A Bayesian approach for modeling non-stationary data streams

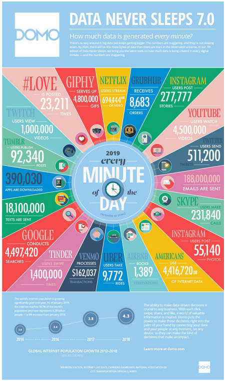
Andrés Masegosa

Department of Mathematics
University of Almería
Spain

Part of this work jointly made with with Thomas D. Nielsen (AAU), Helge Langseth (NTNU) and Antonio Salmeron (UAL).

Introduction

(Ditzler et al. 2015)



Data Streams

- Most of the generated data is in the form of data stream.
- Information processed by the brain is a data stream.
- Data streams usually are **non-stationary**.

See what was trending in 2017 - Global

Searches		People		Global News	
1	Hurricane Irma	1	Matt Lauer	1	Hurricane Irma
2	iPhone 8	2	Meghan Markle	2	Bitcoin
3	iPhone X	3	Nadia Toffa	3	Las Vegas Shooting
4	Matt Lauer	4	Harvey Weinstein	4	North Korea
5	Meghan Markle	5	Kevin Spacey	5	Solar Eclipse

See what was trending in 2018 - Global

Searches		News		People	
1	World Cup	1	World Cup	1	Meghan Markle
2	Avicii	2	Hurricane Florence	2	Demi Lovato
3	Mac Miller	3	Mega Millions Result	3	Sylvester Stallone
4	Stan Lee	4	Royal Wedding	4	Logan Paul
5	Black Panther	5	Election Results	5	Khloé Kardashian

Definition of a Non-stationary Data Stream

- We have a collection of time-indexed samples.

$$\{\mathbf{x}_1, \dots, \mathbf{x}_t\}$$

(Ditzler et al. 2015)

Definition of a Non-stationary Data Stream

- We have a collection of time-indexed samples.

$$\{\mathbf{x}_1, \ldots, \mathbf{x}_t\}$$

- Each $\mathbf{x}_t$ is usually composed by a bunch of data samples.

(Ditzler et al. 2015)

Definition of a Non-stationary Data Stream

- We have a collection of time-indexed samples.

$$\{\mathbf{x}_1, \ldots, \mathbf{x}_t\}$$

- Each $\mathbf{x}_t$ is usually composed by a bunch of data samples.

- The *data generating distribution* $\pi_t(\mathbf{x})$ changes from one time step to another,

$$\mathbf{x}_t \sim \pi_t(\mathbf{x})$$
$$\pi_t(\mathbf{x}) \neq \pi_{t+1}(\mathbf{x})$$
$$KL(\pi_t(\mathbf{x})||\pi_{t+1}(\mathbf{x})) \leq \epsilon$$

(Ditzler et al. 2015)

Definition of a Non-stationary Data Stream

- We have a collection of time-indexed samples.

$$\{\mathbf{x}_1, \ldots, \mathbf{x}_t\}$$

- Each $\mathbf{x}_t$ is usually composed by a bunch of data samples.

- The *data generating distribution* $\pi_t(\mathbf{x})$ changes from one time step to another,

$$\mathbf{x}_t \sim \pi_t(\mathbf{x})$$
$$\pi_t(\mathbf{x}) \neq \pi_{t+1}(\mathbf{x})$$
$$KL(\pi_t(\mathbf{x})||\pi_{t+1}(\mathbf{x})) \leq \epsilon$$

- We do not have **i.i.d.** data.

(Ditzler et al. 2015)

Learning from a non-stationary data stream

- **Problem I**: How to handle an **endless** data set.

(Sugiyama et al. 2007)

Learning from a non-stationary data stream

- **Problem I**: How to handle an **endless** data set.
- **Problem II**: **Training Distribution $\neq$ Test Distribution**.
  - Minimize the empirical loss,

$$\arg\min_{\theta} \mathbb{E}_{\hat{\pi}}[\ell(h(\mathbf{x}, \theta), \mathbf{y})]$$

  - ... but my goal is to minimize,

$$\arg\min_{\theta} \mathbb{E}_{\pi_T}[\ell(h(\mathbf{x}, \theta), \mathbf{y})]$$

  - And $\pi_T \neq \hat{\pi}$ ($\hat{\pi}$ is the empirical distribution of the training data).

(Sugiyama et al. 2007)

Bayesian modeling of Non-stationary Data Streams

- We assume we have a model for the data.

$$\pi(\mathbf{x}) \quad \approx \quad p(\mathbf{x}|\theta)$$

- We assume we have a model for the data.

$$\pi(\mathbf{x}) \quad \approx \quad p(\mathbf{x}|\theta)$$

- ... and a **prior distribution**,

$$\theta \quad \sim \quad p(\theta)$$

- We assume we have a model for the data.

$$\pi(\mathbf{x}) \quad \approx \quad p(\mathbf{x}|\theta)$$

- ... and a **prior distribution**,

$$\theta \quad \sim \quad p(\theta)$$

- **Bayesian recursive updating** naturally deals with data stream,

$$p(\theta|\mathbf{x}_{1:t}) = \frac{1}{Z} p(\mathbf{x}_t|\theta) p(\theta|\mathbf{x}_{1:t-1})$$

- We assume we have a model for the data.

$$\pi_t(\mathbf{x}) \approx p(\mathbf{x}|\theta_t)$$

- We assume we have a model for the data.

$$\pi_t(\mathbf{x}) \approx p(\mathbf{x}|\theta_t)$$

- ... and a **parameter transition distribution**,

$$\begin{aligned} \theta_1 &\sim p(\theta) \\ \theta_{t+1} &\sim p(\theta|\theta_t) \end{aligned}$$

- We assume we have a model for the data.

$$\pi_t(\mathbf{x}) \approx p(\mathbf{x}|\theta_t)$$

- ... and a **parameter transition distribution**,

$$\begin{aligned} \theta_1 &\sim p(\theta) \\ \theta_{t+1} &\sim p(\theta|\theta_t) \end{aligned}$$

- **Bayesian recursive updating** naturally deals with data stream,

$$p(\theta_t|\mathbf{x}_{1:t}) = \frac{1}{Z} p(\mathbf{x}_t|\theta_t) \int p(\theta_t|\theta_{t-1}) p(\theta_{t-1}|\mathbf{x}_{1:t-1}) d\theta_{t-1}$$

- We assume we have a model for the data.

$$\pi_t(\mathbf{x}) \approx p(\mathbf{x}|\theta_t)$$

- ... and a **parameter transition distribution**,

$$\begin{aligned} \theta_1 &\sim p(\theta) \\ \theta_{t+1} &\sim p(\theta|\theta_t) \end{aligned}$$

- **Bayesian recursive updating** naturally deals with data stream,

$$p(\theta_t|\mathbf{x}_{1:t}) = \frac{1}{Z} p(\mathbf{x}_t|\theta_t) \int p(\theta_t|\theta_{t-1}) p(\theta_{t-1}|\mathbf{x}_{1:t-1}) d\theta_{t-1}$$

- Standard Bayesian updating is special case when

$$p(\theta|\theta_t) = \delta(\theta - \theta_t)$$

Problem

- How to define $p(\theta|\theta_{t-1})$: problem dependent, conjugate restrictions, etc.

Problem

- How to define $p(\theta|\theta_{t-1})$: problem dependent, conjugate restrictions, etc.
- ... and how to compute,

$$p(\theta_t|\mathbf{x}_{1:t}) = \frac{1}{Z} p(\mathbf{x}_t|\theta_t) \int p(\theta_t|\theta_{t-1}) p(\theta_{t-1}|\mathbf{x}_{1:t-1}) d\theta_{t-1}$$

A Bayesian approach to non-stationary data streams

Problem

- How to define $p(\theta|\theta_{t-1})$: problem dependent, conjugate restrictions, etc.
- ... and how to compute,

$$p(\theta_t|\mathbf{x}_{1:t}) = \frac{1}{Z}p(\mathbf{x}_t|\theta_t)\int p(\theta_t|\theta_{t-1})p(\theta_{t-1}|\mathbf{x}_{1:t-1})d\theta_{t-1}$$

- Literature is full of **ad-hoc** examples (e.g. Hidden Markov Models, Dynamic LDA models, etc.)

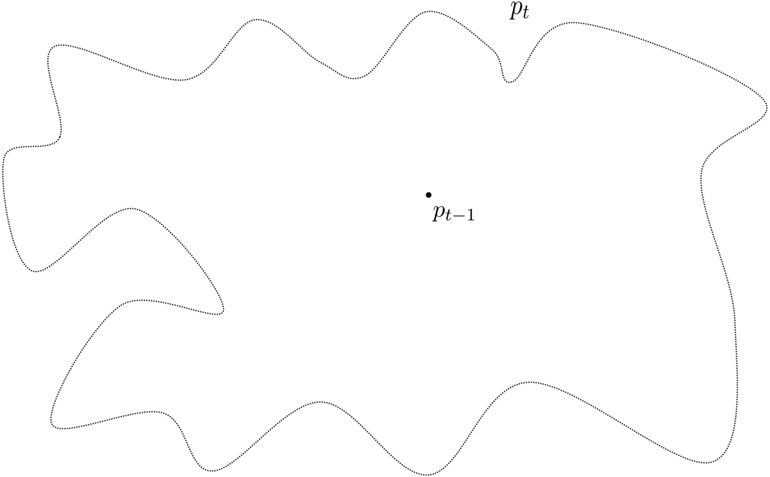
General Solution

- Define a **general family** of parameter transition distributions.
- **Integrates** easily in (approximate) Bayesian inference methods.

Implicit Transition Models

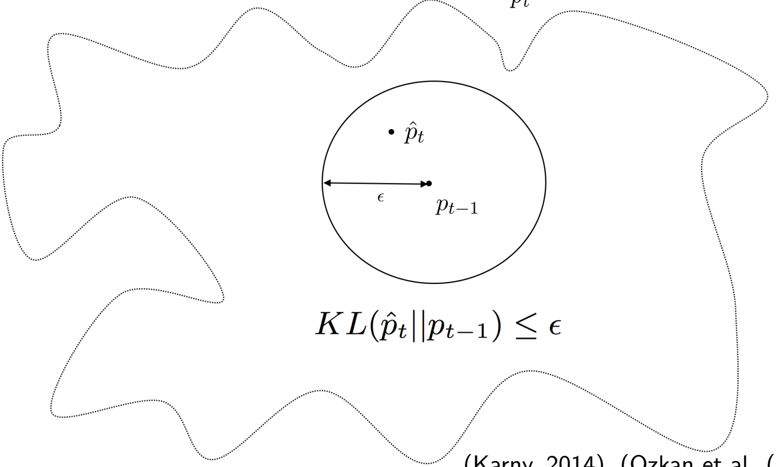
(Karny, 2014), (Ozkan et al. (2013))

$$p(\theta_t|\mathbf{x}_{1:t}) = \frac{1}{Z}p(\mathbf{x}_t|\theta_t)\underbrace{\int p(\theta_t|\theta_{t-1})\overbrace{p(\theta_{t-1}|\mathbf{x}_{1:t-1})}^{p_{t-1}}d\theta_{t-1}}_{\hat{p}_t}$$



Implicit Transition Models

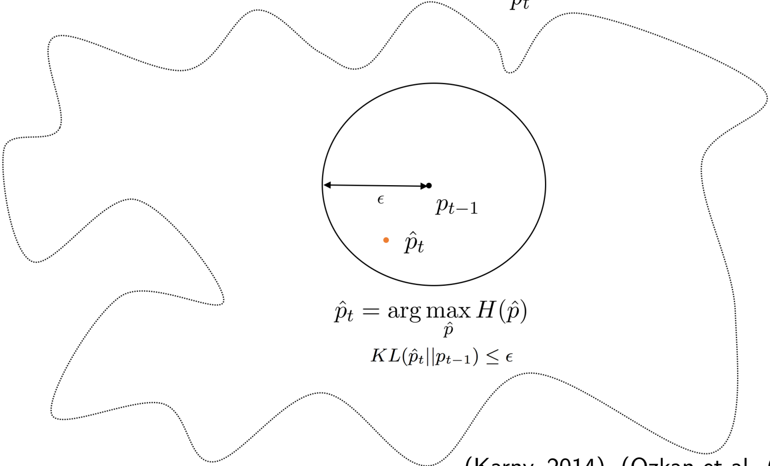
$$p(\theta_t|\mathbf{x}_{1:t}) = \frac{1}{Z} p(\mathbf{x}_t|\theta_t) \underbrace{\int p(\theta_t|\theta_{t-1}) \overbrace{p(\theta_{t-1}|\mathbf{x}_{1:t-1})}^{p_{t-1}} d\theta_{t-1}}_{\hat{p}_t}$$



(Karny, 2014), (Ozkan et al. (2013))

Implicit Transition Models

$$p(\theta_t|\mathbf{x}_{1:t}) = \frac{1}{Z} p(\mathbf{x}_t|\theta_t) \underbrace{\int p(\theta_t|\theta_{t-1}) \overbrace{p(\theta_{t-1}|\mathbf{x}_{1:t-1})}^{p_{t-1}} d\theta_{t-1}}_{\hat{p}_t}$$



$$\hat{p}_t = \arg\max_{\hat{p}} H(\hat{p})$$

$$KL(\hat{p}_t || p_{t-1}) \leq \epsilon$$

(Karny, 2014), (Ozkan et al. (2013))

Bayesian Updating under Implicit Transition Models

- **Closed-form solution** (up to normalization constant):

$$\hat{p}_t \quad \propto \quad p(\theta|\mathbf{x}_{1:t-1})^{\rho} p(\theta)^{1-\rho}$$

with $\rho \in [0, 1]$.

(Karny, 2014), (Ozkan et al. (2013))

Bayesian Updating under Implicit Transition Models

- **Closed-form solution** (up to normalization constant):

$$\hat{p}_t \quad \propto \quad p(\theta|\mathbf{x}_{1:t-1})^{\rho} p(\theta)^{1-\rho}$$

with $\rho \in [0, 1]$.

- Bayesian updating **simplifies** to,

$$p(\theta|\mathbf{x}_{1:t}, \rho) \quad = \quad \frac{1}{Z} p(\mathbf{x}_t|\theta) p(\theta|\mathbf{x}_{1:t-1}, \rho)^{\rho} p(\theta)^{1-\rho}$$

(Karny, 2014), (Ozkan et al. (2013))

Bayesian Updating under Implicit Transition Models

- **Closed-form solution** (up to normalization constant):

$$\hat{p}_t \quad \propto \quad p(\theta|\mathbf{x}_{1:t-1})^{\rho} p(\theta)^{1-\rho}$$

  with $\rho \in [0, 1]$.

- Bayesian updating **simplifies** to,

$$p(\theta|\mathbf{x}_{1:t}, \rho) \quad = \quad \frac{1}{Z} p(\mathbf{x}_t|\theta) p(\theta|\mathbf{x}_{1:t-1}, \rho)^{\rho} p(\theta)^{1-\rho}$$

- ρ is a **forgetting factor** (induced by ϵ)
  - $\rho = 1$ implies standard Bayesian updating.
  - $\rho = 0$ implies discard all past data.

(Karny, 2014), (Ozkan et al. (2013))

Bayesian Updating under Implicit Transition Models

- The ρ-**posterior** can be expressed as :

$$p(\theta|\mathbf{x}_{1:T}, \rho) \quad = \quad \frac{1}{Z} p(\theta) \prod_{t=1}^{T} p(\mathbf{x}_t|\theta)^{w_t}$$

where $w_t = \rho^{T-t}$.

Novel

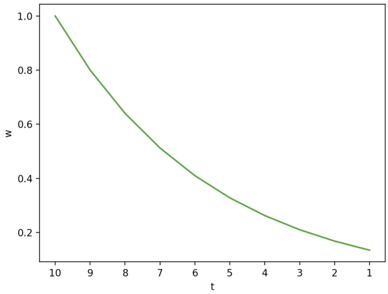
Bayesian Updating under Implicit Transition Models

- The ρ-**posterior** can be expressed as :

$$p(\theta|\mathbf{x}_{1:T}, \rho) \quad = \quad \frac{1}{Z} p(\theta) \prod_{t=1}^{T} p(\mathbf{x}_t|\theta)^{w_t}$$

where $w_t = \rho^{T-t}$.

- **Exponentially down-weight old data samples**.



Novel

Connection with Exponential forgetting

- The log-posterior equals **Exponential Forgetting** with a log-loss,

$$\ln p(\theta|\mathbf{x}_{1:T}, \rho) = \ln p(\theta) + \sum_{t=1}^{T} w_t \ln p(\mathbf{x}_t|\theta) - \ln Z$$

Novel

Connection with Exponential forgetting

- The log-posterior equals **Exponential Forgetting** with a log-loss,

$$\ln p(\theta|\mathbf{x}_{1:T}, \rho) = \ln p(\theta) + \sum_{t=1}^{T} w_t \ln p(\mathbf{x}_t|\theta) - \ln Z$$

- $w_t = \rho^{T-t}$ with $\rho \in [0, 1]$ being the **forgetting factor**.

Novel

Connection with Exponential forgetting

- The log-posterior equals **Exponential Forgetting** with a log-loss,

$$\ln p(\theta|\mathbf{x}_{1:T}, \rho) = \ln p(\theta) + \sum_{t=1}^{T} w_t \ln p(\mathbf{x}_t|\theta) - \ln Z$$

- $w_t = \rho^{T-t}$ with $\rho \in [0, 1]$ being the **forgetting factor**.
  - For $0 < \rho < 1$, it approximates a **sliding window** of size,

$$\lim_{T\to\infty} \sum_{t=1}^{T} w_t = \lim_{T\to\infty} \sum_{t=1}^{T} \rho^{T-t} = \frac{1}{1-\rho}$$

Novel

Connection with Exponential forgetting

- The log-posterior equals **Exponential Forgetting** with a log-loss,

$$\ln p(\theta|\mathbf{x}_{1:T}, \rho) = \ln p(\theta) + \sum_{t=1}^{T} w_t \ln p(\mathbf{x}_t|\theta) - \ln Z$$

- $w_t = \rho^{T-t}$ with $\rho \in [0, 1]$ being the **forgetting factor**.
  - For $0 < \rho < 1$, it approximates a **sliding window** of size,

$$\lim_{T\to\infty} \sum_{t=1}^{T} w_t = \lim_{T\to\infty} \sum_{t=1}^{T} \rho^{T-t} = \frac{1}{1-\rho}$$

Adaption to non-stationarity by exponentially down-weighting past data

Novel

- **Variational Inference tell us that** ,

$$\arg\max_q \mathbb{E}_q[\sum_t \ln p(\mathbf{x}_t|\theta)] - KL\big(q(\theta)||p(\theta)\big)$$

Novel

- **Variational Inference tell us that** ,

$$\arg\max_q \mathbb{E}_q[\sum_t \ln p(\mathbf{x}_t|\theta)] - KL\big(q(\theta)||p(\theta)\big)$$

- is the **Bayesian posterior**,

$$q(\theta) = p(\theta|\mathbf{x}_{1:t}) = \frac{1}{Z}p(\beta)\prod_t p(\mathbf{x}_t|\theta)$$

Novel

- The ρ-**posterior**,

$$p(\theta|\mathbf{x}_{1:T}, \rho) = \frac{1}{Z} p(\theta) \prod_t p(\mathbf{x}_t|\theta)^{w_t}$$

Novel

- The ρ-**posterior**,

$$p(\theta|\mathbf{x}_{1:T}, \rho) = \frac{1}{Z} p(\theta) \prod_t p(\mathbf{x}_t|\theta)^{w_t}$$

- It can be characterized as the one which maximizes,

$$\arg\max_q \mathbb{E}_q[\sum_t w_t \ln p(\mathbf{x}_t|\theta)] - KL\big(q(\theta)||p(\theta)\big)$$

where $w_t = \rho^{T-t}$.

Novel

- The ρ-**posterior**,

$$\arg\max_q \mathbb{E}_q[\frac{1}{T}\sum_t w_t \ln p(\mathbf{x}_t|\theta)] - \frac{1}{T}KL\big(q(\theta)||p(\theta)\big)$$

Novel

- The ρ-**posterior**,

$$\arg\max_q \mathbb{E}_q[\frac{1}{T}\sum_t w_t \ln p(\mathbf{x}_t|\theta)] - \frac{1}{T}KL\big(q(\theta)||p(\theta)\big)$$

- **Exponential forgetting (and Max-Entropy Implicit Transition Models)** assumes,

$$w_t = \rho^{T-t}$$

Novel

- The ρ-**posterior**,

$$\arg\max_q \mathbb{E}_q[\frac{1}{T}\sum_t w_t \ln p(\mathbf{x}_t|\theta)] - \frac{1}{T}KL\big(q(\theta)||p(\theta)\big)$$

- **Exponential forgetting (and Max-Entropy Implicit Transition Models)** assumes,

$$w_t = \rho^{T-t} = \frac{\pi_T\big(\mathbf{x}_t\big)}{\hat{\pi}\big(\mathbf{x}_t\big)}$$

Novel

A Covariate-shifted Posterior distribution

- The ρ-**posterior**,

$$\arg\max_q \mathbb{E}_q[\frac{1}{T}\sum_t w_t \ln p(\mathbf{x}_t|\theta)] - \frac{1}{T}KL\big(q(\theta)||p(\theta)\big)$$

- **Exponential forgetting (and Max-Entropy Implicit Transition Models)** assumes,

$$w_t = \rho^{T-t} = \frac{\pi_T(\mathbf{x}_t)}{\hat{\pi}(\mathbf{x}_t)} = \frac{\pi_T(\mathbf{x}_t)}{\hat{\pi}_t(\mathbf{x}_t)}$$

Novel

- The ρ-**posterior**,

$$\arg\max_q \mathbb{E}_q[\frac{1}{T}\sum_t w_t \ln p(\mathbf{x}_t|\theta)] - \frac{1}{T}KL\big(q(\theta)||p(\theta)\big)$$

- **Exponential forgetting (and Max-Entropy Implicit Transition Models)** assumes,

$$w_t = \rho^{T-t} = \frac{\pi_T(\mathbf{x}_t)}{\hat{\pi}(\mathbf{x}_t)} = \frac{\pi_T(\mathbf{x}_t)}{\hat{\pi}_t(\mathbf{x}_t)}$$

  - and,

$$KL(\pi_t||\pi_{t+1}) = \int \pi_t(\mathbf{x}) \ln \frac{\pi_t(\mathbf{x})}{\pi_{t+1}(\mathbf{x})} d\mathbf{x} = \ln\frac{1}{\rho}$$

Novel

A Covariate-shifted Posterior distribution

- The ρ-**posterior**,

$$\arg\max_q \mathbb{E}_q[\frac{1}{T}\sum_t w_t \ln p(\mathbf{x}_t|\theta)] - \frac{1}{T}KL\big(q(\theta)||p(\theta)\big)$$

- **Exponential forgetting (and Max-Entropy Implicit Transition Models)** assumes,

$$w_t = \rho^{T-t} = \frac{\pi_T(\mathbf{x}_t)}{\hat{\pi}(\mathbf{x}_t)} = \frac{\pi_T(\mathbf{x}_t)}{\hat{\pi}_t(\mathbf{x}_t)}$$

  - and,

$$KL(\pi_t||\pi_{t+1}) = \int \pi_t(\mathbf{x}) \ln \frac{\pi_t(\mathbf{x})}{\pi_{t+1}(\mathbf{x})} d\mathbf{x} = \ln \frac{1}{\rho}$$

- **Importance Sampling** approach applied by Covariate-Shift methods.
  - A method to account for the mismatch between training and test distribution.

Novel

A Covariate-shifted Posterior distribution

- The ρ-**posterior**,

$$\arg\max_q \mathbb{E}_q[\frac{1}{T}\sum_t \frac{\pi_T(\mathbf{x}_t)}{\hat{\pi}(\mathbf{x}_t)} \ln p(\mathbf{x}_t|\theta)] - \frac{1}{T} KL\big(q(\theta)||p(\theta)\big)$$

Novel

- The ρ-**posterior**,

$$\arg\max_q \mathbb{E}_q[\frac{1}{T}\sum_t \frac{\pi_T(\mathbf{x}_t)}{\hat{\pi}(\mathbf{x}_t)} \ln p(\mathbf{x}_t|\theta)] - \frac{1}{T} KL\big(q(\theta)||p(\theta)\big)$$

- aims to maximize,

$$\arg\max_q \mathbb{E}_q[E_{\pi_T}[\ln p(\mathbf{x}|\theta)]] - \frac{1}{T} KL\big(q(\theta)||p(\theta)\big),$$

it is optimal if $supp(\pi_T) \subseteq supp(\hat{\pi})$.

A Covariate-shifted Posterior distribution.

Novel

- The ρ-**posterior**,

$$\arg\max_q \mathbb{E}_q[\frac{1}{T}\sum_t \frac{\pi_T(\mathbf{x}_t)}{\hat{\pi}(\mathbf{x}_t)} \ln p(\mathbf{x}_t|\theta)] - \frac{1}{T} KL\big(q(\theta)||p(\theta)\big)$$

- aims to maximize,

$$\arg\max_q \mathbb{E}_q[E_{\pi_T}[\ln p(\mathbf{x}|\theta)]] - \frac{1}{T} KL\big(q(\theta)||p(\theta)\big),$$

it is optimal if $supp(\pi_T) \subseteq supp(\hat{\pi})$.

A Covariate-shifted Posterior distribution.

Implicit Transiton Models

- They are **generally applicable**.
- They have a clear interpretation in terms of **covariate-shift adaptation**.
- They can be **easily integrated within a variational framework**.

Novel

How to choose ρ?

(Masegosa et al. 2017)

Hierarchical Power Priors (Masegosa et al. 2017)

- Bayesian treatment of the **forgetting factor** ρ.

$$\rho \sim TruncatedExponential(\gamma)$$

How to choose ρ?

Hierarchical Power Priors (Masegosa et al. 2017)

- Bayesian treatment of the **forgetting factor** ρ.

$$\rho \sim TruncatedExponential(\gamma)$$

- Ad-hoc variational scheme for **conjugate exponential models**:

$$\arg\min_{\lambda_t} KL\big(q(\theta, \rho|\lambda_t)||p(\theta, \rho|\mathbf{x}_1, \ldots, \mathbf{x}_t)\big)$$

(Masegosa et al. 2017)

How to choose ρ?

Hierarchical Power Priors (Masegosa et al. 2017)

- Bayesian treatment of the **forgetting factor** ρ.

$$\rho \sim TruncatedExponential(\gamma)$$

- Ad-hoc variational scheme for **conjugate exponential models**:

$$\arg\min_{\lambda_t} KL\big(q(\theta, \rho|\lambda_t)||p(\theta, \rho|\mathbf{x}_1, \ldots, \mathbf{x}_t)\big)$$

- Broadly applicable to many models:
  - Mixture of Gaussians, LDA, Probabilistic PCA, Matrix Factorization, HMM, etc

(Masegosa et al. 2017)

Experiments

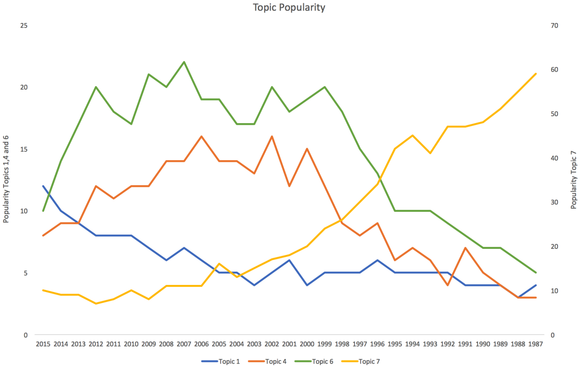
(Masegosa et al. 2017)

How to choose ρ?

Data set	Bayesian Updating	PVB (1)	PVB (2)	PVB (3)	PVB (4)	Fixed Forgetting Rate $\rho = 0.9$	Fixed Forgetting Rate $\rho = 0.99$	Our Approach Single ρ	Our Approach Mutiple ρs
Electricity	-44.91	-51.01	-52.19	-51.11	-61.70	-43.92	-44.80	-40.05	**-40.02**
GPS	-1.98	-2.10	-2.77	-1.97	-4.49	-1.94	-1.97	-1.97	**-1.86**
Finance	-19.84	-22.29	-22.57	-20.40	-20.73	**-19.05**	-19.78	-19.83	-19.83
NIPS	-4.07	-4.04*	-4.21*	-4.01	-4.12	-4.02	-4.06	-4.01	**-4.00**

Table: Aggregated test marginal log-likelihood.

- Adaptive forgetting mechanisms are usually needed.
- HPP with multiple ρ is the most robust approach.
- Non-stationary usually affect only a part of the model.



Topic 1	Topic 4	Topic 6	Topic 7
network	data	inference	input
networks	kernel	distribution	output
training	learning	posterior	networks
image	features	variational	units
learning	points	sampling	system
layer	feature	log	neurons
input	sample	gaussian	fig
model	kernels	bayesian	model
images	dataset	models	cells
output	samples	data	neuron

(Masegosa et al. 2017)

Future Work

Future Work

1. Adapt this scheme to **Non-Stationary Deep Bayesian Bandits**.
    - **Non-linear** relationship between the context and the reward.
    - The reward distribution is **non-stationary**.

1. Adapt this scheme to **Non-Stationary Deep Bayesian Bandits**.
    - **Non-linear** relationship between the context and the reward.
    - The reward distribution is **non-stationary**.
2. Learning **deep neural networks** from non-stationary data streams.