Detecting and Adapting to Irregular Distribution Shifts in Bayesian Online Learning

Aodong Li[1] **Alex Boyd**[2] **Padhraic Smyth**[1,2] **Stephan Mandt**[1,2]
[1]Department of Computer Science [2]Department of Statistics
University of California, Irvine
{aodongl1, alexjb, mandt}@uci.edu smyth@ics.uci.edu

Abstract

We consider the problem of online learning in the presence of distribution shifts that occur at an unknown rate and of unknown intensity. We derive a new Bayesian online inference approach to simultaneously infer these distribution shifts and adapt the model to the detected changes by integrating ideas from change point detection, switching dynamical systems, and Bayesian online learning. Using a binary 'change variable,' we construct an informative prior such that–if a change is detected–the model partially erases the information of past model updates by tempering to facilitate adaptation to the new data distribution. Furthermore, the approach uses beam search to track multiple change-point hypotheses and selects the most probable one in hindsight. Our proposed method is model-agnostic, applicable in both supervised and unsupervised learning settings, suitable for an environment of concept drifts or covariate drifts, and yields improvements over state-of-the-art Bayesian online learning approaches.

1 Introduction

Deployed machine learning systems are often faced with the problem of distribution shift, where the new data that the model processes is systematically different from the data the system was trained on [Zech et al., 2018, Ovadia et al., 2019]. Furthermore, a shift can happen anytime after deployment, unbeknownst to the users, with dramatic consequences for systems such as self-driving cars, robots, and financial trading algorithms, among many other examples.

Updating a deployed model on new, representative data can help mitigate these issues and improve general performance in most cases. This task is commonly referred to as *online* or *incremental learning*. Such online learning algorithms face a tradeoff between remembering and adapting. If they adapt too fast, their performance will suffer since adaptation usually implies that the model loses memory of previously encountered training data (which may still be relevant to future predictions). On the other hand, if a model remembers too much, it typically has problems adapting to new data distributions due to its finite capacity.

The tradeoff between adapting and remembering can be elegantly formalized in a Bayesian online learning framework, where a prior distribution is used to keep track of previously learned parameter estimates and their confidences. For instance, variational continual learning (VCL) [Nguyen et al., 2018] is a popular framework that uses a model's previous posterior distribution as the prior for new data. However, the assumption of such continual learning setups is usually that the data distribution is stationary and not subject to change, in which case adaptation is not an issue.

This paper proposes a new Bayesian online learning framework suitable for non-stationary data distributions. It is based on two assumptions: (i) distribution shifts occur irregularly and must be inferred from the data, and (ii) the model requires not only a good mechanism to aggregate data but

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

vs. orange as "less likely"). While hypothesis 1 assumes a single distribution shift (initially blue), hypothesis 2 (initially orange) assumes two shifts. We see that hypothesis 1 is initially more likely, but gets over-ruled by the better hypothesis 2 later (note the color swap at step 23).

4.2 Baselines

In our supervised experiments (Section 4.3 and Section 4.4), we compared VBS against adaptive methods, Bayesian online learning baselines, and independent batch learning baselines.[8] Among the adaptive methods, we formulated a supervised learning version of Bayesian online change-point detection (BOCD) [Adams and MacKay, 2007].[9] We also implemented Bayesian forgetting (BF) [Kurle et al., 2020] with convolutional neural networks for proper comparisons. Bayesian online learning baselines include variational continual learning (VCL) [Nguyen et al., 2018] and Laplace propagation (LP) [Smola et al., 2003, Nguyen et al., 2018]. Finally, we also adopt a trivial baseline of learning independent regressors/classifiers on each batch in both a Bayesian and non-Bayesian fashion. For VBS and BOCD we always report the most dominant hypothesis. In unsupervised learning experiments, we compared against the online version of word2vec [Mikolov et al., 2013] with a diffusion prior, dynamic word embeddings [Bamler and Mandt, 2017].

4.3 Bayesian Linear Regression Experiments

As a simple first version of VBS, we tested an online linear regression setup for which the posterior can be computed analytically. The analytical solution removes the approximation error of the variational inference procedure as well as optimization-related artifacts since closed-form updates are available. Detailed derivations are in Supplement D.

Real Datasets with Concept Shifts. We investigated three real-world datasets with *concept shifts*:

- **Malware** This dataset is a collection of 100K malignous and benign computer programs, collected over 44 months [Huynh et al., 2017]. Each program has 482 counting features and a real-valued probability $p \in [0, 1]$ of being malware. We linearly predicted the log-odds.
- **SensorDrift** A collection of chemical sensor readings [Vergara et al., 2012]. We predicted the concentration level of gas *acetaldehyde*, whose 2,926 samples and 128 features span 36 months.
- **Elec2** The dataset contains the electricity price over three years of two Australian states [Harries and Wales, 1999]. While the original problem formulation used a majority vote to generate 0-1 binary labels on whether the price increases or not, we averaged the votes out into real-valued probabilities and predicted the log-odds instead. We had 45,263 samples and 14 features.

At each step, only one data sample is revealed to the regressor. We evaluated all methods with one-step-ahead absolute error[10] and computed the mean cumulative absolute error (MCAE) at every step. In Table 1, we didn't report the variance of MCAEs since there is no stochastic optimization noise. Table 1 shows that VBS has the best average of MCAEs among all methods. We also reported the running performance in Supplement G.2, where other experimental details are available as well.

Basketball Player Tracking. We explored a collection of basketball player movement trajectories.[11] Each trajectory has wide variations in player velocities. We treated the trajectories as time series and used a Bayesian transition matrix to predict the next position $\mathbf{x}_{t+1}$ based on the current position $\mathbf{x}_t$. This matrix is learned and adapted on the fly for each trajectory.

We first investigated the effect of the temperature parameter β in our approach. To this end, we visualized the detected change points on an example trajectory. We used VBS (K=1, greedy search) and compared different values of β in Fig. 2 (b). The figure shows that the larger β, the more change

[8] As a reminder, a "batch" at discrete time t is the dataset available for learning; on the other hand, a "mini-batch" is a small set of data used for computing gradients for stochastic gradient-based optimization.

[9] 1) Although BOCD is mostly applied for unsupervised learning, its application in supervised learning and its underlying model's adaptation to change points are seldom investigated. 2) When the model is non-conjugate, such as Bayesian neural networks, we approximate the log evidence $\log p(y|x)$ by the evidence lower bound.

[10] We measured the error in the probability space for classification problems (Malware and Elec2) and the error in the data space for regression problems (SensorDrift).

[11] `https://github.com/linouk23/NBA-Player-Movements`

Table 1: Evaluation of Different Datasets

MODELS	CIFAR-10 (ACCURACY)↑	SVHN (ACCURACY)↑	MALWARE (MCAE 10^{-2})↓	SENSORDRIFT (MCAE 10^{-2})↓	ELEC2 (MCAE 10^{-2})↓	NBAPLAYER (LOGLIKE 10^{-2})↑
VBS (K=6)*	**69.2±0.9**	**89.6±0.5**	**11.61**	**10.53**	7.28	**29.49±3.12**
VBS (K=3)*	68.9±0.9	89.1±0.5	11.65	10.71	7.28	**29.22±2.63**
VBS (K=1)*	68.2±0.8	88.9±0.5	11.65	10.86	**7.27**	**29.25±2.59**
BOCD (K=6)♯	65.6±0.8	88.2±0.5	12.93	24.34	12.49	22.96±7.42
BOCD (K=3)♯	67.3±0.8	88.8±0.5	12.74	24.31	12.49	20.93±7.83
BF¶	**69.8±0.8**	**89.9±0.5**	11.71	11.40	13.37	24.17±2.29
VCL†	66.7±0.8	88.7±0.5	13.27	24.90	16.59	3.48±25.53
LP‡	62.6±1.0	82.8±0.9	13.27	24.90	16.59	3.48±25.53
IB§	63.7±0.5	85.5±0.7	16.6	27.71	12.48	-44.87±16.88
IB§ (BAYES)	64.5±0.3	87.8±0.1	16.6	27.71	12.48	-44.87±16.88

* PROPOSED, ♯ [ADAMS AND MACKAY, 2007], ¶ [KURLE ET AL., 2020]
† [NGUYEN ET AL., 2018], ‡ [SMOLA ET AL., 2003], § INDEPENDENT BATCH

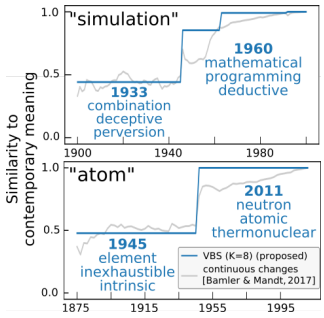

Figure 3: Sparse word meaning changes in "simulation" and "atom".

points are detected; the smaller β, the detected change points get sparser, i.e., β determines the model's sensitivity to changes. This observation confirms the assumption that β controls the assumed strength of distribution shifts.

In addition, the result also implies the robustness of poorly selected βs. When facing an abrupt change in the trajectory, the regressor has two adapt options based on different βs – make a single strong adaptation or make a sequence of weak adaptations – in either case, the model ends up adapting itself to the new trajectory. In other words, people can choose different β for a specific environment, with a trade-off between adaptation speed and the erased amount of information.

Finally, regarding the quantitative results, we evaluated all methods with the time-averaged predictive log-likelihood on a reserved test set in Table 1. Our proposed methods yield better performance than the baselines. In Supplement F, we provide more results of change point detection.

4.4 Bayesian Deep Learning Experiments

Our larger-scale experiments involve Bayesian convolutional neural networks trained on sequential batches for image classification using CIFAR-10 [Krizhevsky et al., 2009] and SVHN [Netzer et al., 2011]. Every few batches, we manually introduce *covariate shifts* through transforming all images globally by combining rotations, shifts, and scalings. Each transformation is generated from a fixed, predefined distribution (see Supplement G.3). The experiment involved 100 batches in sequence, where each batch contained a third of the transformed datasets. We set the temperature $\beta = 2/3$ and set the CELBO temperature $T = 20,000$ (in Eq. 6) for all supervised experiments.

Table 1 shows the performances of all considered methods and both data sets, averaged across all of the 100 batches. Within their confidence bounds, VBS and BF have comparable performances and outperform the other baselines. We conjecture that the strong performance of BF can be attributed to the fact that our imposed changes are still relatively evenly spaced and regular. The benefit of beam search in VBS is evident, with larger beam sizes consistently performing better.

4.5 Unsupervised Experiments

Our final experiment focused on unsupervised learning. We intended to demonstrate that VBS helps uncover interpretable latent structure in high-dimensional time series by detecting change points. We also showed that the detected change points help reduce the storage size and maintain salient features.

Towards this end, we analyzed the semantic changes of individual words over time in an unsupervised setup. We used Dynamic Word Embeddings (DWE) [Bamler and Mandt, 2017] as our base model. The model is an online version of Word2Vec [Mikolov et al., 2013]. Word2Vec projects a vocabulary into an embedding space and measures word similarities by cosine distance in that space. DWE further imposes a time-series prior on the embeddings and tracks them over time. For our proposed approach, we augmented DWE with VBS, allowing us to detect the changes of words meaning.

We analyzed three large time-stamped text corpora, all of which are available online. Our first dataset is the Google Books corpus [Michel et al., 2011] in n-grams form. We focused on 1900 to 2000

with sub-sampled 250M to 300M tokens per year. Second, we used the Congressional Records dataset [Gentzkow et al., 2018], which has 13M to 52M tokens per two-year period from 1875 to 2011. Third, we used the UN General Debates corpus [Jankin Mikhaylov et al., 2017], which has about 250k to 450k tokens per year from 1970 to 2018.

Our first experiments demonstrate VBS provides more interpretable step-wise word meaning shifts than the continuous shifts (DWE). Due to page limits, in Fig. 3 we selected two example words and their three nearest neighbors in the embedding space at different years. The evolving nearest neighbors reflect a semantic change of the words. We plotted the most likely hypothesis of VBS in blue and the continuous-path baseline (DWE) in grey. While people can roughly tell the change points from the continuous path, the changes are surrounded by noisy perturbations and sometimes submerged within the noise. VBS, on the other hand, makes clear decisions and outputs explicit change points. As a result, VBS discovers that the word "atom" changes its meaning from "element" to "nuclear" in 1945–the year when two nuclear bombs were detonated; word "simulation" changes its dominant context from "deception" to "programming" with the advent of computers in the 1950s. Besides interpretable changes points, VBS provides multiple plausible hypotheses (Supplement G.4).

Our second experiments exemplify the usefulness of the detected *sparse* change points, which lead to sparse segments of embeddings. The usefulness comes in two folds: 1) while alleviating the burden of the disk storage by storing one value for each segment, 2) the sparsity preserves the salient features of the original model. To illustrate these two aspects, we design a document dating task that exploits the probabilistic interpretation of word embeddings. The idea is to assign a test document to the year whose embeddings provide the highest likelihood. In Figure 2 (c), we measure the model sparsity on the x-axis with the average updated embeddings per step (The maximum is 10000, which is the vocabulary size). The feature preservation ability is measured by document dating accuracy on the y-axis. We adjust the prior log-odds ξ_0 (Eq. 6) to have successive models with different change point sparsity and then measure the dating accuracy. We also designed an oracle baseline named "binning" (grey, Supplement G.4). For VBS, we show the dominant hypothesis (blue) as well as the subleading hypotheses (orange). The most likely hypothesis of VBS outperforms the baseline, leading to higher document dating precision at much smaller disk storage.

5 Discussion

Beyond Gaussian Posterior Approximations. While the Gaussian approximation is simple and is widely used (and broadly effective) in practice in Bayesian inference [e.g., Murphy [2012], pp.649-662], our formulation does not rule out the extensions to exponential families. $\boldsymbol{\tau}_t$ in Eq. 2 could be generalized by reading off sufficient statistics of the previous approximate posterior. To this end, we need a sufficient statistic that is associated with some measure of entropy or variance that we broaden after each detected change. For example, the Gamma distribution can broaden its scale, and for the categorical distribution, we can increase its entropy/temperature. More intricate (e.g. multimodal) possible alternatives for posterior approximation are also possible, for example, Gaussian mixtures.

6 Conclusions

We introduced variational beam search: an approximate inference algorithm for Bayesian online learning on non-stationary data with irregular changes. Our approach mediates the tradeoff between a model's ability to memorize past data while still being able to adapt to change. It is based on a Bayesian treatment of a given model's parameters and aimed at tuning them towards the most recent data batch while exploiting prior knowledge from previous batches. To this end, we introduced a sequence of a discrete change variables whose value controlled the way we regularized the model. For no detected change, we regularized the new learning task towards the previously learned solution; for a detected change, we broadened the prior to give room for new data evidence. This procedure is combined with beam search over the discrete change variables. In different experiments, we showed that our proposed model (1) achieved lower error in supervised setups, and (2) revealed a more interpretable and compressible latent structure in unsupervised experiments.

Broader Impacts. As with many machine learning algorithms, there is a danger that more automation could potentially result in unemployment. Yet, more autonomous adaptation to changes will enhance the safety and robustness of deployed machine learning systems, such as self-driving cars.

Acknowledgements

We gratefully acknowledge extensive contributions from Robert Bamler (previously UCI, now at the University of Tübingen), which were indispensable to this work.

This material is based upon work supported by the National Science Foundation under the CAREER award 2047418 and grant numbers 1633631, 1928718, 2003237, and 2007719; by the National Science Foundation Graduate Research Fellowship under grant number DGE-1839285; by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001120C0021; by an Intel grant; and by grants from Qualcomm. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation, nor do they reflect the views of DARPA. Additional revenues potentially related to this work include: research funding from NSF, NIH, NIST, PCORI, and SAP; fellowship funding from HPI; consulting income from Amazon.com.

References

John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS medicine*, 15(11):e1002683, 2018.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua V Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, 2019.

Cuong V Nguyen, Yingzhen Li, Thang D Bui, and Richard E Turner. Variational continual learning. In *International Conference on Learning Representations*, 2018.

Antti Honkela and Harri Valpola. On-line variational bayesian learning. In *4th International Symposium on Independent Component Analysis and Blind Signal Separation*, 2003.

Masa-Aki Sato. Online model selection based on the variational bayes. *Neural computation*, 13(7): 1649–1681, 2001.

Thore Graepel, Joaquin Quinonero Candela, Thomas Borchert, and Ralf Herbrich. Web-scale bayesian click-through rate prediction for sponsored search advertising in microsoft's bing search engine. In *ICML*, 2010.

James McInerney, Rajesh Ranganath, and David Blei. The population posterior and bayesian modeling on streams. *Advances in neural information processing systems*, 28:1153–1161, 2015.

Lucas Theis and Matt Hoffman. A trust-region method for stochastic variational inference with applications to streaming data. In *International Conference on Machine Learning*, pages 2503–2511. PMLR, 2015.

R Kulhavỳ and Martin B Zarrop. On a general concept of forgetting. *International Journal of Control*, 58(4):905–924, 1993.

Richard Kurle, Botond Cseke, Alexej Klushyn, Patrick van der Smagt, and Stephan Günnemann. Continual learning with bayesian neural networks for non-stationary data. In *International Conference on Learning Representations*, 2020.

Sebastian Farquhar and Yarin Gal. A unifying bayesian view of continual learning. *Third workshop on Bayesian Deep Learning (NeurIPS 2018)*, 2018.

Jonathan Schwarz, Daniel Altman, Andrew Dudzik, Oriol Vinyals, Yee Whye Teh, and Razvan Pascanu. Towards a natural benchmark for continual learning. In *NeurIPS Workshop on Continual Learning*, 2018.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114 (13):3521–3526, 2017.

Michalis K Titsias, Jonathan Schwarz, Alexander G de G Matthews, Razvan Pascanu, and Yee Whye Teh. Functional regularisation for continual learning with gaussian processes. In *International Conference on Learning Representations*, 2019.

Soochan Lee, Junsoo Ha, Dongsu Zhang, and Gunhee Kim. A neural dirichlet process mixture model for task-free continual learning. In *International Conference on Learning Representations*, 2020.

Dushyant Rao, Francesco Visin, Andrei Rusu, Razvan Pascanu, Yee Whye Teh, and Raia Hadsell. Continual unsupervised representation learning. In *Advances in Neural Information Processing Systems*, pages 7647–7657, 2019.

Daniel Barry and John A Hartigan. Product partition models for change point problems. *The Annals of Statistics*, pages 260–279, 1992.

Paul Fearnhead. Exact bayesian curve fitting and signal segmentation. *IEEE Transactions on Signal Processing*, 53(6):2160–2166, 2005.

Paul Fearnhead and Zhen Liu. Online inference for multiple changepoint problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):589–605, 2007.

Ryan Prescott Adams and David JC MacKay. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*, 2007.

Xiang Xuan and Kevin Murphy. Modeling changing dependency structure in multivariate time series. In *Proceedings of the 24th international conference on Machine learning*, pages 1055–1062, 2007.

Yao Xie, Jiaji Huang, and Rebecca Willett. Change-point detection for high-dimensional time series with missing data. *IEEE Journal of Selected Topics in Signal Processing*, 7(1):12–27, 2012.

Yunus Saatçi, Ryan D Turner, and Carl Edward Rasmussen. Gaussian process change point models. In *ICML*, 2010.

Jeremias Knoblauch and Theodoros Damoulas. Spatio-temporal bayesian on-line changepoint detection with model selection. In *International Conference on Machine Learning*, pages 2718–2727. PMLR, 2018.

Ryan Turner, Steven Bottone, and Clay Stanek. Online variational approximations to non-exponential family change point models: With application to radar tracking. In *Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 1*, pages 306–314, 2013.

Jeremias Knoblauch, Jack Jewson, and Theodoros Damoulas. Doubly robust bayesian inference for non-stationary streaming data with β-divergences. volume 31, 2018.

Michalis K Titsias, Jakub Sygnowski, and Yutian Chen. Sequential changepoint detection in neural networks with checkpoints. *arXiv preprint arXiv:2010.03053*, 2020.

Scott Linderman, Matthew Johnson, Andrew Miller, Ryan Adams, David Blei, and Liam Paninski. Bayesian learning and inference in recurrent switching linear dynamical systems. In *Artificial Intelligence and Statistics*, pages 914–922, 2017.

Philip Becker-Ehmck, Jan Peters, and Patrick Van Der Smagt. Switching linear dynamics for variational bayes filtering. In *International Conference on Machine Learning*, pages 553–562, 2019.

Kevin P Murphy. *Machine learning: A probabilistic perspective*. MIT press, 2012.

David Barber. *Bayesian reasoning and machine learning*. Cambridge University Press, 2012.

Chris Bracegirdle and David Barber. Switch-reset models: Exact and approximate inference. In *Proceedings of The Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 190–198, 2011.

Martin J Wainwright and Michael I Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305, 2008.

Matthew D Hoffman and David M Blei. Structured stochastic variational inference. In *Artificial Intelligence and Statistics*, 2015.

Cheng Zhang, Judith Bütepage, Hedvig Kjellström, and Stephan Mandt. Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):2008–2026, 2018.

Alexander J Smola, Vishy Vishwanathan, and Eleazar Eskin. Laplace propagation. In *NIPS*, pages 441–448, 2003.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

Robert Bamler and Stephan Mandt. Dynamic word embeddings. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 380–389. JMLR. org, 2017.

Ngoc Anh Huynh, Wee Keong Ng, and Kanishka Ariyapala. A new adaptive learning algorithm and its application to online malware detection. In *International Conference on Discovery Science*, pages 18–32. Springer, 2017.

Alexander Vergara, Shankar Vembu, Tuba Ayhan, Margaret A Ryan, Margie L Homer, and Ramón Huerta. Chemical gas sensor drift compensation using classifier ensembles. *Sensors and Actuators B: Chemical*, 166:320–329, 2012.

Michael Harries and New South Wales. Splice-2 comparative evaluation: Electricity pricing. 1999.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, et al. Quantitative analysis of culture using millions of digitized books. *science*, 331(6014):176–182, 2011.

Matthew Gentzkow, JM Shapiro, and Matt Taddy. Congressional record for the 43rd-114th congresses: Parsed speeches and phrase counts. In *URL: https://data. stanford. edu/congress text*, 2018.

Slava Jankin Mikhaylov, Alexander Baturo, and Niheer Dasandi. United Nations General Debate Corpus, 2017. URL `https://doi.org/10.7910/DVN/0TJX8Y`.