APPENDIX 1 - PROOF OF STREAMING VARIATIONAL OBJECTIVES

We will outline the derivations of streaming variational objectives starting from the traditional Bayesian updating framework and the streaming Bayesian updating.

Streaming Variational Objective with Traditional Bayesian Updating

Let us first rewrite conventional Bayesian updating in terms of likelihood, prior and marginal probability of data.

$$p(\mathrm{z}|\mathrm{c}_b \dots \mathrm{c}_1) = p(\mathrm{c}_b|\mathrm{z})p(\mathrm{z}|\mathrm{c}_{b-1}\dots\mathrm{c}_1)/p(\mathrm{c}_b) \tag{1}$$

Consider the KL-divergence between an appropriate family of distribution $q_\theta(.)$ and posterior $p(\mathrm{z}|\mathrm{c}_b \dots \mathrm{c}_1)$ estimated using Bayesian updating. Recall that $q_\theta(.)$ is parameterized by θ.

$$\begin{aligned}
\mathrm{D}_{KL}[q_\theta(\mathrm{z})||p(\mathrm{z}|\mathrm{c}_b\dots\mathrm{c}_1)] &= \int_{-\infty}^{+\infty} q_\theta(\mathrm{z}) \ln \frac{q_\theta(\mathrm{z})}{p(\mathrm{z}|\mathrm{c}_b\dots\mathrm{c}_1)} \mathrm{dz} \\
&= \int_{-\infty}^{+\infty} q_\theta(\mathrm{z}) \ln \frac{q_\theta(\mathrm{z})}{p(\mathrm{c}_b|\mathrm{z})p(\mathrm{z}|\mathrm{c}_{b-1}\dots\mathrm{c}_1)} \mathrm{dz} + \ln p(\mathrm{c}_b), \quad \text{from eq. 1} \\
&= -\mathcal{L}(\theta;\mathrm{c}_b\dots\mathrm{c}_1) + \ln p(\mathrm{c}_b)
\end{aligned}$$

We will derive the streaming variational objective with traditional Bayesian updating by considering the variational lower bound $\mathcal{L}(\theta;\mathrm{c}_b\dots\mathrm{c}_1)$ separately while assuming $p(\mathrm{z}|\mathrm{c}_{b-1}\dots\mathrm{c}_1) \simeq q_{\theta_{b-1}}(\mathrm{z})$.

$$\begin{aligned}
\mathcal{L}(\theta;\mathrm{c}_b\dots\mathrm{c}_1) &= -\int_{-\infty}^{+\infty} q_\theta(\mathrm{z}) \ln \frac{q_\theta(\mathrm{z})}{p(\mathrm{c}_b|\mathrm{z})q_{\theta_{b-1}}(\mathrm{z})} \mathrm{dz} \\
&= \int_{-\infty}^{+\infty} q_\theta(\mathrm{z}) \ln p(\mathrm{c}_b|\mathrm{z}) \mathrm{dz} - \int_{-\infty}^{+\infty} q_\theta(\mathrm{z}) \ln \frac{q_\theta(\mathrm{z})}{q_{\theta_{b-1}}(\mathrm{z})} \mathrm{dz} \\
&= \mathrm{E}[\ln p(\mathrm{c}_b|\mathrm{z})] - \mathrm{D}_{KL}[q_\theta(\mathrm{z})||q_{\theta_{b-1}}(\mathrm{z})],
\end{aligned}$$

Streaming Variational Objective with Proposed Bayesian Updating

Let us first rewrite the proposed Bayesian updating with the scaling function S_b to scale the likelihood of batch c_b instead of simply using the number of batches. Since $S_b \in \mathcal{R}^+$, we substitute the product of likelihoods term with a likelihood raised to the power of S_b.

$$p(\mathrm{z}| < \mathrm{c}_1 \dots c_b >) \simeq p(\mathrm{c}_b|\mathrm{z})^{\mathrm{S}_b} p(\mathrm{z})^* / p(< \mathrm{c}_1 \dots c_b >) \tag{2}$$

Analogous to the previous proof, consider the KL-divergence between an appropriate family of distribution $q_\theta(.)$ and posterior $p(< \mathrm{z}|\mathrm{c}_b \dots \mathrm{c}_1 >)$ estimated using Bayesian updating.

$$\begin{aligned}
\mathrm{D}_{KL}[q_\theta(\mathrm{z})||p(\mathrm{z}| < \mathrm{c}_b\dots\mathrm{c}_1 >)] &= \int_{-\infty}^{+\infty} q_\theta(\mathrm{z}) \ln \frac{q_\theta(\mathrm{z})}{p(\mathrm{z}| < \mathrm{c}_b\dots\mathrm{c}_1 >)} \mathrm{dz} \\
&= \int_{-\infty}^{+\infty} q_\theta(\mathrm{z}) \ln \frac{q_\theta(\mathrm{z})}{p(\mathrm{c}_b|\mathrm{z})^{\mathrm{S}_b} p(\mathrm{z})^*} \mathrm{dz} + \ln p(< \mathrm{c}_b\dots\mathrm{c}_1 >), \quad \text{eq.2} \\
&= -\mathcal{L}(\theta; c_b, \mathrm{S}_b) + \ln p(< \mathrm{c}_b\dots\mathrm{c}_1 >)
\end{aligned}$$

Let us consider the variational lower bound $\mathcal{L}(\theta; c_b, \mathrm{S}_b)$ of the proposed Bayesian updating.

$$\begin{aligned}
\mathcal{L}(\theta; c_b, \mathrm{S}_b) &= -\int_{-\infty}^{+\infty} q_\theta(\mathrm{z}) \ln \frac{q_\theta(\mathrm{z})}{p(\mathrm{c}_b|\mathrm{z})^{\mathrm{S}_b} p(\mathrm{z})^*} \mathrm{dz} \\
&= \mathrm{S}_b \int_{-\infty}^{+\infty} q_\theta(\mathrm{z}) \ln p(\mathrm{c}_b|\mathrm{z}) \mathrm{dz} - \int_{-\infty}^{+\infty} q_\theta(\mathrm{z}) \ln \frac{q_\theta(\mathrm{z})}{p(\mathrm{z})^*} \mathrm{dz} \\
&= \mathrm{S}_b \times \mathrm{E}[\ln p(\mathrm{c}_b|\mathrm{z})] - \mathrm{D}_{KL}[q_\theta(\mathrm{z})||p(\mathrm{z})^*]
\end{aligned}$$

We have derived the proposed streaming variational objective from considering the KL-divergence between an suitable family of distribution $q(.)$ and the proposed posterior $p(\mathrm{z}| < \mathrm{c}_1 \ldots c_b >)$.

APPENDIX 2- BLACK-BOX INFERENCE OF VI, SVI AND PVI

To conduct a fair evaluation, this section derives black-box inference for VI, SVI (Hoffman et al., 2013) and PVI (McInerney et al., 2015) objectives following same approach employed by SSVB and BB-SVB.

A variational gradient estimator (VGE) can be constructed by differentiating the ELBO w.r.t. to the variational parameters θ (Hoffman et al., 2013; Ranganath et al., 2014; Kingma and Welling, 2013; Paisley et al., 2012) as shown below.

$$\nabla_\theta \mathcal{L}(\theta;\mathrm{x}) = \nabla_\theta \mathrm{E}[\log p(\mathrm{x}|\mathrm{z})] - \nabla_\theta \mathrm{D}_{KL}[q_\theta(\mathrm{z})||p(\mathrm{z})] \tag{3}$$

The VGE in equation 3 uses the full dataset to evaluate the gradient in a single iteration. The usual approach to construct a stochastic variational gradient estimator (SVGE) for randomly sampled mini-batches from a dataset with N data-points requires scaling the likelihood term by $\frac{\mathrm{N}}{\mathrm{M}}$ (Hoffman et al., 2013; Kucukelbir et al., 2017). Thus, the likelihood is scaled to as it is computed using the full dataset suppressing the overwhelming priors or in this instance the overwhelming KL divergence term. We obtain SVGE for mini-batches randomly sampled from the full dataset as follows.

$$\nabla_\theta \mathcal{L}(\theta;\mathrm{x}) \simeq \nabla_\theta \mathcal{L}(\theta;\mathrm{c}_b,N,M) = \frac{N}{M}\nabla_\theta \mathrm{E}[\log p(\mathrm{c}_b|\mathrm{z})] - \nabla_\theta \mathrm{D}_{KL}[q_\theta(\mathrm{z})||p(\mathrm{z})] \tag{4}$$

Algorithm 1: Black-Box Variational Inference - VI

```
Inputs : x, p(z)
Initialize : $\theta$
repeat
    $g \leftarrow \nabla_\theta \mathcal{L}(\theta;\mathrm{x})$ (VGE Eq. 3)
    $\theta \leftarrow$ Update parameters using gradients $g$
until $\theta$ converges;
return $\theta$
```

Algorithm 2: Black-Box Stochastic Variational Inference - SVI

```
Inputs : $\mathrm{c}_1 \ldots \mathrm{c}_b$, p(z), N, M
Initialize : $\theta$
foreach $c_i \in c_1 \ldots c_b$ do
    for $t \in 1 : T$ do
        $g \leftarrow \nabla_\theta \mathcal{L}(\theta;\mathrm{c}_i,\mathrm{N},\mathrm{M})$ (SVGE Eq. 4)
        $\theta \leftarrow$ Update parameters using gradients $g$)
    end
end
return $\theta$
```

We can optimize the VGE and SVGE following *reparameterization VI* to construct the black-box inference for VI and SVI. Accordingly, algorithms 1 and 2 respectively present the black-box VI and black-box SVI. Furthermore, we derive black-box PVI by replacing N with an additional hyperparameter α to control the posterior variance as proposed by McInerney et al. (2015).

Recall that the KL divergence term serves as the regularization to the posteriors, thus altering α also adjusts the regularization to the posteriors in addition to controlling the posterior variance. Therefore, estimating the optimal α also resembles finding the ideal regularization to posteriors.

Accordingly, we have obtained black-box counterparts of VI, SVI and PVI following their original objectives in this section. We will be using them throughout our experiments in contrast with SSVB and BB-SVB.

APPENDIX 3 - SVB WITH BLACK-BOX INFERENCE

This section demonstrates the deficiencies in extending the SVB (Broderick et al., 2013) to the black-box VI techniques. As seen in the literature, the black-box inference techniques are mostly motivated by the ability to perform gradient descent updates on the variational objective (Kucukelbir et al., 2017; Ranganath et al., 2014; Kingma and Welling, 2013; Zhang et al., 2018). Therefore, we use black-box VI presented in algorithm 1 as the offline approximation primitive of SVB. We analyze the estimated posteriors using SVB against the posterior approximated by SVI (algorithm 2) for a simple logistic regression task. We perform single-pass updates on each data points from 1e3 generated data-points with five regressors. Figure A1 illustrates the approximated posteriors for the five regression coefficients ($d = 1:5$) after each two hundred data points.

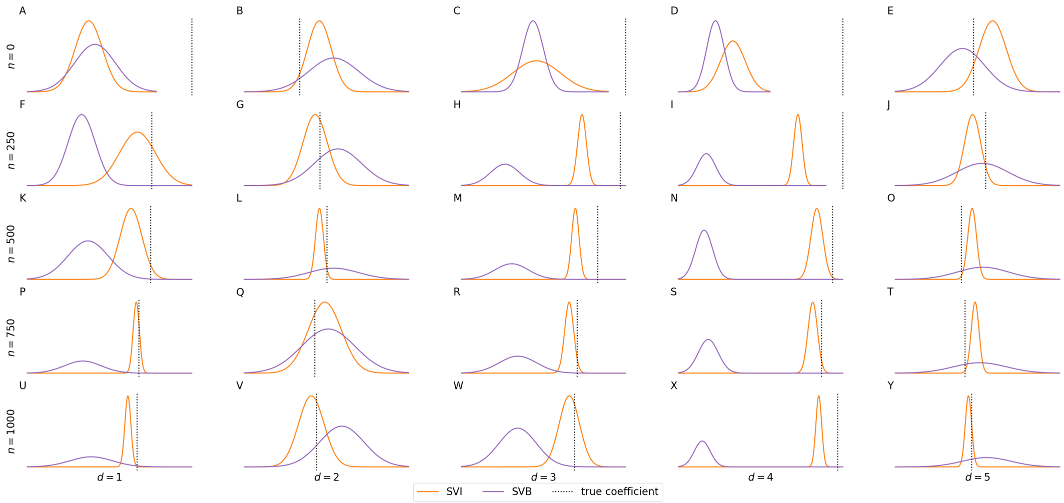

Figure A1. Convergence of posteriors estimated using traditional SVB with black-box inference primitives and SVI

The posteriors estimated using SVB are either failing to converge to the true coefficients or suffering from a high variance when using BB-VI as the approximation inference primitive. This is mainly due to the properties of the steepest descent; for each mini-batch, it initiates the stochastic search from a new random point, which results in a much slower and poor convergence for SVB framework. Since SVB is not extendable as an efficient black-box inference alternative, we do not consider SVB in our analysis. Accordingly, we consider PVI and SVI as the existing state-of-the-art to perform black-box inference with data streams.

APPENDIX 4 - MULTINOMIAL LOGISTIC REGRESSION

Let us consider b$^{\text{th}}$ mini-batch with M data-points $\mathrm{x} = \{x_i\}_{i=1}^{M}$ where each sample x_i is D-dimensional. The targets $\mathrm{y} = \{y_i\}_{i=1}^{M}$ consist K-dimensional vectors representing probability of each class given the respective x_i. Then likelihood presented in equation 5 describes the data generated i.i.d., where $h(.)$ denotes the Softmax function.

$$p(\mathrm{y}|\mathrm{x},\mathrm{w}) = \prod_{i=1}^{M} \mathcal{C}at(y_i|h(x_i.\mathrm{w})) \tag{5}$$

The inference process is expected to approximate the posterior of the coefficient matrix w that is parameterized by μ and σ^2. Therefore, the prior $p(w_{jk})$ and posterior $q(w_{jk})$ corresponding to the j$^{\text{th}}$ predictor and the k$^{\text{th}}$ class can be defined as follows.

$$p(w_{jk}) = \mathcal{N}(\bar{\mu}_{jk}, \bar{\sigma}_{jk}^2) \qquad q(w_{jk}) = \mathcal{N}(\mu_{jk}, \sigma_{jk}^2) \tag{6}$$

This concludes the probabilistic model for multinomial logistic regression. By considering the above probabilistic model, we write the relevant objectives for SSVB and BB-SVB, and the existing state-of-the-art, PVI and SVI. We have illustrated one such objective below which is derived to implement SSVB.

$$\mathcal{L}(\mu,\sigma^2;\mathrm{y},\mathrm{x},\mathrm{S}_b) = \mathrm{S}_b \times \sum_{i-1}^{M} \mathrm{E}[\log p(y_i|h(x_i.w)] - \sum_{j=1}^{D}\sum_{k=1}^{K} \mathrm{D}_{KL}[q_{\mu,\sigma^2}(w_{jk})||p_{(\bar{\mu},\bar{\sigma}^2)}(w_{jk})] \tag{7}$$

Here the parameters to $p(w_{jk})$, $\bar{\mu}_{jk}$ and $\bar{\sigma}^2_{jk}$ are respectively the expectation of the posterior estimated with $b-1^{\mathrm{th}}$ batch and the initial uncertainty assigned to the posteriors.

APPENDIX 5 - CLASSIFICATION FINAL F1 SCORES

	20News	**MNIST**	**Otto Products**
SSVB	0.8236 ± 0.0014	0.8932 ± 0.0014	0.8226 ± 0.0034
BB-SVB	0.8235 ± 0.0027	0.8874 ± 0.0148	0.8226 ± 0.0034
PVI	0.7749 ± 0.0110	0.8906 ± 0.0066	0.8222 ± 0.0007
SVI	0.7123 ± 0.0148	0.8845 ± 0.0097	0.8041 ± 0.0049
AROW	-	0.8970 ± 0.0014	0.8004 ± 0.0018
PA	0.7965 ± 0.0237	0.8792 ± 0.0084	0.8241 ± 0.0325
SGD	0.7659 ± 0.0278	0.8722 ± 0.0082	0.8268 ± 0.0197

Table A1. Final F1 scores using with-hold set for multi-class classification

Table A1 presents the f1 scores computed using the with-hold set for each multiclass classification dataset. These values are computed once the model is updated using all the data-points in the training set during the experiment 1 phase 1.

APPENDIX 6 - LINEAR MIXED-EFFECTS MODEL

Consider a mini-batch having M observations $\mathrm{y} = \{y_i\}_{i=1}^{M}$ corresponding to D-dimensional fixed-effect predictors $\mathrm{X} = \{X_i\}_{i=1}^{M}$ collected sequentially from $\mathcal{C}$ subjects that is described by the random effects vector u. Assuming that the fixed effect predictors and the observations follow a linear relationship we can denote the i^{th} observation y_i as follows.

$$y_i = X_i\beta + Z_i u + \varepsilon_i \tag{8}$$

In equation 8 the observations y_i are described as a combination of the fixed effects β with random effects u. Fixed effect β is a D dimensional vector that consists of regression coefficient for the D predictors X_i, whereas random effect u is $\mathcal{C}$ dimensional vector corresponding to the random effects for $\mathcal{C}$ subjects. Random effects design vector Z_i is typically a one-hot encoded vector indicating the source of the observation y_i out of $\mathcal{C}$ subjects to assign the corresponding random effect from u. Error term ε_i represents the noise in the each observation y_i.

The likelihood of the observations y can be express as the conditional probability shown in 9 which is assumed to be corrupted by i.i.d. Gaussian noise with unknown variance σ^2.

$$p(\mathrm{y}|\mathrm{x},\beta,b) = \prod_{i=1}^{M} \mathcal{N}(y_i|X_i\beta + Z_i u,\sigma^2) \tag{9}$$

In our implementations of LME, we consider both β and u as random variables, thus coefficients of fixed effect predators and random effects are respectively given Gaussian and Multivariate Gaussian priors as illustrated in equation 10. The respective posteriors are approximated to the same distributions as their priors.

$$p(\beta_d) = \mathcal{N}(\mu_d, \sigma_d^2) \qquad\qquad p(u) = \mathcal{N}_C(\mu, \Sigma) \tag{10}$$

We can derive the online inference objective similar to equation 7. Figure A2 illustrates the inference network implemented for LME model.

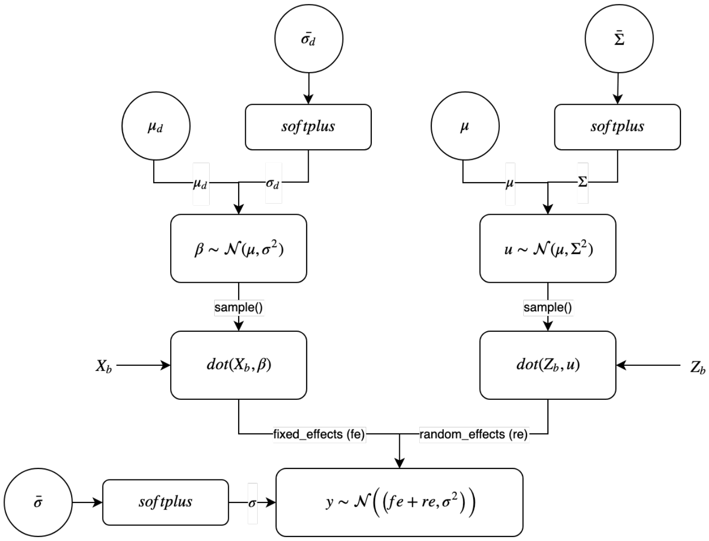

Figure A2. Forward Propagation of the Inference Network Implemented for LME Model

REFERENCES

Broderick, T., Boyd, N., Wibisono, A., Wilson, A. C., and Jordan, M. I. (2013). Streaming variational bayes. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 26*, pages 1727–1735. Curran Associates, Inc.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *J. Mach. Learn. Res.*, 14(1):1303–1347.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *CoRR*, abs/1312.6114.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. M. (2017). Automatic differentiation variational inference. *J. Mach. Learn. Res.*, 18(1):430–474.

McInerney, J., Ranganath, R., and Blei, D. M. (2015). The population posterior and bayesian modeling on streams. In *NIPS*.

Paisley, J. W., Blei, D. M., and Jordan, M. I. (2012). Variational bayesian inference with stochastic search. In *ICML*.

Ranganath, R., Gerrish, S., and Blei, D. (2014). Black Box Variational Inference. In Kaski, S. and Corander, J., editors, *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pages 814–822, Reykjavik, Iceland. PMLR.

Zhang, C., Butepage, J., Kjellstrom, H., and Mandt, S. (2018). Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence*.