Information-Theoretic Policy Extraction from Partial Observations

Tom Lefebvre

Abstract—We investigate the problem of extracting a control policy from a single or multiple partial observation sequences. Therefore we cast the problem as a Controlled Hidden Markov Model. We then sketch two information-theoretic approaches to extract a policy which we refer to as A Posteriori Control Distributions. The performance of both methods is investigated and compared empirically on a linear tracking problem.

I. INTRODUCTION

Bayesian estimation [1], [2], [3] is a well-known and well-studied paradigm to treat generic estimation problems involving time-series data. The problem is stated in a probabilistic context representing the system as a Hidden Markov Model (HMM) (Fig. 1). Typically a series of measurements is available from which the most likely hidden state trajectory needs to be recovered or, when the application requires so, distributions that characterise the probability of that state. The type of probabilistic state-space models that enjoy our interest and their filtering and smoothing solutions have numerous applications, for example, in navigation, tracking, and audio- and biomedical signal processing [4], [5], [6].

For some applications the system and therefore the model can easily be extended to incorporate some external stimulus that influences the dynamics of the system. An example of such an extension is the case of a feedback policy.

The problem of extracting a policy from demonstrations has been investigated by the Reinforcement Learning (RL) community where it is referred to as inverse RL (IRL). There the policy is modelled as a realisation of an optimal control problem, yet capable of reproducing the behavioural features of the measurements? See for example Fig. 2.

A. Preliminaries

1) CHMMs: We consider controlled Hidden Markov Models (CHMMs) as a trivial extension of regular HMMs obtained by extending the model with a probabilistic control variable. Apart from time dependent state- and observation variables, Z ∈ Rn is, in a particular we define control variables $u ∈ R^m$. See Fig. 1 for a graphical representation of the probabilistic model. The graph model is characterised by a set of conditional probabilities, M, describing the probabilistic dynamics of the system M contains an initial state distribution $x_0 ∼ p(x_0)$, a transition probability $x_t ∼ p(x_{t+1}|x_t, u_t)$, an emission probability $z_t ∼ p(z_t|x_t, u_t)$ and a probabilistic policy $u_t ∼ p(u_t|x_t)$. Note that a CHMM is therefore not equivalent to a Partially Observable Markov Decision Process. Process for notational brevity we further introduce the variable tuple $ξ = (x, u) ∈ R^n$ and the formatted notation $X_t = \{x_0, ..., x_t\}$ and $X_t = \{x_1, ..., x_T\}$. Following the structural properties of the graphical model the joint probabilities $p(ζ_T)$ and $p(ζ_T, Z_T)$ can be factorized

\[
p(ζ_T) = \prod_{t=0}^{T} p(x_{t+1}|x_t, u_t) p(u_t|x_t)
\]

\[
p(ζ_T, Z_T) = p(ζ_T) \prod_{t=0}^{T} p(z_t|x_t, u_t)
\]

The model is Markovian which implies that past measurements do not contain more information about the transition, control and emission probabilities than the present state-action itself. Mathematically this property manifests as the conditional independence on historical measurements (not necessarily future measurements). We have that $p(x_t|ξ_{t-1}, Z_{t-1}) = p(x_t|ξ_{t-1})$, $p(u_t|x_t, Z_{t-1}) = p(u_t|x_t)$ and $p(z_t|ξ_t, Z_t) = p(z_t|ξ_t)$. When the system manifests into a path $(ξ, Z_T)$, it does so according to the probabilities characterised by M.

2) The I- and M-projection: Probabilistic inference refers to the process of reasoning with incomplete information according to rational principles. Inference principles therefore determine how to update a prior belief into a posterior belief when new information becomes available. Bayesian inference can be used to process information that is represented by the outcome of experiments, i.e. empirical evidence. The I- and M-projection are information-theoretic concepts that can be used to process information represented by constraints that affect the belief space. We refer to such evidence as structural.
The concepts are based on the relative entropy $\mathbb{D}[\pi \parallel \rho] = \mathbb{E}_\pi[\log \frac{\pi}{\rho}]$ between distributions $\pi \in \mathcal{P}$ and $\rho \in \mathcal{P}$ where $\mathcal{P}$ denotes the space of all probability distributions. The relative entropy is a measure of the inefficiency of assuming that the distribution is $\rho$ when the true distribution is $\pi$. The relative entropy is always positive, zero only for $\pi \equiv \rho$ and asymmetric. According to the principle of maximum entropy advocated by Jaynes amongst others [8], the relative entropy must be minimized if we want to end up with some form of new information (usually an expectation, i.e. $\mathbb{E}_\pi[f] = \mu$) into the prior $\rho$. The I-projection and M-projection behave differently [7].

### I-projection

The I-projection and maximum entropy principle are equivalent with $\pi^*$ the I-projection of $\rho$ onto $\mathcal{P}^*$.

$$
\pi^* = \arg \min_{\pi \in \mathcal{P}^*} \mathbb{D}[\pi \parallel \rho]
$$

### M-projection

The M-projection is the reciprocal of the I-projection with $\pi^*$ the M-projection of $\rho$ onto $\mathcal{P}^*$.

$$
\pi^* = \arg \min_{\pi \in \mathcal{P}^*} \mathbb{D}[\rho \parallel \pi]
$$

The set $\mathcal{P}^* \subset \mathcal{P}$ represents the constrained belief space. As the relative entropy is not symmetric in its arguments, the I-projection and the M-projection behave differently [1].

#### B. Vanilla APCD

We will refer to any of the proposed policy extraction methods as A Posteriori Control Distributed (APCDs). Our first strategy to derive an APCD follows a standard Bayesian approach. We investigate its properties and explore whether the first strategy to derive an APCD follows a standard Bayesian approach. We investigate its properties and explore whether the first strategy to derive an APCD follows a standard Bayesian approach. We investigate its properties and explore whether the first strategy to derive an APCD follows a standard Bayesian approach. We investigate its properties and explore whether the first strategy to derive an APCD follows a standard Bayesian approach. We investigate its properties and explore whether the first strategy to derive an APCD follows a standard Bayesian approach. We investigate its properties and explore whether the first strategy to derive an APCD follows a standard Bayesian approach. We investigate its properties and explore whether the first strategy to derive an APCD follows a standard Bayesian approach. We investigate its properties and explore whether the first strategy to derive an APCD follows a standard Bayesian approach. We investigate its properties and explore whether the first strategy to derive an APCD follows a standard Bayesian approach.

1) **Bayesian approach**: We propose to synthesise an APCD by conditioning the posterior marginal, $p(\xi_t|\mathcal{Z}_T)$, on the state, i.e. $p(u_t|x_t, \mathcal{Z}_T)$. As a result of the Markov assumption the distribution must be equivalent to $p(u_t|x_t, \mathcal{Z}_t)$. This resonates with the common sense that once we arrive at some state, $x_t$, we can only hope to reproduce measurements $Z_t$ but have no more effect on the preceding measurements $Z_{t-1}$. Consequently the V-APCD is subject to interesting structures.

Instead of relying on the Bayesian smoothing equations to compute $p(\xi_t|\mathcal{Z}_T)$ and then $p(u_t|x_t, \mathcal{Z}_T)$ we use Bayes’ rule to decompose the V-APCD as follows

$$
p(u_t|x_t, \mathcal{Z}_t) = \rho(u_t|x_t) \frac{p(\mathcal{Z}_t|\xi_t)}{p(\mathcal{Z}_t|x_t)}
$$

which reduces the problem to finding efficient expressions for the probabilities $p(\mathcal{Z}_t|x_t)$ and $p(\mathcal{Z}_t|\xi_t)$. The latter is a generalisation of the backward filtering distribution for CHMMs. The former can be derived from there.

$$
p(\mathcal{Z}_t|x_t) = \int p(u_t|x_t)p(\mathcal{Z}_t|\xi_t)du_t = \mathbb{E}_{\rho(u_t|x_t)}[p(\mathcal{Z}_t|\xi_t)]
$$

One now easily verifies that the distribution $p(\mathcal{Z}_t|\xi_t)$ is governed by a backward recursive expression.

$$
p(\mathcal{Z}_t|\xi_t) = p(z_t|\xi_t) \int p(x_{t+1}|\xi_t)p(\mathcal{Z}_{t+1}|x_{t+1})dx_{t+1} = p(z_t|\xi_t)\mathbb{E}_{p(x_{t+1}|\xi_t)}[p(\mathcal{Z}_{t+1}|x_{t+1})]
$$

Both the problem statements themselves, as well as the backward recursive calculation procedure, hint at a connection with the theory of optimal control and dynamic programming. For given $\mathcal{Z}_T$ we can define the functions $Q^*_t(\xi_t) = l(\mathcal{Z}_t|\xi_t)$, $V^*_t(x_t) = l(\mathcal{Z}_t|x_t)$ where we introduced the likelihood $l(\cdot) \equiv -\log p(\cdot)$ and defined $\pi^*_t(x_t) = p(u_t|x_t, \mathcal{Z}_t)$. It follows that

$$
Q^*_t(\xi_t) = l(z_t|\xi_t) - \log \mathbb{E}_{\rho(x'_t|\xi_t)}[\exp(-V^*_{t+1}(x'_t))]
$$

$$
V^*_t(x_t) = -\log \mathbb{E}_{\rho(x'_t|x_t)}[\exp(-Q^*_t(\xi_t))]
$$

whereas the V-APCD is given by

$$
\pi^*_t(u_t|x_t) = \rho(u_t|x_t)\exp(V^*_t(x_t) - Q^*_t(\xi_t))
$$

These expressions exhibit remarkable similarities with the risk-sensitive optimal control framework [9].

2) **Information-theoretic approach**: First we raise the question whether the V-APCD renders some objective function optimal? In particular we show that the V-APCD is governed by the M-projection of $p(\mathbb{E}_{\rho}|\mathcal{Z}_T)$ onto the probability space spanned by $p(\mathbb{E}_{\pi}; \mathcal{Z}_T)$ where $p(\mathbb{E}_{\pi}; \mathcal{Z}_T)$ is defined as the joint distribution obtained by administering some distributional policy sequence $\mathcal{Z}_T$ instead of the probabilistic control model $\rho_T(x)$. Note that therefore $p(\mathbb{E}_{\rho}|\mathcal{Z}_T)$ is conditioned on the models $\rho_T$. To keep notation light this subtlety was not included. We can further extend our investigation by considering a sum over $N$ independent measurement sequences $\{\mathcal{Z}_T^n\}$. We consider the following variational optimization problem which can be manipulated into $T - 1$ separate problems.

$$
\arg \min_{\mathbb{E}_{\rho} \in \mathcal{P}} \sum_n \mathbb{D}[p(\mathbb{E}_{\rho}|\mathcal{Z}_T^n) \parallel p(\mathbb{E}_{\rho}; \mathcal{Z}_T^n)]
$$

$$
= \arg \max_{\mathbb{E}_{\rho} \in \mathcal{P}} \sum_n \int p(\mathbb{E}_{\rho}; \mathcal{Z}_T^n) \log \frac{p(\mathbb{E}_{\rho}; \mathcal{Z}_T^n)}{p(\mathbb{E}_{\rho}|\mathcal{Z}_T^n)} d\mathbb{E}_{\rho}
$$

$$
= \arg \min_{\mathbb{E}_{\rho} \in \mathcal{P}} \sum_n \sum_i \int p(\mathbb{E}_{\rho}; \mathcal{Z}_T^n) \log \pi^*_i(u_t|x_t) d\mathbb{E}_{\rho}
$$

Taking into account the normalization condition it follows that the solution is governed by the V-APCD for $N = 1$ and by a mixture of individual V-APCDs for $N > 1$.

$$
\pi^*_t(u_t|x_t) = \frac{p(\xi_t|\mathcal{Z}_T^n)}{\sum_n p(x_t|\mathcal{Z}_T^n)} = \frac{p(x_t|\mathcal{Z}_t, \mathcal{Z}_T^n)}{\sum_n p(x_t|\mathcal{Z}_T^n)} p(u_t|x_t, \mathcal{Z}_t)
$$

### C. Natural APCD

Now we raise the obvious question whether the reciprocal I-projection also generates an APCD? According to the principle of Maximum Entropy it follows that then we minimize the inefficiency of assuming the prior control distribution $\rho_T(u_t|x_t)$ whilst the true control distribution is given by the posterior $\pi^*_t(u_t|x_t)$. This generates a related APCD that collapses onto the V-APCD under some conditions. We are inclined to refer to the APCD derived here as the natural APCD (N-APCD). Our use of terminology is inspired by vanilla and natural gradients in policy search where the natural gradient follows a relative entropy constraint corresponding the I-projection [10].

Here we consider the variational optimization problem

$$
\min_{\mathbb{E}_{\rho} \in \mathcal{P}} \sum_n \mathbb{D}[p(\mathbb{E}_{\rho}|\mathcal{Z}_T^n) \parallel p(\mathbb{E}_{\rho}; \mathcal{Z}_T^n)]
$$

which can be recast as

$$
\min_{\mathbb{E}_{\rho} \in \mathcal{P}} \mathbb{E}_{p(\mathbb{E}_{\rho}|\mathcal{Z}_T^n)} \left[ \log \frac{p(\mathbb{E}_{\rho}|\mathcal{Z}_T^n)}{p(\mathbb{E}_{\rho}; \mathcal{Z}_T^n)} - \frac{1}{N} \sum_n \log p(\mathcal{Z}_T^n|\mathcal{Z}_T) \right]
$$
This problem can be treated as an optimal control problem as it exhibits an optimal substructure which we may exploit to solve it. Inspired by the structural findings in sec. [II-B1] let us define the following V-function

\[ V^*(x_t) = \min_{\pi_t \in P} \mathbb{E}_{p(x_{t+1}|x_t, \pi_t)} \left[ \log \frac{p_t}{\pi_t} - \frac{1}{N} \sum_n \log p(Z_t^n|\Xi_t) \right] \]

which satisfies the recursion

\[ V^*_t(x_t) = \min_{\pi_t \in P} \mathbb{E}_{\pi_t(u_t|x_t)} \left[ \log \frac{p_t(x_t|u_t)}{\pi_t(u_t|x_t)} + Q^*_t(\xi_t) \right] \]

where

\[ Q^*_t(\xi_t) = \frac{1}{N} \sum_{n} l(x_t^n|\xi_t) + \mathbb{E}_{p(x_{t+1}|\xi_t)}[V^*_t(x_{t+1})] \]  

(3)

Variational optimization then yields the N-APCD

\[ V^*_t(x_t) = -\log \mathbb{E}_{p(x_t|u_t)} [\exp(-Q^*_t(\xi_t))] \]

\[ \pi^*_t(x_t|u_t) = \rho_t(u_t|x_t) \exp(V^*_t(x_t) - Q^*_t(\xi_t)) \]  

(4)

On first notice the solution of the vanilla and natural problems are equivalent for \( N = 1 \). At least the recursive expressions for the APCDs and V-functions are. The subtle difference lies in the definition of the Q-functions. Comparing definitions for \( Q^*_t(\xi) \) and \( Q^*_t(\xi) \) reveals that the former additionally transforms the expectation in agreement with the probability-likelihood transformation (i.e. \( l = -\log p) \) so that the apparent addition in likelihood space in fact amounts to a multiplication in probability space. In contrast the natural approach is to carry out the computation in likelihood space. The difference is rendered irrelevant for deterministic dynamics.

D. Exact solutions

The V-ACDP and N-ACDP defined in (2) and (4) render explicit solutions for linear-Gaussian models. We consider \( x_0 \sim \mathcal{N}(\mu_0, \Sigma_0), x_{t+1} = F_t x_t + f_t + q_t, u_t = K_t x_t + k_t + s_t \)

and \( z_t = G_t x_t + g_t + r_t \) where \( q_t \sim \mathcal{N}(0, Q_t), s_t \sim \mathcal{N}(0, S_t) \) and \( r_t \sim \mathcal{N}(0, R_t) \). This linear-Gaussian context it is reasonable to assume that \( \pi^*_t = \mathcal{N}(u_t|K_t x_t + k_t, \Sigma_t) \) and that both \( Q^*_t \) and \( V^*_t \) are quadratic in their arguments, with \( * \in \{\star, \bullet\} \). Further we use the symbol \( \times \) in to constants that do not affect the computations. As such

\[ V^*_t(x_t) = \frac{1}{2} \begin{bmatrix} 1 & x_t \end{bmatrix}^T \begin{bmatrix} V_{xx,t} & V_{x,t} \\ V_{x,t} & V_{t} \end{bmatrix} \begin{bmatrix} 1 \\ x_t \end{bmatrix} \]

\[ Q^*_t(\xi_t) = \frac{1}{2} \begin{bmatrix} 1 & \xi_t \end{bmatrix}^T \begin{bmatrix} Q_{\xi\xi,t} & Q_{\xi,t} \\ Q_{\xi,t} & Q_{t} \end{bmatrix} \begin{bmatrix} 1 \\ \xi_t \end{bmatrix} \]

Then provided that we have gained access to values for \( Q^*_t \) and \( Q^*_t \) and observing that whether the expressions for \( V^*_t, \pi^*_t, V^*_t \) and \( \pi^*_t \) respectively defined in (1), (3) and (4), are equivalent, we can derive expressions that hold irrespective whether we consider the vanilla or natural APCD.

Since \( \pi^*_t(u_t|x_t) \propto \rho_t(u_t|x_t) \exp(-Q^*_t(\xi_t)) \) one verifies that

\[ k_t^\star = \Sigma_t^* (S_t^{-1} k_t - Q_{u,t}^*) \]

\[ K_t^\star = \Sigma_t^* (S_t^{-1} K_t - Q_{x,t}^*) \]

\[ \Sigma_t^* = (S_t^{-1} + Q_{u,t}^*)^{-1} \]

and since \( \exp(-V^*_t(x_t)) \pi^*_t(u_t|x_t) = \rho(u_t|x_t) \exp(-Q^*_t(\xi_t)) \) for any \( u_t \) including 0 it also follows that

\[ V^*_{x,t} = Q^*_{x,t} + K_t^\top S_t^{-1} K_t - K_t^\top \Sigma_t^{-1} k_t^\star \]

\[ V^*_{x,t} = Q^*_{x,t} + K_t^\top S_t^{-1} K_t - K_t^\top \Sigma_t^{-1} K_t^\star \]

Only the update for the \( Q^*\)-function differs between the vanilla and natural APCDs. We can derive expression based on the definitions in (1) and (3) respectively. Therefore we further introduce a quadratic expression for \( I(z_t|\xi_t) \). Note that the parameters depend on \( G_{t}, u_t \) and \( R_t \).

\[ I(z_t|\xi_t) = R_t(\xi_t) = \frac{1}{2} \begin{bmatrix} 1 & \xi_t \end{bmatrix} \begin{bmatrix} R_{\xi\xi,t} & R_{\xi,t} \\ R_{\xi,t} & R_t \end{bmatrix} \begin{bmatrix} 1 \\ \xi_t \end{bmatrix} \]

1) V-APCD: For the vanilla \( Q^*\)-function we find

\[ Q^*_{\xi,t} = R_{\xi,t} + F_t^\top (V_{x,t}^{-1} + Q_t)^{-1} (V_{x,t}^{-1} V_{x,t} + f_t) \]

\[ Q^*_{\xi,t} = R_{\xi,t} + F_t^\top (V_{x,t}^{-1} + Q_t)^{-1} F_t \]

This expression holds both for single as well as multiple measurement sequences. In the latter case \( R_t(\xi_t) \) is given by the average as seen in (3). This means that in either case the N-APCD is given by a linear-Gaussian policy. For any nonlinear model the solution is intractable but might be retrieved locally by iterating the linear solutions.

III. Experiments

In this section we document numerical experiments to validate the algorithms proposed in sec. [II]. All experiments were implemented using Matlab. Each experiment was executed on a single 2.10GHz Intel Xeon Gold 6130 processor.

A. Problem definition

We consider a force controlled planar mass with Brownian input noise (see Fig. [2]). The covariance of the input noise is spawned by cascading the \( \text{rand} \) and \( \text{sprandsym} \) commands generating correlated white noise. The random seed is multiplied with 10 and \( 10^2 \) in its respective dimensions to realise anisotropy. The system is discretised using a sample period \( \Delta t = 2 \cdot 10^{-3}s \). We define a path tracking problem over the horizon \( T = 2s \). The system is controlled using a Linear Quadratic Exponential Regulator (LQER) (11) which minimizes the objective defined below. Here \( p_t, p_t^* \) and \( v_t \) define the position, desired position and velocity of the particle respectively. The LQER is chosen over the standard LQR because the LQER can take into account the anisotropic related input noise, otherwise the control would be uncorrelated. We set \( R_p = 10^4 \) and \( R_p = R_v = 1 \) and \( \lambda = 10^{-4} \). It is well known that the solution is given by a time dependent linear policy, i.e. \( u_t = k_{tq} + K_{tq} x_t \). When simulating the system we use exact state observations though the system is tracked using position measurements for post-processing. The covariance of the measurement noise was determined using the same procedure as for the input noise, though here the random seed was multiplied with \( 10^{-1} \) in either of its dimensions. Finally we assume that the system is initialised with zero mean and uncorrelated white noise with \( \sigma_0^2 = 10^{-1} \).

\[ \min_{\xi_t} \mathbb{E}_{p(\xi_t|x_t)} \left[ \frac{1}{2} \sum_{i=0}^n \|p_t - p_i\|^2_{u_n} + \|v_t - v_i\|^2_{u_n} + \|u_t - u_i\|^2_{u_n} \right] \]
B. Results

We verify the capacity of the V-APCD and N-APCD to reconstruct the underlying policy \({\{x^\text{ler}, u^\text{ler}\}}\) using the APCD mean as a proxy for the LQER. Thus we interpret the APCD covariance as a measure for our epistemic uncertainty about the APCD mean. For the N-APCD this results into a linear feedback policy. For the V-APCD this results into a Gaussian mixture of \(N\) linear feedback policies. The initial policy \(\rho_0\) is characterised as a linear Gaussian policy with zero mean and uncorrelated covariance with magnitude \(\sigma^2\). Due to the relative sensitivity of the LQER policy to the control parameters we do not quantify the reconstruction of the LQER itself but compare the performance of the APCDs with respect to the control objective defined above. In our experiments we vary two hyperparameters, in particular the magnitude \(\sigma^2\) and the number of measurement sequences \(N\) picked randomly from half of \(M = 10^2\) individual experiments. The APCDs are then validated on the same \(10^2\) experiments using the same in- and output noise. To counteract the influence of the specific \(\sigma^2\) from half the \(M\) sequences on the reconstruction, we verify \(P\) unique but random permutations so that the probability of never having included a specific sequence is less than \(1\%\).

Fig. 3 visualizes the performance of both the V- and N-APCD respectively for \(\sigma^2 = 10^4\) and \(N = 3\). It is interesting to note that either APCD is capable of reconstructing the behavioural features of the LQER to visual satisfaction. One can verify that the V-APCD acts as a combination of 3 individual policies where the acting policy is determined by the measurement sequences that best explains the current state according to \(p(x_t|Z^n_{\tau})\). On the contrary the N-APCD averages out the contribution of each measurement sequence in likelihood space. As can be seen, for smaller \(N\) this results into a slight misalignment of the reconstructed desired behaviour and the true reference path. Fig. 4 documents results for varying \(\sigma^2\) and \(N\). Depending on \(N\) the experiment was repeated \(P\) times with a unique combination of sequences.

IV. Conclusion

In this work we discussed two strategies to extract a control policy from a single or multiple partial observation sequences.

1The number of unique combinations when picking \(N\) sequences from \(M\) is \(B = \binom{M}{N}\). The number of unique combinations that does not include a specific sequence is \(A = \binom{M-1}{N}\). The probability of picking a combinations that does not include a specific sequence on the \((p-1)\text{th}\) try equals \(q(p) = \frac{1}{p}\). The probability of picking \(P\) unique combinations that do not include a specific sequence therefore equals \(f(P) = \prod_{p=1}^{P} q(p)\). Based on this result we can compute \(P\) so that \(f(P) \leq \frac{1}{10}\) for given \(N\). See inline figure.

An interesting feature of our approach is that we can process the measurements directly into a control policy. We envision applications in imitation learning and learning by demonstration in particular but also transfer learning. To justify the latter statement we emphasize that it is not required that the system dynamics for which the policy is reconstructed must be the same as the system dynamics underlying the demonstrations. This would allow for example to reconstruct certain behavioural features demonstrated by a human with a robot that exhibits a different morphology.

In future work it would be interesting to extend the treatment to nonlinear probabilistic state-space models as well as to further explore the theoretical connection with the generic representation learning problem.

REFERENCES

[1] T. Barfoot. *State estimation for robotics*. Cambridge University Press, 2017.
[2] K. Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.
[3] S. Särkkä. *Bayesian filtering and smoothing*. Number 3. Cambridge University Press, 2013.
[4] P. Hiltunen, S. Särkkä, I. Nissilä, A. Lajunen, and J. Lampinen. State space regularization in the nonstationary inverse problem for diffuse optical tomography. *Inverse Problems*, 27(2):025009, 2011.
[5] V. Grancharov, J. Samuelsson, and B. Kleijn. On causal algorithms for speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(3):764–773, 2006.
[6] Mika P. Tarvainen, Juana K. Hiltunen, Perttu O. Ranta-Aho, and Pasi A. Karjalainen. Estimation of nonstationary eeg with kalman smoother approach: an application to event-related synchronization (ers). In *IEEE Transactions on Biomedical Engineering*, 51(3):516–524, 2004.
[7] T. M. Cover and J. A. Thomas. *Elements of Information Theory* (Wiley Series in Telecommunications and Signal Processing). Wiley-Interscience, USA, 2006.
[8] J. Largeault. Papers on probability, statistics, and statistical physics, 1986.
[9] P. Whittle. *Optimal control : basics & beyond*. Chichester : Wiley, 1996.
[10] S.-I. Amari. *Natural gradient works efficiently in learning*. *Neural computation*, 10(2):251–276, 1998.
[11] David Jacobson. Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games. *IEEE Transactions on Automatic control*, 18(2):124–131, 1973.