

Efficient Total Probability Prediction via Convex Optimization and Optimal Transport

Sanggyun Kim
Dept. of Bioengineering
University of California: San Diego
La Jolla, CA 92037
Email: s2kim@ucsd.edu

Diego Mesa
Dept. of Bioengineering
University of California: San Diego
La Jolla, CA 92037
Email: damesa@eng.ucsd.edu

Todd Coleman
Dept. of Bioengineering
University of California: San Diego
La Jolla, CA 92037
Email: tpcoleman@ucsd.edu

Abstract—In this paper, we consider state space modeling for sequential continuous estimation. We consider the one-step prediction update, which transforms our previous belief state (posterior distribution of the previous state) to new belief state (posterior distribution of the current state). We demonstrate a recursive algorithm for updating the latent state at every time by avoiding intractable integral or Gaussian approximation. The construction of the desired map is pursued through the optimal transportation theory, and we demonstrate that for the large class of log-concave state transition functions, the one-step prediction problem for continuous hidden variable is solvable through convex optimization.

Index Terms—machine learning, KL divergence, optimal transport, state space models, convex optimization

I. INTRODUCTION

State space models involve sequential estimation of an unknown-changing quantity. It is well known that in many applications, we need to deal with problem to estimate an unknown quantity that is changing over time from sequential data. Sequential data includes the measurements of either time series such as speech or non-time series such as DNA sequence. Independent and identically distributed (i.i.d.) assumption on a set of measurements simplifies the underlying mathematics of statistical inference. However, it would fail to exploit the sequential pattern of the data. Typically, this problem has been addressed using sliding window techniques. More systematic methods have been also developed by defining a probabilistic model that captures the time evolution and measurement process [1], [2].

The basic question that we have in this situation is how to predict the future outcome given the previous observations. Intuitively, we can expect the recent observations make a greater contribution to the next outcome than more historical observations. This leads to developing Markov models where the next outcome depends on the most recent observations but not on all the past observations. The state space model (SSM) provides us a general framework to consider this problem by introducing the latent variables (states) [3]. It describes the probabilistic dependence between the latent state variable and the observation. The hidden Markov model (HMM) corresponds to the SSM where the latent process $(X_i)_{i \geq 1}$ is a Markov process and $(Y_i)_{i \geq 1}$ are conditionally independent given X_i [4]. The Kalman filter enables an efficient way to

perform posterior updates for linear Gaussian models where $(X_i)_{i \geq 1}$ is a Gauss-Markov process and the likelihood model is linear additive Gaussian system [5]. In this paper, we consider the Hidden Markov models where each X_i has a continuous distribution but no Gaussian assumptions are made.

One of the primary goals in using the SSM is to *recursively* update our belief state, which is quantified as the posterior distribution of the latent state at time t given the observations up to time t . Usually, a linear-Gaussian SSM is considered so that the latent variables as well as the observations are multivariate Gaussian distributions whose means are linear functions of the states at the last iteration. Although the Gaussian assumption provides us efficient algorithms for learning and inference, this has significant limitations to be applied to various applications. For example, it implies the posterior distribution is Gaussian, which is often a poor approximation. In this paper, we propose an efficient approach to compute the evolution of the distribution of the latent state (our belief), conditioned on all the observation at that time without Gaussian approximation. The proposed method constructs an optimal map to perform the one-step prediction update and transform the previous belief $(P_{X_i|Y_{1:i}})$ to the current belief $(P_{X_{i+1}|Y_{1:i}})$ through optimal transport theory [6], [7], [8]. When the state transition probability distribution is the log-concave, the proposed method can be phrased as a convex problem. The class of the log-concave is quite broad, which includes the uniform, exponential, Gaussian, logistic regression function, etc. Then we show that the construction of the nonlinear optimal can be solved through the convex optimization problem. Thus the belief update problem in the SSM, called the one-step prediction problem, can also be efficiently solved through convex optimization.

The rest of the paper is organized as follows. Section II describes a general push-forward theorem to transform a distribution to another distribution. Section III briefly explains the SSM and its one-step prediction problem. Section IV describes how the one-step prediction problem can be efficiently solved through optimal transport theory and convex optimization. Section IV-A implements the proposed algorithm using the polynomial chaos expansion. Section V concludes the paper with discussion.

II. GENERAL PUSH-FORWARD THEOREM

Before we go into the details of the recursive algorithm for the SSM, we present a general push-forward theorem to transform a distribution P to another distribution Q .

Consider a set $W \subset \mathbb{R}^d$ for some d . Define the space of all probability measures on W as $\mathcal{P}(W)$.

Definition II.1 (Push-forward). *Given $P \in \mathcal{P}(W)$ and $Q \in \mathcal{P}(W)$, we say that the map $S : W \rightarrow W$ pushes P to Q (denoted as $S_{\#}P = Q$) if a random variable W with distribution P results in $Z \triangleq S(W)$ having distribution Q .*

We say that $S : W \rightarrow W$ is a ‘diffeomorphism’ on W if S is invertible and both S and S^{-1} are differentiable. Denote the set of all diffeomorphisms on W as $\mathcal{S}(W)$. With this, we have the following lemma from standard probability:

Lemma II.2. *Consider a diffeomorphism $S \in \mathcal{S}(W)$ and $P, Q \in \mathcal{P}(W)$ that both have densities p, q with respect to the Lebesgue measure. Then $S_{\#}P = Q$ if and only if*

$$p(u) = q(S(u)) |\det(J_S(u))| \quad \text{for all } u \in W \quad (1)$$

where $|\det(J_S(u))|$ is the absolute value of the determinant of the Jacobian of S at u , that is, $J_S(u) \equiv S'(u)$.

From now let us denote S to push forward P to Q as S^* . Then we can consider an arbitrary map S , which pushes some other P to Q as shown in Fig. 1.

We then search over all possible maps S that pushes forward some \tilde{P}_S to Q . That is, finding a differentiable map S to push forward P to Q is equivalent to search over all the maps S and stopping when $\tilde{P}_S = P$. Based on the Kullback-Leibler divergence, we have the following optimization problem to search the optimal map S^* :

$$S^* = \arg \min_S D(P \|\tilde{P}_S) \quad (2)$$

$$= \arg \min_S \mathbb{E}_P \left[\log \frac{p(u)}{\tilde{p}_S(u)} \right] \quad (3)$$

$$= \arg \min_S \mathbb{E}_P \left[\log \frac{p(u)}{q(S(u)) |\det(J_S(u))|} \right] \quad (4)$$

$$= \arg \min_S \mathbb{E}_P [\log p(u) - \log q(S(u)) - \log |\det(J_S(u))|] \quad (5)$$

$$= \arg \max_S \mathbb{E}_P [\log q(S(u)) + \log |\det(J_S(u))|] .(6)$$

The essence of this problem is at the core of optimal transport theory [8], [9], and in the most general sense is intractable. For example, here we have the issue of trying to taze out maps, for which $\det(J_S(u)) > 0$, and others, for which $\det(J_S(u)) < 0$. Both classes of maps are equally as good; however, from an optimization viewpoint, the latter leads to non-convexity in an optimization problem. As such, we plan to try to find only one such solution satisfying the Jacobian equation, for which it can be efficiently found.

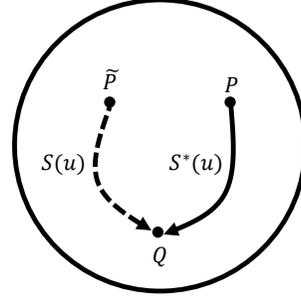


Fig. 1. Transformation from a distribution P to Q . There are maps S^* that push forward P to Q . Using an arbitrary map S will always push forward some \tilde{P}_S to the target Q , as shown above. But the \tilde{P}_S does not need to be the P we started from.

We instead consider finding maps, for which $\det(J_S(u)) > 0$, so we consider the following (implied) modified Jacobian equation:

$$\tilde{p}_S(u) = q(S(u)) \det(J_S(u)) \quad (7)$$

where $\det(J_S(u)) > 0$. Then we define a map S as:

Definition II.3. *We say a map S is a monotonic diffeomorphism, if it satisfies the followings:*

- S is differentiable,
- S^{-1} exists and is differentiable,
- $J_S(u)$ is positive definite, denoted as $J_S(u) \succ 0$, for all u .

Let us define \mathcal{MD} to be the set of all monotonic diffeomorphisms on \mathbb{R}^d . This immediately leads to the following optimization problem:

$$S^* = \arg \min_{S \in \mathcal{MD}} D(P \|\tilde{P}_S) \quad (8)$$

$$= \arg \max_{S \in \mathcal{MD}} \mathbb{E}_P [\log q(S(u)) + \log \det(J_S(u))] (9)$$

To emphasize, this means we are guaranteed to find a map S that pushes P to Q while searching over only monotonic diffeomorphisms. We now note that for a large class of P and Q problems, we can find S^* by an *efficiently* solvable convex optimization problem. Because $\log \det(\cdot)$ is strictly concave over the space of positive definite matrices, this means that if $\log q(u)$ is concave, then $\log q(S(u)) + \log \det(J_S(u))$ is concave over all S .

Theorem II.4. *If $\log q(u)$ is concave in u , then finding a map S that pushes P to Q is a convex optimization problem over \mathcal{MD} , the space of monotonic diffeomorphism.*

The proof of this follows because convex combinations of positive definite matrices are positive definite, and because the $\log \det$ operator is strictly concave over the space of positive definite matrices. This is the key general push forward theorem that we will make use of in the following sections. In essence, as long as P is easy to sample from and Q is a member of the log-concave family of distributions, then we can *efficiently*

solve a convex optimization problem to find S^* . When P is represented a prior distribution and Q represented a posterior distribution, it was demonstrated that this framework could apply to perform the efficient Bayesian inference through convex optimization [10], [11].

III. STATE SPACE MODEL

The SSM aims to estimate the hidden state given the observation. In a general formulation, let \mathbf{x}_t denote the state at time t and \mathbf{y}^t denote a set of observations up to time t . The posterior probability of the state \mathbf{x}_t given the observation \mathbf{y}^t is given by

$$p(\mathbf{x}_t|\mathbf{y}^t) = \frac{p(\mathbf{x}_t, \mathbf{y}^t)}{p(\mathbf{y}^t)} \quad (10)$$

$$= \frac{p(\mathbf{x}_t|\mathbf{y}^{t-1})p(\mathbf{y}_t|\mathbf{x}_t, \mathbf{y}^{t-1})}{p(\mathbf{y}_t|\mathbf{y}^{t-1})} \quad (11)$$

$$= \frac{p(\mathbf{x}_t|\mathbf{y}^{t-1})p(\mathbf{y}_t|\mathbf{x}_t, \mathbf{y}^{t-1})}{p(\mathbf{y}_t|\mathbf{y}^{t-1})} \quad (12)$$

where $p(\mathbf{y}_t|\mathbf{x}_t, \mathbf{y}^{t-1})$ represents the observational distribution. The one-step prediction density defined by the *Chapman-Kolmogorov* equation is

$$p(\mathbf{x}_t|\mathbf{y}^{t-1}) = \int p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{y}^{t-1})d\mathbf{x}_{t-1}, \quad (13)$$

which includes two integrand components: $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ that represents the state transition probability and $p(\mathbf{x}_{t-1}|\mathbf{y}^{t-1})$ that the posterior distribution from the last iteration step.

By inserting the equation (13) into the equation (12), we get the recursive algorithm for the evolution of the posterior distribution of the latent variable \mathbf{x}_t as follows:

$$\begin{aligned} & p(\mathbf{x}_t|\mathbf{y}^t) \\ &= \frac{p(\mathbf{y}_t|\mathbf{x}_t, \mathbf{y}^{t-1})}{p(\mathbf{y}_t|\mathbf{y}^{t-1})} \int p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{y}^{t-1})d\mathbf{x}_{t-1} \end{aligned} \quad (14)$$

The equation (14) denotes the exact posterior distribution of the hidden variable, \mathbf{x}_t . It allows us to compute the dynamic evolution of the hidden state given all the observations up to that time t . However, the integration in (14) is usually intractable except for certain cases such as Gaussian systems. There have been developed several ways to compute the recursive equation in (14), and they depends on the trade-off between computation and accuracy. Typically, Gaussian approximation on the posterior in (14) is made for computational efficiency, however, the Gaussian approximation can deviate from the true distribution, when the posterior is a distribution with multiple modes or asymmetric distribution. In this paper, we propose an efficient one-step prediction method to transform a posterior distribution at the last time, $t-1$, into a posterior distribution at the current time, t through a nonlinear high-dimensional map through the optimal transport theory. The construction of the nonlinear map can be efficiently solvable through the convex optimization by making a mild assumption on the state transition distributions.

IV. OPTIMAL TRANSPORT AND ONE-STEP PREDICTION UPDATE

Imagine that X lies in a continuum. Suppose the random process $(X_t : t \geq 1)$ is a Markov process. Assume \mathbf{X}_{t-1} is drawn according to $p_{\mathbf{X}_{t-1}|\mathbf{Y}^{t-1}}$, and \mathbf{X}_t is drawn according to $q_{\mathbf{X}_t|\mathbf{Y}^{t-1}}$. Let us denote \mathbf{x}_{t-1} as u and \mathbf{x}_t as u' for the notational simplicity, and then $Q(u'|u)$ represents the state transition rule, $P_{\mathbf{X}_t|\mathbf{X}_{t-1}=u}$. By the integral in the equation (14), q is given by

$$q(u') = \int_{u \in X} p(u)Q(u'|u)du. \quad (15)$$

Under an arbitrary map S , by inserting the equation (15) into the equation (7), we get

$$\tilde{p}_S(v) = q(S(v)) \det(J_S(v)) \quad (16)$$

$$= \left[\int_{u \in X} p(u)Q(S(v)|u)du \right] \det(J_S(v)). \quad (17)$$

Thus we have that

$$\begin{aligned} \log \frac{p(v)}{\tilde{p}_S(v)} &= \log p(v) - \log \left[\int_{u \in X} p(u)Q(S(v)|u)du \right] \\ &\quad - \log \det(J_S(v)) \\ &= \log p(v) - \log \beta_{S(v)} - \log \det(J_S(v)) \end{aligned}$$

where

$$\beta_a \triangleq \int_{u \in X} p(u)Q(a|u)du. \quad (18)$$

Thus, it is our objective to find an optimal to minimize the KL-divergence between the true P and the designed \tilde{P}_S as the following:

$$\begin{aligned} S^* &= \arg \min_{S \in \mathcal{S}(X)} D(P \parallel \tilde{P}_S) \\ &= \arg \max_{S \in \mathcal{S}(X)} \int_{v \in X} [\log \beta_{S(v)} + \log \det(J_S(v))] p(v)dv \end{aligned} \quad (19)$$

We now take an aside and represent $\log \beta_{S(v)}$ in a variational form. Assume $S(v) = a$ for some a . Define $p(u|a)$ as

$$\nu(u|a) = \frac{p(u)Q(a|u)}{\beta_a}. \quad (20)$$

And note that

$$\log \beta_a = \log \beta_a - \max_{z \in \mathcal{P}(X)} D(z \parallel \nu(\cdot|a)) \quad (21)$$

$$= \max_{z \in \mathcal{P}(X)} \log \beta_a - D(z \parallel \nu(\cdot|a)) \quad (22)$$

where $z^*(u)$ above satisfies $z^*(u) = \nu(u|a)$. Carrying on, we can simply the above as

$$\log \beta_a = \max_{z \in \mathcal{P}(X)} \int_{u \in X} z(u) \log \beta_a du \quad (23)$$

$$+ \int_{u \in X} z(u) \log \frac{\nu(u|a)}{z(u)} du \quad (24)$$

$$= \max_{z \in \mathcal{P}(X)} \int_{u \in X} z(u) \log \frac{p(u)Q(v|u)}{z(u)} du. \quad (25)$$

Thus by defining

$$\begin{aligned} & \Gamma(z, v, S) \\ &= \int_{u \in X} z(u) \log \frac{p(u)Q(S(v)|u)}{z(u)} du + \log \det (J_S(v)) \end{aligned} \quad (26)$$

and we have that

$$\log \beta_{S(v)} + \log \det (J_S(v)) = \max_{z \in \mathcal{P}(X)} \Gamma(z, v, S) \quad (27)$$

Thus by combining the above equation with (19), we have

$$\begin{aligned} S^* &= \arg \max_{S \in \mathcal{S}(X)} \int_{v \in X} [\log \beta_{S(v)} + \log \det (J_S(v))] p(v) dv \\ &= \arg \max_{S \in \mathcal{S}(X)} \max_{z \in \mathcal{P}(X)} \int_{v \in X} \Gamma(z, v, S) p(v) dv. \end{aligned} \quad (28)$$

Thus we have that

$$\begin{aligned} S^* &= \arg \max_{S \in \mathcal{S}(X)} \int_{v \in X} \left[\max_{z \in \mathcal{P}(X)} \Gamma(z, v, S) \right] p(v) dv \quad (29) \\ &\underset{V_i \text{ i.i.d. } \sim p(v)}{\approx} \arg \max_{S \in \mathcal{S}(X)} \frac{1}{N} \sum_{i=1}^N \max_{z^{V_i} \in \mathcal{P}(X)} \Gamma(z^{V_i}, V_i, S). \end{aligned} \quad (30)$$

We can formulate this as an alternating minimization algorithm to generate a series of z_k and S_k objects by fixing one, optimizing the other, and going back and forth [12].

- 1) Let $k = 1$ and assign S_k to be the identity map.
- 2) Draw $(V_i)_{i=1 \dots n}$ i.i.d. from $p(v)$
- 3) (**'E'-like step**). For each $v = V_1, \dots, v = V_n$, define

$$\begin{aligned} z_k^v(\cdot) &= \arg \max_{z \in \mathcal{P}(X)} \Gamma(z, v, S_k) \\ &= \nu(\cdot | S_k(v)) \end{aligned}$$

- 4) (**'M'-like step**). Define

$$\begin{aligned} S_{k+1} &= \arg \max_{S \in \mathcal{S}(X)} \frac{1}{N} \sum_{i=1}^N \Gamma(z_k^{V_i}, V_i, S) \\ &= \arg \max_{S \in \mathcal{S}(X)} \frac{1}{N} \sum_{i=1}^N \left[\int_{u \in X} z_k^{V_i}(u) \log Q(S(v)|u) du \right. \\ &\quad \left. + \log \det (J_S(v)) \right] \\ &= \arg \max_{S \in \mathcal{S}(X)} \frac{1}{MN} \sum_{i=1}^N \sum_{j=1}^M [\log Q(S(v)|u_{i,j}) \\ &\quad + \log \det (J_S(v))] \end{aligned}$$

where for a fixed i , $(u_{i,j})_{j=1, \dots, M}$ are drawn i.i.d. from $z_k^{V_i}(\cdot)$.

- 5) Let $k = k + 1$ and go to 2.

A. Implementation

Note that the optimization problem in **'M'-like step** is a search over a space of functions, and is thus in general not computationally feasible. Rather than optimizing over functions, we can search for coefficients of orthogonal basis functions by approximating any $S \in \mathcal{S}(X)$ as a linear combination of basis functions. That is, we can represent any $S \in \mathcal{S}(X)$ as:

$$S(v) = \sum_{j \in \mathcal{J}} g_j \phi^{(j)}(v) \quad (31)$$

where $\phi^{(j)}(v) \in \mathbb{R}$ are d -variate bases and $g_j \in \mathbb{R}^d$ are basis coefficients, with d being the dimension of W . One natural way to do this, for example if $X \subset \mathbb{R}$, is to perform a polynomial chaos expansion (PCE) [13], [14] where

$$\int_{v \in X} \phi^{(i)}(v) \phi^{(j)}(v) p(v) dv = \begin{cases} C_i, & i = j \\ 0, & i \neq j. \end{cases}$$

For example, if $X = [-1, 1]$ and P is uniformly distributed, then $\phi^{(j)}(v)$ are the Legendre polynomials. If $X = \mathbb{R}$ and P is Gaussian, then $\phi^{(j)}(v)$ are the Hermite polynomials [13].

Remark 1. In principle, any basis of polynomials, for which the truncated expansion of functions is dense in the space of all functions on X , suffices. Using the PCE where orthogonality is measured with respect to the prior, means that computing conditional expectations and other calculations can be done only with linear algebra.

Now define $K = |\mathcal{J}|$ and we have that for $X \subset \mathbb{R}$:

$$F = [g_1, \dots, g_K], \quad d \times K \quad (32)$$

$$A(v) = [\phi^{(1)}(v), \dots, \phi^{(K)}(v)]^T, \quad K \times 1 \quad (33)$$

$$S(v) = FA(v), \quad d \times 1 \quad (34)$$

$$J_A(v) = \left[\frac{\partial \phi^{(i)}}{\partial x_j}(v) \right]_{i,j}, \quad K \times d \quad (35)$$

$$J_S(v) = FJ_A(v), \quad d \times d. \quad (36)$$

Note that as $K \rightarrow \infty$, we develop a richer and richer set of candidate functions representative of $\mathcal{S}(X)$. Extension to $X \subset \mathbb{R}^d$ can be done using tensor products where

$$\phi^{(j)}(v) = \prod_{a=1}^d \psi_{j_a}(v_a)$$

where ψ_{j_a} is a univariate polynomial of order j_a , and $j \rightarrow (j_1, \dots, j_d)$ is defined in a standard diagonal manner. For example, if $d = 2$, then we have $j \rightarrow (j_1, j_2)$ constructed as:

$$\begin{aligned} 0 &\rightarrow (0, 0), & 1 &\rightarrow (0, 1), & 2 &\rightarrow (1, 0), \\ 3 &\rightarrow (0, 2), & 4 &\rightarrow (1, 1), & 5 &\rightarrow (2, 0), \end{aligned}$$

and so

$$\begin{aligned} \phi^{(0)}(v) &= \psi_0(v_1)\psi_0(v_2), & \phi^{(1)}(v) &= \psi_0(v_1)\psi_1(v_2), \\ \phi^{(2)}(v) &= \psi_1(v_1)\psi_0(v_2), & \phi^{(3)}(v) &= \psi_0(v_1)\psi_2(v_2), \\ \phi^{(4)}(v) &= \psi_1(v_1)\psi_1(v_2), & \phi^{(5)}(v) &= \psi_2(v_1)\psi_0(v_2). \end{aligned}$$

It should be noted that with PCE, an orthogonal basis with respect to the prior measure is used. As mentioned above, there are well-known polynomial families that have this property with respect to well-known distributions. However, in general, constructing an orthogonal basis with respect to a general distribution is highly non-trivial and is the subject of intense research in the statistical community [15]. To emphasize, the

construction is not necessarily dependent on the choice of PCE as the representation of the map.

With the linear basis expansion, we have that $\log \det (J_{S_y}(x)) \equiv \log \det (FJ_A(x))$. By approximating the set of all functions using truncated PCE, we now define an optimization problem that is in essence the same problem in (9) as follows:

$$F^* = \arg \max_{F \in \mathbb{R}^{d \times \kappa}} \frac{1}{MN} \sum_{i=1}^N \sum_{j=1}^M \left[\log Q(FA(v)|u_{i,j}) + \log \det (FJ_A(v)) \right] \quad (37a)$$

s.t. $FJ_A(v) \succ 0$ (37b)

where for a fixed i , $(u_{i,j})_{j=1,\dots,M}$ are drawn i.i.d. from $z_k^{V_i}(\cdot)$. This optimization problem can be easily implemented with convex optimization software such as CVX in MATLAB, CVXPY, etc. [16].

This leads to our following key theorem:

Theorem IV.1. *If $p(\cdot)$ is log-concave, if $Q(\cdot|u)$ is log-concave for each u , and $Q(u'|\cdot)$ is log-concave for each u' , then performing the one-step prediction update can be solved with convex optimization.*

Proof. If $p(\cdot)$ and $Q(u'|\cdot)$ are log-concave, then the E-step corresponds to finding the posterior distribution for Bayesian inference under log-concavity of the posterior distribution ν . As such, we can implement our previously developed convex optimization methodology discussed in [11] [10]. If $Q(\cdot|u)$ is log-concave, then the M-step given by (37) can also be performed with convex optimization. Because this is a sequence of alternating KL divergence minimizations over convex sets, we have that this algorithm is guaranteed to converge to its optimal solution, from the i-Projection framework of Csiszar [17, Thm 3]. \square

V. DISCUSSION

This methodology allows for full implementation of the nonlinear filter for state-space modeling in general conditions that extend beyond simple Gaussian state space models. For example, many point process filtering techniques [18], [19] use a linear Gaussian state space model for X and a point process generalized linear model for the point process Y . This framework, which is efficient and guaranteed to converge, can be used instead to provide more accurate quantification of uncertainty. In future work, we will compare our methods to recently developed particle filtering algorithms for non-Gaussian, non-linear state space estimation applications [20].

REFERENCES

[1] S. S. Haykin, *Adaptive filter theory*. Pearson Education India, 2008.
 [2] J. Durbin and S. J. Koopman, *Time series analysis by state space methods*. Oxford University Press, 2012, no. 38.
 [3] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*. MIT press, 2009.
 [4] L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[5] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Fluids Engineering*, vol. 82, no. 1, pp. 35–45, 1960.
 [6] G. Monge, *Mémoire sur la théorie des déblais et des remblais*. De l'Imprimerie Royale, 1781.
 [7] L. Kantorovich, "On mass transportation," in *Dokl. Akad. Nauk. SSSR*, vol. 37, 1942, pp. 227–229.
 [8] C. Villani, *Topics in optimal transportation*. AMS, 2003.
 [9] —, *Optimal transport: old and new*. Springer, 2008, vol. 338.
 [10] S. Kim, C. J. Quinn, N. Kiyavash, and T. P. Coleman, "Dynamic and succinct statistical analysis of neuroscience data," *Proceedings of the IEEE*, vol. 102, pp. 683–698, 2014.
 [11] M. R. M. D. Kim, S. and T. Coleman, "Efficient bayesian inference methods via convex optimization and optimal transport," in *ISIT 2013*, pp. 2259–2263.
 [12] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–38, 1977.
 [13] D. Xiu and G. Karniadakis, "The Wiener-Askey polynomial chaos for stochastic differential equations," *SIAM journal on scientific computing*, vol. 24, no. 2, pp. 619–644, 2003.
 [14] O. G. Ernst, A. Mugler, H. Starkloff, and E. Ullman, "On the convergence of generalized polynomial chaos expansions," *ESAIM: Mathematical Modeling and Numerical Analysis*, vol. 46, pp. 317–339, 2012.
 [15] F. S. Hover and M. S. Triantafyllou, "Application of polynomial chaos in stability and control," *Automatica*, vol. 42, no. 5, pp. 789–795, 2006.
 [16] M. Grant, S. Boyd, and Y. Ye, "Cvx: Matlab software for disciplined convex programming," 2008.
 [17] I. Csiszar, G. Tusnády *et al.*, "Information geometry and alternating minimization procedures," *Statistics and decisions*, 1984.
 [18] J. W. Pillow, Y. Ahmadian, and L. Paninski, "Model-based decoding, information estimation, and change-point detection techniques for multi-neuron spike trains," *Neural Computation*, vol. 23, no. 1, pp. 1–45, 2011.
 [19] U. Eden, L. Frank, R. Barbieri, V. Solo, and E. Brown, "Dynamic analysis of neural encoding by point process adaptive filtering," *Neural Computation*, vol. 16, no. 5, pp. 971–998, 2004.
 [20] R. Van Der Merwe, A. Doucet, N. De Freitas, and E. Wan, "The unscented particle filter," in *NIPS*, 2000, pp. 584–590.