REPORT DOCUMENTATION PAGE

PLEASE DO NOT RETURN YOUR FORM TO THE A8OVE ORGANIZATION.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
02 FEB 10	FINAL REPORT	01 FEB 07 TO 01 FEB 10

4. TITLE AND SUBTITLE
ADAPATIVE MULTI-SENSOR INTERROGATION OF TARGETS EMBEDDED IN COMPLEX ENVIRONMENTS

5a. CONTRACT NUMBER
FA9550-07-1-0455

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
PROF LAWRENCE CARIN

5d. PROJECT NUMBER
2311

5e. TASK NUMBER
NX

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
DUKE UNIVERSITY, ELECTRICAL & COMPUTER ENGINEERING DEPT.
129 HUDSON HALL, SCIENCE DRIVE
BOX 90291, DURHAM, NC 27708-0291

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFOSR/NL
875 NORTH RANDOLPH STREET
SUITE 325, ROOM 3112
ARLINGTON, VA 2203-1768

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE. DISTRIBUTION IS UNLIMITED

13. SUPPLEMENTARY NOTES

14. ABSTRACT
This project is critical for distributed sensor and communication network operation where there are multiple sensors whose data must be combined in a computationally efficient manner across a distributed network. This methodology is required in current Air Force C2/ISR networks as requirements for distributed platform data increase as with existing and future Airborne Networks.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER *(Include area code)*

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

DTIC Data

Purchase Request Number:	FQ8671-0900002
BPN:	F1ATA08235B008
Proposal Number:	06-NM-145
Research Title:	ADAPTIVE MULTI-SENSOR INTERROGATION OF TARGETS EMBEDDED IN COMPLEX ENVIRONMENTS
Type Submission:	Final Report
Inst. Control Number:	FA9550-07-1-0455P00002
Institution:	DUKE UNIVERSITY
Primary Investigator:	Professor Lawrence Carin
Invention Ind:	none
Project/Task:	2311N / X
Program Manager:	Bob Bonneau

Objective:

There are five objectives to this proposal that can be used as a core set of theoretical approaches for content based data refinement in distributed and networked sensors using Markov decision theory. The first is to exploit information from previous sensing and learning, the second is to address incomplete multi-sensor data, the third is to develop POMDP and reinforcement learning sensor-query algorithms, the fourth is to develop information-theoretic algorithms for acquisition of imperfect labels, and the fifth is the development of semi-supervised algorithms.

Approach:

The approach is to develop state-of-the-art information management and process integration theory for Air Force sensing and networking applications. The objective of this program is to combine sensing and decision theory techniques for successive refinement of data as it is integrated across groups of sensors. The fundamental process enables both supervised and unsupervised learning algorithms across raw data that may be combined in both stationary and non-stationary conditions. This type of data refinement is critical to efficient refinement of data from distributed networked sensor systems for interpretation by both machines and humans in a low latency and computationally efficient. This approach will enable efficient information theoretic models that incorporate subspace methods for selective data detection, discrimination, and identification.

Progress:

Year: 2007 **Month:** 05

Not required at this time.

Year: 2008 **Month:** 06

Annual Accomplishments: Two research directions have been pursued on this project. First, hierarchical statistical algorithms, based on extensions of the Dirichlet process, have been developed to integrate general multi-modality networked sensor data. The algorithms have been fully developed, and demonstrated on DoD sensor data. In addition, in situ compressive sensing (CS) has been developed, in which the complex-media Green's function has been employed to constitute the CS projections. The underlying theory has been developed, with initial results on measured and simulated sensor data.

Year: 2009 **Month:** 06

Have developed a new framework for networked POMDP policy design, for sensing and communication systems. The most innovative aspect of this is that it is based on a reinforcement learning construct, in which the optimal policy is

DTIC Data

Progress:

Year: 2009 **Month:** 06

learned based on experience with the networked environment. The networked POMDP is linked via the Dirichlet process.

Year: 2010 **Month:** 06 **Final**

In the context of reinforcement learning for networked POMDPs, two principal contributions have been made: (i) a new algorithm has been developed to balance exploration and exploitation, and (ii) a new local partition process (LPP) framework has been developed to appropriate share information across a network. As a second accomplishment, we have developed a new class of time-evolving topic models, appropriate for analysis of network traffic.

ADAPTIVE MULTI-SENSOR INTERROGATION OF TARGETS EMBEDDED IN COMPLEX ENVIRONMENTS

Grant Number: FA9550-07-1-0455

Final Report, 1 February 2010

Lawrence Carin
Department of Electrical and Computer Engineering
Duke University
Durham, NC 27708-0291, USA

CONTENTS

Abstract

This report summarizes three areas of research investigated under the Air Force grant: (*i*) balancing exploration and exploitation when performing reinforcement learning in POMDPs; (*ii*) proper sharing of information when performing POMDP-based reinforcement learning on a network; and (*iii*) topic modeling for time-evolving systems, with the latter now being transitioned to cybersecurity.

For research thrust (*i*), a fundamental objective in reinforcement learning is the maintenance of a proper balance between exploration and exploitation. This problem becomes more challenging when the agent can only partially observe the states of its environment. In this project we propose a dual-policy method for jointly learning the agent behavior and the balance between exploration exploitation, in partially observable environments. The method subsumes traditional exploration, in which the agent takes actions to gather information about the environment, and active learning, in which the agent queries an oracle for optimal actions (with an associated cost for employing the oracle). The form of the employed exploration is dictated by the specific problem. Theoretical guarantees are provided concerning the optimality of the balancing of exploration and exploitation. The effectiveness of the method is demonstrated by experimental results on benchmark problems.

For research thrust (*ii*), the Dirichlet process (DP) has proven a powerful nonparametric prior in multi-task reinforcement learning (MTRL). A drawback of the DP prior is that it either encourages global clustering based on all parameters, or it encourages independent local clustering based on subsets of parameters. In this report we generalize the MTRL framework by employing the nonparametric dependent *local partition process* (LPP) as a prior to promote simultaneous local and global clustering. We provide theoretical analysis of the correlated local clustering structure induced by the LPP and show the structure facilitates information-sharing between partially similar RL tasks. We develop the LPP-based MTRL framework assuming the environment in each RL task is an unknown partially observable Markov decision process, and we provide experimental results to demonstrate the advantage of the LPP-based MTRL.

For research thrust (*ii*), we consider the problem of inferring and modeling topics in a sequence of documents with known publication dates. The documents at a given time are each characterized by a topic, and the topics are drawn from a mixture model. The proposed model infers the change in the topic mixture weights as a function of time. The details of this general framework may take different forms, depending on the specifics of the model. For the examples considered here we examine base measures based on independent multinomial-Dirichlet measures for representation of topic-dependent word counts. The form of the hierarchical model allows efficient variational Bayesian (VB) inference, of interest for large-scale problems. We demonstrate results and make comparisons to the model when the dynamic character is removed, and also compare to latent Dirichlet allocation (LDA) and topics over time (TOT). We consider a database of NIPS papers as well as the United States presidential State of the Union addresses from 1790 to 2008. This is a demonstration of the technology, which is now being transitioned to time-evolving data from a computer network.

I. Exploring and Exploiting in POMDPs

A fundamental challenge facing reinforcement learning (RL) algorithms is to maintain a proper balance between exploration and exploitation. The policy designed based on previous experiences is by construction constrained, and may not be optimal as a result of inexperience. Therefore, it is desirable to take actions with the goal of enhancing experience. Although these actions may not necessarily yield optimal *near-term* reward toward the ultimate goal, they could, over a long horizon, yield improved *long-term* reward. The fundamental challenge is to achieve an optimal balance between exploration and exploitation; the former is performed with the goal of enhancing experience and preventing premature convergence to suboptimal behavior, and the latter is performed with the goal of employing available experience to define perceived optimal actions.

For a Markov decision process (MDP), the problem of balancing exploration and exploitation has been addressed successfully by the E^3 [1], [2] and R-max [3] algorithms. Many important applications, however, have environments whose states are not completely observed, leading to partially observable MDPs (POMDPs). Reinforcement learning in POMDPs is challenging, particularly in the context of balancing exploration and exploitation. Recent work targeted on solving the exploration vs. exploitation

problem is based on an augmented POMDP, with a product state space over the environment states and the unknown POMDP parameters [4]. This, however, entails solving a complicated planning problem, which has a state space that grows exponentially with the number of unknown parameters, making the problem quickly intractable in practice. To mitigate this complexity, active learning methods have been proposed for POMDPs, which borrow similar ideas from supervised learning, and apply them to selectively query an oracle (domain expert) for the optimal action [5]. Active learning has found success in many collaborative human-machine tasks where expert advice is available.

In this report we propose a dual-policy approach to balance exploration and exploitation in POMDPs, by simultaneously learning two policies with partially shared internal structure. The first policy, termed the *primary policy*, defines actions based on previous experience; the second policy, termed the *auxiliary policy*, is a meta-level policy maintaining a proper balance between exploration and exploitation. We employ the regionalized policy representation (RPR) [6] to parameterize both policies, and perform Bayesian learning to update the policy posteriors. The approach applies in either of two cases: (*i*) the agent explores by randomly taking the actions that have been insufficiently tried before (traditional exploration), or (*ii*) the agent explores by querying an oracle for the optimal action (active learning). In the latter case, the agent is assessed a query cost from the oracle, in addition to the reward received from the environment. Either (*i*) or (*ii*) is employed as an exploration vehicle, depending upon the application.

The dual-policy approach possesses interesting convergence properties, similar to those of E^3 [2] and Rmax [3]. However, our approach assumes the environment is a POMDP while E^3 and Rmax both assume an MDP environment. Another distinction is that our approach learns the agent policy directly from episodes, without estimating the POMDP model. This is in contrast to E^3 and Rmax (both learn MDP models) and the active-learning method in [5] (which learns POMDP models).

II. Regionalized Policy Representation

We first provide a brief review of the regionalized policy representation, which is used to parameterize the primary policy and the auxiliary policy as discussed above. The material in this section is taken from [6], with the proofs omitted here.

Definition 2.1: A regionalized policy representation is a tuple $(\mathcal{A}, \mathcal{O}, \mathcal{Z}, W, \mu, \pi)$. The $\mathcal{A}$ and $\mathcal{O}$ are respectively a finite set of actions and observations. The $\mathcal{Z}$ is a finite set of belief regions. The W is the belief-region transition function with $W(z, a, o', z')$ denoting the probability of transiting from z to z' when taking action a in z results in observing o'. The μ is the initial distribution of belief regions with $\mu(z)$ denoting the probability of initially being in z. The π are the region-dependent stochastic policies with $\pi(z, a)$ denoting the probability of taking action a in z.

We denote $\mathcal{A} = \{1, 2, \ldots, |\mathcal{A}|\}$, where $|\mathcal{A}|$ is the cardinality of $\mathcal{A}$. Similarly, $\mathcal{O} = \{1, 2, \ldots, |\mathcal{O}|\}$ and $\mathcal{Z} = \{1, 2, \ldots, |\mathcal{Z}|\}$. We abbreviate $(a_0, a_1, \ldots, a_T)$ as $a_{0:T}$ and similarly, $(o_1, o_2, \ldots, a_T)$ as $o_{1:T}$ and $(z_0, z_1, \ldots, z_T)$ as $z_{0:T}$, where the subscripts indexes discrete time steps. The history $h_t = \{a_{0:t-1}, o_{1:t}\}$ is defined as a sequence of actions performed and observations received up to t. Let $\Theta = \{\pi, \mu, W\}$ denote the RPR parameters. Given h_t, the RPR yields a joint probability distribution of $z_{0:t}$ and $a_{0:t}$ as follows

$$p(a_{0:t}, z_{0:t}|o_{1:t}, \Theta) = \mu(z_0)\pi(z_0, a_0)\prod_{\tau=1}^{t} W(z_{\tau-1}, a_{\tau-1}, o_\tau, z_\tau)\pi(z_\tau, a_\tau) \tag{1}$$

By marginalizing $z_{0:t}$ out in (11), we obtain $p(a_{0:t}|o_{1:t}, \Theta)$. Furthermore, the history-dependent distribution of action choices is obtained as follows:

$$p(a_\tau|h_\tau, \Theta) = p(a_{0:\tau}|o_{1:\tau}, \Theta)[p(a_{0:\tau-1}|o_{1:\tau-1}, \Theta)]^{-1}$$

which gives a stochastic policy for choosing the action a_τ. The action choice depends solely on the historical actions and observations, with the unobservable belief regions marginalized out.

A. Learning Criterion

Bayesian learning of the RPR is based on the experiences collected from the agent-environment interaction. Assuming the interaction is episodic, i.e., it breaks into subsequences called episodes [7], we represent the experiences by a set of episodes.

Definition 2.2: An episode is a sequence of agent-environment interactions terminated in an absorbing state that transits to itself with zero reward. An episode is denoted by $(a_0^k r_0^k o_1^k a_1^k r_1^k \cdots o_{T_k}^k a_{T_k}^k r_{T_k}^k)$, where the subscripts are discrete times, k indexes the episodes, and o, a, and r are respectively observations, actions, and immediate rewards.

Definition 2.3: (The RPR Optimality Criterion) Let $\mathcal{D}^{(K)} = \{(a_0^k r_0^k o_1^k a_1^k r_1^k \cdots o_{T_k}^k a_{T_k}^k r_{T_k}^k)\}_{k=1}^K$ be a set of episodes obtained by an agent interacting with the environment by following policy Π to select actions, where Π is an arbitrary stochastic policy with action-selecting distributions $p^\Pi(a_t|h_t) > 0$, $\forall$ action a_t, $\forall$ history h_t. The RPR optimality criterion is defined as

$$\widehat{V}(\mathcal{D}^{(K)};\Theta) \stackrel{def.}{=} \frac{1}{K}\sum_{k=1}^{K}\sum_{t=0}^{T_k}\gamma^t r_t^k \frac{\prod_{\tau=0}^{t} p(a_\tau^k|h_\tau^k,\Theta)}{\prod_{\tau=0}^{t} p^\Pi(a_\tau^k|h_\tau^k)} \tag{2}$$

where $h_t^k = a_0^k o_1^k a_1^k \cdots o_t^k$ is the history of actions and observations up to time t in the k-th episode, $0 < \gamma < 1$ is the discount, and Θ denotes the RPR parameters.

Throughout the report, we call $\widehat{V}(\mathcal{D}^{(K)};\Theta)$ the empirical value function of Θ. It is proven in [6] that $\lim_{K\to\infty}\widehat{V}(\mathcal{D}^{(K)};\Theta)$ is the expected sum of discounted rewards by following the RPR policy parameterized by Θ for an infinite number of steps. Therefore, the RPR resulting from maximization of $\widehat{V}(\mathcal{D}^{(K)};\Theta)$ approaches the optimal as K is large (assuming $|Z|$ is appropriate). In the Bayesian setting discussed below, we use a noninformative prior for Θ, leading to a posterior of Θ peaked at the optimal RPR, therefore the agent is guaranteed to sample the optimal or a near-optimal policy with overwhelming probability.

B. Bayesian Learning

Let $G_0(\Theta)$ represent the prior distribution of the RPR parameters. We define the posterior of Θ as

$$p(\Theta|\mathcal{D}^{(K)}, G_0) \stackrel{def.}{=} \widehat{V}(\mathcal{D}^{(K)};\Theta)G_0(\Theta)[\widehat{V}(\mathcal{D}^{(K)})]^{-1} \tag{3}$$

where $\widehat{V}(\mathcal{D}^{(K)}) = \int \widehat{V}(\mathcal{D}^{(K)};\Theta)G_0(\Theta)d\Theta$ is the marginal empirical value. Note that $\widehat{V}(\mathcal{D}^{(K)};\Theta)$ is an empirical value function, thus (13) is a non-standard use of Bayes rule. However, (13) indeed gives a distribution whose shape incorporates both the prior and the empirical information.

Since each term in $\widehat{V}(\mathcal{D}^{(K)};\Theta)$ is a product of multinomial distributions, it is natural to choose the prior as a product of Dirichlet distributions,

$$G_0(\Theta) = p(\mu|\upsilon)p(\pi|\rho)p(W|\omega) \tag{4}$$

where $p(\mu|\upsilon) = \mathrm{Dir}\big(\mu(1),\cdots,\mu(|\mathcal{Z}|)\big|\upsilon\big)$, $p(\pi|\rho) = \prod_{i=1}^{|\mathcal{Z}|}\mathrm{Dir}\big(\pi(i,1),\cdots,\pi(i,|\mathcal{A}|)\big|\rho_i\big)$, $p(W|\omega) = \prod_{a=1}^{|\mathcal{A}|}\prod_{o=1}^{|\mathcal{O}|}\prod_{i=1}^{|\mathcal{Z}|}\mathrm{Dir}\big(W(i,a,o,1),\cdots,W(i,a,o,|\mathcal{Z}|)\big|\omega_{i,a,o}\big)$; $\rho_i = \{\rho_{i,m}\}_{m=1}^{|\mathcal{A}|}$, $\upsilon = \{\upsilon_i\}_{i=1}^{|\mathcal{Z}|}$, and $\omega_{i,a,o} = \{\omega_{i,a,o,j}\}_{j=1}^{|\mathcal{Z}|}$ are hyper-parameters. With the prior thus chosen, the posterior in (13) is a large mixture of Dirichlet products, and therefore posterior analysis by Gibbs sampling is inefficient. To overcome this, we employ the variational Bayesian technique [8] to obtain a variational posterior by maximizing a lower bound to $\ln\int \widehat{V}(\mathcal{D}^{(K)};\Theta)G_0(\Theta)d\Theta$,

$$\mathrm{LB}(\{q_t^k\}, g(\Theta)) = \ln\int\widehat{V}(\mathcal{D}^{(K)};\Theta)G_0(\Theta)d\Theta - \mathrm{KL}(\{q_t^k(z_{0:t}^k)g(\Theta)\}||\{\nu_t^k p(z_{0:t}^k,\Theta|a_{0:t}^k,o_{1:t}^k)\})$$

where $\{q_t^k\}, g(\Theta)$ are variational distributions satisfying $q_t^k(z_{0:t}^k) \geq 1$, $g(\Theta) \geq 1$, $\int g(\Theta)d\Theta = 1$, and $\frac{1}{K}\sum_{k=1}^{K}\sum_{t=1}^{T_k}\sum_{z_0^k,\cdots,z_t^k=1}^{|\mathcal{Z}|} q_t^k(z_{0:t}^k) = 1$; $\nu_t^k = \frac{\gamma^t r_t^k p(a_{0:t}^k|o_{1:t}^k)}{\prod_{\tau=0}^{t} p^\Pi(a_\tau^k|h_\tau^k)\widehat{V}(\mathcal{D}^{(K)})}$ and $\mathrm{KL}(q\|p)$ denotes the Kullback-Leibler (KL) distance between probability measure q and p.

The factorized form $\{q_t(z_{0:t})g(\Theta)\}$ represents an approximation of the weighted joint posterior of Θ and z's when the lower bound reaches the maximum, and the corresponding $g(\Theta)$ is called the variational approximate posterior of Θ. The lower bound maximization is accomplished by solving $\{q_t(z_{0:t})\}$ and $g(\Theta)$ alternately, keeping one fixed while solving for the other. The solutions are summarized in Theorem 2.4; the proof is in [6].

Theorem 2.4: Given the initialization $\widehat{\rho} = \rho$, $\widehat{\upsilon} = \upsilon$, $\widehat{\omega} = \omega$, iterative application of the following updates produces a sequence of monotonically increasing lower bounds $\mathrm{LB}(\{q_t^k\}, g(\Theta))$, which converges to a maxima. The update of $\{q_t^k\}$ is

$$q_z^k(z_{0:t}^k) = \sigma_t^k p(z_{0:t}^k | a_{0:t}^k, o_{1:t}^k, \widetilde{\Theta})$$

where $\widetilde{\Theta} = \{\widetilde{\pi}, \widetilde{\mu}, \widetilde{W}\}$ is a set of under-normalized probability mass functions, with $\widetilde{\pi}(i,m) = e^{\psi(\widehat{\rho}_{i,m}) - \psi(\sum_{m=1}^{|\mathcal{A}|} \widehat{\rho}_{i,m})}$, $\widetilde{\mu}(i) = e^{\psi(\widehat{\upsilon}_i) - \psi(\sum_{i=1}^{|\mathcal{Z}|} \widehat{\upsilon}_i)}$, and $\widetilde{W}(i,a,o,j) = e^{\psi(\widehat{\omega}_{i,a,o,j}) - \psi(\sum_{j=1}^{|\mathcal{A}|} \widehat{\omega}_{i,a,o,j})}$, and ψ is the digamma function. The $g(\Theta)$ has the same form as the prior G_0 in (14), except that the hyper-parameter are updated as

$$\begin{aligned}
\widehat{\upsilon}_i &= \upsilon_i + \textstyle\sum_{k=1}^{K} \sum_{t=0}^{T_k} \sigma_t^k \phi_{t,0}^k(i) \\
\widehat{\rho}_{i,a} &= \rho_{i,a} + \textstyle\sum_{k=1}^{K} \sum_{t=0}^{T_k} \sum_{\tau=0}^{t} \sigma_t^k \phi_{t,\tau}^k(i) \delta(a_\tau^k, a) \\
\widehat{\omega}_{i,a,o,j} &= \omega_{i,a,o,j} + \textstyle\sum_{k=1}^{K} \sum_{t=0}^{T_k} \sum_{\tau=1}^{t} \sigma_t^k \xi_{t,\tau-1}^k(i,j) \delta(a_{\tau-1}^k, a) \delta(o_\tau^k, o)
\end{aligned}$$

where $\xi_{t,\tau}^k(i,j) = p(z_\tau^k = i, z_{\tau+1}^k = j | a_{0:t}^k, o_{1:t}^k, \widetilde{\Theta})$, $\phi_{t,\tau}^k(i) = p(z_\tau^k = i | a_{0:t}^k, o_{1:t}^k, \widetilde{\Theta})$, and

$$\sigma_t^k = \left[\gamma^t r_t^k p(a_{0:t}^k | o_{1:t}^k, \widetilde{\Theta})\right] \left[\textstyle\prod_{\tau=0}^{t} p^{\Pi}(a_\tau^k | h_\tau^k) \widehat{V}(\mathcal{D}^{(K)} | \widetilde{\Theta})\right]^{-1} \tag{5}$$

III. Dual-RPR

Assume that the agent uses the RPR described in Section VIII to govern its behavior in the unknown POMDP environment (the primary policy). Bayesian learning employs the empirical value function $\widehat{V}(\mathcal{D}^{(K)}; \Theta)$ in (12) in place of a likelihood function, to obtain the posterior of the RPR parameters Θ. The episodes $\mathcal{D}^{(K)}$ may be obtained from the environment by following an arbitrary stochastic policy Π with $p^{\Pi}(a|h) > 0$, $\forall\, a$, $\forall\, h$. Although any such Π guarantees optimality of the resulting RPR, the choice of Π affects the convergence speed. A good choice of Π avoids episodes that do not bring new information to improve the RPR, and thus the agent does not have to see all possible episodes before the RPR becomes optimal.

In batch learning, all episodes are collected before the learning begins, and thus Π is pre-chosen and does not change during the learning [6]. In online learning, however, the episodes are collected during the learning, and the RPR is updated upon completion of each episode. Therefore there is a chance to exploit the RPR to avoid repeated learning in the same part of the environment. The agent should recognize belief regions it is familiar with, and exploit the existing RPR policy there; in belief regions inferred as new, the agent should explore. This balance between exploration and exploitation is performed with the goal of accumulating a large long-run reward.

We consider online learning of the RPR (as the primary policy) and choose Π as a mixture of two policies: one is the current RPR Θ (exploitation) and the other is an exploration policy Π_e. This gives the action-choosing probability $p^{\Pi}(a|h) = p(y=0|h)p(a|h,\Theta,y=0) + p(y=1|h)p(a|h,\Pi_e,y=1)$, where $y=0$ ($y=1$) indicates exploitation (exploration). The problem of choosing good Π then reduces to a proper balance between exploitation and exploration: the agent should exploit Θ when doing so is highly rewarding, while following Π_e to enhance experience and improve Θ.

An *auxiliary RPR* is employed to represent the policy for balancing exploration and exploitation, i.e., the history-dependent distribution $p(y|h)$. The auxiliary RPR shares the parameters $\{\mu, W\}$ with the primary RPR, but with $\pi = \{\pi(z,a) : a \in \mathcal{A}, z \in \mathcal{Z}\}$ replaced by $\lambda = \{\lambda(z,y) : y = 0 \text{ or } 1, z \in \mathcal{Z}\}$, where

$\lambda(z, y)$ is the probability of choosing exploitation ($y = 0$) or exploration ($y = 1$) in belief region z. Let λ have the prior

$$p(\lambda|u) = \prod_{i=1}^{|\mathcal{Z}|} \text{Beta}\Big(\lambda(i,0), \lambda(i,1)\Big|u_0, u_1\Big). \tag{6}$$

In order to encourage exploration when the agent has little experience, we choose $u_0 = 1$ and $u_1 > 1$ so that, at the beginning of learning, the auxiliary RPR always suggests exploration. As the agent accumulates episodes of experience, it comes to know a certain part of the environment in which the episodes have been collected. This knowledge is reflected in the auxiliary RPR, which, along with the primary RPR, is updated upon completion of each new episode.

Since the environment is a POMDP, the agent's knowledge should be represented in the space of belief states. However, the agent cannot directly access the belief states, because computation of belief states requires knowing the true POMDP model, which is not available. Fortunately, the RPR formulation provides a compact representation of $\mathcal{H} = \{h\}$, the space of histories, where each history h corresponds to a belief state in the POMDP. Within the RPR formulation, $\mathcal{H}$ is represented internally as the set of distributions over belief regions $z \in \mathcal{Z}$, which allows the agent to access $\mathcal{H}$ based on a subset of samples from $\mathcal{H}$. Let $\mathcal{H}_{\text{known}}$ be the part of $\mathcal{H}$ that has become known to the agent, i.e., the primary RPR is optimal in $\mathcal{H}_{\text{known}}$ and thus the agent should begin to exploit upon entering $\mathcal{H}_{\text{known}}$. As will be clear below, $\mathcal{H}_{\text{known}}$ can be identified by $\mathcal{H}_{\text{known}} = \{h : p(y = 0|h, \Theta, \lambda) \approx 1\}$, if the posterior of λ is updated by

$$\widehat{u}_{i,0} = u_0 + \sum_{k=1}^{K} \sum_{t=0}^{T_k} \sum_{\tau=0}^{t} \sigma_t^k \phi_{t,\tau}^k(i), \tag{7}$$

$$\widehat{u}_{i,1} = \max\left(\eta, u_1 - \sum_{k=1}^{K} \sum_{t=0}^{T_k} \sum_{\tau=0}^{t} y_t^k \gamma^t c\, \phi_{t,\tau}^k(i)\right), \tag{8}$$

where η is a small positive number, and σ_t^k is the same in (5) except that r_t^k is replaced by m_t^k, the meta-reward received at t in episode k. We have $m_t^k = r_{\text{meta}}$ if the goal is reached at time t in episode k, and $m_t^k = 0$ otherwise, where $r_{\text{meta}} > 0$ is a constant. When Π_{e} is provided by an oracle (active learning), a query cost $c > 0$ is taken into account in (8), by subtracting c from u_1. Thus, the probability of exploration is reduced each time the agent makes a query to the oracle (i.e., $y_t^k = 1$). After a certain number of queries, $\widehat{u}_{i,1}$ becomes the small positive number η (it never becomes zero due to the max operator), at which point the agent stops querying in belief region $z = i$.

In (7) and (8), exploitation always receives a "credit", while exploration never receives credit (exploration is actually discredited when Π_e is an oracle). This update makes sure that the chance of exploitation monotonically increases as the episodes accumulate. Exploration receives no credit because it has been pre-assigned a credit (u_1) in the prior, and the chance of exploration should monotonically decrease with the accumulation of episodes. The parameter u_1 represents the agent's prior for the amount of needed exploration. When $c > 0$, u_1 is discredited by the cost and the agent needs a larger u_1 (than when $c = 0$) to obtain the same amount of exploration. The fact that the amount of exploration monotonically increases with u_1 implies that, one can always find a large enough u_1 to ensure that the primary RPR is optimal in $\mathcal{H}_{\text{known}} = \{h : p(y = 0|h, \Theta, \lambda) \approx 1\}$. However, an unnecessarily large u_1 makes the agent over-explore and leads to slow convergence. Let $u_1^{\min}$ denote the minimum u_1 that ensures optimality in $\mathcal{H}_{\text{known}}$. We assume $u_1^{\min}$ exists in the analysis below. The possible range of $u_1^{\min}$ is examined in the experiments.

IV. Optimality and Convergence Analysis

Let M be the true POMDP model. We first introduce an equivalent expression for the empirical value function in (12),

$$\widehat{V}(\mathcal{E}_T^{(K)}; \Theta) = \sum_{\mathcal{E}_T^{(K)}} \sum_{t=0}^{T} \gamma^t r_t p(a_{0:t}, o_{1:t}, r_t | y_{0:t} = 0, \Theta, M), \tag{9}$$

where the first summation is over all elements in $\mathcal{E}_T^{(K)} \subseteq \mathcal{E}_T$, and $\mathcal{E}_T = \{(a_{0:T}, o_{1:T}, r_{0:T}) : a_t \in \mathcal{A}, o_t \in \mathcal{O}, t = 0, 1, \cdots, T\}$ is the complete set of episodes of length T in the POMDP, with no repeated elements.

The condition $y_{0:t} = 0$, which is an an abbreviation for $y_\tau = 0 \; \forall \; \tau = 0, 1, \cdots, t$, indicates that the agent always follows the RPR (Θ) here. Note $\widehat{V}(\mathcal{E}_T^{(K)}; \Theta)$ is the empirical value function of Θ defined on $\mathcal{E}_T^{(K)}$, as is $\widehat{V}(\mathcal{D}^{(K)}; \Theta)$ on $\mathcal{D}^{(K)}$. When $T = \infty$ [1], the two are identical up to a difference in acquiring the episodes: $\mathcal{E}_T^{(K)}$ is a simple enumeration of distinct episodes while $\mathcal{D}^{(K)}$ may contain identical episodes. The multiplicity of an episode in $\mathcal{D}^{(K)}$ results from the sampling process (by following a policy to interact with the environment). Note that the empirical value function defined using $\mathcal{E}_T^{(K)}$ is interesting only for theoretical analysis, because the evaluation requires knowing the true POMDP model, not available in practice. We define the optimistic value function

$$\widehat{V}_f(\mathcal{E}_T^{(K)}; \Theta, \lambda, \Pi_e) = \sum_{\mathcal{E}_T^{(K)}} \sum_{t=0}^{T} \gamma^t \sum_{y_0, \cdots, y_t = 0}^{1} \left(r_t + (R_{\max} - r_t) \vee_{\tau=0}^{t} y_\tau\right) p(a_{0:t}, o_{1:t}, r_t, y_{0:t} | \Theta, \lambda, M, \Pi_e) \tag{10}$$

where $\vee_{\tau=0}^{t} y_\tau$ indicates that the agent receives r_t if and only if $y_\tau = 0$ at all time steps $\tau = 1, 2, \cdots, t$; otherwise, it receives $R_{\max}$ at t, which is an upper bound of the rewards in the environment. Similarly we can define $\widehat{V}(\mathcal{D}^{(K)}; \Theta, \lambda, \Pi_e)$, the equivalent expression for $\widehat{V}_f(\mathcal{E}_T^{(K)}; \Theta, \lambda, \Pi_e)$. The following lemma is proven in the Appendix.

Lemma 4.1: Let $\widehat{V}(\mathcal{E}_T^{(K)}; \Theta)$, $\widehat{V}_f(\mathcal{E}_T^{(K)}; \Theta, \lambda, \Pi_e)$, and $R_{\max}$ be defined as above. Let $P_{\text{explpore}}(\mathcal{E}_T^{(K)}, \Theta, \lambda, \Pi_e)$ be the probability of executing the exploration policy Π_e at least once in some episode in $\mathcal{E}_T^{(K)}$, under the auxiliary RPR (Θ, λ) and the exploration policy Π_e. Then

$$P_{\text{explpore}}(\mathcal{E}_T^{(K)}, \Theta, \lambda, \Pi_e) \geq \frac{1-\gamma}{R_{\max}} |\widehat{V}(\mathcal{E}_T^{(K)}; \Theta) - \widehat{V}_f(\mathcal{E}_T^{(K)}; \Theta, \lambda, \Pi_e)|.$$

Proposition 4.2: Let Θ be the optimal RPR on $\mathcal{E}_\infty^{(K)}$ and Θ^* be the optimal RPR in the complete POMDP environment. Let the auxiliary RPR hyper-parameters (λ) be updated according to (7) and (8), with $u_1 \geq u_1^{\min}$. Let Π_e be the exploration policy and $\epsilon \geq 0$. Then either (a) $\widehat{V}(\mathcal{E}_\infty; \Theta) \geq \widehat{V}(\mathcal{E}_\infty; \Theta^*) - \epsilon$, or (b) the probability that the auxiliary RPR suggests executing Π_e in some episode unseen in $\mathcal{E}_\infty^{(K)}$ is at least $\frac{\epsilon(1-\gamma)}{R_{\max}}$.

Proof: It is sufficient to show that if (a) does not hold, then (b) must hold. Let us assume $\widehat{V}(\mathcal{E}_\infty; \Theta) < \widehat{V}(\mathcal{E}_\infty; \Theta^*) - \epsilon$. Because Θ is optimal in $\mathcal{E}_\infty^{(K)}$, $\widehat{V}(\mathcal{E}_\infty^{(K)}; \Theta) \geq \widehat{V}(\mathcal{E}_\infty^{(K)}; \Theta^*)$, which implies $\widehat{V}(\mathcal{E}_\infty^{(\backslash K)}; \Theta) < \widehat{V}(\mathcal{E}_\infty^{(\backslash K)}; \Theta^*) - \epsilon$. where $\mathcal{E}_\infty^{(\backslash K)} = \mathcal{E}_\infty \setminus \mathcal{E}_\infty^{(K)}$. We show below that $\widehat{V}_f(\mathcal{E}_\infty^{(\backslash K)}; \Theta, \lambda, \Pi_e) \geq \widehat{V}(\mathcal{E}_\infty^{(\backslash K)}; \Theta^*)$ which, together with Lemma 4.1, implies

$$\begin{aligned} P_{\text{explpore}}(\mathcal{E}_\infty^{(\backslash K)}, \Theta, \lambda, \Pi_e) &\geq \frac{1-\gamma}{R_{\max}} \left[\widehat{V}_f(\mathcal{E}_\infty^{(\backslash K)}; \Theta, \lambda, \Pi_e) - \widehat{V}(\mathcal{E}_\infty^{(\backslash K)}; \Theta)\right] \\ &\geq \frac{1-\gamma}{R_{\max}} \left[\widehat{V}(\mathcal{E}_\infty^{(\backslash K)}; \Theta^*) - \widehat{V}(\mathcal{E}_\infty^{(\backslash K)}; \Theta)\right] \geq \frac{\epsilon(1-\gamma)}{R_{\max}} \end{aligned}$$

We now show $\widehat{V}_f(\mathcal{E}_\infty^{(\backslash K)}; \Theta, \lambda, \Pi_e) \geq \widehat{V}(\mathcal{E}_\infty^{(\backslash K)}; \Theta^*)$. By construction, $\widehat{V}_f(\mathcal{E}_\infty^{(\backslash K)}; \Theta, \lambda, \Pi_e)$ is an optimistic value function, in which the agent receives $R_{\max}$ at any time t unless if $y_\tau = 0$ at $\tau = 0, 1, \cdots, t$. However, $y_\tau = 0$ at $\tau = 0, 1, \cdots, t$ implies that $\{h_\tau : \tau = 0, 1, \cdots, t\} \subset \mathcal{H}_{\text{known}}$, By the premise, λ is updated according to (7) and (8) and $u_1 \geq u_1^{\min}$, therefore Θ is optimal in $\mathcal{H}_{\text{known}}$ (see the discussions following (7) and (8)), which implies Θ is optimal in $\{h_\tau : \tau = 0, 1, \cdots, t\}$. Thus, the inequality holds. Q.E.D.

Proposition 4.2 shows that whenever the primary RPR achieves less accumulative reward than the optimal RPR by ϵ, the auxiliary RPR suggests exploration with a probability exceeding $\epsilon(1-\gamma)R_{\max}^{-1}$. Conversely, whenever the auxiliary RPR suggests exploration with a probability smaller than $\epsilon(1-\gamma)R_{\max}^{-1}$, the primary RPR achieves ϵ-near optimality. This ensures that the agent is either receiving sufficient rewards or it is performing sufficient exploration.

[1] An episode almost always terminates in finite time steps in practice and the agent stays in the absorbing state with zero reward for the remaining infinite steps after an episode is terminated [7]. The infinite horizon is only to ensure theoretically all episodes have the same horizon length.

V. Experimental Results

Our experiments are based on Shuttle, a benchmark POMDP problem [9], with the following setup. The primary policy is a RPR with $|\mathcal{Z}| = 10$ and a prior in (14), with all hyper-parameters initially set to one (which makes the initial prior non-informative). The auxiliary policy is a RPR sharing $\{\mu, W\}$ with the primary RPR and having a prior for λ as in (6). The prior of λ is initially biased towards exploration by using $u_0 = 1$ and $u_1 > 1$. We consider various values of u_1 to examine the different effects. The agent performs online learning: upon termination of each new episode, the primary and auxiliary RPR posteriors are updated by using the previous posteriors as the current priors. The primary RPR update follows Theorem 2.4 with $K = 1$ while the auxiliary RPR update follows (7) and (8) for λ (it shares the same update with the primary RPR for μ and W). We perform 100 independent Monte Carlo runs. In each run, the agent starts learning from a random position in the environment and stops learning when K_{total} episodes are completed. We compare various methods that the agent uses to balance exploration and exploitation: (*i*) following the auxiliary RPR, with various values of u_1, to adaptively switch between exploration and exploitation; (*ii*) randomly switching between exploration and exploitation with a fixed exploration rate P_{explore} (various values of P_{explore} are examined). When performing exploitation, the agent follows the current primary RPR (using the Θ that maximizes the posterior); when performing exploration, it follows an exploration policy Π_e. We consider two types of Π_e: (*i*) taking random actions and (*ii*) following the policy obtained by solving the *true* POMDP using PBVI [10] with 2000 belief samples. In either case, $r_{\text{meta}} = 1$ and $\eta = 0.001$. In case (*ii*), the PBVI policy is the oracle and incurs a query cost c.

We report: (*i*) the sum of discounted rewards accrued within each episode during learning; these rewards result from both exploitation and exploration. (*ii*) the quality of the primary RPR upon termination of each learning episode, represented by the sum of discounted rewards averaged over 251 episodes of following the primary RPR (using the standard testing procedure for Shuttle: each episode is terminated when either the goal is reached or a maximum of 251 steps is taken); these rewards result from exploitation alone. (*iii*) the exploration rate P_{explore} in each learning episode, which is the number of time steps at which exploration is performed divided by the total time steps in a given episode. In order to examine the optimality, the rewards in (*i*)-(*ii*) has the corresponding optimal rewards subtracted, where the optimal rewards are obtained by following the PBVI policy; the difference are reported, with zero difference indicating optimality and minus difference indicating sub-optimality. All results are averaged over the 100 Monte Carlo runs. The results are summarized in Figure 1 when Π_e takes random actions and in Figure 2 when Π_e is an oracle (the PBVI policy).

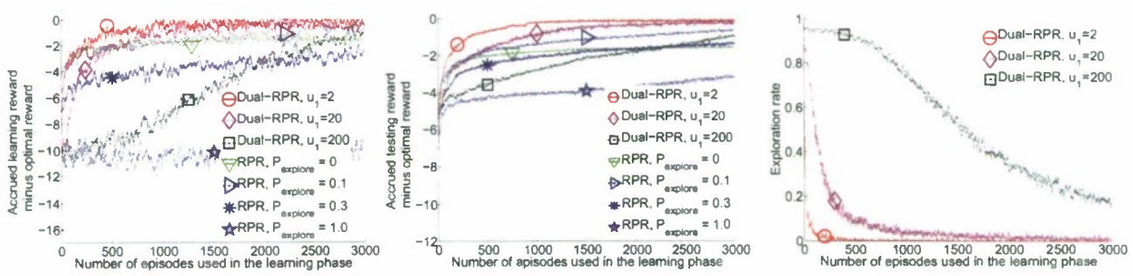


Fig. 1. Results on Shuttle with a random exploration policy, with $K_{\text{total}} = 3000$. Left: accumulative discounted reward accrued within each learning episode, with the corresponding optimal reward subtracted. Middle: accumulative discounted rewards averaged over 251 episodes of following the primary RPR obtained after each learning episode, again with the corresponding optimal reward subtracted. Right: the rate of exploration in each learning episode. All results are averaged over 100 independent Monte Carlo runs.

It is seen from Figure 1 that, with random exploration and $u_1 = 2$, the primary policy converges to optimality and, accordingly, P_{explore} drops to zero, after about 1500 learning episodes. When u_1 increases

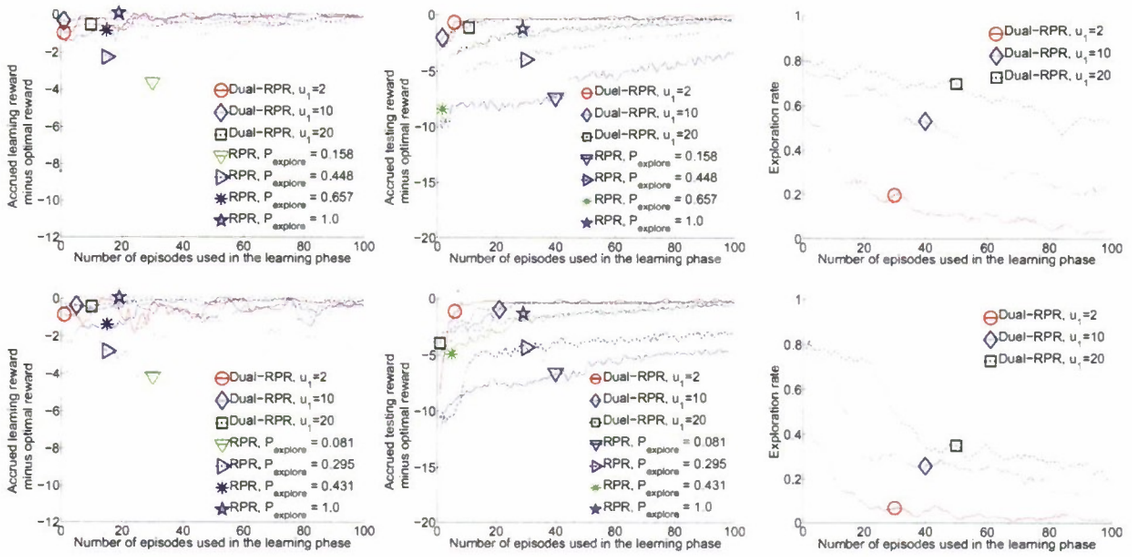


Fig. 2. Results on Shuttle with an oracle exploration policy incurring cost $c = 1$ (top row) and $c = 3$ (bottom row), and $K_{\text{total}} = 100$. Each figure in a row is a counterpart of the corresponding figure in Figure 1, with the random Π_e replaced by the oracle Π_e. See the captions there for details.

to 20, the convergence is slower: it does not occur (and $P_{\text{explore}} > 0$) until after abound 2500 learning episodes. With u_1 increased to 200, the convergence does not happen and $P_{\text{explore}} > 0.2$ within the first 3000 learning episodes. These results verify our analysis in Section III and IV: (*i*) the primary policy improves as P_{explore} decreases; (*ii*) the agent explores when it is not acting optimally and it is acting optimally when it stops exploring; (*iii*) there exists finite u_1 such that the primary policy is optimal if $P_{\text{explore}} = 0$. Although $u_1 = 2$ may still be larger than $u_1^{\min}$, it is small enough to ensure convergence within 1500 episodes. We also observe from Figure 1 that: (*i*) the agent explores more efficiently when it is adaptively switched between exploration and exploitation by the auxiliary policy, than when the switch is random; (*ii*) the primary policy cannot converge to optimality when the agent never explores; (*iii*) the primary policy may converge to optimality when the agent always takes random actions, but it may need infinite learning episodes to converge.

The results in Figure 2, with Π_e being an oracle, provide similar conclusions as those in Figure 1 when Π_e is random. However, there are two special observations from Figure 2: (*i*) P_{explore} is affected by the query cost c: with a larger c, the agent performs less exploration. (*ii*) the convergence rate of the primary policy is not significantly affected by the query cost. The reason for (*ii*) is that the oracle always provides optimal actions, thus over-exploration does not harm the optimality; as long as the agent takes optimal actions, the primary policy continually improves if it is not yet optimal, or it remains optimal if it is already optimal.

VI. Summary of Balancing Exploration & Exploitation

We have presented a dual-policy approach for jointly learning the agent behavior and the optimal balance between exploitation and exploration, assuming the unknown environment is a POMDP. By identifying a known part of the environment in terms of histories (parameterized by the RPR), the approach adaptively switches between exploration and exploitation depending on whether the agent is in the known part. We have provided theoretical guarantees for the agent to either explore efficiently or exploit efficiently. Experimental results show good agreement with our theoretical analysis and that our approach finds the optimal policy efficiently. Although we empirically demonstrated the existence of a small u_1 to ensure

efficient convergence to optimality, further theoretical analysis is needed to find $u_1^{\min}$, the tight lower bound of u_1, which ensures convergence to optimality with just the right amount of exploration (without over-exploration). Finding the exact $u_1^{\min}$ is difficult because of the partial observability. However, it is hopeful to find a good approximation to $u_1^{\min}$. In the worst case, the agent can always choose to be optimistic, like in E^3 and Rmax. An optimistic agent uses a large u_1, which usually leads to over-exploration but ensures convergence to optimality.

APPENDIX

Proof of Lemma 4.1: We expand (10) as,

$$\begin{aligned}\widehat{V}_f(\mathcal{E}_T^{(K)};\Theta,\lambda,\Pi_e) = &\textstyle\sum_{\mathcal{E}_T^{(K)}}\sum_{t=0}^{T}\gamma^t r_t p(a_{0:t},o_{1:t},r_t|y_{0:t}=0,\Theta,M)p(y_{0:t}=0|\Theta,\lambda)\\ &+\textstyle\sum_{\mathcal{E}_T^{(K)}}\sum_{t=0}^{T}\gamma^t R_{\max}\sum_{y_{0:t}\neq 0}p(a_{0:t},o_{1:t},r_t|y_{0:t},\Theta,M,\Pi_e)p(y_{0:t}|\Theta,\lambda)\end{aligned}$$

where $y_{0:t}$ is an an abbreviation for $y_\tau = 0\ \forall\ \tau = 0,\cdots,t$ and $y_{0:t}\neq 0$ is an an abbreviation for $\exists\ 0\leq\tau\leq t$ satisfying $y_\tau\neq 0$. The sum $\sum_{\mathcal{E}_T^{(K)}}$ is over all episodes in $\mathcal{E}_T^{(K)}$. The difference between (9) and (11) is

$$\begin{aligned}&|\widehat{V}(\mathcal{E}_T^{(K)},\Theta)-\widehat{V}(\mathcal{E}_T^{(K)};\Theta,\lambda)| = \Big|\textstyle\sum_{\mathcal{E}_T^{(K)}}\sum_{t=0}^{T}\gamma^t r_t p(a_{0:t},o_{1:t},r_t|y_{0:t}=0,\Theta,M)(1-p(y_{0:t}=0|\Theta,\lambda))\\ &\qquad -\textstyle\sum_{\mathcal{E}_T^{(K)}}\sum_{t=0}^{T}\gamma^t R_{\max}\sum_{y_{0:t}\neq 0}p(a_{0:t},o_{1:t},r_t|y_{0:t},\Theta,M,\Pi_e)p(y_{0:t}|\Theta,\lambda)\Big|\\ &= \Big|\textstyle\sum_{\mathcal{E}_T^{(K)}}\sum_{t=0}^{T}\gamma^t r_t p(a_{0:t},o_{1:t},r_t|y_{0:t}=0,\Theta,M)\sum_{y_{0:t}\neq 0}p(y_{0:t}|\Theta,\lambda)\\ &\qquad -\textstyle\sum_{\mathcal{E}_T^{(K)}}\sum_{t=0}^{T}\gamma^t R_{\max}\sum_{y_{0:t}\neq 0}p(a_{0:t},o_{1:t},r_t|y_{0:t},\Theta,M,\Pi_e)p(y_{0:t}|\Theta,\lambda)\Big|\\ &= \Big|\sum_{\mathcal{E}_T^{(K)}}\sum_{t=0}^{T}\gamma^t r_t\sum_{y_{0:t}\neq 0}\Big[p(a_{0:t},o_{1:t},r_t|y_{0:t}=0,\Theta,M)-\frac{R_{\max}}{r_t}p(a_{0:t},o_{1:t},r_t|y_{0:t},\Theta,M,\Pi_e)\Big]p(y_{0:t}|\Theta,\lambda)\Big|\\ &\leq \textstyle\sum_{\mathcal{E}_T^{(K)}}\sum_{t=0}^{T}\gamma^t R_{\max}\sum_{y_{0:t}\neq 0}p(y_{0:t}|\Theta,\lambda) = \sum_{\mathcal{E}_T^{(K)}}\sum_{t=0}^{T}\gamma^t R_{\max}(1-p(y_{0:t}=0|\Theta,\lambda))\\ &\leq \textstyle\sum_{\mathcal{E}_T^{(K)}}(1-p(y_{0:T}=0|\Theta,\lambda))\sum_{t=0}^{T}\gamma^t R_{\max} \leq \frac{R_{\max}}{1-\gamma}\sum_{\mathcal{E}_T^{(K)}}(1-p(y_{0:T}=0|\Theta,\lambda))\end{aligned}$$

where $\sum_{y_{0:t}\neq 0}$ is a sum over all sequences $\{y_{0:t}:\exists 0\leq\tau\leq t$ satisfying $y_\tau\neq 0\}$. Q.E.D.

VII. NETWORKED POMDPS AND SHARING INFORMATION

Reinforcement learning (RL) typically requires a large quantity of trial-and-error searches (data) to discover the long-term consequences of actions in an unknown dynamic environment [7]. When the environment is not fully observable, the situation becomes more severe, since the agent needs more data to reason about the state uncertainty in addition to the consequences of each state-action pair. Therefore, it is important to utilize as much prior knowledge as possible in reinforcement learning, to promote parsimony in data usage.

To be specific, let the unknown environment be characterized by a partially observable Markov decision process (POMDP), the states of which the agent infers through observations that are probabilistically dependent on the states. This gives rise to the belief state, a probability distribution of the states conditioned on all observed data up to the moment. The belief state is a sufficient statistic summarizing all the information required to make the decision about the action at any given moment [11]. To compute belief states, however, the agent must assume complete knowledge of the environment (i.e., must know the underlying POMDP model), which is not available by the assumption of reinforcement learning.

Methods of addressing reinforcement learning in POMDPs are generally divided into two categories: model-based and model-free [12]. A model-based method first seeks to learn the underlying POMDP model of the dynamic environment in question, and then applies POMDP planning algorithms to find the optimal policy. A model-free method directly finds the policy, avoiding the intermediate step of learning the underlying POMDP.

As pointed out in [12], model-free methods are computationally advantageous, but they cannot take advantage of prior knowledge about the environment, as their model-based counterparts do. The latter is true because it is generally difficult to establish exact correspondence between a POMDP and its policy, and therefore the knowledge for a specific POMDP cannot easily be transferred into the knowledge for its policy. The difficulty, however, is alleviated in multi-task reinforcement learning (MTRL) [13], in which one is interested in the prior knowledge that one environment is similar to another. This type of *relational* knowledge transfers readily from POMDPs to their optimal policies, because similar POMDPs will accordingly have similar optimal policies. The key is then to infer which environments are similar and how many clusters (classes of environments) are present. This is accomplished in [13] by using a nonparametric Dirichlet process (DP) [14] prior imposed on the policies across the environments. With the experiences from multiple environments, the DP prior encourages the environments to form *appropriate* clusters, so that data are shared within each cluster to enhance the cumulative information and improve policy learning. It is noteworthy that the computational advantages of model-free methods are magnified in the MTRL setting, because they avoid solving a POMDP planning problem for each cluster of environments, repeatedly whenever the clusters are updated.

Model-free methods rely on an appropriate way of representing the policy, based directly on the available observed information. The MTRL framework in [13] is based on the *regionalized policy representation* (RPR), proposed there to yield an efficient parametrization for the conditional distribution of action choices given historical actions and observations. The RPR is amenable to a Bayesian formulation and the Dirichlet process prior can be employed to promote clustering of the RL tasks.

A drawback of the DP prior is that it either encourages global clustering based on the complete set of parameters, or it encourages independent local clustering based on disjoint subsets of parameters; however, it does not encourage an appropriate balance of both. On one hand, global clustering enforces two partly similar tasks to either share information inappropriately or not share information at all. On the other hand, independent local clustering yields unnecessary local clusters, increasing the burden on data usage. In this project we aim to address this problem by employing a nonparametric dependent *local partition process* (LPP) [15] in place of the DP. A major advantage arising from this replacement is that the LPP allows simultaneous local and global clustering, and therefore it provides an effective vehicle for sharing information between partially similar tasks.

This aspect of the project has two major contributions. The first is the proposed LPP-based MTRL framework, which includes the DP-based framework in [13] as a special case, and the associated learning algorithm and experimental studies. The second principal contribution is the theoretical analysis of the LPP, which extends the results in [15] from two subsets of the parameters to an arbitrary number of subsets. Our theoretical analysis provides further insights into the LPP, both in the general sense and for our specific problem. In addition, we also provide analysis justifying the LPP as a relational prior for model-free RL.

VIII. Regionalized Policy Representation

Definition 8.1: [13] A regionalized policy representation is a tuple $(\mathcal{A}, \mathcal{O}, \mathcal{Z}, W, \mu, \pi)$, where $\mathcal{A}$, $\mathcal{O}$, and $\mathcal{Z}$ are respectively a finite set of actions, observations, and belief regions. The W is the belief-region transition function, with $W(z, a, o', z')$ denoting the probability of transiting from z to z' when taking action a in z results in observing o'. The μ is the initial distribution of belief regions, with $\mu(z)$ denoting the probability of initially being in z. The π are the region-dependent stochastic policies, with $\pi(z, a)$ denoting the probability of taking action a in z.

The history $h_t = \{a_{0:t-1}, o_{1:t}\}$ is a sequence of actions performed and observations received up to t. Let $\Theta = \{\pi, \mu, W\}$ denote the set of RPR parameters. The number of parameters is given by $|\Theta| = |\pi| + |\mu| + |W| = |\mathcal{Z}| + |\mathcal{A}||\mathcal{Z}| + |\mathcal{A}||\mathcal{O}||\mathcal{Z}|^2$. The RPR expresses the joint probability distribution of $z_{0:t}$ and $a_{0:t}$ as

$$p(a_{0:t}, z_{0:t}|o_{1:t}, \Theta) = \mu(z_0)\pi(z_0, a_0)$$

$$\times \prod_{\tau=1}^{t} W(z_{\tau-1}, a_{\tau-1}, o_\tau, z_\tau)\pi(z_\tau, a_\tau) \tag{11}$$

By marginalizing $z_{0:t}$ out in (11), one obtains $p(a_{0:t}|o_{1:t}, \Theta)$, which can be used to yield $p(a_t|h_t, \Theta)$.

Assuming episodic agent-environment interactions [7], the RPR is learned using a set of episodes, where an episode of length T is denoted by $(a_0^k r_0^k o_1^k a_1^k r_1^k \cdots o_{T_k}^k a_{T_k}^k r_{T_k}^k)$, with k the index.

Definition 8.2: [13] Let $\mathcal{D}^{(K)} = \{(a_0^k r_0^k o_1^k a_1^k r_1^k \cdots o_{T_k}^k a_{T_k}^k r_{T_k}^k)\}_{k=1}^K$ be a set of episodes obtained by an agent interacting with the environment by following policy Π to select actions, where Π is an arbitrary stochastic policy with action-selecting distributions $p^\Pi(a_t|h_t) > 0$, $\forall$ action a_t, $\forall$ history h_t. The empirical value function is defined as

$$\widehat{V}(\mathcal{D}^{(K)}; \Theta) \stackrel{def.}{=} \frac{1}{K} \sum_{k=1}^{K} \sum_{t=0}^{T_k} \gamma^t r_t^k \frac{\prod_{\tau=0}^{t} p(a_\tau^k|h_\tau^k, \Theta)}{\prod_{\tau=0}^{t} p^\Pi(a_\tau^k|h_\tau^k)} \tag{12}$$

where $h_t^k = a_0^k o_1^k a_1^k \cdots o_t^k$ is the history of actions and observations up to time t in the k-th episode, $0 < \gamma < 1$ is the discount, and Θ denotes the RPR parameters.

It is proven in [13] that, as $K \to \infty$, the limit of $\widehat{V}(\mathcal{D}^{(K)}; \Theta)$ is the expected sum of discounted rewards by following the RPR policy parameterized by Θ for an infinite number of steps.

Let $G_0(\Theta)$ represent the prior distribution of the RPR parameters. The posterior of Θ is defined as

$$p(\Theta|\mathcal{D}^{(K)}, G_0) \stackrel{def.}{=} \frac{\widehat{V}(\mathcal{D}^{(K)}; \Theta) G_0(\Theta)}{\int \widehat{V}(\mathcal{D}^{(K)}; \Theta) G_0(\Theta) d\Theta} \tag{13}$$

where $\widehat{V}(\mathcal{D}^{(K)}) = \int \widehat{V}(\mathcal{D}^{(K)}; \Theta) G_0(\Theta) d\Theta$ is the marginal empirical value. Since each term in $\widehat{V}(\mathcal{D}^{(K)}; \Theta)$ is a product of multinomial distributions, it is natural to choose the prior as a product of Dirichlet distributions,

$$G_0(\Theta) = p(\mu|\upsilon) p(\pi|\rho) p(W|\omega) \tag{14}$$

where $p(\mu|\upsilon) = \mathrm{Dir}\big(\mu(1), \cdots, \mu(|\mathcal{Z}|)\big|\upsilon\big)$, $p(\pi|\rho) = \prod_{i=1}^{|\mathcal{Z}|} \mathrm{Dir}\big(\pi(i,1), \cdots, \pi(i,|\mathcal{A}|)\big|\rho_i\big)$, $p(W|\omega) = \prod_{a=1}^{|\mathcal{A}|} \prod_{o=1}^{|\mathcal{O}|} \prod_{i=1}^{|\mathcal{Z}|} \mathrm{Dir}\big(W(i,a,o,1), \cdots, W(i,a,o,|\mathcal{Z}|)\big|\omega_{i,a,o}\big)$; $\rho_i = \{\rho_{i,m}\}_{m=1}^{|\mathcal{A}|}$, $\upsilon = \{\upsilon_i\}_{i=1}^{|\mathcal{Z}|}$, and $\omega_{i,a,o} = \{\omega_{i,a,o,j}\}_{j=1}^{|\mathcal{Z}|}$ are hyper-parameters.

IX. Reinforcement Learning in Multiple Environments

We consider M environments indexed by $m = 1, 2, \cdots, M$, each characterized by an unknown POMDP with the same action set $\mathcal{A}$ and observation set $\mathcal{O}$. Though the environments may apparently look different from each other, it is often the case in practice that they fall into clusters such that those in the same cluster share fundamental common characteristics. Assume that, from each environment m, we have collected a set of episodes denoted as $\mathcal{D}^{(K_m)} = \{(a_0^{m,k} r_0^{m,k} o_1^{m,k} a_1^{m,k} r_1^{m,k} \cdots o_{T_{m,k}}^{m,k} a_{T_{m,k}}^{m,k} r_{T_{m,k}}^{m,k})\}_{k=1}^{K_m}$, where a subscript or superscript m indicates the environment from which the episodes originated.

One may pursue various paradigms to learn the RPR policies for the M environments. At one extreme, one may perform single-task reinforcement learning (STRL), i.e., employing $\mathcal{D}^{(K_m)}$ to obtain a distinct RPR policy for the m-th environment, for any $m \in \{1, 2, \cdots, M\}$. At the other extreme, one may aggregate the episodes across the environments to form a pool $\cup_{m=1}^{M} \mathcal{D}^{(K_m)}$, which is then employed to get one RPR for all environments. Clearly, STRL treats the environments as independent to each other while pool-based reinforcement learning (PBRL) treats the environments as identical.

Between the two extremes is MTRL, in which one partitions the environments into clusters based on an appropriate similarity measure. Given the partition, one performs PBRL within each cluster or, equivalently, performs STRL by treating each cluster as a task. In Bayesian learning, the similarity measure is implicitly prescribed by the Bayesian prior, which induces probabilistic task clusters. Different Bayesian priors induce

different task clusters. By changing the priors, one obtains a wide spectrum of MTRL algorithms, bridging the gap between STRL and PBRL.

The success of MTRL hinges on the choice of the Bayesian prior. The key is that the similarity measure prescribed by the prior should distinguish the differences between tasks while being able to also find similarities at the proper level of fidelity. In other words, a good prior should provide a reasonable balance between capturing the common characteristics among the tasks and respecting the idiosyncracies of each individual task. This motivates use of a dependent local partition prior, that promotes correlated local task clusters, allows a flexible similarity pattern that accounts for common as well as idiosyncratic aspects among the tasks, and thus makes the information sharing more efficient.

X. Multi-task Reinforcement Learning via Correlated Local Task Clusters

Introducing notation, we let Θ_m denote the RPR parameters for the m-th environment, with the number of belief regions $|\mathcal{Z}|$ independent of m. Recalling the environments have the same $\mathcal{A}$ and $\mathcal{O}$ and $|\Theta|$ is a function of $(|\mathcal{Z}|, |\mathcal{A}|, |\mathcal{O}|)$, we have $|\Theta_m| = d_{RPR}$, where d_{RPR} is constant. We further assume the elements in Θ_m follow the same order across $m = 1, 2, \ldots, M$. Let $\{1, 2, \cdots, d_{RPR}\}$ be partitioned into J nonempty disjoint subsets denoted by $\{\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_J\}$. Hence $\mathcal{I}_j$ indexes the j-th part of an RPR in any environment. We define Θ_{mj} as the subset of Θ_m such that the elements of Θ_{mj} are indexed by $\mathcal{I}_j$ in Θ_m. Thus Θ_m is accordingly partitioned into J disjoint nonempty subsets $\{\Theta_{m1}, \cdots, \Theta_{mJ}\}$, and the consistency among the partitions in different environments is ensured by using the same $\{\mathcal{I}_j\}$.

The partition $\{\Theta_{mj}\}$ should be constructed to facilitate local information sharing among the environments. For example, one may let $\Theta_{m1} = \mu_m$, $\Theta_{m2} = \pi_m(z = 1, :)$, $\cdots$, $\Theta_{m,|\mathcal{Z}|+1} = \pi_m(z = |\mathcal{Z}|, :)$, $\Theta_{m,|\mathcal{Z}|+2} = W_m(:, a = 1, o = 1, :)$, $\cdots$, $\Theta_{m,J} = W_m(:, a = |\mathcal{A}|, o = |\mathcal{O}|, :)$, where $J = |\mathcal{A}||\mathcal{O}| + |\mathcal{Z}| + 1$. In this partition, each subset of parameters play a specific role in the RPR policy, for example, action selection in a particular belief region. This encourages environments to share the same subset of RPR parameters when they have similar goal states.

A. The Dependent Local Partition Prior

With an appropriate partition of RPR parameters, one could place a local DP prior on each subset in the partition to encourage *local* information-sharing among the environments,

$$\Theta_{mj} \sim \widetilde{G}_j,\ \widetilde{G}_j \sim DP(\alpha G_{0j}),\ m = 1, \cdots, M,\ j = 1, \cdots, J \tag{15}$$

where G_{0j} is the j-th marginal of a probability measure G_0. Each DP partitions the environments into clusters, with information shared within each cluster for a particular subset of Θ. This is appealing when the environments are only partly similar. The drawback, however, is that it ignores the correlation between subsets of Θ, and is prone to generate an unnecessarily large number of clusters.

Alternatively, one may place a DP prior on all components of Θ to encourage global information-sharing,

$$\Theta_m \sim \overline{G},\ \overline{G} \sim DP(\alpha G_0),\ m = 1, 2, \cdots, M \tag{16}$$

A clear drawback is that, under the *global* DP prior, any two partly similar environments will be forced into the same cluster or they will be allocated to different clusters; in the former case, the idiosyncratic subsets of Θ will be learned inappropriately due to the wrong information-sharing, while useful information is forfeited for the related subsets in the latter case.

The dependent local partition process (LPP) [15] is a nonparametric Bayesian prior imposing that when two environments share a certain subset of RPR parameters, say $\mathcal{I}_j$, they are encouraged to share other subsets $\{\mathcal{I}_{j'} : j' \neq j\}$. Such a prior promotes correlated local clusters to capture the dependence between RPR parameters. Formally, we specify the LPP prior on $\{\Theta_m\}$ as follows:

$$\Theta_{mj} \sim \eta_j \delta_{\overline{\Theta}_{mj}} + (1 - \eta_j)\delta_{\widetilde{\Theta}_{mj}},\ \eta_j \sim \mathrm{Be}(1, \beta),\ j = 1, \cdots, J$$

$$(\overline{\Theta}_{m1},\cdots,\overline{\Theta}_{mJ})\sim\overline{G},\ \overline{G}\sim DP(\alpha G_0),\ m=1,\cdots,M$$
$$\widetilde{\Theta}_{mj}\sim\widetilde{G}_j,\ \widetilde{G}_j\sim DP(\alpha G_{0j}),\ m=1,\cdots,M,\ j=1,\cdots,J \tag{17}$$

where G_0 is the base probability measure and $\alpha,\beta>0$. Following [15], we denote the LPP prior as $LPP(\alpha,\beta,G_0)$.

We have specified the LPP in (17) differently than in [15], to make it easier to discern the structure. It is clear form (17) that the LPP reduces to the global DP in (16) or the local DPs in (15), when β takes extreme values. As $\beta\to 0$, $\Theta_{mj}=\overline{\Theta}_{mj}$, with $\overline{\Theta}_m$ drawn from $DP(\alpha G_0)$. As $\beta\to\infty$, $\Theta_{mj}=\widetilde{\Theta}_{mj}$ is drawn from $DP(\alpha G_{0j})$. Thus, $LPP(\alpha,\beta\to 0,G_0)$ is reduced to $DP(\alpha G_0)$, which is a global DP imposed on $\{\Theta_m\}$, and $LPP(\alpha,\beta\to\infty,G_0)$ is reduced to J independent DPs, $\{DP(\alpha G_{0j})\}_{j=1}^J$, where $DP(\alpha G_{0j})$ is a local DP independently imposed on Θ_{mj}, with the local DP base G_{0j} being the j-th marginal of the global DP base G_0. With $0<\beta<\infty$, the density of Θ_{mj} is a mixture of two point masses, respectively centered at a sample from the global DP and a sample from the j-th local DP, thus the LPP generally combines the global DP and independent local DPs.

The random probability measures $\overline{G}$ and $\{\widetilde{G}_j\}$ can be explicitly expressed by the stick-breaking construction of [16],

$$\overline{G}=\sum_{i=1}^{\infty}\lambda_{0i}\delta_{\overline{\Theta}_i^*},\ \widetilde{G}_j=\sum_{i=1}^{\infty}\lambda_{ji}\delta_{\widetilde{\Theta}_{ij}^*},\ j=1,\cdots,J$$
$$\lambda_{ji}=\lambda_{ji}^*\prod_{l<i}(1-\lambda_{jl}),\ \lambda_{ji}^*\sim \mathrm{Be}(1,\alpha),\ j=0,1,\cdots,J$$
$$\overline{\Theta}_i^*\sim G_0,\ (\widetilde{\Theta}_{i1}^*,\cdots,\widetilde{\Theta}_{iJ}^*)\sim G_0,\ i=1,2,\cdots \tag{18}$$

which will be used to derive the Gibbs sampler for posterior inference.

B. Analyzing the LPP Clustering Mechanism

The expressions in (17) and (18) provide insight into the clustering mechanism of the LPP, which we analyze below. It is seen that the LPP promotes clustering through the discrete random measures $\overline{G}$ and $\{\widetilde{G}_j\}$, where $\overline{G}$ is drawn from the global DP and is responsible for global clustering of $\{\Theta_m\}_{m=1}^M$, while $\widetilde{G}_j$ is drawn from the j-th local DP and responsible for local clustering of $\{\Theta_{mj}\}_{m=1}^M$. The proportions in the LPP are clearly seen from (17), which shows that, a sample Θ_{mj} drawn from the LPP has two choices: it enters some global cluster, along with $\{\Theta_{mj'}:j'\neq j\}$, with an average probability of $\frac{1}{1+\beta}$; it enters a local cluster, independently of $\{\Theta_{mj'}:j'\neq j\}$, with an average probability of $\frac{\beta}{1+\beta}$. The simultaneous global and local clustering yields correlated local clusters.

Analytic expressions have been given in [15] for $p(\Theta_{mj}=\Theta_{m'j})$ and $p(\Theta_{mj}=\Theta_{m'j},\Theta_{mj'}=\Theta_{m'j'})$, $\forall\, m\neq m', j\neq j'$, which yield Proposition 10.1.

Proposition 10.1: Let $\Theta_m,\Theta_{m'}\overset{i.i.d.}{\sim}G$ with $G\sim LPP(\alpha,\beta,G_0)$. Denote $C_1(\alpha,\beta)=p(\Theta_{mj'}=\Theta_{m'j'})$ and $C_2(\alpha,\beta)=p(\Theta_{mj'}=\Theta_{m'j'}|\Theta_{mj}=\Theta_{m'j})$. Then

$$\frac{C_2(\alpha,\beta)}{C_1(\alpha,\beta)}=1+\frac{4\alpha}{(\beta^2+\beta+2)^2}$$

It is clear that $C_2(\alpha,\beta)>C_1(\alpha,\beta)$, because $\alpha>0$. Thus knowledge that tasks m and m' are in the same cluster for $\mathcal{I}_j$ strictly increases the probability that these tasks are in the same cluster for $\mathcal{I}_{j'}$. In this sense, we say the local cluster for $\mathcal{I}_j$ is positiveley correlated with the local cluster for $\mathcal{I}_{j'}$.

The analysis based on Proposition 10.1 considers only two subsets of Θ and does not reveal the correlation among $n>2$ subsets. In what follows, we extend the analysis to $2\leq n\leq J$ subsets. Our analysis begins with Lemma 10.2, where we provide an analytic formula for the joint probability that two tasks are in the same cluster for n distinct subsets of Θ. The lemma is proven in the Appendix.

Lemma 10.2: Let $\{j_1, j_2, \cdots, j_n\} \subset \{1, 2, \cdots, J\}$ have distinct elements. Let $\Theta_m = (\Theta_{m1}, \Theta_{m2}, \cdots, \Theta_{mJ})$ and $\Theta_{m'} = (\Theta_{m'1}, \Theta_{m'2}, \cdots, \Theta_{m'J})$ be i.i.d. drawn from G with $G \sim LPP(\alpha, \beta, G_0)$, then

$$\begin{aligned} &p(\Theta_{mj_1} = \Theta_{m'j_1}, \Theta_{mj_2} = \Theta_{m'j_2}, \cdots, \Theta_{mj_n} = \Theta_{m'j_n}) \\ &= \frac{1}{(1+\alpha)^{n+1}(2+\beta)^n} \left\{ \left[\frac{2(1+\alpha)}{1+\beta} + \beta \right]^n + \alpha\beta^n \right\} \end{aligned}$$

It is easy to verify that the two formulae in [15] are special cases of the formula in Lemma 10.2, corresponding to $n = 1$ and $n = 2$, respectively. Lemma 10.2 provides complete information for the correlations among different subsets of Θ when considering the clustering of two tasks. Here, we are particularly interested in the additional change of the probability of $\Theta_{mj_1} = \Theta_{m'j_1}$ when one observes the new local cluster $\Theta_{mj_{n+1}} = \Theta_{m'j_{n+1}}$, given that one has already observed $n-1$ previous local clusters $\{\Theta_{mj_k} = \Theta_{m'j_k}, k = 2, 3, \cdots, n\}$ before observing the new one.

Proposition 10.3: Let $\Theta_m, \Theta_{m'} \overset{i.i.d.}{\sim} G$ with $G \sim LPP(\alpha, \beta, G_0)$. Then it holds

$$\begin{aligned} &C_n(\alpha, \beta) \\ &\overset{Def.}{=} p(\Theta_{mj_1}{=}\Theta_{m'j_1} | \Theta_{mj_2}{=}\Theta_{m'j_2}, \cdots, \Theta_{mj_n}{=}\Theta_{m'j_n}) \\ &< p(\Theta_{mj_1}{=}\Theta_{m'j_1} | \Theta_{mj_2}{=}\Theta_{m'j_2}, \cdots, \Theta_{mj_{n+1}}{=}\Theta_{m'j_{n+1}}) \\ &= C_{n+1}(\alpha, \beta) \end{aligned}$$

Proof. Since both sides are positive, we need only to prove that the ration of the right side to the left side is larger than one. Denote $\zeta = \frac{2(1+\alpha)}{1+\beta} + \beta$. The ratio, using Lemma 10.2, is

$$\begin{aligned} &\frac{\zeta^{n+1} + \alpha\beta^{n+1}}{\zeta^n + \alpha\beta^n} \frac{\zeta^{n-1} + \alpha\beta^{n-1}}{\zeta^n + \alpha\beta^n} \\ &= \frac{\zeta^{2n} + \alpha^2\beta^{2n} + \zeta^{n+1}\alpha\beta^{n-1} + \zeta^{n-1}\alpha\beta^{n+1}}{\zeta^{2n} + \alpha^2\beta^{2n} + 2\zeta^n\alpha\beta^n} > 1, \end{aligned}$$

because

$$\frac{\zeta^{n+1}\alpha\beta^{n-1} + \zeta^{n-1}\alpha\beta^{n+1}}{2\zeta^n\alpha\beta^n} = \frac{1}{2}\left(\frac{\zeta}{\beta} + \frac{\beta}{\zeta}\right) > 1,$$

where the last inequality is arrived using the facts that $\frac{\zeta}{\beta} > 1$ and $x + \frac{1}{x} > 2$ for $x > 1$. Q.E.D.

It is clear from Proposition 10.1 that $C_2 < C_3 < \cdots < C_n < C_{n+1} < \cdots < C_J$, which shows that the LPP prior cumulatively increases the probability of $\Theta_{mj_1} = \Theta_{mj_1}$, when tasks m and m' are observed to cluster for an increasing number of other subsets of Θ. In other words, each observation, say $\Theta_{mj_k} = \Theta_{mj_k}$ ($k > 1$), increases the probability of $\Theta_{mj_1} = \Theta_{mj_1}$ on the basis of the increases brought by the previous observations $\{\Theta_{mj_i} = \Theta_{mj_i} : i = 2, \cdots, k-1\}$. It is noted that Proposition 10.1 has shown that $C_1 < C_2$, which along with Proposition 10.3, establishes that $\{C_n : n \geq 1\}$ is a strict monotonically-increasing sequence. Therefore the two propositions provide a complete picture of the positive correlation between the local cluster formed on a single subset of Θ and the local clusters formed on multiple other subsets of Θ.

C. The Relevance to MTRL

The correlation analysis above shows that the LPP has a more flexible clustering structure than either a global DP as in (16) or multiple local DPs as in (15), which allows the LPP to capture a richer set of similarity patterns among the tasks and, accordingly, to make information sharing more effective. The positive correlation is particularly appealing in our present case, in which an RPR policy is sought in each environment to accomplish the task of accruing long term reward. Each subset of RPR parameters assume a particular responsibility in the task and two different subsets of parameters may need to coordinate with each other to make an overall functioning policy.

Consider, for instance, several subsets of RPR parameters, respectively performing action selection in distinct and yet related belief regions. The actions in these regions must coordinate to produce a sequence of actions that lead to the desired consequence. If, indeed, two environments have some similar belief regions and the same consequence (hence policy) is desired with respect to these regions, then the action selection in these regions must be shared across the environments. Independent local sharing by independent DPs ignores these relations and leads to inefficient information usage. On the other hand, complete global sharing by DP is inappropriate for partially similar environments.

In the MTRL we consider here, the LPP is imposed on the RPRs, not on the associated environments (POMDPs). When the prior knowledge is about the environments, how does one transform it into the knowledge about the RPRs? If we are specifying the knowledge about each individual environment, then we indeed need to transform it into the knowledge about each associated RPR. However, the prior we are trying to impose is about the *relations* between the environments. For the relational prior, we do not need the transform itself; we only need to require that the transform is continuous, in the sense that, if two POMDPs (say m and m') are similar for part i, then their corresponding RPRs are similar for part j.

To examine whether such a requirement is satisfied in our case, we recall that each POMDP is a belief-state MDP (Markov decision process) and that the corresponding RPR is defined in terms of the regions in the belief-state space [13]. The locality with respect to (w.r.t.) the POMDP states corresponds to the locality w.r.t. to the belief-states, which then transfer to the locality w.r.t. the belief-regions in the RPR, if the belief regions form a Markov partition of the belief-state space [11]. When the last condition is not satisfied exactly, the locality correspondence between POMDP and RPR is approximate. However, this does not affect our method too much, given that the RPR yields a stochastic policy and that the information sharing here is probabilistic under a nonparametric Bayesian prior. The advantage of our method is still prominent, as demonstrated by the experimental results.

D. Posterior Inference

We first introduce some latent variables. For $j = 1, \cdots, J$, let $s_{mj} \in \{0, 1\}$ with $s_{mj} = 1$ denoting $\Theta_{mj} = \overline{\Theta}_{mj}$ and $s_{mj} = 0$ denoting $\Theta_{mj} = \widetilde{\Theta}_{mj}$. Let $\varphi_{m0} \in \{1, 2, \cdots, \infty\}$ index the global cluster that task m is allocated to. Let $\varphi_{mj} \in \{1, 2, \cdots, \infty\}$ index the local cluster that task m is allocated to, concerning the j-th subset of Θ. It is easy to see that $p(s_{mj} = 1) = \eta_j$, and $p(\varphi_{mj} = i) = \lambda_{ji}$ for $j = 0, 1, \cdots, J$ and $i = 1, 2, \cdots, \infty$.

We are interested in the posterior of $\{\Theta_m\}_{m=1}^M$, given the LPP prior specified in (17) and the episodes $\cup_{m=1}^M \mathcal{D}^{(K_m)}$. To allow inference of the LPP parameters, we put a Gamma prior on each, i.e., we assume *a priori* that $\alpha \sim \mathrm{Ga}(a_\alpha, b_\alpha)$ and $\beta \sim \mathrm{Ga}(a_\beta, b_\beta)$. We employ a hybrid approach to posterior inference, based on the stick-breaking construction in (18). Specifically, we employ the slice sampler in [17] to perform conditional Gibbs sampling of $\{u_{mj}\} \cup \{s_{mj}\} \cup \{\lambda^*_{ji}\} \cup \{\varphi_{mj}\} \cup \{\eta_j\} \cup \{\alpha, \beta\}$ given $\{\overline{\Theta}^*_i\} \cup \{\widetilde{\Theta}^*_i\}$, where $\{u_{mj}\}$ are auxiliary latent variables conditional on which the infinite mixtures in $\overline{G}$ and $\{\widetilde{G}_j\}$ become finite. Given the Gibbs samples, we then employ the variational Bayesian (VB) algorithm in [13] to infer $\{\overline{\Theta}^*_i\} \cup \{\widetilde{\Theta}^*_i\}$. The steps of the hybrid Gibbs-variational approach are summarized as follows.

Step 1. Draw $u_{mj} \sim \mathrm{Unif}(0, \lambda_{j\varphi_{mj}})$, $j = 0, 1, \ldots, J$.

Step 2. Draw $s_{mj} \sim \mathrm{Ber}(p_{mj})$, with

$$p_{mj} = \frac{\eta_j \widehat{V}(\mathcal{D}^{(K_m)}; \Theta_{m(s_{mj}=1)})}{\eta_j \widehat{V}(\mathcal{D}^{(K_m)}; \Theta_{m(s_{mj}=1)}) + (1-\eta_j)\widehat{V}(\mathcal{D}^{(K_m)}; \Theta_{m(s_{mj}=0)})}$$

where $\widehat{V}(\cdot)$ is the empirical value function as defined in Definition 8.2, $\Theta_{m(s_{mj}=1)}$ is Θ_m with $\Theta_{mj} = \overline{\Theta}^*_{\varphi_{m0}j}$ and $\Theta_{m(s_{mj}=0)}$ is Θ_m with $\Theta_{mj} = \widetilde{\Theta}^*_{\varphi_{mj}j}$.

Step 3. Draw λ^*_{ji} from the conditional density

$$p(\lambda^*_{ji}|\cdots) \propto (1-\lambda^*_{ji})^{(\alpha-1)} \prod_{m=1}^{M} \prod_{j=0}^{J} \mathbb{I}(\lambda^*_{j\varphi_{mj}} \prod_{l<\varphi_{mj}} (1-\lambda^*_{jl}) > u_{mj})$$

Let $\varphi^*_j = \max\{\varphi_{mj} : m = 1, \ldots, M\}$. It is easy to show $p(\lambda^*_{ji}|\cdots) = \mathrm{Be}(\lambda^*_{ji}|1, \alpha)$ for $i > \varphi^*_j$ while, for $i \leq \varphi^*_j$, $p(\lambda^*_{ji}|\cdots) \propto \mathrm{Be}(\lambda^*_{ji}|1, \alpha)\mathbb{I}(d^L_{ji} < \lambda^*_i < d^R_{ji})$ with

$$d^L_{ji} = \max_m \left\{ \frac{u_{mj}}{\prod_{l<i}(1-\lambda^*_{jl})} : \varphi_{mj} = i \right\}$$
$$d^R_{ji} = 1 - \max_m \left\{ \frac{u_{mj}}{\lambda^*_{j\varphi_{mj}} \prod_{l<\varphi_{mj}, l\neq i}(1-\lambda^*_{jl})} : \varphi_{mj} > i \right\}$$

Step 4. Draw φ_{m0} and φ_{mj} according to

$$p(\varphi_{m0}|\cdots) \propto \mathbb{I}(\varphi_{m0} \in \Gamma_{m0}) \widehat{V}(\mathcal{D}^{(K_m)}; \Theta_{m(s_{mj}=1)})$$
$$p(\varphi_{mj}|\cdots) \propto \mathbb{I}(\varphi_{mj} \in \Gamma_{mj}) \widehat{V}(\mathcal{D}^{(K_m)}; \Theta_{m(s_{mj}=0)})$$

where, for $j = 0, 1, \cdots, J$, $\Gamma_{mj} = \{i : \lambda_{ji} > u_{mj}\}$.

Step 5. Draw $\eta_j \sim \mathrm{Be}(1 + \sum_m s_{mj}, \beta + \sum_m (1 - s_{mj}))$.

Step 6. Draw $\beta \sim \mathrm{Ga}(a_\beta + J, b_\beta - \sum_{j=1}^{J} \log(1 - \eta_j))$.

Step 7. Draw $\alpha \sim \mathrm{Ga}(a_\alpha + \sum_{j=0}^{J} \varphi^*_j, b_\alpha - \sum_{j=0}^{J} \sum_{i=1}^{\varphi^*_j} \log(1 - \lambda^*_{ji}))$, with $\{\varphi^*_j\}$ given in Step 3.

Step 8. For $j = 1, 2, \cdots, J$, $i = 1, 2, \cdots$, do the following. If $\{m : s_{mj} = 1, \varphi_{m0} = i\}$ is nonempty, infer $\overline{\Theta}^*_{ij}$ by applying the VB algorithm to $\cup_{s_{mj}=1, \varphi_{m0}=i} \mathcal{D}^{(K_m)}$ with the prior G_{0j}; if $\{m : s_{mj} = 0, \varphi_{mj} = i\}$ is nonempty, infer $\widetilde{\Theta}^*_{ij}$ by applying the VB algorithm to $\cup_{s_{mj}=0, \varphi_{mj}=i} \mathcal{D}^{(K_m)}$ with the prior G_{0j}.

XI. Experimental Results

We consider the ten maze navigation tasks in [13]. Of the ten environments, the first four, the following three, and the last three, are respectively duplicates of the gird-world (a), (b), and (c) in Figure 3. The grid-worlds (a) and (c) are partly similar for the cells enclosed by dashed lines. We use these ground truths in analyzing the sharing mechanism later.

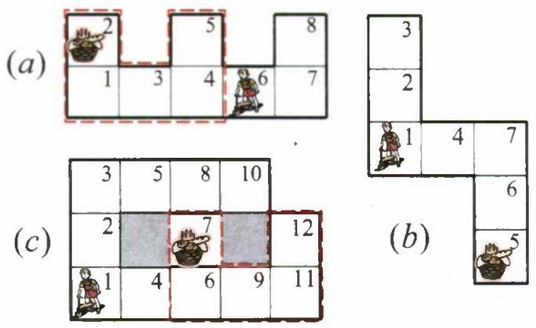


Fig. 3. The three distinct grid-world environments considered in [13], with the goal indicated by a basket. The goal states are fully observable.

We follow the experimental setup used in [13] to replicate the results for comparison. In particular, we set $|\mathcal{Z}| = 6$ for all RPRs and perform off-line learning, assuming the episodes $\{\mathcal{D}^{(K_m)}\}_{m=1}^{10}$ are collected beforehand, by taking random actions or querying a PBVI expert, the selection between the two manifested with probability of 0.5. The Gamma hyper-parameters for α and β in the LPP are set to $a_\alpha = b_\alpha = a_\beta = b_\beta = 1$. The same base measure G_0 is used for the LPP and all DPs (global and local), and the form of G_0 is given in (14) with all Dirichlet hyper-parameters set to one.

We compare the proposed LPP-based MTRL method to the following methods: STRL, PBRL, DP-based MTRL, and IDP-based MTRL, where IDP is a set of independent DPs with each associated with a subset of Θ. The DP is implemented by the LPP with β set to 10^{-20} and the IDP is implemented by the LPP with β set to 10^{20}.

The results are reported in Figure 4. It is seen that the LPP-based MTRL earns significantly larger rewards than the DP-based MTRL. The improvements are attributed to the higher goal rates, since the LPP actually takes a larger number of steps to reach the goal. The improvements are most significant when the number of episode is small (< 10 here). The IDP-based MTRL performs most poorly among the methods.

The performance of each method can be explained by the sharing patterns it infers, which we visualize using Hinton diagrams [13], shown in Figure 5 for $K = 3$ episodes (top row) and $K = 120$ episodes (bottom row). The block (i, j) is each diagram displays the frequency the tasks i and j are assigned to the same cluster in the last 1000 iterations of Gibbs sampling. It is seen that DP infers three global clusters, respectively corresponding to the three grid-worlds in Figure 3. By contrast, the LPP combines grid-worlds (a) and (c) into a single global cluster, to capture the similar parts enclosed by the dashed lines shown in Figure 3, with the differences between (a) and (c) distinguished by splitting them in local clusters.

An example is given in the fifth column of Figure 5, where it is seen that the LPP tends to split grid-worlds (a) and (c) for $W(:, a, o', :)$, which involves the belief-region transitions when walking west leads to observing walls on the south and north. It is clear from Figure 3 that such an (action, observation) pair exclusively leads towards the goal in grid-world (a) while it may also move away from the goal in grid-world (c). By locally splitting them, the LPP encourages respective appropriate transitions in the two grid-worlds.

The diagrams show that the IDP-based MTRL has a strong tendency of isolating tasks, which is detrimental to information sharing and explains its poor performances. For example, the third column of Figure 5 shows the local sharing patterns involving the belief-region transitions when walking north leads to seeing the goal. Clearly, this (action, observation) pair leads to large rewards in both grid-worlds (a) and (c) and hence local sharing between them is helpful here. It is seen that the IDP needs a large amount of episodes ($K = 120$) to infer this, while the LPP infers this using only three episodes.

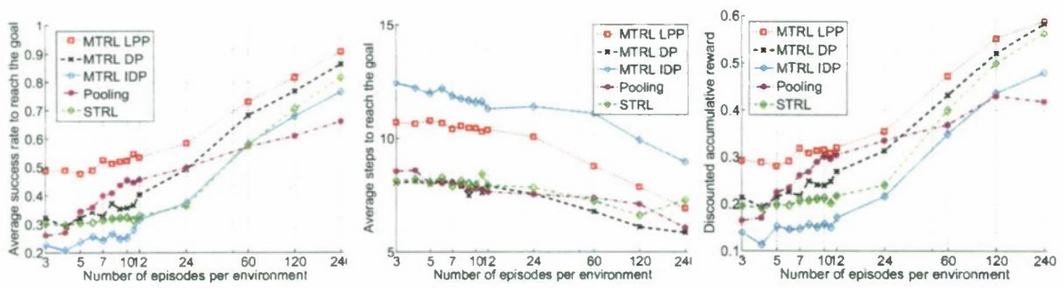


Fig. 4. Performance comparison on the ten maze navigation tasks, as a function of the number of episodes per environment used by the algorithms. Left: Average success rate for the agent to reach the goal within 15 steps. Middle: Average steps taken by the agent to reach the goal. Right: Discounted cumulative reward with the discount $\gamma = 0.95$.

XII. Summary on Networked POMDPs

We have presented a new framework for multi-task reinforcement learning, based on simultaneous global and local information-sharing imposed by the LPP. We have extended the second-order analysis in [15] to higher-order analysis involving an arbitrary number of subsets of parameters. Our analysis provide further insights into the clustering structure under the LPP. Experimental results demonstrate the new MTRL

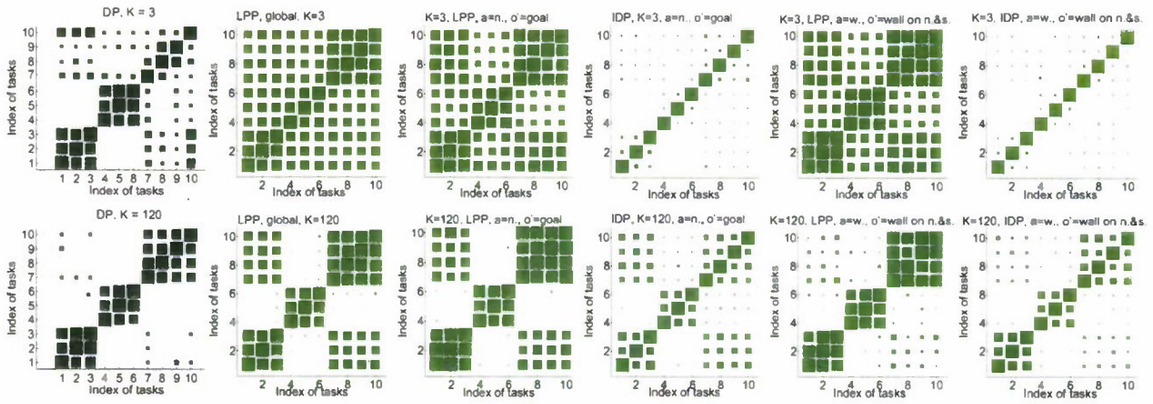


Fig. 5. The between-task sharing patterns, represented by Hinton diagrams, inferred by DP (column 1), IDPs (columns 4 & 6), and LPP (column 2 for the global sharing; columns 3 & 5 for the local sharing), where K is the number of episodes per environment ($K = 3$ for the first row and $k = 120$ for the second row.). The local sharing patterns are compared for $W(:, a, o', :)$, with a="walk north" and o'="goal" in columns 3-4, and a="walk west" and o'="walls on the north and south" in columns 5-6. Note the goal states are fully observable.

method yields significant performance improvements, relative to previous published results. Future work includes extension of the method to online learning and the study of exploitation vs exploration within the MTRL framework.

APPENDIX

Proof of Lemma 10.2: Let $s_{mj} \in \{0, 1\}$ with $s_{mj} = 1$ denoting $\Theta_{mj} = \overline{\Theta}_{mj}$ and $s_{mj} = 0$ denoting $\Theta_{mj} = \widetilde{\Theta}_{mj}$.

$$
\begin{aligned}
& p(\Theta_{mj_1} = \Theta_{m'j_1}, \Theta_{mj_2} = \Theta_{m'j_2}, \cdots, \Theta_{mj_n} = \Theta_{m'j_n}) \\
\stackrel{a}{=} & \int \Big\{ \prod_{k=1}^{n} \sum_{s_{mj_k}=0}^{1} \sum_{s_{m'j_k}=0}^{1} p(\Theta_{mj_k} = \Theta_{m'j_k} | s_{mj_k}, s_{m'j_k}) \\
& \qquad \times p(s_{mj_k}, s_{m'j_k} | \eta_{j_k}) p(\eta_{j_k}) \Big\} d\eta_{j_1} \cdots d\eta_{j_n} \\
\stackrel{b}{=} & \int p(\eta_{j_1}) \cdots p(\eta_{j_n}) \Big\{ \frac{1}{1+\alpha} \prod_{k=1}^{n} \Big[\eta_{j_k}^2 + \frac{(1-\eta_{j_k})^2}{1+\alpha} \Big] \\
& \qquad + \Big(1 - \frac{1}{1+\alpha} \Big) \prod_{k=1}^{n} \frac{(1-\eta_{j_k})^2}{1+\alpha} \Big\} d\eta_{j_1} \cdots d\eta_{j_n} \\
\stackrel{c}{=} & \frac{1}{1+\alpha} \Big[\frac{2}{(1+\beta)(2+\beta)} + \frac{\beta}{(1+\alpha)(2+\beta)} \Big]^n \\
& \qquad + \frac{\alpha}{1+\alpha} \Big[\frac{\beta}{(1+\alpha)(2+\beta)} \Big]^n \\
\stackrel{d}{=} & \frac{1}{(1+\alpha)^{n+1}(2+\beta)^n} \Big\{ \Big[\frac{2(1+\alpha)}{1+\beta} + \beta \Big]^n + \alpha \beta^n \Big\}
\end{aligned}
$$

Equation (a) follows because s_{mj} is independent of $s_{mj'}$ and η_j is independent of $\eta_{j'}$, $\forall\, j' \neq j$. Equation (b) is arrived based on that $p(\theta = \theta') = \frac{1}{1+\alpha}$, $\forall\, \theta, \theta' \stackrel{i.i.d.}{\sim} DP(\alpha P_0)$ [18], and that $p(\Theta_{mj_k} = \Theta_{m'j_k}) = 0$ whenever one of them is from the global DP and the other from the j_k-th local DP, and that $p(\overline{\Theta}_{mj_1} = \overline{\Theta}_{m'j_1}, \cdots, \overline{\Theta}_{mj_n} = \overline{\Theta}_{m'j_n}) = \frac{1}{1+\alpha}$. To reach equation (c), one calculates the moments of $\eta_{j_k} \sim \mathrm{Be}(1, \beta)$. Q.E.D.

XIII. REVIEW OF TOPIC MODELING

Topic models attempt to infer sets of words from text data that together form meaningful contextual and semantic relationships. Finding these groups of words, known as topics, allows effective clustering, searching, sorting, and archiving of a corpus of documents. If we assume the bag-of-words structure, *i.e.*, that words are exchangeable and independent, then there are in general two ways to consider a collection

of documents. Factor models such as probabilistic Latent Semantic Indexing (pLSI) [19], Latent Dirichlet Allocation (LDA) [20] and Topics over Time (TOT) [21] assume that each word in a given document is drawn from a mixture model whose components are topics. Other models assume that words in a sentence or even in an overall document are drawn simultaneously from one topic [22], [23]. In [22], the authors propose modeling topics of words as a Markov chain, with successive sentences modeled as being likely to share the same topic. Since topics are hidden, learning and inferring the model are done using tools from hidden Markov models. Whether one draws a topic for every word or considers all words within a sentence/document as being generated by a common topic, documents are represented as counts over the dictionary, and topics are represented as multinomial distributions over the dictionary. This approach to topic representation is convenient, as the Dirichlet distribution is the conjugate prior to the multinomial. However, because the distribution over the dictionary must be normalized, problems can occur if a previously unknown word is encountered, as can often happen when using a trained model on an unknown testing set.

A new factor model has been proposed [24] that represents each integer word count from the term-document matrix as a sample from an independent poisson distribution. This model, called GaP for gamma-poisson, factorizes the sparse term-document matrix into the product of an expected-counts matrix and a theme probability matrix. Note that the GaP model is equivalent to placing a multinomial-Dirichlet implementation over the dictionary, so that one can model both the relative word frequencies and the overall word count. One may use the poisson-gamma characterization as a starting point to building a dynamic topic model by using a closely-related approach to [25]. Using an independent distribution for each word is attractive, as it addresses the problem of adding unknown words to the dictionary. Further, since each word is allowed to evolve independently, this approach leads to a more flexible model than using a traditional multinomial-Dirichlet structure. We build upon this construct in the model presented here.

The main focus of this component of the project is on development of a hierarchical Bayesian model for characterizing documents with known time stamp. Each document is assumed to have an associated topic, and all documents at a given time are assumed to have topics that are drawn from a mixture model; the mixture weights in this model evolve with time. This framework imposes the idea that documents that appear at similar times are likely to be drawn from similar mixtures of topics. To achieve this goal, we develop a simplified form of the dynamic hierarchical Dirichlet process (dHDP) [26]. Inference is performed efficiently via a variational Bayesian analysis [8].

Our model differs from other time-evolving topic models [27], [28] in that our topics do not evolve over time; what changes in time are the mixing weights over topics, while the overall set of topics are kept unchanged. Specific topics are typically localized over a period of time, with new dominant topics spawned after other topics diminish in importance (the temporally localized topics may alternatively be viewed as a time evolution of a single topic [28], but such single-topic evolution is not considered here).

XIV. Review of Semi-Parametric Statistical Modeling

The Dirichlet process (DP) is a semi-parametric measure for development of general mixture models (in principle, in terms of an infinite number of mixture components). Let H be a measure and α is a non-negative real number. A draw from a Dirichlet process parameterized by α and H is denoted $G \sim DP(\alpha, H)$. Sethuraman [29] introduced the stick-breaking representation of a DP draw:

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*},$$
$$\pi_k = V_k \prod_{l=1}^{k-1}(1 - V_l), \quad V_k \overset{i.i.d.}{\sim} Beta(1, \alpha), \quad \theta_k^* \overset{i.i.d.}{\sim} H, \tag{19}$$

where $\delta_{\theta_k^*}$ is a point measure concentrated at θ_k^* (each θ_k^* is termed an atom), and $Beta(1, \alpha)$ is a beta distribution with shape parameter α. Note that G is almost surely discrete, with this playing a key role in the utility of DP for clustering. To simplify notation below, an infinite probability vector π constructed as above is denote $\pi \sim Stick(\alpha)$.

Suppose the data of interest are divided into different sets, and each data set is termed a "task" for analysis. For clustering of T tasks the DP imposes the belief that when two tasks are associated with the

same cluster, all data within the tasks are shared. This may be too restrictive in some applications and has motivated the hierarchical Dirichlet process (HDP) [30]. We denote the data in task t as $\{\boldsymbol{x}_{t,i}\}_{i=1}^{N_t}$, where N_t is the number of data in the task. The HDP may be represented as

$$\begin{aligned} \boldsymbol{x}_{t,i} &\sim f(\theta_{t,i}); \quad i = 1, 2, ..., N_t; \quad t = 1, 2, ..., T, \\ \theta_{t,i} &\sim G_t; \quad i = 1, 2, ..., N_t; \quad t = 1, 2, ..., T, \\ G_t &\sim DP(\alpha, G); \quad t = 1, 2, ..., T, \\ G &\sim DP(\gamma, H), \end{aligned} \tag{20}$$

where $f(\theta)$ represents the specific parametric model under consideration. Because the task-dependent DPs share the same (discrete) base G, all $\{G_t\}_{t=1}^{T}$ share the same set of mixture atoms, with different mixture weights. The measures $\{G_t\}_{t=1,T}$ are *jointly* drawn from an HDP:

$$\{G_1, \cdots, G_T\} \sim HDP(\alpha, \gamma, H). \tag{21}$$

The HDP assumes the T tasks are exchangeable; however, there are many applications for which it is desirable to remove this exchangeability assumption. Models such as the kernel stick breaking process [31], [32], the generalized product partition model [33], the correlated topic model [34] and the dynamic DP [35] are techniques that impose structure on the dependence of the tasks (removing exchangeability). Some of these models rely on modifying the mixing weights to impose dependence on location [31], [32] or covariate [33], while others impose sequential time dependence on the structure of consecutive tasks (see [35]).

We again consider T tasks, but now index t explicitly denotes the sequential time of data production/collection. To address the sequential nature of the time blocks, [26] imposes a dynamic HDP (dHDP)

$$G_t = w_t D_t + (1 - w_t) G_{t-1}; \quad t = 2, ..., T, \tag{22}$$

where $\{G_1, D_2, \cdots, D_T\} \sim HDP(\alpha, \gamma, H)$. The parameter $w_t \in [0, 1]$ is drawn from a beta distribution $Beta(a_0, b_0)$, and it controls the degree of innovation in G_t relative to G_{t-1}. The DP and HDP are limiting cases of this model:

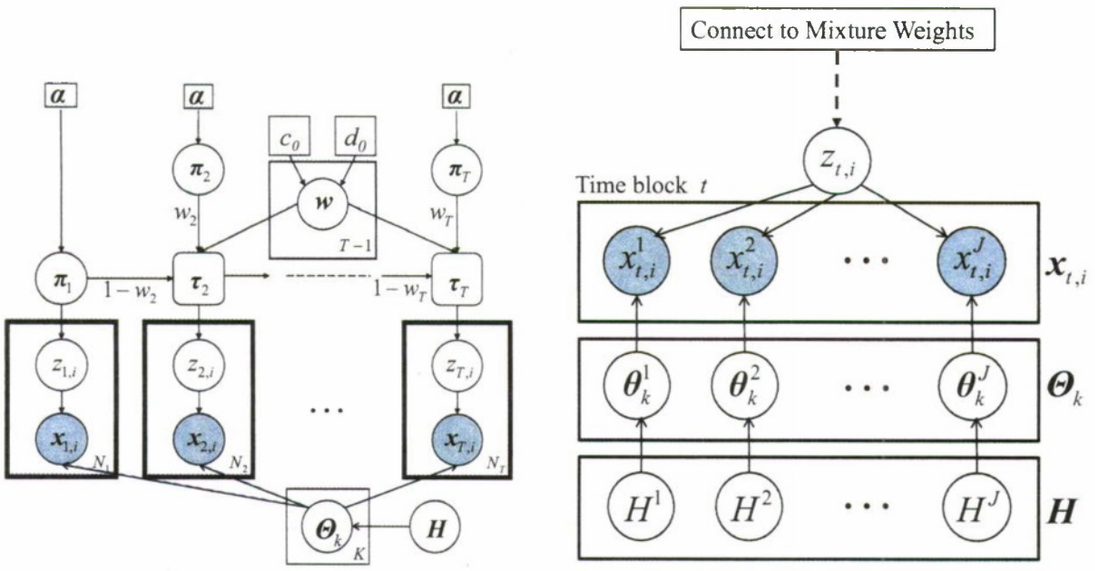
- when $w_t \to 0$, $G_t \to G_{t-1}$ and there is no innovation, resulting in a common set of mixture weights for all time blocks (DP);
- when $w_t \to 1$, $G_t \to D_t$, where the new innovation distribution D_t controls the sharing mechanism, resulting in each time block having a unique set of mixing weights (HDP).

It is important to restate that dHDP does not assume the mixture components evolve over time, only the mixing weights. The mixture components are shared explicitly across all time blocks. This is fundamentally different from other models that impose temporal dependence through component evolution [28], [27], this allowing a unique and independent set of mixing weights for each block.

XV. Semi-Parametric Dynamic Topic Model

A. Model construction

Consider a collection of documents with known time stamps, with time evolving from $t = 1, \cdots, T$. At any particular time we have N_t such independent documents. The total set of documents over all time is represented as $\{\boldsymbol{x}_{t,i}\}_{i=1\,t=1}^{N_t\ \ T}$, where $\boldsymbol{x}_{t,i}$ represents a vector of word counts associated with document i at time t. In the form of the model presented here, we are only interested in the number of times a given word is present in a particular document; the set of J unique words in the collection forms a dictionary. Each document is assumed characterized by a single topic, and at time t the topics across all documents are assumed drawn from a mixture model. In the proposed model the mixture weights on the topics are assumed to evolve with time (analogous to as implemented in the dHDP [26] discussed above). The assumption that each document is characterized by a single topic may seem restrictive; however, we observe in Section XVIII that for our motivating example this assumption is reasonable.



(a) (b)

Fig. 6. Dynamic Dirichlet topic model (dDTM). (a) graphical representation of the model, (b) expanded representation of the product measure aspect of the model.

To constitute a model with a time-evolving mixture of topics, we seek a simplified representation of the dHDP. Specifically, the proposed topic model, termed dDTM for dynamic Dirichlet topic model, is represented as

$$\begin{aligned}
x^j_{t,i}|z_{t,i} &\sim F(\boldsymbol{\theta}^j_{z_{t,i}}); \quad j = 1, \ldots, J. \\
z_{t,i}|\boldsymbol{\tau}_t &\sim Mult(\boldsymbol{\tau}_t); \quad t = 2, \ldots, T, \\
z_{1,i}|\boldsymbol{\pi}_1 &\sim Mult(\boldsymbol{\pi}_1), \\
\boldsymbol{\tau}_t &= (1 - w_t)\boldsymbol{\tau}_{t-1} + w_t\boldsymbol{\pi}_t; \quad t = 2, \ldots, T, \\
\boldsymbol{\pi}_t &\sim Dir(\boldsymbol{\alpha}); \quad t = 1, \ldots, T, \\
w_t &\sim Beta(c_0, d_0); \quad t = 2, \ldots, T, \\
\boldsymbol{\theta}^j_k &\sim H^j; \quad j = 1, \ldots, J; \quad k = 1, \ldots, K,
\end{aligned} \tag{23}$$

Note that $\boldsymbol{\tau}_1 = \boldsymbol{\pi}_1$. The factorized structure $\boldsymbol{H} = \prod_{j=1}^{J} H^j$ is similar to [24], which allows insertion of new words with time.

Although perhaps not apparent at this point, for large K the proposed model is closely related to dHDP; this is analyzed in detail below. The model is represented graphically in Fig. 6(a), and in Fig. 6(b) we illustrate how a single mixture component is drawn, with the parametric model of each dimension drawn independently from its respective prior.

The form of the parametric model $F(\cdot)$ in (23) may vary depending on the application; in the work presented here it corresponds to a multinomial-Dirichlet model. We consider the number of times a word is present in a given document; to do this, $F(\cdot)$ is defined as a multinomial distribution and consequently, to preserve the conjugacy requirements, each H^j is a Dirichlet distribution.

B. Relationship to dHDP

We now make explicit the relationship between dHDP [26] and dDTM represented in (23). Recall that the draws $\{G_1, D_2, \cdots, D_T\} \sim HDP(\alpha, \gamma, H)$ may be constructed as [30]

$$\begin{aligned} G_1 &= \sum_{k=1}^{\infty} \pi_{1,k} \delta_{\boldsymbol{\Theta}_k} \\ D_t &= \sum_{k=1}^{\infty} \pi_{t,k} \delta_{\boldsymbol{\Theta}_k} \quad , \quad t = 2, \cdots, T \\ \boldsymbol{\pi}_t &\sim DP(\alpha, \boldsymbol{v}) \quad , \quad t = 1, \cdots, T \\ \boldsymbol{v} &\sim Stick(\gamma) \\ \boldsymbol{\Theta}_k &\sim \boldsymbol{H} \quad , \quad k = 1, \cdots, \infty \end{aligned} \tag{24}$$

The draw $\boldsymbol{\pi}_t \sim DP(\alpha, \boldsymbol{v})$ may be represented in stick-breaking form, with the kth component of $\boldsymbol{\pi}_t$ constructed as $\pi_{t,k} = \sum_{j=1}^{\infty} w_{t,j} \delta(Y_{t,j} = k)$, with $\boldsymbol{w}_t \sim Stick(\alpha)$, $Y_{t,j} \sim Mult(\boldsymbol{v})$; $\delta(Y_{t,j} = k)$ equals one if $Y_{t,j} = k$, and its zero otherwise. We may also truncate the draw $\boldsymbol{v} \sim Stick(\alpha)$ to K sticks (denoted $\boldsymbol{v}_K \sim Stick_K(\alpha)$), for large K [36]. Using these representations, the overall HDP construction, when truncated to K topics (atoms), may be represented as

$$\begin{aligned} G_1 &= \sum_{k=1}^{K} \pi_{1,k} \delta_{\boldsymbol{\Theta}_k} \\ D_t &= \sum_{k=1}^{K} \pi_{t,k} \delta_{\boldsymbol{\Theta}_k} \quad , \quad t = 2, \cdots, T \\ \pi_{t,k} &= \sum_{j=1}^{\infty} w_{t,j} \delta(Y_{t,j} = k) \quad ; \quad k = 1, \cdots, K; \quad t = 1, \cdots, T \\ \boldsymbol{w}_t &\sim Stick(\alpha) \quad , \quad t = 1, \cdots, T \\ Y_{t,j} &\sim Mult(\boldsymbol{v}_K) \quad ; \quad j = 1, \cdots, \infty \quad ; \quad t = 1, \cdots, T \\ \boldsymbol{v}_K &\sim Stick_K(\gamma) \\ \boldsymbol{\Theta}_k &\sim \boldsymbol{H} \quad , \quad k = 1, \cdots, K \end{aligned} \tag{25}$$

Note that we truncate $Stick(\gamma)$ to K sticks, but do *not* truncate $Stick(\alpha)$. Additionally, $Y_{t,j} \in \{1, \cdots, K\}$, with the particular value of $Y_{t,j}$ depending on which component is selected from the multinomial.

To appreciate the relationship between dHDP and the proposed dDTM, note that (23) corresponds to drawing atoms/topics at time t from the finite mixture model $G_t = w_t D_t + (1 - w_t) G_{t-1}$, with

$$\begin{aligned} G_1 &= \sum_{k=1}^{K} \pi_{1,k} \delta_{\boldsymbol{\Theta}_k} \\ D_t &= \sum_{k=1}^{K} \pi_{t,k} \delta_{\boldsymbol{\Theta}_k} \quad , \quad t = 2, \cdots, T \\ \boldsymbol{\pi}_t &\sim Dir(\alpha/K, \cdots, \alpha/K) \quad , t = 1, \cdots, T \\ \boldsymbol{\Theta}_k &\sim \boldsymbol{H} \quad , \quad k = 1, \cdots, K \end{aligned} \tag{26}$$

Recall that Sethuraman demonstrated [29] that a draw $\boldsymbol{\pi} \sim Dir(\alpha \boldsymbol{g}_0)$, where $\boldsymbol{g}_0$ is a K-dimensional

probability vector and $\alpha > 0$, may be constructed as

$$\begin{aligned} \pi_k &= \sum_{j=1}^{\infty} w_j \delta(Y_j = k) \quad , \quad k = 1, \cdots, K \\ \boldsymbol{w} &\sim Stick(\alpha) \\ Y_j &\sim Mult(\boldsymbol{g}_0) \quad , \quad j = 1, \cdots, \infty \end{aligned} \tag{27}$$

with π_k representing the kth component of $\boldsymbol{\pi}$. Using Sethuraman's stick-breaking representation of the Dirichlet distribution in (26), the proposed dDTM is constructed as

$$\begin{aligned} G_1 &= \sum_{k=1}^{K} \pi_{1,k} \delta_{\boldsymbol{\Theta}_k} \\ D_t &= \sum_{k=1}^{K} \pi_{t,k} \delta_{\boldsymbol{\Theta}_k} \quad , \quad t = 2, \cdots, T \\ \pi_{t,k} &= \sum_{j=1}^{\infty} w_j \delta(Y_{t,j} = k) \quad ; \quad k = 1, \cdots, K; \quad t = 1, \cdots, T \\ \boldsymbol{w}_t &\sim Stick(\alpha) \quad , t = 1, \cdots, T \\ Y_{t,j} &\sim Mult(1/K, \cdots, 1/K) \quad ; \quad j = 1, \cdots, \infty; \quad t = 1, \cdots, T \\ \boldsymbol{\Theta}_k &\sim \boldsymbol{H} \quad , \quad k = 1, \cdots, K \end{aligned} \tag{28}$$

The truncated dHDP model in (22) draws $\{G_1, D_2, \cdots, D_T\}$ from (25), assuming $Stick(\gamma)$ is truncated to K sticks [36]. By contrast, within dDTM the measures $\{G_1, D_2, \cdots, D_T\}$ are drawn from (28). In the former the random variables $Y_{t,j}$ are drawn from $\boldsymbol{v}_K$, which is in turn drawn from the truncated stick-breaking process $Stick_K(\gamma)$; in the latter we simply set $\boldsymbol{v}_K = (1/K, \cdots, 1/K)$ and remove the parameter γ altogether. It is felt that this relatively small change does not significantly affect the expressibility of the proposed prior. Within the proposed model the weights w_t explicitly impose temporal relationships between the topics (documents at proximate times are more likely to share the same topics).

The above discussion also demonstrates that considering the Dirichlet distribution $Dir(\alpha/K, \cdots, \alpha/K)$ with large K is analogous (but distinct from) a truncated stick-breaking representation. In this sense, the proposed model is non-parametric, in that setting a large K allows the model to infer the proper number of topics from the data, analogous to studies of the truncated stick-breaking representation [36]. Setting a large K (*e.g.*, $K = 50$ in the examples below), does not imply that we believe that there are actually K topics, since from (27) only a relatively small set of components in $\boldsymbol{\pi}_t$ will have appreciable amplitude (the same type of motivation for the stick-breaking view of DP and HDP). As in other non-parametric methods, the proposed model infers a distribution on the proper number of topics, based on the data.

We also emphasize that the stick-breaking representation of a draw from a Dirichlet distribution has been introduced above to make the connection between the proposed model and a truncated representation of dHDP. However, when actually performing inference, it is often simpler to just draw directly from $Dir(\alpha/K, \cdots, \alpha/K)$. However, this issue is revisited in the Conclusions.

C. Limiting cases

In Section XIV we considered dHDP under limiting cases of w_t, and we do so here for the proposed dDTM in (23). In the limit $w_t \rightarrow 0$, the dDTM parameters are drawn at all time from the same measure $G_1 = \sum_{k=1}^{K} \pi_{1,k} \delta_{\boldsymbol{\Theta}_k}$ with $\boldsymbol{\pi}_1 \sim Dir(\alpha/K, \cdots, \alpha/K)$ and $\boldsymbol{\Theta}_k \sim \boldsymbol{H}$. Therefore, in the limit $K \rightarrow \infty$ and $w_t \rightarrow 0$ the topic-model parameters for dDTM are drawn from $DP(\alpha, \boldsymbol{H})$, as is the case for dHDP when $w_t \rightarrow 0$. Since K is finite in dDTM, the limit $w_t \rightarrow 0$ yields a model similar to LDA [20] (in LDA one performs a point estimate for α, while here α is set).

In the limit $w_t \rightarrow 1$, at time t the dDTM model parameters are drawn from $G_t = \sum_{k=1}^{K} \pi_{t,k} \delta_{\boldsymbol{\Theta}_k}$, again with $\boldsymbol{\Theta}_k \sim \boldsymbol{H}$, and with each $\boldsymbol{\pi}_t \overset{iid}{\sim} Dir(\alpha/K, \cdots, \alpha/K)$. Thus the $\{G_t\}_{t=1,T}$ all share the same

atoms (topics), with distinct t-dependent probability weights $\boldsymbol{\pi}_t$. The dHDP model has a similar limit when $w_t \to 1$, with the weights drawn $\boldsymbol{\pi}_t \overset{iid}{\sim} DP(\alpha, \boldsymbol{v})$ for $\boldsymbol{v} \sim Stick(\gamma)$. In both cases the atoms/topics are shared across all time, with different mixture weights. The dHDP arguably allows for more modeling flexibility, through the parameter γ, while dDTM yields a simpler model with very similar structural form.

XVI. Model Properties

To examine properties of the model in (23), we consider the discrete indicator's space $I = \{1, 2, \ldots, K\}$ with $k \in I$ indicating one of the K mixing components of the model. Therefore, we can write

$$\begin{aligned} \boldsymbol{\tau}_t(I)|\boldsymbol{\tau}_{t-1}, w_t &= (1-w_t)\boldsymbol{\tau}_{t-1}(I) + w_t\boldsymbol{\pi}_t(I) \\ &= \boldsymbol{\tau}_{t-1}(I) + \Delta_t(I), \end{aligned} \tag{29}$$

where $\Delta_t(I) = w_t(\boldsymbol{\pi}_t(I) - \boldsymbol{\tau}_{t-1}(I))$ is the random deviation from $\boldsymbol{\tau}_{t-1}(I)$ to $\boldsymbol{\tau}_t(I)$.

Theorem 1. The mean and the variance of the random deviation Δ_t are controlled by the innovating weight w_t and model parameter $\boldsymbol{\alpha} = [\alpha/K, \ldots, \alpha/K]$:

$$\begin{aligned} E\{\Delta_t(I)|\boldsymbol{\tau}_{t-1}, w_t, \boldsymbol{\alpha}\} &= w_t(E(\boldsymbol{\pi}_t(I)) - \boldsymbol{\tau}_{t-1}(I)) \\ &= w_t([\frac{1}{K}, \ldots \frac{1}{K}] - \boldsymbol{\tau}_{t-1}(I)). \end{aligned} \tag{30}$$

$$V\{\Delta_t(I)|\boldsymbol{\tau}_{t-1}, w_t, \boldsymbol{\alpha}\} = \frac{w_t^2}{K}[\frac{1}{\alpha+1}(1-\frac{1}{K}), \ldots, \frac{1}{\alpha+1}(1-\frac{1}{K})], \tag{31}$$

where we observe two limiting cases:

- when $w_t \to 0$, $\boldsymbol{\tau}_t = \boldsymbol{\tau}_{t-1}$.
- when $\boldsymbol{\tau}_{t-1} \to [\frac{1}{K}, \ldots, \frac{1}{K}]$, $E\{\boldsymbol{\tau}_t(I)|\boldsymbol{\tau}_{t-1}, w_t\} = \boldsymbol{\tau}_{t-1}(I)$.

Theorem 2. The correlation coefficient between two adjacent distributions $\boldsymbol{\tau}_{t-1}$ and $\boldsymbol{\tau}_t$ for $t = 2, \ldots, T$ is

$$\begin{aligned} Corr(\tau_{t-1,k}, \tau_{t,k}) &= \frac{E\{\tau_{t-1,k}\tau_{t,k}\} - E\{\tau_{t-1,k}\}E\{\tau_{t,k}\}}{\sqrt{V\{\tau_{t-1,k}\}V\{\tau_{t,k}\}}} \\ &= (1-w_t)\sqrt{\frac{\sum_{l=1}^{t-1} w_l^2 \prod_{q=l+1}^{t-1}(1-w_q)^2}{\sum_{l=1}^{t} w_l^2 \prod_{q=l+1}^{t}(1-w_q)^2}}, \end{aligned} \tag{32}$$

for any $k \in I$. The proofs of these theorems are provided in Appendix A.

To compare the similarity of two adjacent tasks/documents, the two theorems yield insights through the mean and variance of the random deviation and the correlation coefficient which can be estimated from (32), using the posterior expectation of $\boldsymbol{w}$. Although dDTM represents a simplification of the dHDP framework [26], the sharing properties are similar. The proofs to both theorems are summarized in Appendix A.

XVII. Variational Bayes Inference

To motivate the theory of variational inference, we first recognize that the equality

$$\int_{\boldsymbol{O}} Q(\boldsymbol{O}) \ln \frac{Q(\boldsymbol{O})}{P(\boldsymbol{O}|\boldsymbol{X})P(\boldsymbol{X})} d\boldsymbol{O} = \int_{\boldsymbol{O}} Q(\boldsymbol{O}) \ln \frac{Q(\boldsymbol{O})}{P(\boldsymbol{X}|\boldsymbol{O})P(\boldsymbol{O})} d\boldsymbol{O}. \tag{33}$$

can be rewritten as

$$\ln P(\boldsymbol{X}) = \mathcal{L}(Q) + KL(Q||P), \tag{34}$$

where $\boldsymbol{O}$ represents the model latent parameters $\boldsymbol{O} = \{\{\boldsymbol{\Theta}_k\}_{k=1}^K, \boldsymbol{z}, \boldsymbol{d}, \{\boldsymbol{\pi}_t\}_{t=1}^T, \boldsymbol{w}\}$, $\boldsymbol{X}$ the observed data, $Q(\boldsymbol{O})$ some yet to be determined approximating density and

$$\mathcal{L}(Q) = \int_{\boldsymbol{O}} Q(\boldsymbol{O}) \ln \frac{P(\boldsymbol{X}|\boldsymbol{O})P(\boldsymbol{O})}{Q(\boldsymbol{O})} d\boldsymbol{O}, \quad KL(Q||P) = \int_{\boldsymbol{O}} Q(\boldsymbol{O}) \ln \frac{Q(\boldsymbol{O})}{P(\boldsymbol{O}|\boldsymbol{X})} d\boldsymbol{O}. \tag{35}$$

For inference purposes, instead of drawing $z_{t,i} \sim Mult(\tau_t)$, we use an extra variable $d_{t,i}$ indicating the task/document we are drawing the mixing weights $\tau_{d_{t,i}}$ from; for each document-dependent $\boldsymbol{x}_{t,i}$ we first draw the task indicator variable $d_{t,i}$ from a stick-breaking construction and then the corresponding topic indicator $z_{t,i}$ as follows:

$$z_{t,i} \sim Mult(\boldsymbol{\pi}_{d_{t,i}}), \quad d_{t,i} \sim Mult(\boldsymbol{V}),$$

$$V_q = w_q \prod_{l=1}^{q-1}(1 - w_l), \quad w_q \sim Beta(1, d_0), \tag{36}$$

where $Beta(1, d_0)$ corresponds to $Beta(c_0, d_0)$ in (23), with $c_0 = 1$.

Therefore, the joint distribution of the indicator variables $\boldsymbol{d}$ and $\boldsymbol{z}$ can be written as follows:

$$\begin{aligned} p(d_{t,i} = v, z_{t,i} = k | \boldsymbol{\pi}_t, \boldsymbol{\Theta}_k, \boldsymbol{x}_{t,i}) &\propto (\prod_{i=1}^{N_t} p(\boldsymbol{x}_{t,i} | \boldsymbol{\Theta}_{z_{t,i}=k}, \boldsymbol{\pi}_t)) p(z_{t,i} = k | d_{t,i}, \boldsymbol{\pi}_t, \boldsymbol{\Theta}_k) p(d_{t,i} = v) \\ &= (\prod_{i=1}^{N_t} \prod_{j=1}^{J} Mult(\boldsymbol{\theta}^j_{z_{t,i}=k}) \pi_{t,k} w_v \prod_{l=1}^{v-1}(1 - w_l), \end{aligned} \tag{37}$$

where N_t is the total number of documents in block t, and $x^j_{t,i}$ corresponds to word j in $\boldsymbol{x}_{t,i}$.

Our desire is to best approximate the true posterior $P(\{\boldsymbol{\Theta}_k\}_{k=1}^K, \boldsymbol{z}, \boldsymbol{d}, \{\boldsymbol{\pi}_t\}_{t=1}^T, \boldsymbol{w} | \boldsymbol{X})$ by minimizing $KL(Q||P)$, and this is accomplished by maximizing $\mathcal{L}(Q)$. In doing so, we assume that $Q(\boldsymbol{O})$ can be factorized, meaning

$$Q(\boldsymbol{O}) = Q(\{\boldsymbol{\Theta}_k\}_{k=1}^K, \boldsymbol{z}, \boldsymbol{d}, \{\boldsymbol{\pi}_t\}_{t=1}^T, \boldsymbol{w}) = Q(\{\boldsymbol{\Theta}_k\}_{k=1}^K) Q(\boldsymbol{z}) Q(\boldsymbol{d}) Q(\{\boldsymbol{\pi}_t\}_{t=1}^T) Q(\boldsymbol{w}). \tag{38}$$

A general method for writing inference for conjugate-exponential Bayesian networks, as outlined in [37], is as follows: for a given node in a graph, write out the posterior as though everything were known, take the natural algorithm, the expectation with respect to all unknown parameters and exponentiate the result. Since it requires computational resources comparable to the expectation-maximization (EM) algorithm, variational inference is fast relative to Markov chain Monte Carlo (MCMC) [26] methods (based on empirical studies for this particular application, and depending on what level of convergence MCMC is run to).

A. VB-E step

For the VB-E step, we calculate the variational expectation with respect to all unknown model parameters $\boldsymbol{\Theta}_k$, $\boldsymbol{\pi}_t$ and w_t. The variational equations of the model parameters $\boldsymbol{\Theta}_k$, $\boldsymbol{\pi}_t$ and w_t are shown below; their derivation is summarized in Appendix C. The analysis yields

$$\begin{aligned} \widetilde{\boldsymbol{\Theta}}_k &= exp[\sum_{j=1}^{J} \sum_{m=1}^{M} (\sum_{T_k} \langle x^j_{t,i} \rangle + \beta/M - 1) \ln(\theta^{j,m}_k)], \\ \widetilde{\boldsymbol{\pi}_t} &= exp[\sum_{i=1}^{N} \sum_{k=1}^{K} (\langle \ln \pi_{t,k} \rangle + \langle \ln w_t \rangle + \sum_{q=1}^{t-1} \langle \ln(1 - w_q) \rangle) + \sum_{k=1}^{K} (\alpha/K - 1) \langle \ln \pi_{t,k} \rangle], \\ \widetilde{w}_t &= exp[\sum_{i=1}^{N_t} \sum_{k=1}^{K} (\langle \ln w_t \rangle + \sum_{q=1}^{t-1} \langle \ln(1 - w_q) \rangle) + (d_0 - 1) \langle \ln(1 - w_t) \rangle], \end{aligned} \tag{39}$$

where

$$\begin{aligned} \langle \ln \pi_{t,k} \rangle &= \psi(\langle n_{t,k} \rangle + \alpha) - \psi(N_t + 1), \\ \langle x^j_{t,i} \rangle &= x^j_{t,i} p(x^j_{t,i} | z_{t,i} = k), \end{aligned}$$

$$\langle \ln w_v \rangle = \psi(1+N_v) - \psi(1+d_0+\sum_{l=v}^{T} N_l),$$

$$\langle \ln(1-w_l) \rangle = \psi(d_0+\sum_{m=l+1}^{T} N_m) - \psi(1+d_0+\sum_{m=l}^{T} N_m), \quad (40)$$

with $\psi(\cdot)$ the digamma function, $\langle n_{t,k} \rangle$ the number of words sharing topic k in block t, $\boldsymbol{\beta} = [\beta/M, ..., \beta/M]$ the Dirichet hyper-parameters for the priors on the words distribution, and $m \in \{1, 2., ..., M\}$ a possible outcome of the multinomial distributions on the word counts.

B. VB-M step

Updating the variational posteriors in the VB-M step is performed by updating the sufficient statistics of the model parameters, obtained from the VB-E step. The analysis yields

$$\begin{aligned} Q(\boldsymbol{\Theta}_k) &= \prod_{j=1}^{J} Dir(\beta/M + \langle p_{k,1} \rangle, ..., \beta/M + \langle p_{k,M} \rangle), \\ Q(\boldsymbol{\pi}_t) &= Dir(\alpha/K + \langle n_{t,1} \rangle, ..., \alpha/K + \langle n_{t,K} \rangle), \\ Q(w_t) &= Beta(1 + \langle m_t \rangle, d_0 + \sum_{b=1}^{t-1} \langle m_b \rangle), \end{aligned} \quad (41)$$

where $\langle m_t \rangle = \sum_{k=1}^{K} \langle n_{t,k} \rangle$ and $\langle p_{k,m} \rangle$ is the number of words with outcome m in topic k.

XVIII. Experimental Results

The proposed model is demonstrated on two data sets, each corresponding to a sequence of documents with known time dependence: (i) the NIPS data set [22] containing publications from the NIPS conferences between 1987 and 1999 and (ii) every United States presidential State of the Union Address from 1790-2008.

As comparisons to the dDTM model developed here, we consider LDA [20] and TOT [21], and dDTM with innovation weights set as $\{w_t\}_{t=2}^{T} = 1$ (termed DTM). For the dDTM framework, we initialized the hyper-parameters as follows: the parameter $\alpha = 1$, $c_0 = 1$, $d_0 = 2$, and Dirichlet distributions with uniform parameters $\boldsymbol{\beta} = [\frac{1}{M}, ..., \frac{1}{M}]$ as priors on the words distribution; the integer M defines the number of possible outcomes concerning the occurrence of a given word in a document, and this is detailed below for the particular examples. We ran VB until the relative change in the marginal likelihood bound [38] was less than 0.01%. For the LDA and TOT model initializations, we used exchangeable Dirichlet distributions as priors on word probabilities and initialized the Dirichlet hyper-parameters for the topic mixing weights with $\boldsymbol{\alpha} = [\frac{1}{K}, ..., \frac{1}{K}]$. The truncation level was set to $K = 50$ topics in all four models. For the reasons discussed in Section XV, the dDTM are expected to be insensitive to the setting of K, as long as it is "large enough"; we also performed studies of the below data for $K = 75$ and $K = 100$, with very similar results manifested.

A. NIPS Data Set

The NIPS (Neural Information Processing Systems) data set comprises $1,740$ publications. The total number of unique words was $J = 13,649$. The observation vector $\boldsymbol{x}_{t,i}$ corresponds to the frequency of all words in paper i of the NIPS proceedings from year t. We set the total number of outcomes of the multinomial distributions to $M = 5$; $m = 1$ corresponds to a word occurring zero times, $m = 2$ corresponds to a word occurring once or twice, $m = 3$ corresponds to a word occurring between three-five times, $m = 4$ corresponds to a word occurring between six-ten times, and $m = 5$ corresponds to a

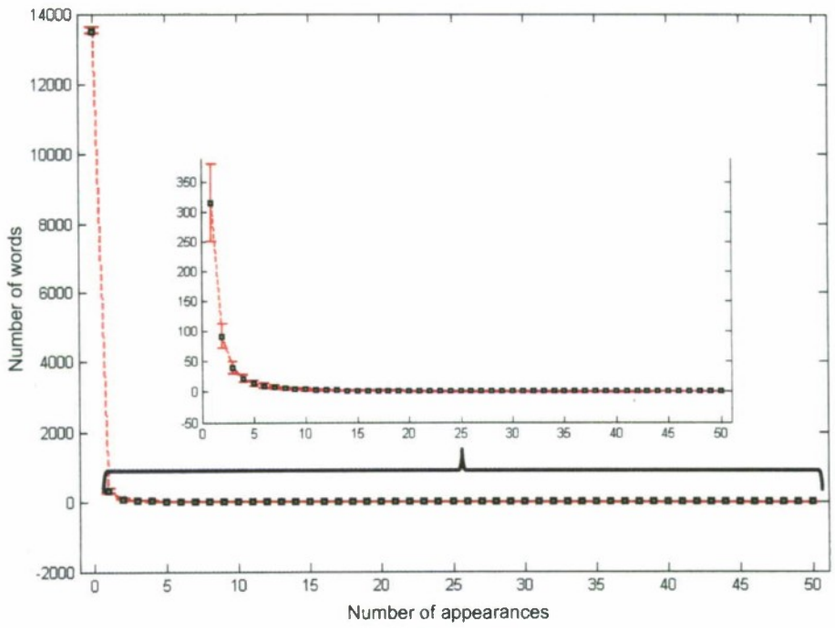


Fig. 7. Histogram of the rate of word appearances in the NIPS data set; the horizontal axis represents the number of times a given word appears in one document, and the vertical axis quantifies the number of times such words occurred across all documents. For example, in an average document, there will be 95 words that appear twice. From this we note that most words rarely occur more than five times in a given document.

word occurring more than ten times in a publication. This decomposition was defined based on examining a histogram of the rate with which any given word appeared in a given publication (see Fig. 7).

We first estimated the dDTM posterior distributions over the entire set of topics; the time evolution of the posterior dDTM probabilities for four representative topics and their ten most probable words, as computed via the posterior updates of words distributions within topics, are shown in Fig. 8; we ran the algorithm 20 times (with different randomly selected initializations) and chose the VB realization with the highest lower bound.

We then selected the years when the four topics represented above reached their highest probability of being drawn and identified associated publications; as we can see in Table I, for a given topic, there is a strong dependency between the most probable words and associated publications, with this proving to be a useful method of searching for papers based on a topic name or topic identifying words. These representative results are interpreted as follows: Topics A and C appear to be related to neural networks and speech processing, which appear to have a diminishing importance with time. By contrast, Topics B and D appear to be related to more statistical approaches, which have an increasing importance with time. The specific topic label is artificially given; it corresponds to one indicator variable in the VB solution.

In our next experiment, we quantitatively compared the dDTM, LDA, TOT and DTM models by computing the *perplexity* of a held-out test set; perplexity [20] is a popular measure used in language modeling, reflecting the difficulty of predicting new unseen documents after learning the model from a training data set. The perplexity results considered here are not the typical held-out at random type, but real prediction where we are using the past to build a model for the future; a lower perplexity score indicates better model performance. The perplexity for a test set of N_{test} documents is defined to be

$$P = exp\{-\frac{\sum_{i=1}^{N_{test}} \ln p(\boldsymbol{x}_{test,i})}{\sum_{i=1}^{N_{test}} n_{test,i}}\}, \tag{42}$$

where $\boldsymbol{x}_{test,i}$ represents the document i in the test set and $n_{test,i}$ is the number of words in document

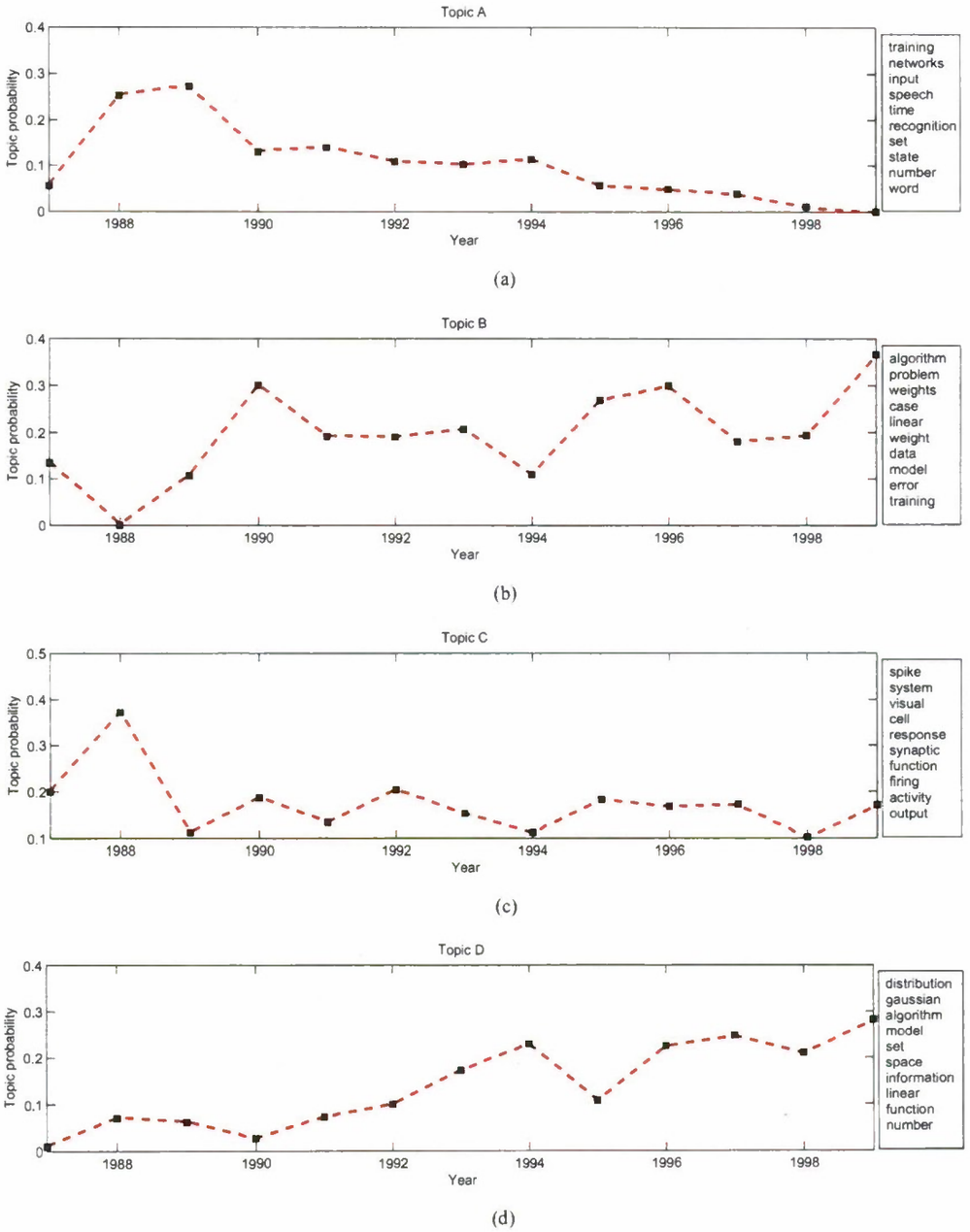


Fig. 8. Posterior topic probabilities distribution and most probable words for NIPS data set, as computed by the dDTM model.

TABLE I

REPRESENTATIVE TOPICS FROM THE NIPS DATABASE, WITH THEIR MOST PROBABLE WORDS AND ASSOCIATED PUBLICATIONS.

Topic A (year 1989)	training networks input speech time recognition set state number word	'A Continuous Speech Recognition System Embedding MLP into HMM' 'Training Stochastic Model Recognition Algorithms as Networks can Lead to Maximum Mutual Information Estimation of Parameters' 'Speaker Independent Speech Recognition with Neural Networks and Speech Knowledge' 'The Cocktail Party Problem: Speech/Data Signal Separation Comparison between Back propagation and SONN'
Topic B (year 1999)	algorithm problem weights case linear weight data model error training	'Model Selection for Support Vector Machines' 'Uniqueness of the SVM Solution' 'Differentiating Functions of the Jacobian with Respect to the Weights' 'Transductive Inference for Estimating Values of Functions'
Topic C (year 1988)	spike system visual cell response synaptic function firing activity output	'Models of Ocular Dominance Column Formation: Analytical and Computational Results' 'Modeling the Olfactory Bulbs Coupled Nonlinear Oscillators' 'A Model for Resolution Enhancement (Hyperacuity) in Sensory Representation' 'A Computationally Robust Anatomical Model for Retinal Directional Selectivity'
Topic D (year 1999)	distribution gaussian algorithm model set space information linear function number	'Local Probability Propagation for Factor Analysis' 'Algorithms for Independent Components Analysis and Higher Order Statistics' 'Correctness of Belief Propagation in Gaussian Graphical Models of Arbitrary Topology' 'Data Visualization and Feature Selection: New Algorithms for Nongaussian Data'

$\boldsymbol{x}_{test,i}$.

In our experiment the role of a document is played by a publication; the perplexity results correspond to a real prediction scenario, where we are using the past to build a model for the future: we held out all the publications from one year for test purposes and trained the models on all the publications from all the years prior to the testing year; as testing years we considered the last five years between 1995 and 1999.

The perplexity for the LDA and TOT models was computed as in [20]; for the dDTM and DTM models it was computed as follows:

$$P_{dDTM} = exp\{-\frac{\sum_{i=1}^{N_{test}} \ln(\sum_{j=1}^{J}\sum_{z_n}\sum_{t=1}^{T} p(x_{test,i}^{j}|z_{test,i},\boldsymbol{\Theta})p(z_{test,i}|\boldsymbol{\tau})p(\boldsymbol{\tau}_{d_{test,i}=t}|\boldsymbol{\alpha})\frac{d_0}{1+d_0}w_t\prod_{l>t}(1-w_l))}{\sum_{i=1}^{N_{test}} n_{test,i}}\},$$

$$P_{DTM} = exp\{-\frac{\sum_{i=1}^{N_{test}} \ln(\sum_{j=1}^{J}\sum_{z_n}\sum_{t=1}^{T} p(x_{test,i}^{j}|z_{test,i},\boldsymbol{\Theta})p(z_{test,i}|\boldsymbol{\tau})p(\boldsymbol{\tau}_{d_{test,i}=t}|\boldsymbol{\alpha})}{\sum_{i=1}^{N_{test}} n_{test,i}}\}, \quad (43)$$

where z is the topic indicator, i is the publication index, d is the block/year indicator, T is the total number of training years, and d_0 is the hyper-parameter of the beta prior distributions $Beta(1, d_0)$ on the innovating weights $\{w_t\}_{t=2}^{T}$. The perplexity computation for the dDTM model is provided in Appendix B.

Figure 9 shows the mean value and standard deviation of the perplexity of dDTM, LDA TOT and DTM models with $K = 50$ topics; we ran 20 VB realizations for the dDTM, LDA and DTM and 20 MCMC realizations (with 1000 iterations each) for the TOT model. We see that the dDTM model slightly outperforms the other models, with the LDA and TOT better than the DTM. The improved performance of dDTM model is due to the time evolving structure; the order of publications plays an important role in predicting new documents, through the innovation weight probability $\boldsymbol{w}$, as can be seen in (43).

While the NIPS database is widely used for topic modeling, the relatively small number of years it entails mitigates interesting analysis of the ability of dDTM to model the time-evolving properties of

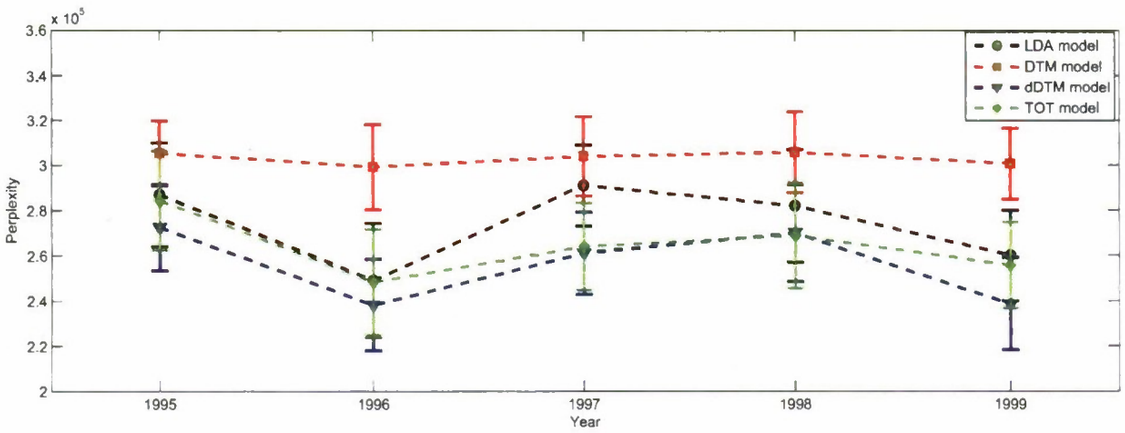


Fig. 9. Perplexity results on the NIPS data set for dDTM, LDA, TOT and DTM: mean value and standard deviation.

documents. This motivates the next example, which corresponds to a yearly database extending over 200 years.

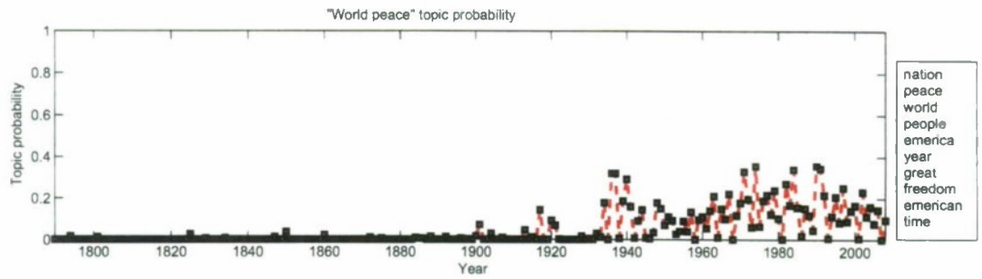
B. State of the Union Data Set

The State of the Union data set comprised 20,431 paragraphs, each with a time stamp from 1790 to 2008. The observation vector $x_{t,i}$ corresponds to the frequency of all words in paragraph i of the State of the Union from year t. In this (motivating) example, "document" i for year t corresponds to paragraph i from the State of the Union for year t. Therefore, the model assumes the State of the Union is represented by a mixture of topics, and within dDTM the mixture weights evolve with time.

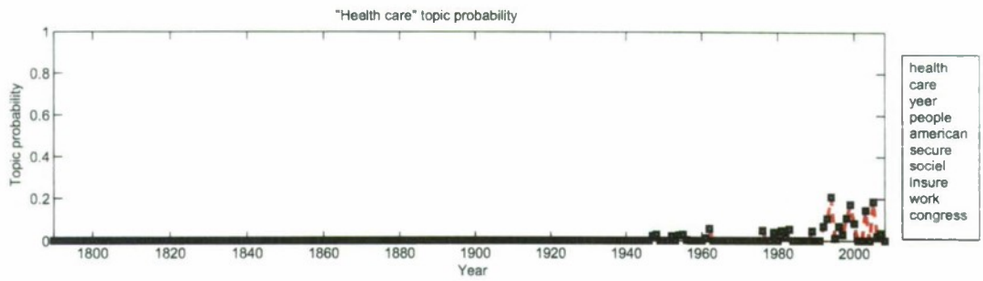
After removing common stop words by referencing a common list which can be found at $http://www.dcs.gla.ac.uk/idom/ir\_resources/linguistic\_utils/stop\_words$, and applying the Porter stemming algorithm [39], the total number of unique words was $J = 747$. In the rare years where two state of the union addresses were given, the address given by the outgoing president was used. Similar to the NIPS data set, each paragraph was represented as a datum, a vector of word counts over the dictionary. However, to match the data structure, we set the number of possible outcomes as $M = 2$, indicating whether a given word was present ($m = 1$) or not ($m = 2$) in a given paragraph. This structure corresponds to a binomial-beta representation of the words distribution, a special case of the multinomial-Dirichlet model used in the NIPS experiment.

For our first experiment we estimated the posterior distributions over the entire set of topics, for each of the three models mentioned above. Results for the dDTM model are shown in Fig. 10 for the time evolution of the posterior dDTM probabilities for five important topics in American history: 'American civil war', 'world peace', 'health care', 'U. S. Navy' and 'income tax'; similar to the NIPS experiment, we ran the algorithm 20 times (with random initialization) and chose the VB realization with the highest lower bound. The topic distributions preserve sharp peaks in time indicating significant information content at particular historical time points. It is important to mention that we have (artificially) named the topics based on their ten most probable words. The corresponding most probable words are shown in the right hand side of each plot. In comparison, the dDTM seems to perform better than LDA, TOT and DTM: 'American civil war' and 'health care' are topics that were not found by LDA, TOT or DTM. The better performance of the dDTM model can be explained by the sharing properties that exist between adjacent blocks, properties controlled by the innovation weight w. Figures 11, 12 and 13 show topic distributions and their associated ten most probable words for the LDA, TOT and DTM models, respectively.

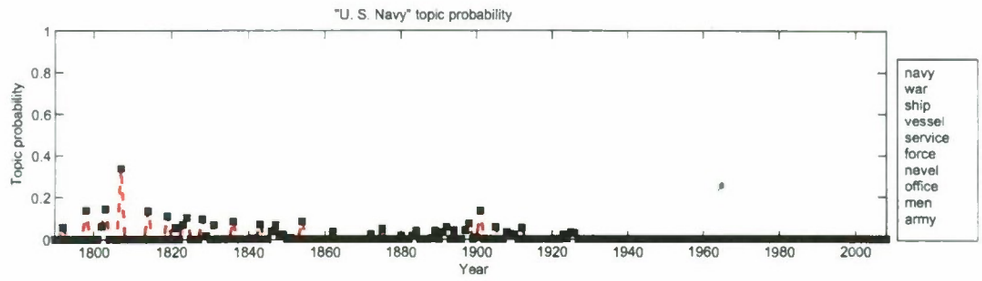
Concerning the interpretation of these results, we note that the US was not a world power until after World War II, consistent with Fig. 10(a). National health care in the US became a political issue in the



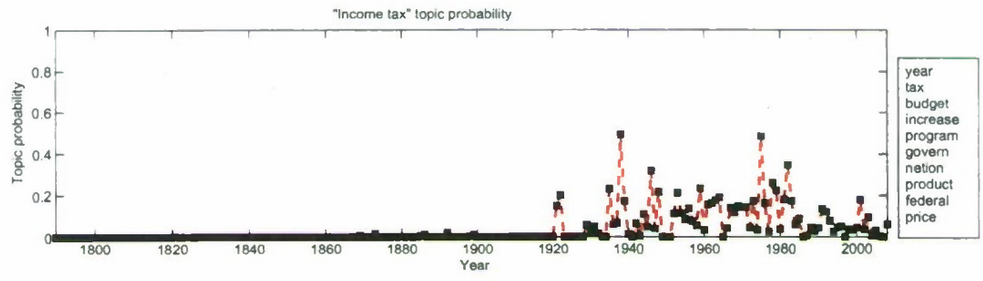
(a)



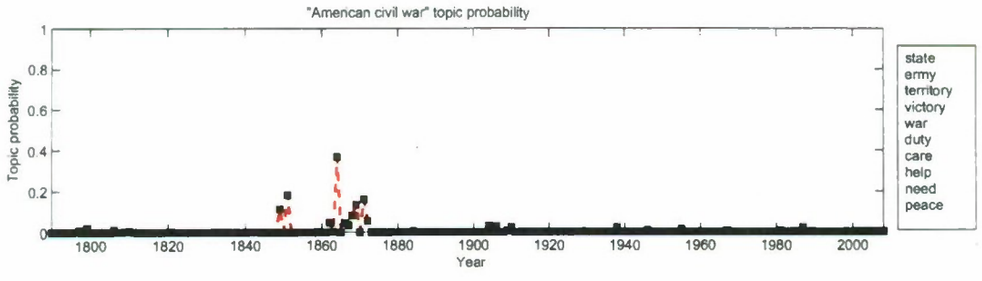
(b)



(c)



(d)



(e)

Fig. 10. dDTM model - topic probabilities distribution and most probable words for State of the Union data set. (a) World peace, (b) health care, (c) U.S. Navy, (d) income tax, (e) Civil War.

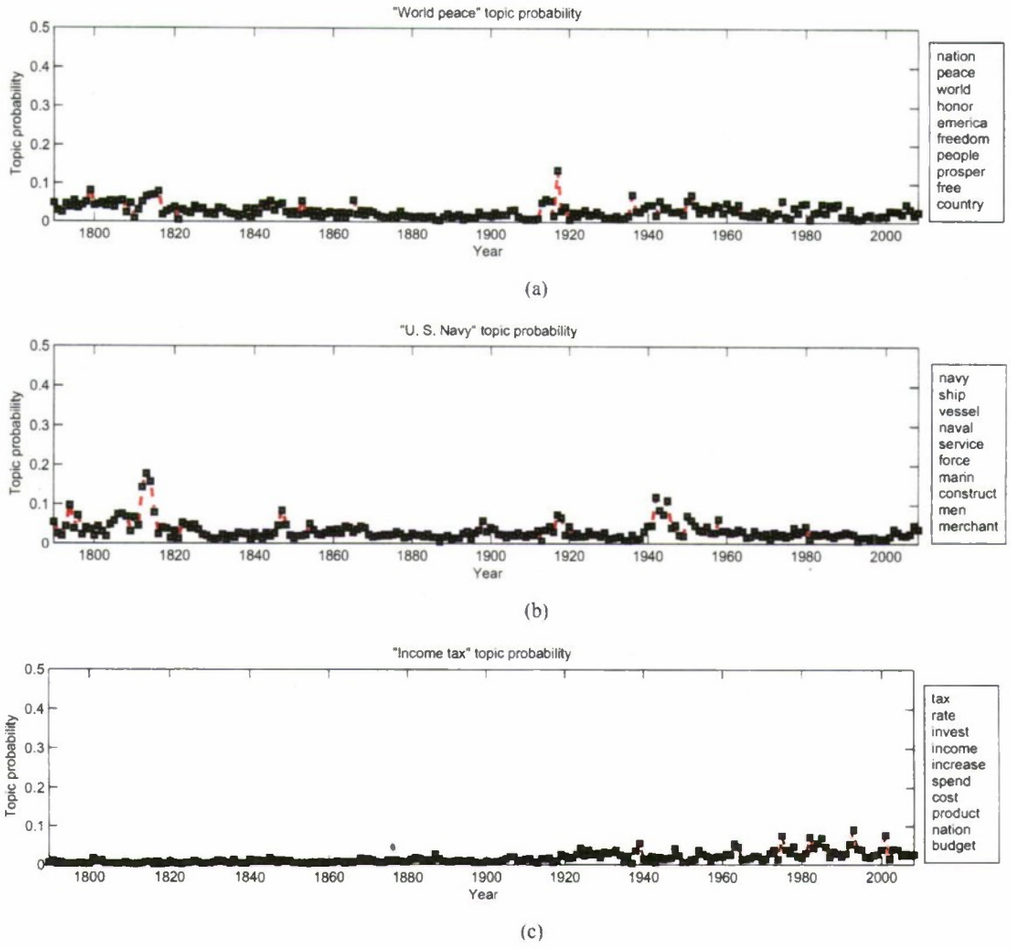


Fig. 11. LDA model - topic probabilities distribution and most probable words. (a) World peace, (b) U.S. Navy, (c) income tax.

early and mid 1990s, and continues such to this day. The US Navy was an important defense issue from the earliest days of the country, particularly in wars with Britain and Spain. With the advent of aircraft, the importance of the navy diminished, while still remaining important today. Concerning Fig. 10(d) on taxation, the first federal laws on federal (national) income tax were adopted by Congress in 1861 and 1862, and the Sixteenth Amendment to the US Constitution (1913) also addressed federal taxation. The heavy importance of this topic around 1920 is attributed to World War I, with this becoming an important issue/topic thereafter (concerning the appropriate tax rate). The US Civil War, which had a heavy focus on "state rights" was of course in the 1860-1865 period, with state rights being a topic of some focus sporadically thereafter.

Another advantage of dDTM over LDA, TOT and DTM is that it allows us to analyze the dynamic evolution of topic mixing weights through innovation probabilities. For that, using the dDTM model, we examined the innovation weight probability w, for each year from 1790 to 2008. Table II shows the years when the mean innovation probability was greater than 0.8, the year-period description and the name of the associated president. As observed during those years, important political events are well identified by dDTM. For each of the innovating years shown in Table II we also estimated the 'most innovative' words with respect to their previous year. For example, we were interested in finding the words that caused innovation during year 1829. For that, we first calculated the distribution of the words within one year, by integrating out the topics; we then estimated the Kullback-Leibler (KL) divergence between the

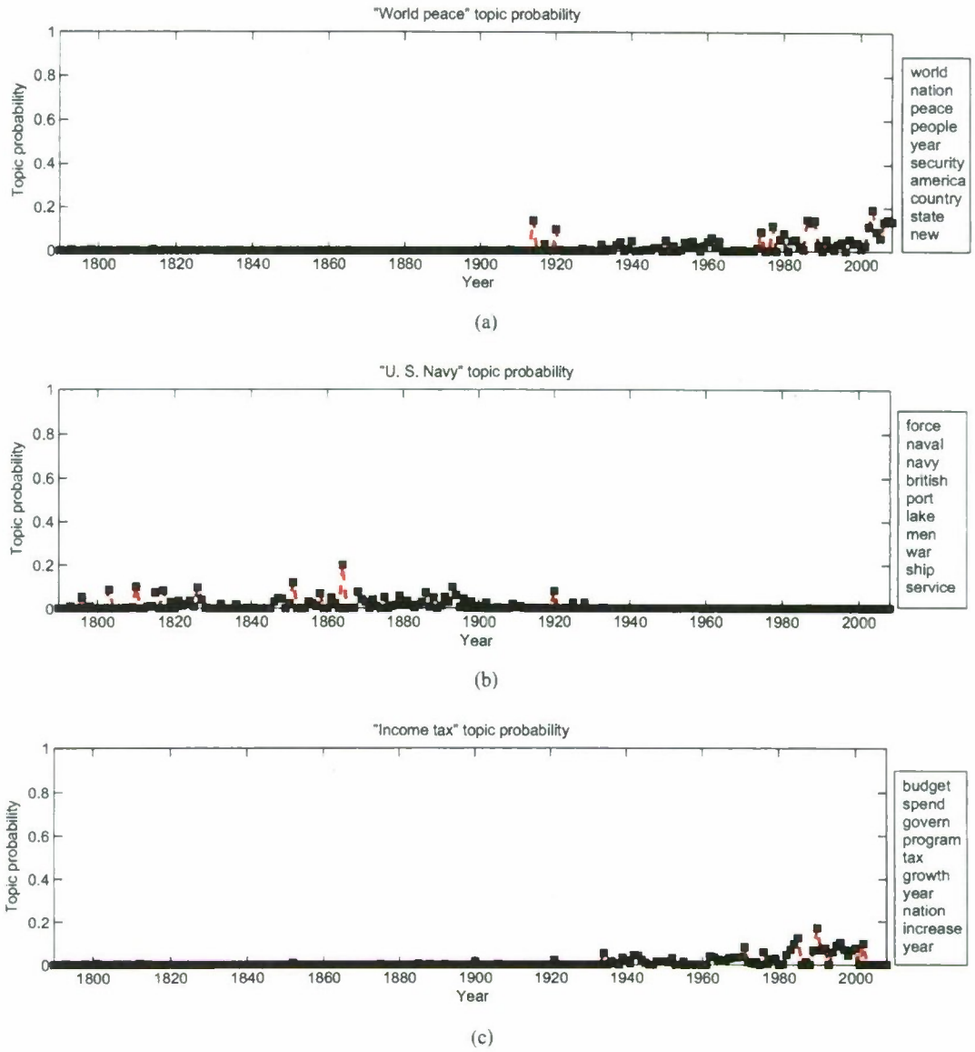


Fig. 12. TOT model - topic probabilities distribution and most probable words. (a) World peace, (b) U.S. Navy, (c) income tax.

probabilities of a given word belonging to two consecutive years, 1828 and 1829. The higher the KL distance is for a given word, the more innovation it produces; the ten most innovative words for each of the years of interest are shown in Table III.

The results in Table II ideally (if dDTM works properly) correspond to periods of significant change in the US. Concerning interpretation of these results, President Jackson was the first non-patrician US president, and he brought about significant change (*e.g.*, he ended the national banking system in the US). The Civil War, World War I, World War II, Vietnam and the end of the Cold War were all significant changes of "topics" within the US. Ronald Reagan also brought a level of conservative government to the US which was a significant change. These key periods, as inferred automatically via dDTM, seem to be in good agreement with historical events in the US.

We also analyzed the ability of dDTM to group paragraphs into topics. We chose two distinguishing years in American history, 1861 (during the American Civil War) and 2002 (post terrorist attacks) and show the most probable three topics as computed via the VB posterior updates and their associated paragraphs (see Tables IV and V). In 1861 the three major topics were 'political situation', 'finances' and 'army',

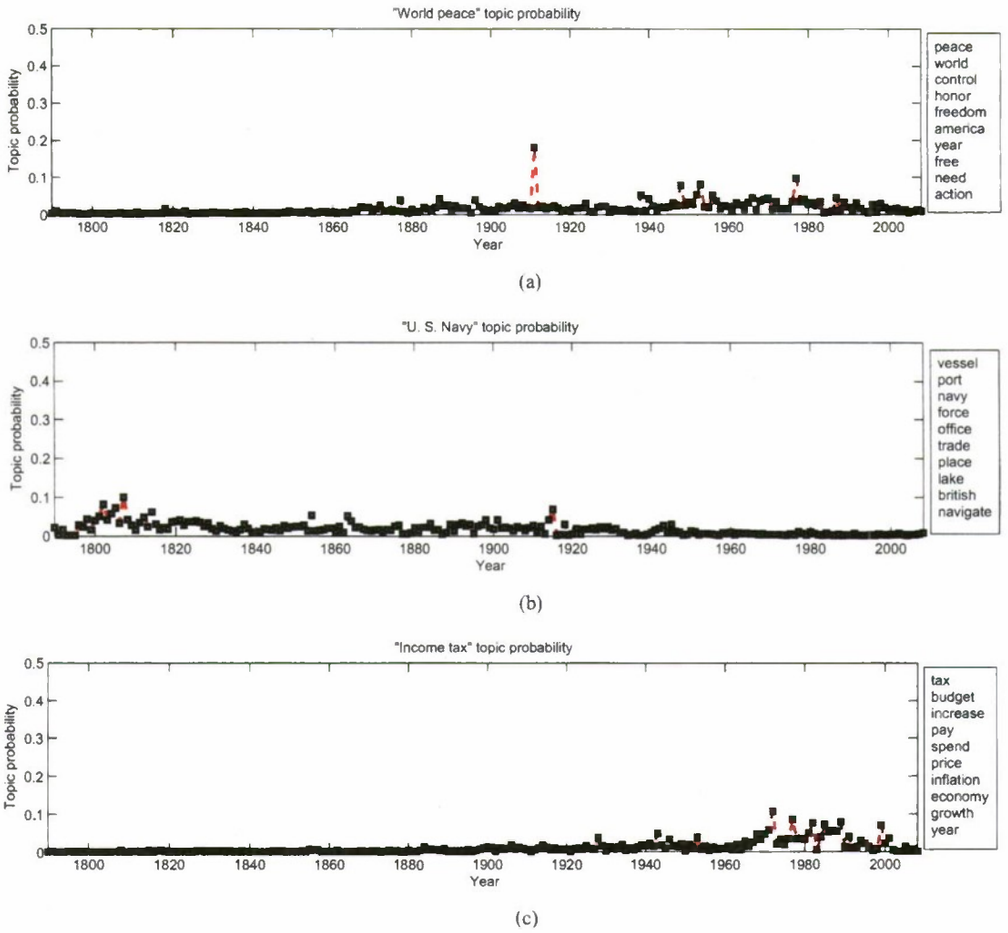


Fig. 13. DTM model - topic probabilities distribution and most probable words. (a) World peace, (b) U.S. Navy, (c) income tax.

TABLE II

YEARS WITH THE MEAN INNOVATION WEIGHT PROBABILITY GREATER THAN 0.8 IN THE DDTM MODEL, YEAR-PERIOD DESCRIPTION AND THE ASSOCIATED PRESIDENT.

Year	Mean innovation weight probability	Period description	President
1829	0.87	Pres. A. Jackson's era	A. Jackson
1831	0.84		
1861	0.82	Civil war	A. Lincoln
1909	0.81	Industrialization	W. H. Taft
1919	0.85	Post "world war I" era	W. Wilson
1938	0.84	Roosevelt's economical recovery	F. D. Roosevelt
1939	0.82	Second world war	F. D. Roosevelt
1965	0.8	Vietnam's war	L. B. Johnson
1981	0.81	R. Reagan's promised economic revival and the recession	R. Reagan
1982	0.89		
1990	0.82	The end of the "cold war"	G. H. W. Bush

TABLE III

MOST INNOVATIVE WORDS IN THE YEARS WITH THE MEAN INNOVATION WEIGHT PROBABILITY GREATER THAN 0.8

1829	1831	1861	1909	1919	1938	1939	1965	1981	1982	1990
Indian	Treaty	State	Unit	Legislation	Federal	Peace	War	Federal	Spend	World
Lew	Unit	Right	Treaty	Army	Tax	Freedom	World	Public	Budget	Peace
Tribe	Claim	War	British	Labor	Budget	America	Help	Program	Agriculture	War
Report	Convent	Increase	Report	Navy	Billion	War	Social	Budget	Senate	Free
Secretary	Prosper	Total	Negotiate	Peace	Deficit	Cut	Nation	Union	Let	Cut
Service	Nation	Power	Spain	District	Increase	God	Care	Increase	Know	Union
Work	Report	June	People	Federal	Fiscal	Spend	Million	Health	House	Strength
Constitute	Negotiate	Total	Subject	Ship	Income	Budget	Parent	Legislation	Tax	Cooper
Construct	Minister	Year	Session	Regulation	Rate	Percent	Peace	Tax	People	Rate
Navy	People	Service	Claim	Law	Spend	Army	Drug	America	Represent	Great

TABLE IV

PARAGRAPH CLUSTERING ANALYSIS FOR YEAR 1861: TOP THREE MOST PROBABLE TOPICS AND THEIR ASSOCIATED PARAGRAPHS.

Topic 40	Topic 41	Topic 22
Nations thus tempted to interfere are not always able to resist the counsels of seeming expediency and ungenerous ambition although measures adopted under such influences seldom fail to be unfortunete and injurious to those adopting them.	The revenue from all sources including loans for the financial year ending on the 10th of June was and the expenditures for the same period including payments on account of the public debt were leaving a balance in the Treasury on the 1st of July of	I respectfully refer to the report of the Secretary of War for information respecting the numerical strength of the Army and for recommendations having in view an increase of its efficiency and the wellbeing of the various branches of the service intrusted to his care.
It is not my purpose to review our discussions with foreign states because whatever might be their wishes or dispositions the integrity of our country and the stability of our Govemment mainly depend not upon them but on the loyalty virtue patriotism and intelligence of the people.	For the first quarter of the financial year ending on the 30th of September the receipts from all sources including the balance of the 1st of July were and the expenses leaving a balance on the 1st of October of	The large eddition to the Regular Army in connection with the defection that has so considerably diminished the number of its officers gives peculiar importance to his recommendation for increasing the corps of cadets to the greatest capecity of the Military Academy.
Some treaties designed chiefly for the interests of commerce and having no grave political importance have been negotiated end will be submitted to the Senate for their consideration.	The revenue from all sources during the fiscal year ending June including the annual permanent eppropriation of for the transportation of free mail matter was being ebout 12 per cent less than the revenue for	It is gratifying to know that the patriotism of the people has proved equal to the occasion end thet the number of troops tendered greatly exceeds the force which Congress authorized me to call into the field.

whereas in 2002 the topics were 'terrorism', 'national budget' and 'overall progress of the country'. In both cases, the algorithm automatically clusters the paragraphs using what appears to be an accurate topic representation.

To show the dynamic structure of dDTM, we selected 2002 as a reference year and its two years before and after as years where topic transition could be manifested. For each of the five years, we estimated the most probable topic and identified its associated paragraphs. As we can see in Table VI, a topic transition is manifested during this time interval: if in 2000, the main topic was 'economy', in the following years attention is paid to 'education', 'terrorism', 'economy' again and 'war in Iraq', respectively. The terrorist attacks on the World Trade Center and on the Pentagon occurred in 2001, manifesting the clear change in the important "topics".

Finally, we again compared dDTM, LDA, TOT and DTM models by computing their perplexities; in this case, the role of a document was represented by a paragraph and, similar to the NIPS experiment, we considered the task of real prediction, by holding out all the paragraphs from one year for test purposes and training the models on all the paragraphs from all the years prior to the testing year; as testing years we considered the ending year of each decade from 1901 to 2000.

Figure 14 shows the mean perplexity of dDTM, LDA, TOT and DTM models with $K = 50$ topics and 10 testing years. We ran 20 VB realizations for the dDTM, LDA and DTM and 20 MCMC realizations (with 1000 iterations each) for the TOT model; the standard deviation values are included as well. We see that the dDTM model consistently performs better than the other models. We also observe that LDA

TABLE V

PARAGRAPH CLUSTERING ANALYSIS FOR YEAR 2002: TOP THREE MOST PROBABLE TOPICS AND THEIR ASSOCIATED PARAGRAPHS.

Topic 19	Topic 2	Topic 39
America has a window of opportunity to extend and secure our present peace by promoting a distinctly American internationalism. We will work with our allies and friends to be e force for good and a champion of freedom. We will work for free markets free trade.	Government cannot be replaced by charities or volunteers. Government should not fund religious activities. But our Nation should support the good works of these good people who are helping their neighbors in need. So I propose allowing all taxpayers whether they itemize or not to deduct their charitable contributions. Estimates show this could encourage as much as one billion a year in new charitable giving money that will save and change lives.	Together we are changing the tone in the Nation's Capital. And this spirit of respect and cooperation is vital because in the end we will be judged not only by what we say or how we say it we will be judged by what were able to accomplish.
Our Nation also needs a clear strategy to confront the threats of this century threats that are more widespread and less certain. They range from terrorists who threaten with bombs to tyrants in rogue nations intent upon developing weapons of mass destruction .To protect our own people our allies and friends we must develop and we must deploy effective missile defenses.	I propose we make a major investment in conservation by fully funding the Land and Water Conservation Fund and our national parks. As good stewards we must leave them better than we found them. So I propose to provide one billion over ten years for the upkeep of these national treasures.	The last time I visited the Capitol I came to take an oath on the steps of this building. I pledged to honor our Constitution and laws and I esked you to join me in setting a tone of civility and respect in Washington. I hope America is noticing the difference because we're making progress.
Yet the cause of freedom rests on more than our ability to defend ourselves and our allies. Freedom is exported every day as we ship goods and products that improve the lives of millions of people. Free trade brings greater political and personal freedom. Each of the previous five Presidents has had the ability to negotiate far reaching trade agreements.	The budget adopts a hopeful new approach to help the poor and the disadvantaged. We must encourage and support the work of charities and faith based and community groups that offer help and love one person at a time. These groups ere working in every neighborhood in America to fight homelessness and addiction and domestic violence to provide a hot meal or a mentor or a safe haven for our children. Govemment should welcome these groups to apply for funds not discriminate against them.	Neither picture is complete in and of itself. Tonight I challenge and invite Congress to work with me to use the resources of one picture to repaint the other to direct the advantages of our time to solve the problems of our people. Some of these resources will come from Government, some but not all.

TABLE VI

DYNAMIC STRUCTURE ANALYSIS FOR YEARS 2000-2004: MOST PROBABLE TOPIC AND ASSOCIATED PARAGRAPHS.

Year 2000 (topic 37)	Year 2001 (topic 12)	Year 2002 (topic 19)	Year 2003 (topic 37)	Year 2004 (topic 34)
We begin the new century with over one million new jobs; the fastest economic growth in more than ten years; the lowest unemployment rates in years; the lowest poverty rates in years; the lowest African American end Hispanic unemployment rates on record. America will echieve the longest period of economic growth in our entire history. We have built e new economy.	A budget's impact is counted in dollers but measured in lives. Excellent schools quality health care e secure retirement e cleaner environment e stronger defense, these ere ell important needs end we fund them. The highest percentage increese in our budget should go to our children's education. Education is my top priority end by supporting this budget you'll meke it yours es well.	Americe hes e window of opportunity to extend end secure our present peace by promoting e distinctly Americen Internationalism. We will work with our ellies and friends to be e force for good end e champion of freedom. We will work for free merkets free trede.	To lift the stenderds of our public schools we achieved historic education reform which must now be carried out in every school end in every classroom so thet every child in America can read end leern end succeed in life. To protect our country we reorgenized our Government end creeted the Department of Homelend Security which is mobilizing egeinst the threets of e new ere. To bring our economy out of recession we delivered the lergest tax relief in e generetion.	We have faced serious chellenges together end now we face e choice. We can go forwerd with confidence end resolve or we can turn back to the dangerous illusion thet terrorists ere not plotting end outlew regimes ere no threet to us. We can press on with economic growth end reforms in education end Medicare or we can turn back to old policies end old divisions.
Our economic revolution hes been metched by e revival of the American spirit crime down by percent to its lowest level in years teen births down yeers in e row edoptions up by percent welfere rolls cut in half to their lowest levels in yeers.	When it comes to our schools dollers elone do not elweys meke the difference. Funding is important end so is reform. So we must tie funding to higher standards end accountability for results.	Our Netion elso needs e cleer strategy to confront the threets of this century threets thet ere more widespreed end less certein. They renge from terrorists who threeten with bombs to tyrents in rogue netions intent upon developing weepons of mess destruction. To protect our own people our ellies end friends we must develop and we must deploy effective missile defenses.	Our first goal is cleer we must heve en economy that grows fast enough to employ every men end woman who seeks e job. After recession terrorist ettacks corporate scandels and stock merket declines our economy is recovering. Yet its not growing fast enough or strongly enough. With unemployment rising our Netion needs more small businesses to open more companies to invest end expand more employers to put up the sign thet says Help Wented.	Heving broken the Baathist regime we face e remnant of violent Saddem supporters. Men who ren eway from our troops in bettle ere now dispersed end etteck from the shedows. These killers joined by foreign terrorists ere e serious continuing denger. Yet we're meking progress egeinst them. The once ell powerful ruler of Ireq wes found in e hole end now sits in e prison cell. The top officiels of the former regime we heve captured or killed. Our forces ere on the offensive leading over patrols e dey end conducting en everege of reids e week.
Eight years ago it was not so cleer to most Americans there would be much to celebrate in the yeer. Then our Netion was gripped by economic distress social decline political gridlock. The title of e bestselling book esked America What Went Wrong.	Schools will be given e reesoneble chance to improve end the support to do so. Yet if they don't if they continue to fail we must give perents end students different options e better public school e private school tutoring or e charter school. In the end every child in a bad situetion must be given e better choice because when it comes to our children failure is simply not en option.	Yet the cause of freedom rests on more than our ebility to defend ourselves end our ellies. Freedom is exported every day es we ship goods end products thet improve the lives of millions of people. Free trade brings greeter political end personel freedom. Each of the previous five Presidents has had the ebility to negotiete fer reaching trade agreements.	A growing economy end e focus on essentiel priorities will be cruciel to the future of Social Security. As we continue to work together to keep Social Security sound end relieble we must offer younger workers e chence to invest in retirement eccounts that they will control end they will own.	As democracy takes hold in Ireq the enemies of freedom will do all in their power to spreed violence end fear. They ere trying to sheke the will of our country end our friends but the United States of Americe will never be intimideted by thugs and essassins. The killers will fail end the Iraqi people will live in freedom.

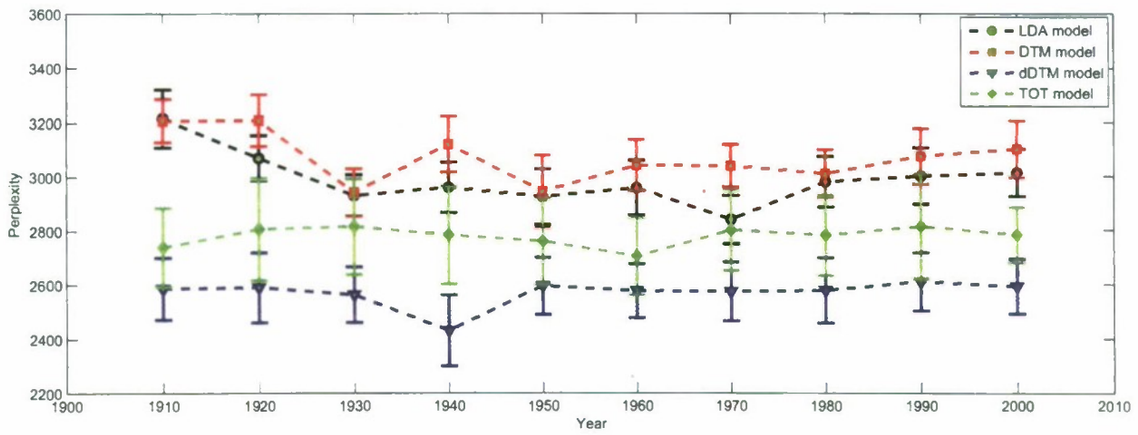


Fig. 14. Perplexity results on the United States presidential State of the Union Address for dDTM, LDA, TOT and DTM: mean value and standard deviation, estimated from 20 randomly initialized VB realizations.

and TOT slightly outperform the DTM model due to the Dirichlet distribution approximations made in the DTM model.

Concerning computational costs, all code was run in Matlab^{TM} on a PC with Intel 2.33GHz processor. For the NIPS data dDTM, LDA, TOT and DTM required (for each VB and MCMC runs) 4 hours and 16 minutes, 3 hours and 22 minutes, 10 hours and 31 minutes, and 3 hours and 45 minutes, respectively. For the State of the Union data these respective times were 25, 22, 104 and 23 minutes. These times are meant to give relative computational costs; none of the software was optimized.

XIX. Topic Modeling Summary

We have developed a novel topic model, the truncated dynamic HDP, or dDTM, to analyze topics associated with documents with known time stamps. The new model allows simple variational Bayesian (VB) inference, yielding fast computation times. The algorithm has been demonstrated on a large database, the US State of the Unions for a 220 year period, and the results seem to be able to highlight significant events in the US history (although it should be emphasized that the authors are not historians, and much further testing and evaluation is required). The algorithm is able to identify important historical topics, as well as periods of time over which significant changes in topics are realized. The model compares favorably with LDA, TOT and a simplified form of dDTM (for which time dependence is ignored).

Concerning future research, other approaches that might be considered for approximate inferences include collapsed sampling [40]. It would be interesting to analyze how these different inferences influence the overall performance of the model. In order to capture semantics evolution with time, one may consider a similar dynamic model for topics themselves. This could be accomplished by allowing the words distributions change in time; for identifiability, constraints could be used so that the majority of words in a topic, and their associated frequencies, remain constant across time. In addition, the evolution of the model occurred in only one dimension (time). There may be problems for which documents may be collected at different geographical locations, for example from different cities across the world. In this case one may have spatial proximity as well as temporal proximity to consider, when considering inter-document relationships. It is of interest to extend the dynamic structure from one dimension to perhaps a graphical structure, where the nodes of the graph may represent space and time.

We also note there may be general interest within topic-model research in representing a draw from a Dirichlet process in the form in (27). While this increases the complexity of the analysis, it has the significant advantage of allowing one to place a Gamma prior on α and perform full VB inference (we no longer have to set α). As discussed in Section XV, α plays an important role in defining the number of expected topics per document (since it controls the number of important mixture weights). One may

place a separate prior on the distinct α associated with each document, so that the number of important topics per document may change. The complication with doing this, rather than just directly drawing from $Dir(\alpha/K, \cdots, \alpha/K)$ is that one must now perform inference on many more parameters (on the sticks of the stick-breaking representation). In some applications such added complexity will be warranted by a desire to infer α in a full VB analysis.

REFERENCES

[1] M. Kearns and D. Koller, "Efficient reinforcement learning in factored mdps," in *Proc. of the Sixteenth International Joint Conference of Artificial Intelligence*, 1999, pp. 740–747.

[2] M. Kearns and S. P. Singh, "Near-optimal performance for reinforcement learning in polynomial time," in *Proc. ICML*, 1998, pp. 260–268.

[3] R. I. Brafman and M. Tennenholtz, "R-max - a general polynomial time algorithm for near-optimal reinforcement learning," *Journal of Machine Learning Research*, vol. 3, no. OCT, pp. 213–231, 2002.

[4] P. Poupart and N. Vlassis, "Model-based bayesian reinforcement learning in partially observable domains," in *International Symposiu on Artificial Intelligence and Mathmatics (ISAIM)*, 2008.

[5] F. Doshi, J. Pineau, and N. Roy, "Reinforcement learning with limited reinforcement: using bayes risk for active learning in pomdps," in *ICML*, 2008.

[6] H. Li, X. Liao, and L. Carin, "Multi-task reinforcement learning in partially observable stochastic environments," *Journal of Machine Learning Research*, vol. 10, pp. 1131–1186, 2009.

[7] R. Sutton and A. Barto, *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press, 1998.

[8] M. J. Beal, "Variational algorithms for approximate bayesian inference," *Gatsby Computational Neuroscience Unit, Ph.D. thesis, University College London*, 2003.

[9] M. Littman, A. Cassandra, and L. Kaelbling, "Learning policies for partially observable environments: scaling up," in *ICML*, 1995.

[10] J. Pineau, G. Gordon, and S. Thrun, "Point-based value iteration: An anytime algorithm for POMDPs," in *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI)*, August 2003, pp. 1025 – 1032.

[11] E. J. Sondik, "The optimal control of partially observable markov processes over the infinite horizon: Discounted costs," *Operations Research*, vol. 26, no. 2, pp. 282–304, 1978.

[12] P. Poupart and N. Vlassis, "Model-based Bayesian reinforcement learning in partially observable domains," in *International Symposium on Artificial Intelligence and Mathematics (ISAIM)*, 2008.

[13] H. Li, X. Liao, and L. Carin, "Multi-task reinforcement learning in partially observable stochastic environments," *Journal of Machine Learning Research*, vol. 10, pp. 1131–1186, 2009.

[14] T. Ferguson, "A Bayesian analysis of some nonparametric problems," *Annals of Statistics*, vol. 1, pp. 209–230, 1973.

[15] D. B. Dunson, "Nonparametric bayes local partition models for random effects," *Biometrika*, vol. 96, no. 2, pp. 249–262, 2009.

[16] J. Sethuraman, "A constructive definition of Dirichlet priors," *Statistica Sinica*, vol. 4, pp. 639–650, 1994.

[17] S. G. Walker, "Sampling the dirichlet mixture model with slices," *Communications in Statistics - Simulation and Computation*, vol. 36, no. 1, pp. 45–54, 2007.

[18] C. E. Antoniak, "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems," *The Annals of Statistics*, vol. 2, no. 6, pp. 1152–1174, November 1974.

[19] T. Hofmann, "Probabilistic latent semantic analysis," *Proceedings of Uncertainty in Artificial Intelligence*, 1999.

[20] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[21] X. Wang and A. McCallum, "Topics over time: a non-markov continuous-time model of topical trends," *The Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 424–433, 2006.

[22] A. Gruber, M. Rosen-Zvi, and Y. Weiss, "Hidden topic markov models," *Artificial Intelligence and Statistics*, 2007.

[23] J. Zhang, Z. Ghahramani, and Y. Yang, "A probabilistic model for online document clustering with application to novelty detection," *Proceedings of Neural Information Processing Systems*, 2004.

[24] J. F. Canny and T. L. Rattenbury, "A dynamic topic model for document segmentation," *Technical Report, Department of Electrical Engineering and Computer Sciences, University of California at Berkeley*, 2006.

[25] M. L. Pennell and D. B. Dunson, "Bayesian semiparametric dynamic frailty models for multiple event time data," *Biometrics*, vol. 6, pp. 1044–1052, 2006.

[26] L. Ren, D. B. Dunson, and L. Carin, "The dynamic hierarchical dirichlet process," *International Conference on Machine Learning*, 2008.

[27] N. Srebro and S. Roweis, "Time-varying topic models using dependent dirichlet processes," *Technical Report, Department of Computer Science, University of Toronto*, 2005.

[28] D. M. Blei and J. D. Lafferty, "Dynamic topic models," *Proceedings of the 23rd International Conference on Machine Learning*, pp. 113–120, 2006.

[29] J. Sethuraman, "A constructive definition of dirichlet priors," *Statistica Sinica*, vol. 4, pp. 639–650, 1994.

[30] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical dirichlet processes," *Journal of the American Statistical Association*, vol. 101, pp. 1566–1582, 2005.

[31] D. B. Dunson and J.-H. Park, "Kernel stick-breaking processes," *Biometrika*, 2007.

[32] Q. An, E. Wang, I. Shterev, L. Carin., and D. B. Dunson, "Hierarchical kernel stick-breaking process for multi-task image analysis," *Proceedings of the 25th International Conference on Machine Learning*, 2008.

[33] J.-H. Park and D. B. Dunson, "Bayesian generalized product partition model," 2006.

[34] D. M. Blei and J. D. Lafferty, "A correlated topic model of science," *The Annals of Applied Statistics*, vol. 1, no. 1, pp. 17–35, 2007.
[35] D. B. Dunson, "Bayesian dynamic modeling of latent trait distributions," *Biostatistics*, vol. 7, pp. 551–568, 2006.
[36] J. Ishwaran and L. James, "Gibbs sampling methods for stick-breaking priors," *Journal of the American Statistical Association*, vol. 96, p. 161174, 2001.
[37] J. Winn and C. M. Bishop, "Variational message passing," *Journal of Machine Learning Research*, vol. 6, pp. 661–694, 2005.
[38] D. M. Blei and M. I. Jordan, "Variational methods for the dirichlet process," *Proceedings of the 21st International Conference on Machine Learning*, 2004.
[39] C. J. van Rijsbergen, S. E. Robertson, and M. F. Porter, "Information retrieval," *Butterworths, London, 2nd edition*, vol. 6, pp. 111–143, 1979.
[40] M. Welling, I. Porteous, and E. Bart, "Infinite state bayes-nets for structured domains," *Proceedings of the International Conference on Neural Information Processing Systems*, 2007.