}_{t-1}) \tag{36}$$

Proof

$$p(\mathbf{x}_t|\mathbf{R}_t) = \int_{\mathbf{x}_{t-1}} p(\mathbf{x}_t, \mathbf{x}_{t-1}|\mathbf{R}_{t-1}, \mathbf{r}_t) = \tag{37}$$

$$= \int_{\mathbf{x}_{t-1}} p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{R}_{t-1}, \mathbf{r}_t) p(\mathbf{x}_{t-1}|\mathbf{R}_{t-1}, \mathbf{r}_t) = \tag{38}$$

$$= \int_{\mathbf{x}_{t-1}} p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{r}_t) p(\mathbf{x}_{t-1}|\mathbf{R}_{t-1}) \tag{39}$$

where in the last line we used:

$$(34) \Rightarrow p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{R}_{t-1}, \mathbf{r}_t) = p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{r}_t)$$
$$(35) \Rightarrow p(\mathbf{x}_{t-1}|\mathbf{R}_{t-1}, \mathbf{r}_t) = p(\mathbf{x}_{t-1}|\mathbf{R}_{t-1})$$

3.1.1 A note on working with $p(\mathbf{x}_t|\mathbf{r}_t)$ and $p(\mathbf{x}_t|\mathbf{x}_{t-1})$

It appears natural to ask whether we can derive the filtering recursions using separate conditionals $p(\mathbf{x}_t|\mathbf{r}_t)$ and $p(\mathbf{x}_t|\mathbf{x}_{t-1})$. However, for discriminative chains, 'explaining away' effects prevent a simple factorization of $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{r}_t)$ into separate observation and dynamic state conditionals. While $\mathbf{x}_{t-1}$ and $\mathbf{r}_t$ are marginally independent, they become conditionally dependent when observing (conditioning on) $\mathbf{x}_t$. Several partial derivations are possible, but none seems to be satisfactory:

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{r}_t) = \frac{p(\mathbf{x}_{t-1}, \mathbf{r}_t|\mathbf{x}_t)p(\mathbf{x}_t)}{p(\mathbf{x}_{t-1}, \mathbf{r}_t)} \tag{40}$$
$$= \frac{p(\mathbf{r}_t|\mathbf{x}_{t-1}, \mathbf{x}_t)p(\mathbf{x}_{t-1}|\mathbf{x}_t)p(\mathbf{x}_t)}{p(\mathbf{x}_{t-1})p(\mathbf{r}_t)} = \tag{41}$$
$$= \frac{p(\mathbf{r}_t|\mathbf{x}_{t-1}, \mathbf{x}_t)p(\mathbf{x}_t|\mathbf{x}_{t-1})}{p(\mathbf{r}_t)} = \tag{42}$$
$$= \frac{p(\mathbf{r}_t|\mathbf{x}_{t-1}, \mathbf{x}_t)}{p(\mathbf{r}_t)} p(\mathbf{x}_t|\mathbf{x}_{t-1}) = \tag{43}$$
$$= \frac{p(\mathbf{r}_t|\mathbf{x}_t, \mathbf{x}_{t-1})}{p(\mathbf{r}_t|\mathbf{x}_t)} \frac{p(\mathbf{x}_t|\mathbf{r}_t)p(\mathbf{x}_t|\mathbf{x}_{t-1})}{p(\mathbf{x}_t)} \tag{44}$$

An alternative derivation gives:

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{r}_t) = \frac{p(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{r}_t)p(\mathbf{x}_t|\mathbf{r}_t)p(\mathbf{r}_t)}{p(\mathbf{x}_{t-1}, \mathbf{x}_t)} \tag{45}$$
$$= p(\mathbf{r}_t) \frac{p(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{r}_t)}{p(\mathbf{x}_{t-1})} \frac{p(\mathbf{x}_t|\mathbf{r}_t)}{p(\mathbf{x}_t|\mathbf{x}_{t-1})} \tag{46}$$

3.2 Smoothing using $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{r}_t)$

$$p(\mathbf{X}_T, \mathbf{R}_T) = \prod_{t=1}^{T} p(\mathbf{r}_t) \prod_{t=2}^{T} p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{r}_t) \tag{47}$$

Proof using (4)
From (35):

$$p(\mathbf{r}_T|\mathbf{X}_{T-1}, \mathbf{R}_{T-1}) = p(\mathbf{r}_T) \tag{48}$$

$$p(\mathbf{x}_T|\mathbf{x}_{T-1}, \mathbf{X}_{T-2}, \mathbf{R}_{T-1}, \mathbf{r}_T) = p(\mathbf{x}_T|\mathbf{x}_{T-1}, \mathbf{r}_T) \tag{49}$$

Then:

$$p(\mathbf{X}_T, \mathbf{R}_T) = p(\mathbf{X}_{T-1}, \mathbf{R}_{T-1})p(\mathbf{r}_T)p(\mathbf{x}_T|\mathbf{x}_{T-1}, \mathbf{r}_T) \tag{50}$$

4 Discriminative Models with Longer Windows of Observations

These are model where each state is conditioned by a k-window of observations in the past. For simplicity we work with running time indexes starting at 1, and assume, without loss of generality, that the k past observations are available when estimation starts.

4.1 Filtering using $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{R}_t^k)$

$$p(\mathbf{x}_t|\mathbf{R}_t) = \int_{\mathbf{x}_{t-1}} p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{R}_t^k) p(\mathbf{x}_{t-1}|\mathbf{R}_{t-1}) \tag{51}$$

Proof follows directly from (37) by this time taking the longer range observation dependency into account:

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{R}_t) = p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{R}_t^k) \tag{52}$$

4.2 Smoothing using $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{R}_t^k)$

$$p(\mathbf{X}_T, \mathbf{R}_T) = \prod_{t=1}^{T} p(\mathbf{r}_t) \prod_{t=2}^{T} p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{R}_t^k) \tag{53}$$

Proof follows directly from (4), similarly with (47) and using (52).

4.3 Mixture of Gaussian Density Propagation

In this section we will give formulas for density propagation in the case where the relevant distributions, *i.e.* the temporal prior and the local state conditionals are represented as mixtures. All cases can be treated similarly, here we give formulas for density propagation using generative models with separate dynamic and observation-based, discriminative style state conditionals §2.1.1, which is more complex to compute. We focus on one Gaussian component in each of the resulting distributions involved. In principle we have $\mathcal{O}(M^3)$ such triples, where M is, say, an upper bound on the number of components in each distribution:

$$p(\mathbf{x}_{t-1}|\mathbf{R}_{t-1}) = \mathcal{G}_{\mathbf{x}_{t-1}}[\mathbf{y}, \mathbf{P}] \tag{54}$$

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{G}_{\mathbf{x}_t}[\mathbf{A}\mathbf{x}_{t-1}, \mathbf{Q}], \mathbf{A} = \frac{d\mathbf{F}}{d\mathbf{x}}(\mathbf{x} = \mathbf{x}_{t-1} = \mathbf{y}) \tag{55}$$

$$p(\mathbf{x}_t|\mathbf{r}_t) = \mathcal{G}_{\mathbf{x}_t}[\mathbf{B}\mathbf{x}_t, \mathbf{Z}], \mathbf{B} = \frac{d\mathbf{H}}{d\mathbf{r}}(\mathbf{r} = \mathbf{r}_t) \tag{56}$$

In the above we will assume that $\mathbf{F}$ and $\mathbf{H}$ are non-linear (*e.g.* regressor) functions. In this case, it may be necessary to do statistical linearization in order to account for the uncertainty of using a linear approximation, *i.e.* the matrices $\mathbf{Q}$ and $\mathbf{Z}$. This can be done using an unscented transform [7]. The integral representing the predicted distribution for a single doublet (out of M^2) of Gaussian mixture components in $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ and $p(\mathbf{x}_{t-1}|\mathbf{R}_{t-1})$ is Gaussian:

$$\int_{\mathbf{x}_{t-1}} p(\mathbf{x}_t|\mathbf{x}_{t-1}) p(\mathbf{x}_{t-1}|\mathbf{R}_{t-1}) = \mathcal{G}_{\mathbf{x}_t}[\mathbf{A}\mathbf{y}, \mathbf{A}\mathbf{P}\mathbf{A}^\top + \mathbf{Q}] \tag{57}$$

The numerator of the filtered distribution $p(\mathbf{x}_t|\mathbf{R}_t)$ is then a product of two Gaussians:

$$\mathcal{G}_{\mathbf{x}_t}[\mathbf{B}\mathbf{x}_t, \mathbf{Z}]\mathcal{G}_{\mathbf{x}_t}[\mathbf{A}\mathbf{y}, \mathbf{A}\mathbf{P}\mathbf{A}^\top + \mathbf{Q}] \tag{58}$$

The product of two Gaussians is again a Gaussian but not normalized *i.e.*:

$$\mathcal{G}_{\mathbf{x}}[\mathbf{a}, \mathbf{A}]\mathcal{G}_{\mathbf{x}}[\mathbf{b}, \mathbf{B}] \propto \mathcal{G}_{\mathbf{x}}[\mathbf{c}, \mathbf{C}] \tag{59}$$

where

$$\mathbf{C} = (\mathbf{A}^{-1} + \mathbf{B}^{-1})^{-1}, \mathbf{c} = \mathbf{C}\mathbf{A}^{-1}\mathbf{a} + \mathbf{C}\mathbf{B}^{-1}\mathbf{b} \tag{60}$$

and the normalization factor, say Z is (where d is the dimensionality of the state $\mathbf{c}$):

$$Z = (2\pi)^{-d/2}|\mathbf{C}|^{1/2}|\mathbf{A}|^{-1/2}|\mathbf{B}|^{-1/2}\exp[-\frac{1}{2}(\mathbf{a}^\top\mathbf{A}^{-1}\mathbf{a} + \mathbf{b}^\top\mathbf{B}^{-1}\mathbf{b} - \mathbf{c}^\top\mathbf{C}^{-1}\mathbf{c})] \tag{61}$$

The denominator of the filtered distribution $p(\mathbf{x}_t|\mathbf{R}_t)$ can be derived similarly to (57).

Given the above, with p and s being distributions for the numerator (M^3 components) and denominator (M^2 components) of $p(\mathbf{x}_t|\mathbf{R}_t)$, respectively, we seek an $\mathcal{O}(M)$ component distribution q_θ, parameterized by θ, that approximates the ratio:

$$p(\mathbf{x}_t|\mathbf{R}_t) = \frac{p}{s} \approx q_\theta \tag{62}$$

This is equivalent to minimizing the KL divergence $KL(\frac{p}{s}||q_{\boldsymbol{\theta}})$, and can be optimized using gradient descent to find θ [9].

References

[1] J. Binder, K. Murphy, and S. Russell. Space-efficient inference in dynamic probabilistic networks. In *International Joint Conference on Uncertainty in Artificial Intelligence*, 1997.

[2] N. Gordon, D. Salmond, and A. Smith. Novel Approach to Non-linear/Non-Gaussian State Estimation. *IEE Proc. F*, 1993.

[3] M. Isard. PAMPAS: Real-valued graphical models for computer vision. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2003.

[4] M. Isard and A. Blake. CONDENSATION – Conditional Density Propagation for Visual Tracking. *International Journal of Computer Vision*, 1998.

[5] M. Jordan, editor. *Learning in graphical models.* MIT Press, 2001.

[6] G. Kitagawa. Monte Carlo Filter and Smoother for Non-Gaussian Nonlinear State Space Models. *J. Comput. Graph. Statist.*, 1996.

[7] R. Merwe, A. Doucet, N. Freitas, and E. Wan. The Unscented Particle Filter. Technical Report CUED/F-INFENG/TR 380, Cambridge University, Department of Engineering, May 2000.

[8] C. Sminchisescu and A. Jepson. Generative Modeling for Continuous Non-Linearly Embedded Visual Inference. In *International Conference on Machine Learning*, pages 759–766, Banff, 2004.

[9] C. Sminchisescu and A. Jepson. Variational Mixture Smoothing for Non-Linear Dynamical Systems. In *IEEE International Conference on Computer Vision and Pattern Recognition*, volume 2, pages 608–615, Washington D.C., 2004.

[10] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas. 3D human motion reconstruction using Bayesian mixtures of experts. A probabilistic discriminative approach. Technical Report CSRG-502, University of Toronto, October 2004.

[11] E. Sudderth, A. Ihler, W. Freeman, and A.Wilsky. Non-parametric belief propagation. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2003.