LOW COMPLEXITY APPROXIMATE BAYESIAN LOGISTIC REGRESSION FOR SPARSE ONLINE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Theoretical results show that Bayesian methods can achieve lower bounds on regret for online logistic regression. In practice, however, such techniques may not be feasible especially if the feature set is very large. Various approximations that, for huge sparse feature sets, diminish the theoretical advantages, must be used. Often, such approximations apply stochastic gradient methods with parameters that must be tuned on some surrogate loss, defeating the theoretical nonparametric advantages of Bayesian methods. The surrogate loss is defined to approximate the mixture, and requires techniques like Monte Carlo sampling that increase the number of computations per example. We propose low complexity (linear overall, fixed per example) analytical approximations for the sparse online logistic regression problem. Unlike variational inference and other methods, our methods lead to analytical closed forms, lowering the practical number of computations. Unlike techniques used for dense features sets, such as Gaussian Mixtures, out methods allow for sparse problems with huge feature sets without increasing complexity. With the analytical closed forms, there is also no need for applying stochastic gradient methods on surrogate losses, and for tuning and balancing learning and regularization parameters of such methods. Empirical results on synthetic data top the performance of the more computationally involved methods. Like such methods, our methods still reveal per feature and per example uncertainty measures.

1 INTRODUCTION

We consider *online* (Anderson, 2008; Bottou, 1998; Shalev-Shwartz et al., 2011) binary logistic regression over a series of rounds $t \in \{1, 2, \ldots, T\}$. At round t, a *sparse* feature vector $\mathbf{x}_t \in [-1, 1]^d$ with $d_t \ll d$ nonzero values, is revealed, and a prediction for the label $y_t \in \{-1, 1\}$ must be generated. The dimension d of the feature vector can be huge (billions). The number d_t of nonzero elements of $\mathbf{x}_t$ at any t is usually in the tens or at most in the hundreds. Logistic regression is used in a huge portion of existing learning problems. It can be used to predict medical risk factors, to predict world phenomena, stock market movements, or click-through-rate in online advertising. The online sparse setup is also very common to these application areas, particularly, if predictions need to be streamed in real time as the model keeps updating from newly seen examples.

The goal of a prediction algorithm is to maximize the prediction probabilities of the labels that are actually observed. Online algorithms sequentially learn parameters associated with the d features. With *stochastic gradient* methods (Bottou, 2010; Duchi et al., 2011), these are weights $w_{i,t}$ (of component $i \in \{1, \cdots, d\}$ at round t) assigned to each of the features. Bayesian methods, on the other hand, attempt to keep track of some distribution over the parameters, and assign an expected *mixture* probability to the generated prediction (Hoffman et al., 2010; Opper & Winther, 1998). In either case, the algorithms attempt to maximize a sequence likelihood probability, consisting of the product of all predictions. Alternatively, logarithmic loss is minimized. A benchmark to measure the performance of an algorithm is its *regret*. The regret is defined as the excess loss attained by an algorithm over that of an algorithm that uses some fixed *comparator* values of $\mathbf{w}^* \triangleq (w_1, w_2, \ldots, w_d)^\mathcal{T}$ ($\mathcal{T}$ denoting transpose). A comparator $\mathbf{w}^*$ that minimizes the cumulative loss can be picked to measure the regret relative to the best possible comparator in some *space of parameter values*.

Several papers (see, e.g., Kakade & Ng (2005); Foster et al. (2018)) demonstrated that, *in theory*, Bayesian methods are capable to achieve regret, logarithmic with the horizon T and linear with d.

Shamir (2020) showed that such methods are capable of achieving lower bounds on the regret for $d = o(T)$. Classical stochastic gradient methods, on the other hand, are usually implemented as *proper* learning algorithms (Foster et al., 2018), that determine $\mathbf{w}_t$ prior to observing $\mathbf{x}_t$. Proper algorithms were shown (Hazan et al., 2014) to be inferior, in the worst case, achieving regret $\Omega(\sqrt{T})$. However, in many cases, gradient based algorithms can still achieve logarithmic regret, depending on the actual data and feature sets, see, e.g., Bach (2010; 2014); Bach & Moulines (2013). Furthermore, recent work Jézéquel et al. (2020) demonstrates non-Bayesian improper algorithms based on gradient methods techniques that achieve better regret than lower bounds in Hazan et al. (2014).

Unfortunately, the superiority of Bayesian methods diminishes by their intractability. Results in Kakade & Ng (2005); Foster et al. (2018); Shamir (2020) demonstrate that with a prior whose covariance is diagonal, with each component distributed either uniformly or Normally with large variance, one can achieve regret lower bounds. However, an online method with high dimensions requires computing the posterior after an example is observed. This posterior becomes the prior for future examples. In sparse problems with huge feature sets, direct computation of this posterior is intractable. Instead, Gaussian approximations that rely on diagonalization of the covariance matrix between features must be used. Because both the normal and the diagonal assumptions are not true for the real posterior at a given round (even if the prior is truely diagonal), they lead to inaccuracies that degrade the optimality of even an optimal prior. The diagonalization can be viewed as a similar mechanism to linearization used in convex optimization, in general, and for *stochastic gradient descent* (SGD), in particular, (see, e.g. Zinkevich (2003)). While it allows for handling each feature independently, it potentially increases the best achievable regret rates.

While there are clearly theoretical advantages to applying Bayesian methods to improve regret in online problems, most of the focus in Bayesian learning has been on applying such methods to generate predictions of the posterior probabilities, and leveraging them to provide model (*epistemic*) uncertainty measurements (Bishop, 2006; Dempster, 1968; Huelsenbeck & Ronquist, 2001; Knill & Richards, 1996). However, the uncertainty of a feature is, in fact, mostly a function the number of examples in which the feature was present; a measure that can be tracked instead of estimated. Methods, such as *Variational Bayesian (VB) Inference* (Bishop, 2006; Blei et al., 2017; Drugowitsch, 2019; Drugowitsch et al., 2019; Ranganath et al., 2014), do attempt to track such measurements through matching the posterior. However, as demonstrated in Rissanen (1984) seminal work, regret minimization is an identical problem to uncertainty reduction, as regret is merely a different representation of uncertainty. As shown in Kakade & Ng (2005); Foster et al. (2018); Shamir (2020), regret can be *universally* minimized over the possible parameter space through a good choice of a prior. Hence, to minimize uncertainty, the role of an approximation is *to preserve the effect of such a prior at least in the region of the distribution that dominates the ultimate posterior at the horizon* T. This may, in fact, be a simpler problem than matching the exact posterior, thus opening possibilities for simpler approximations that may lead to results identical to those of heavy methods as VB.

Variational Bayesian methods are typically used offline to match a the true posterior with a tractable one through upper bounding overall loss. They are computationally involved, requiring either iterative techniques (Bishop, 2006; Blei et al., 2017; Murphy, 2012) like Expectation Maximization (EM) (Dempster et al., 1977; Moon, 1996); or Monte Carlo (MC) sampling, replacing analytical expectation by an empirical one over a randomly drawn set. To converge, MC can be combined with gradient descent, either requiring heavy computations per example, or adapting *stochastic* gradient methods to update posteriors (Broderick et al., 2013; Knowles, 2015; Nguyen et al., 2017a;b). The posterior of an example becomes the prior of the subsequent example. To prevent additional losses of the approximation, the algorithm must converge to the optimal value of the approximate posterior at every example. Thus with stochastic gradient methods, additional losses may be incurred, as the algorithm may not converge at each example, while the target posterior keeps moving with subsequent examples. Furthermore, one of the mere benefits of Bayesian methods is their parameter free (universal) nature (Abdellaoui, 2000; Mcmahan & Streeter, 2012; Orabona & Pál, 2016), that eliminates the need of tuning learning rates, regularization and tradeoffs between components of the loss. This benefit is defeated by combining methods that are governed by such parameters.

Most Bayesian literature addressed the *dense* problem, where $\mathbf{x}_t$ consists of mostly nonzero entries for every t, and *the dimension* d *of the feature space is relatively small*. Techniques, like *Gaussian Mixtures* (Herbrich et al., 2003; Montuelle et al., 2013; 2014; Rasmussen, 2000; Reynolds et al., 2000; Sung, 2004), that may use VB, usually apply matrix computations quadratic in d on the

covariance matrix. In many practical problems, however, a very small feature subset is present in each example. For *categorical* features, only one of the features in the vector is present at any example. Techniques, useful for the low dimensional dense problem, may thus not be practical.

Paper Contributions: We provide a simple analytical Bayesian method for online sparse logistic regression with closed form updates. Empirical results on synthetically generated data demonstrate that the method provides regret rates which in many cases are better than classical methods, and are comparable to or better than computationally involved Bayesian methods, such as VB. As other online Bayesian methods, it measures the evolution of uncertainty on features and incoming examples. As we focus on optimizing regret, our results provide, yet another example, to the connection between uncertainty and regret rates (Rissanen, 1984), and more broadly the *Minimum Description Length (MDL)* principle (Grunwald, 2007; Rissanen, 1978a;b; 1984; 1986; Shamir, 2015; 2020).

We demonstrate that it is sufficient to approximate the component of the posterior which will dominate at the horizon, matching it by a diagonal Gaussian approximation, but avoiding matching (a bound on) the full posterior, as done by VB methods. While matching the posterior may lead to comparable performance, it brings unnecessary complexity to the real goal of preserving the effects of a good prior on the regret. Specifically, in an online setting, matching the full range of the posterior can lead to poor generalization and overfitting if matching tails is at the expense of important parts of the distribution. Unlike VB methods, that approximate an upper bound on the loss and apply expectation on the bound, our marginalization applies updates *directly* on the current approximate of the posterior. Instead of matching a multivariate distribution, as done for dense methods, we leverage sparsity and diagonalization by marginalizing on all features except the one being updated.

Related Work: The simplest single dimensional online logistic regression problem ($d = 1$ and $x_{1,t} = 1, \forall t$) was widely studied. Jeffery's prior, $\rho(\theta) \triangleq 1/\left(\pi\sqrt{\theta(1-\theta)}\right)$, is asymptotically optimal (Clarke & Barron, 1994) (see also, Xie & Barron (1997; 2000); Drmota & Szpankowski. (2004)). The prior can be expressed in terms of *log-odds* weights w as $\rho(w) = e^{w/2}/[\pi\left(1+e^{w}\right)]$. Applying the mixture leads to the *Krichevsky & Trofimov (1981)* add-$1/2$ online estimator $Q\left(y_t|y^{t-1}\right) = \left[n_{t-1}\left(y_t\right) + 0.5\right]/t$, where $n_{t-1}(y_t)$ counts prior occurrences of y_t. We use y^t to express a sequence between 1 and t. Interestingly, applying this prior in a *Follow The Regularized Leader* (FTRL) setting (McMahan, 2011) also leads to the KT estimator. This raised the question whether this approach can be extended to larger dimensions (McMahan & Streeter, 2012). Hazan et al. (2014) showed though that the regret bounds do not generalize to higher dimensions for proper gradient methods based on FTRL methods. The bounds, however, theoretically generalize for Bayesian methods, as shown by Kakade & Ng (2005); Foster et al. (2018); Shamir (2020). However, the prior required to achieve the regret bounds is no longer the Dirichlet-$1/2$ prior. Instead, large variance Gaussian or uniform priors must be used, but they can have diagonal covariance matrix over the dimensions. Peaked priors cannot be used because in the multi-dimensional case, for every feature in a given example, all other features in the example provide a *self excluding log-odds prior*, which can shift the mean of the distribution for the given feature, thus rendering peaked 0-mean priors possibly useless. While wide priors are good theoretically, because of the intractability of the Bayesian mixture integrals, diagonal approximations must be used for high dimensional sparse problems. Such approximations unfortunately degrade the wide priors.

Bayesian methods have been studied extensively for estimating posteriors and uncertainty (Bishop, 2006; Makowski et al., 2002; Sen & Stoffa, 1996). There is ample literature researching such techniques in deep networks (see, e.g., Blundell et al. (2015); Hwang et al. (2020); Kendall & Gal (2017); Lakshminarayanan et al. (2017); Malinin & Gales (2018); Wilson (2020)). Most of the work focuses on the ultimate posterior after the full training dataset has been visited. One attempts to leverage the uncertainty measurements to aid in inference on unseen examples. Techniques like *expectation propagation (EP)* (Minka, 2001; 2013) (see also Bishop (2006); Chu et al. (2011); Cunningham et al. (2011); Graepel et al. (2010)) and VB are used to generate estimates of the posterior. In a dense setup, where there is a relatively small number features (or units in a deep network), Gaussian Mixture models can also be learned, where a jointly Gaussian posterior is learned, usually with some kernel that is used to reduce the dimensionality of the parameters that are actually being trained. Such methods, however, do not fit the sparse online setup.

Variational methods are derived utilizing Jensen's inequality to upper bound loss of expectation by expectation of the negative logarithm of the product of the prior $\rho(\mathbf{w})$ and data likelihood

$P(y^T|\mathbf{x}^T, \mathbf{w})$. Normalizing this joint by the expected label sequence probability gives the posterior $P(\mathbf{w}|y^T, \mathbf{x}^T)$. Then, a posterior $Q(\cdot)$ with a desired form is matched by minimizing the KL divergence $KL(Q||P)$. This minimization breaks into minimizing the sum of the expected loss on y^T w.r.t. $Q(\dot{})$ and the divergence $KL(Q||\rho)$ between the approximated posterior and the true prior. The first term may require techniques like the iterative EM (performing mean field approximation, (Bishop, 2006; Jaakkola & Jordan, 1998)), or Monte Carlo sampling, to be approximated. Gradient methods can also minimize $KL(Q||P)$. In the sparse setup, it is standard to assume a diagonal $Q(\cdot)$. In an online setting, the process can be iterated over the examples (or over mini-batches), where the posterior at time t becomes the prior at $t+1$. Computing the approximate posterior may be very expensive if done for every example. SGD can be used with Monte Carlo sampling, but that would incur additional losses, as the posterior changes between successive examples. Like VB, EP minimizes the opposite divergence $KL(P||Q)$ between the posterior and its approximate.

2 PRELIMINARIES

Let $\rho_t(\mathbf{w})$ be the prior on the weights at round t, where we start by initializing some $\rho_1(\mathbf{w})$. We will assume that $\rho(\cdot)$ is approximated by a diagonal covariance Gaussian, with means $\mu_{i,t}$ and variances $\sigma^2_{i,t}$ for component i at time t. Leveraging results in Kakade & Ng (2005); Foster et al. (2018); Shamir (2020), if we restrict $w_i \in [-B, B]$, a uniform prior over this interval or a normal prior with standard deviation proportional to B can be picked. (To approximate a Dirichlet-$1/2$, 0-mean normal prior with variance 2π can be used.) Observing sparse $\mathbf{x}_t$, the prediction for y_t is given by

$$p_t \triangleq P(y_t|\mathbf{x}_t) = \int_{\mathbf{w}} p(y_t|\mathbf{x}_t, \mathbf{w})\rho_t(\mathbf{w})d\mathbf{w} \triangleq \int_{\mathbf{w}} p_t(y_t, \mathbf{w}|\mathbf{x}_t)d\mathbf{w}, \tag{1}$$

where for binary logistic regression, the probability of the label given the example and weights is given by the Sigmoid of the label weighted dot product of the example and weights

$$p(y_t|\mathbf{x}_t, \mathbf{w}) \triangleq \frac{1}{1+\exp\left(-y_t\mathbf{x}_t^{\mathcal{T}}\mathbf{w}\right)} \triangleq \text{Sigma}\left(y_t\mathbf{x}_t^{\mathcal{T}}\mathbf{w}\right). \tag{2}$$

The expected prediction p_t in (1) marginalizes out the weights $\mathbf{w}$ according to the prior $\rho_t(\cdot)$ from the joint probability of $\mathbf{w}$ and y_t. The prediction p_t is a function also of all prior pairs sequence $\{\mathbf{x}^{t-1}, y^{t-1}\}$ though the prior $\rho_t(\cdot)$. After observing y_t, we try to match a (diagonal) posterior $Q(\cdot)$ to the weights that will equal the next round's prior

$$\rho_{t+1}(\mathbf{w}) \triangleq Q_t(\mathbf{w}) \approx \frac{p(y_t|\mathbf{x}_t, \mathbf{w})\rho_t(\mathbf{w})}{P(y_t|\mathbf{x}_t)} = \frac{p(y_t|\mathbf{x}_t, \mathbf{w})\rho_t(\mathbf{w})}{p_t}. \tag{3}$$

Using $\mathcal{S}_T \triangleq \{\mathbf{x}^T, y^T\}$, the logarithmic loss incurred by approximation $Q(\cdot)$ on the sequence of predictions is $L(\mathcal{S}_T, Q) \triangleq -\sum_{t=1}^T \log p_t$. Let $\mathbf{w}^*$ be some fixed comparator in the parameter values' space. Then, the *regret* of approximation $Q(\cdot)$ relative to comparator $\mathbf{w}^*$ is given by

$$\mathcal{R}(\mathcal{S}_T, Q, \mathbf{w}^*) \triangleq L(\mathcal{S}_T, Q) - L(\mathcal{S}_T, \mathbf{w}^*) = -\sum_{t=1}^T \left[\log p_t + \log(1+\exp(-y_t\mathbf{x}_t^{\mathcal{T}}\mathbf{w}^*)\right]. \tag{4}$$

The regret can measure the excess loss relative to the best possible $\mathbf{w}^*$ comparator, if it is chosen.

3 MARGINALIZED BAYESIAN GAUSSIAN APPROXIMATION

Gaussian Approximation of a Sigmoid: The relation between the logistic distribution and the Normal one was well studied in the statistics literature (see, e.g., Bishop (2006); Murphy (2012)). The Sigmoid function in (2) can be viewed as a Cumulative Distribution Function (CDF), which can be approximated by a normal CDF $\Phi(z)$ (The inverse of $\Phi(\cdot)$ is the *Probit* function.) The derivative of the Sigmoid function is the 0-mean *Logistic* Probability Density Function (PDF). Matching the PDFs, we have $e^w/(1+e^w)^2 \approx 1/\sqrt{2\pi\sigma^2}\exp\left\{-w^2/2\sigma^2\right\}$. This yields that the Sigmoid function

can be approximated by a 0-mean Gaussian CDF with variance $8/\pi$. Using the standard 0-mean normal $\Phi(\cdot)$ function, the argument is scaled by the inverse of the standard deviation $\sqrt{\pi/8}$, giving

$$\text{Sigma}(w) \triangleq \frac{1}{1+e^{-w}} \approx \Phi\left(\sqrt{\frac{\pi}{8}} \cdot w\right). \tag{5}$$

Approximation approach and some notation: With the diagonal and Gaussian assumptions, for each sparse example (with only $d_t \ll d$ nonzero entries in $\mathbf{x}_t$), we can assume that we have a single normal random variable, whose mean is the $\mathbf{x}_t$ weighted mean, and whose variance is the quadratically weighted sum of variances. Denote the example total weight, mean, and variance by

$$w_t \triangleq \sum_{i=1}^{d} x_{i,t} w_{i,t}, \quad \mu_t \triangleq \sum_{i=1}^{d} x_{i,t} \cdot \mu_{i,t}, \quad \sigma_t^2 \triangleq \sum_{i=1}^{d} x_{i,t}^2 \cdot \sigma_{i,t}^2 \tag{6}$$

(where the diagonalization assumption is important for the simplicity of the approximation of σ_t^2).

Since we consider a sparse problem, there is benefit to breaking the dependencies between features present in a given example and updating each independently. We can achieve that by marginalizing the prior at t over all other features. Because we assume all features are jointly independent Gaussians, we can break the joint prior into a product of two components; one, the marginal of the feature, and the other the marginal of all other features together, i.e., the *self excluding prior*. To match the posterior, we then marginalize on the latter, and match a single dimensional posterior for each feature. We define the self excluding prior for feature i at time t, its mean and variance as

$$w_{-i,t} = \sum_{j=1}^{d} x_{j,t} w_{j,t} - x_{i,t} w_{i,t} = \sum_{j \neq i} x_{j,t} w_{j,t}; \quad \mu_{-i,t} \triangleq \mu_t - x_{i,t}\mu_{i,t}; \quad \sigma_{-i,t}^2 \triangleq \sigma_t^2 - x_{i,t}^2 \sigma_{i,t}^2. \tag{7}$$

Prediction: With the *probit* approximation in (5) and the single dimensional variable w_t, we can compute p_t in (1), replacing $p(y_t|\mathbf{x}_t, \mathbf{w})$ in (2) by a normal CDF. Approximating this integral (see, e.g. Murphy (2012), Section 8.4.4.2, and Bishop (2006)) gives

$$p_t \approx \text{Sigma}\left(\frac{y_t \mu_t}{\sqrt{1 + \frac{\pi}{8}\sigma_t^2}}\right). \tag{8}$$

This result demonstrates how the prediction variance shrinks the prediction towards probability 0.5.

Marginalization: Given the diagonalization assumption, the prior at t can be expressed as $\rho_t(\mathbf{w}) = \rho_{i,t}(w_i) \cdot \rho_{-i,t}(w_{-i})$, where $\rho_{-i,t}(\cdot)$ is the prior on the self excluding prior of w_i. Hence,

$$p(y_t, \mathbf{w}|\mathbf{x}_t) = p(y_t|\mathbf{x}_t, \mathbf{w})\rho_{i,t}(w_i)\rho_{-i,t}(w_{-i}). \tag{9}$$

Marginalizing on w_{-i} gives

$$p(y_t, w_i|\mathbf{x}_t) = \rho_{i,t}(w_i) \int_{-\infty}^{\infty} p(y_t|\mathbf{x}_t, \mathbf{w})\rho_{-i,t}(w_{-i}) dw_{-i} \triangleq \rho_{i,t}(w_i) I_{w_{-i},t}. \tag{10}$$

The inner integral, which marginalizes over w_{-i} with its prior $\rho_{-i,t}(w_{-i})$, can be approximated by

$$\begin{aligned}
I_{w_{-i},t} &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma_{-i,t}^2}} \exp\left(-\frac{(w_{-i} - \mu_{-i,t})^2}{2\sigma_{-i,t}^2}\right) \cdot \text{Sigma}\left[y_t(x_{i,t}w_i + w_{-i})\right] dw_{-i} \\
&\overset{(a)}{\approx} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma_{-i,t}^2}} \exp\left(-\frac{(w_{-i} - \mu_{-i,t})^2}{2\sigma_{-i,t}^2}\right) \cdot \Phi\left[\sqrt{\frac{\pi}{8}} y_t(x_{i,t}w_i + w_{-i})\right] dw_{-i} \\
&\overset{(b)}{=} \int_{-\infty}^{\infty} \phi(z) \cdot \Phi\left[\sqrt{\frac{\pi}{8}} y_t(x_{i,t}w_i + \mu_{-i,t} + \sigma_{-i,t}z)\right] dz \\
&\overset{(c)}{=} \Phi\left(\frac{\sqrt{\frac{\pi}{8}} y_t(\mu_{-i,t} + x_{i,t}w_i)}{\sqrt{1 + \frac{\pi}{8}\sigma_{-i,t}^2}}\right) \\
&\overset{(d)}{=} \text{Sigma}\left(\frac{y_t(\mu_{-i,t} + x_{i,t}w_i)}{\sqrt{1 + \frac{\pi}{8}\sigma_{-i,t}^2}}\right).
\end{aligned} \tag{11}$$

Step (a) follows, again, from the approximation in (5). For (b), we apply the change of variables $z = (w_{-i} - \mu_{-i,t})/\sigma_{-i,t}$, where $\phi(\cdot)$ is the standard Gaussian PDF. The integral in (b) gives $\Phi\left(\frac{a}{\sqrt{1+b^2}}\right)$, with $a = \sqrt{\frac{\pi}{8}} y_t(\mu_{-i,t} + x_{i,t} w_i)$ and $b^2 = \frac{\pi}{8}\sigma^2_{-i,t}$ to lead to (c). Finally, the approximation in (5) is used to go back from a Normal CDF to a Sigmoid in (d).

Posterior: The posterior on w_i is given by plugging (11) into (10) normalizing by p_t given in (1).

$$\rho_{i,t+1}(w_i) = Q_{i,t}(w_i) \approx \frac{1}{p_t} \cdot \rho_{i,t}(w_i) \cdot \text{Sigma}\left(\frac{y_t(\mu_{-i,t} + x_{i,t} w_i)}{\sqrt{1 + \frac{\pi}{8}\sigma^2_{-i,t}}}\right). \tag{12}$$

The approximation on the right implies matching the current true posterior with the ith component of the approximate posterior $Q(\cdot)$. It can be simplified to

$$\frac{1}{\sigma_{i,t+1}} \exp\left(-\frac{(w_i - \mu_{i,t+1})^2}{2\sigma^2_{i,t+1}}\right) \approx \frac{1}{p_t \sigma_{i,t}} \exp\left(-\frac{(w_i - \mu_{i,t})^2}{2\sigma^2_{i,t}}\right) \cdot \text{Sigma}\left(\frac{y_t(\mu_{-i,t} + x_{i,t} w_i)}{\sqrt{1 + \frac{\pi}{8}\sigma^2_{-i,t}}}\right). \tag{13}$$

Approximations: Because the functional form of the posterior is not Gaussian, there are multiple ways to fit a Gaussian. We review alternatives in Appendix B. However, we want to ensure that the regions of the true posterior we are most likely to converge to at the horizon are not scaled down too much, as this will incur additional loss. It is thus desirable to match the peak of the true posterior with the peak of the approximation. One method is to match both the location and hight of the peak. The other, Laplace approximation (Bishop, 2006), matches the location and curvature at the peak. Both methods give the same approximate for $\mu_{i,t+1}$, but a somewhat different one for $\sigma^2_{i,t+1}$.

To give $\mu_{i,t+1}$, we find w_i that maximizes the r.h.s. of (13), or minimizes its negative logarithm. Let

$$p_{i,t+} \triangleq \text{Sigma}\left(\frac{y_t\left(\mu_{-i,t} + x_{i,t}\mu_{i,t+1}\right)}{\sqrt{1 + \frac{\pi}{8}\sigma^2_{-i,t}}}\right) = \left[1 + \exp\left(-\frac{y_t\left(\mu_{-i,t} + x_{i,t}\mu_{i,t+1}\right)}{\sqrt{1 + \frac{\pi}{8}\sigma^2_{-i,t}}}\right)\right]^{-1} \tag{14}$$

be almost p_t in (8), except that $\mu_{i,t}$ is replaced by $\mu_{i,t+1}$ and σ^2_t by $\sigma^2_{-i,t}$, i.e., $p_{i,t+}$ is the probability predicted for y_t if we update $\mu_{i,t}$ and shrink the prediction as function of the self excluding variance. The minimization gives

$$\mu_{i,t+1} = \mu_{i,t} + \frac{y_t x_{i,t} \sigma^2_{i,t}}{\sqrt{1 + \frac{\pi}{8}\sigma^2_{-i,t}}} \cdot \left(1 - p_{i,t+}\right). \tag{15}$$

Eq. (15) can be solved iteratively, where the Newton's method can be used, as described in Appendix A. The solution for $\mu_{i,t+1}$ can also be expressed in terms of the r *generalized Lambert W function* (Corless et al., 1996; Mezo & Baricz, 2015).

Alternatively, to avoid multiple iterations per update when using Newton's method, we can use a Taylor series approximation of $1 - p_{i,t+}$ around $1 - p_{i,t}$, where

$$p_{i,t} \triangleq \text{Sigma}\left(\frac{y_t\left(\mu_{-i,t} + x_{i,t}\mu_{i,t}\right)}{\sqrt{1 + \frac{\pi}{8}\sigma^2_{-i,t}}}\right) = \text{Sigma}\left(\frac{y_t \mu_t}{\sqrt{1 + \frac{\pi}{8}\sigma^2_{-i,t}}}\right). \tag{16}$$

Like $p_{i,t+}$, $p_{i,t}$ is not p_t. Instead, it is the probability of label y_t as projected by the means of the weights at t, shrunk as a function of the self excluding variance of feature i instead of the full example predicted variance. More importantly, it is a function of parameters before the update at $t+1$ is applied, giving a closed form solution, instead of an iterative one. Applying first order approximation we have

$$\mu_{i,t+1} = \mu_{i,t} + \frac{y_t x_{i,t} \sigma^2_{i,t}\left(1 - p_{i,t}\right)}{\sqrt{1 + \frac{\pi}{8}\sigma^2_{-i,t}}\left[1 + \frac{1}{1 + \frac{\pi}{8}\sigma^2_{-i,t}} y_t^2 x^2_{i,t} \sigma^2_{i,t}\left(1 - p_{i,t}\right) p_{i,t}\right]}. \tag{17}$$

Second or higher orders approximations can also be applied, but may not be necessary, as the first order approximation already gives identical performance to the iterative method.

After updating $\mu_{i,t+1}$, we can apply (15) to update $\sigma_{i,t+1}$. Plugging (15) in (13), we solve for $\sigma_{i,t+1}$,

$$\sigma_{i,t+1} = \frac{p_t \sigma_{i,t}}{p_{i,t+}} \cdot \exp\left\{ \frac{(\mu_{i,t+1} - \mu_{i,t})^2}{2\sigma_{i,t}^2} \right\} = \frac{p_t \sigma_{i,t}}{p_{i,t+}} \cdot \exp\left\{ \frac{y_t^2 x_{i,t}^2 \sigma_{i,t}^2}{2(1 + \frac{\pi}{8}\sigma_{-i,t}^2)} \cdot (1 - p_{i,t+})^2 \right\}. \quad (18)$$

Alternatively to (18), Laplace approximation can be used by finding the second derivative of the negative logarithm of the posterior, giving

$$\sigma_{i,t+1}^2 = \left[\frac{1}{\sigma_{i,t}^2} + \frac{y_t^2 x_{i,t}^2}{1 + \frac{\pi}{8}\sigma_{-i,t}^2} \cdot p_{i,t+} \cdot (1 - p_{i,t+}) \right]^{-1}. \quad (19)$$

The procedures described are summarized in Algorithm 1. As described in Appendix B, different methods and approximations can be used. We can update the mean vector and covariance of d_t features that occur at t, without marginalization, discarding off-diagonal covariance terms. One can choose any two points on the true posterior and match $Q(\cdot)$ on these. Least squares can be used with several points that represent a region of w_i, for which the posterior at T is likely to have most mass. VB and EP can be applied. Specifically, single dimensional VB on the marginalized posterior can be applied to minimize $KL(Q_{i,t}(w_i)||p(w_i|\mathbf{x}_t, y_t))$, where $p(w_i|\mathbf{x}_t, y_t)$ is the true posterior on the r.h.s. of (12) given by $p(y_t, w_i|\mathbf{x}_t)/p_t$, where $p(y_t, w_i|\mathbf{x}_t)$ is in (10). This still requires expensive Monte Carlo estimation of expectations with an iterative Newton method. Without marginalization, however, a similar VB approach would require even more expensive Monte Carlo sampling or iterative mean field approximation EM to converge on all components. As empirical results, in the next section, show, there is no performance advantage to applying any of the more involved methods.

Algorithm 1 Marginalized Bayesian Gaussian Approximation

procedure MARGINALIZED BAYESIAN GAUSSIAN APPROXIMATION(Parameters: μ_0, σ_0^2)
  $\forall i \in 1, \ldots, d$; $\mu_{i,1} \leftarrow \mu_0$, $\sigma_{i,1}^2 \leftarrow \sigma_0^2$.
  for t=1,2,...,T **do**
    Get $\mathbf{x}_t$.
    Compute μ_t, σ_t^2 with (6).
    Generate p_t for $y_t \in \{-1, 1\}$ with (8).
    Observe y_t.
    for $i : x_{i,t} \neq 0$ **do**
      Compute $p_{i,t}$ $p_{i,t+}$ with (16) and (14), respectively, using $\mu_{i,t+1} = \mu_{i,t}$.
      Iterate on (15) and (14) with Newton's method, or use (17) to update $\mu_{i,t+1}$.
      Update $\sigma_{i,t+1}^2$ with either (18) or (19)
    end for
  end for
end procedure

4 NUMERICAL RESULTS

We simulated data by setting d features with "true" log-odds weights that were drawn randomly with some prior (or in several sets, each of which with its own prior). At t, $d_t \ll d$ features were selected in random from the set of d features, where different rules were used for different simulations to draw the d_t features. In the fully random case, we set a random fraction $\alpha = d_t/d$ parameter, and each feature was activated with probability α. We either used binary features $x_{i,t} \in \{0, 1\}$, or also drew $x_{i,t}$ randomly in $[0, 1]$. For categorical features, the d features were partitioned into categories, and for every example, one or a few features from each category were selected at random, with different types of randomness, including fast decaying distributions with long tails. The "true" log-odds of the features drawn at t were added up, weighted by $x_{i,t}$, and converted to probability with the Sigmoid function. The probability was used to randomly draw y_t. We used Algorithm 1, as well as other algorithms, described in Appendix B, to sequentially predict y_t and update. For gradient methods, we used *Stochastic Gradient Descent (SGD)* with *AdaGrad* scheduling (Duchi et al., 2011).

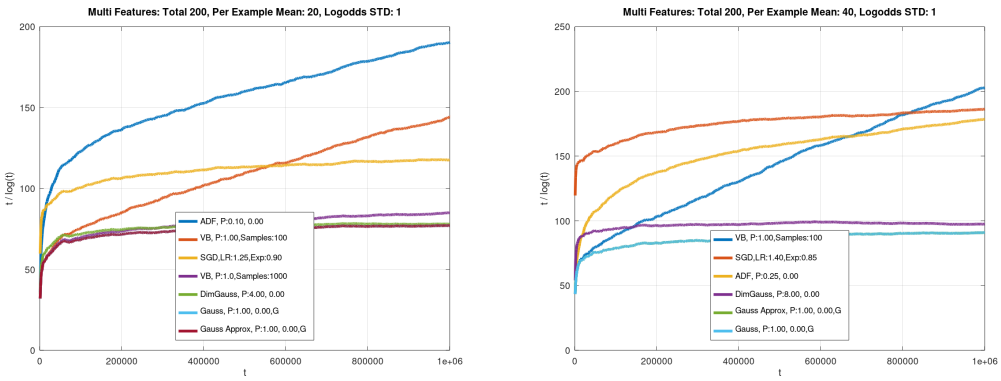

Figure 1: Regret normalized by $\log t$ vs. round t for different methods for randomly drawn d binary features, expected d_t/d, and standard deviation of "true" log-odds noted in each graph.

We ran grids of algorithm hyper-parameters for all algorithms to find optimal ones, and we show results for these parameters for all algorithms. Since we know the "true" weights, we use them as $\mathbf{w}^*$. Results describe a time series on (4), where summation is on τ up to example t at time t. We normalize the regret up to t by $\log t$. If the algorithm exhibits logarithmic regret, such normalized regret converges to a constant horizontal curve. Results are shown for Algorithm 1 with its different variations (labeled by *Gauss* for updates with (15), and by *Gauss Approx* with (17)). Labels also designate the prior used (σ_0^2 and μ_0), and which variance approximation was used (G for (18) and L for (19)). Reference results are shown for SGD, multi-dimensional Gaussian approximation update (DimGauss), EP; using *Assumed Density Filtering (ADF)*, and marginalized VB (VBApprox).

Fig. 1 shows normalized regret for two different "true" data configurations described at the top of each graph with random binary features. More detailed results on multiple configurations are shown in Appendix C. Unfortunately, unlike theoretical results (Shamir, 2020), the prior has to fit the data for all methods of Gaussian approximation for good regret. However, if we choose a prior that matches the "true" prior, regret rates matching the lower bounds of $0.5 \log T$ per parameters are achieved consistently with Algorithm 1 on the various configurations. The same results are achieved whether we use the iterative version in (15) or its simpler approximation (17), and whether the variance is updated by (18) or (19). Therefore, it is much more efficient to use (17) for updating $\mu_{i,t}$ with one of the updates of $\sigma_{i,t}^2$. Using vector updates with Laplace approximation is inferior to the marginalization approach because there are no correlations in the real data. The prior needed for the high dimensional update is larger than the "true" prior, and grows with the fraction α of features in an example. While we can find SGD hyper-parameters that seem to exhibit logarithmic regret for each configuration, these converge to regrets that are larger than that of Algorithm 1 and increase with more active features. Marginalized VB requires 1000 MC random samples to approach the performance of Algorithm 1, where even with 100 samples there is substantial degradation. It thus runs much slower. EP seems to be unable to achieve logarithmic regret altogether in most cases.

5 Conclusions

We introduced a simple Bayesian mixture diagonal Gaussian approximation method based on marginalization for sparse online logistic regression that attempts to retain the affects of a good prior around the optimal weights. The method does not require the complexities of standard Bayesian methods, as VB, but was empirically shown to achieve regret rates as good and even better. With proper priors, it matches regret lower bounds. With the strong relation between regret and uncertainty, this approach gives good uncertainty estimates. The methodology was proposed for logistic regression, but can be extended to other settings, from Gaussian regression and beyond.

REFERENCES

Mohammed Abdellaoui. Parameter-free elicitation of utility and probability weighting functions. *Management science*, 46(11):1497–1512, 2000.

Terry Anderson. *The theory and practice of online learning*. Athabasca University Press, 2008.

F. Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414, 2010.

F. Bach. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *Journal of Machine Learning Research*, 15(1):595–627, 2014.

F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $o(1/n)$. In *Advances in neural information processing systems*, pp. 773–781, 2013.

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2006.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. *arXiv preprint arXiv:1505.05424*, 2015.

L. Bottou. Large-scale machine learning with stochastic gradient descent. In *COMPSTAT*, 2010.

Léon Bottou. Online learning and stochastic approximations. *On-line learning in neural networks*, 17(9):142, 1998.

Tamara Broderick, Nicholas Boyd, Andre Wibisono, Ashia C Wilson, and Michael I Jordan. Streaming variational bayes. In *Advances in neural information processing systems*, pp. 1727–1735, 2013.

Wei Chu, Martin Zinkevich, Lihong Li, Achint Thomas, and Belle Tseng. Unbiased online active learning in data streams. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 195–203, 2011.

Bertrand S Clarke and Andrew R Barron. Jeffreys' prior is asymptotically least favorable under entropy risk. *Journal of Statistical planning and Inference*, 41(1):37–60, 1994.

R. Corless, G. Gonnet, D. Hare, D. Jeffrey, and D. Knuth. On the lambert w function. *Adv. Computational Mathematics*, 5:329–359, 1996.

John P Cunningham, Philipp Hennig, and Simon Lacoste-Julien. Gaussian probabilities and expectation propagation. *arXiv preprint arXiv:1111.6832*, 2011.

Arthur P Dempster. A generalization of bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(2):205–232, 1968.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1): 1–22, 1977.

M. Drmota and W. Szpankowski. Precise minimax redundancy and regrets. *IEEE Trans. Inf. Theory*, IT-50:2686–2707, 2004.

J. Drugowitsch. Variational bayesian inference for linear and logistic regression. *arXiv:1310.5438v4*, 2019.

Jan Drugowitsch, André G Mendonça, Zachary F Mainen, and Alexandre Pouget. Learning optimal decisions with confidence. *Proceedings of the National Academy of Sciences*, 116(49):24872–24880, 2019.

J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, Feb. 2011.

Dylan J Foster, Satyen Kale, Haipeng Luo andMehryar Mohri, and Karthik Sridharan. Logistic regression: The importance of being improper. In *COLT - Conference on Learning Theory*, 2018.

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. Omnipress, 2010.

P. D. Grunwald. *The Minimum Description Length Principle*. MIT Press, 2007.

E. Hazan, T. Koren, and K. Y. Levy. Logistic regression: Tight bounds for stochastic and online optimization. In *The 27th Conference on Learning Theory, COLT 2014*, pp. 197–209. MIT press, 2014.

Ralf Herbrich, Neil D Lawrence, and Matthias Seeger. Fast sparse gaussian process methods: The informative vector machine. In *Advances in neural information processing systems*, pp. 625–632, 2003.

Matthew Hoffman, Francis R. Bach, and David M. Blei. Online learning for latent dirichlet allocation. In *Advances in Neural Information Processing Systems 23*, pp. 856–864. Curran Associates, Inc., 2010.

John P Huelsenbeck and Fredrik Ronquist. Mrbayes: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8):754–755, 2001.

Seong Jae Hwang, Ronak R Mehta, Hyunwoo J Kim, Sterling C Johnson, and Vikas Singh. Sampling-free uncertainty estimation in gated recurrent units with applications to normative modeling in neuroimaging. In *Uncertainty in Artificial Intelligence*, pp. 809–819. PMLR, 2020.

Tommi S Jaakkola and Michael I Jordan. Improving the mean field approximation via the use of mixture distributions. In *Learning in graphical models*, pp. 163–173. Springer, 1998.

Rémi Jézéquel, Pierre Gaillard, and Alessandro Rudi. Efficient improper learning for online logistic regression. In *COLT - Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2085–2108. PMLR, 09–12 Jul 2020.

S. M. Kakade and A. Y. Ng. Online bounds for bayesian algorithms. In *NIPS*, pp. 641–648, 2005.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, pp. 5574–5584, 2017.

David C Knill and Whitman Richards. *Perception as Bayesian inference*. Cambridge University Press, 1996.

David A Knowles. Stochastic gradient variational bayes for gamma approximating distributions. *arXiv preprint arXiv:1509.01631*, 2015.

R. E. Krichevsky and V. K. Trofimov. he performance of universal encoding. *IEEE Trans. Inform. Theory*, IT-27(2):199?207, Mar. 1981.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in neural information processing systems*, pp. 6402–6413, 2017.

David Makowski, Daniel Wallach, and Marie Tremblay. Using a bayesian approach to parameter estimation; comparison of the glue and mcmc methods. *Agronomie*, 22(2):191–203, 2002.

Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. In *Advances in Neural Information Processing Systems*, pp. 7047–7058, 2018.

Brendan Mcmahan and Matthew Streeter. No-regret algorithms for unconstrained online convex optimization. In *Advances in neural information processing systems*, pp. 2402–2410, 2012.

H. B. McMahan. Follow-the-regularized-leader and mirror descent: Equivalence theorems and L1 regularization. In *AISTATS*, 2011.

H. B. McMahan and M. J. Streeter. Open problem: Better bounds for online logistic regression. In *Journal of Machine Learning Research-Proceedings Track, 23*, 2012.

I. Mezo and A. Baricz. On the generalization of the lambert w function. *arXiv:1408.3999v2*, 2015.

Thomas P Minka. Expectation propagation for approximate bayesian inference. *arXiv preprint arXiv:1301.2294*, 2013.

Thomas Peter Minka. *A family of algorithms for approximate Bayesian inference*. PhD thesis, Massachusetts Institute of Technology, 2001.

Lucie Montuelle, Erwan Le Pennec, and Serge Cohen. Gaussian mixture regression model with logistic weights, a penalized maximum likelihood approach. *arXiv preprint arXiv:1304.2696*, 2013.

Lucie Montuelle, Erwan Le Pennec, et al. Mixture of gaussian regressions model with logistic weights, a penalized maximum likelihood approach. *Electronic Journal of Statistics*, 8(1):1661–1695, 2014.

Todd K Moon. The expectation-maximization algorithm. *IEEE Signal processing magazine*, 13(6): 47–60, 1996.

K. P. Murphy. *Machine Learning A Probabilistic Perspective*. MIT Press, Cambridge, Massachusetts, 2012.

Cuong V Nguyen, Thang D Bui, Yingzhen Li, and Richard E Turner. Online variational bayesian inference: Algorithms for sparse gaussian processes and theoretical bounds. 2017a.

Cuong V Nguyen, Yingzhen Li, Thang D Bui, and Richard E Turner. Variational continual learning. *arXiv preprint arXiv:1710.10628*, 2017b.

Manfred Opper and Ole Winther. A bayesian approach to on-line learning. *On-line learning in neural networks*, pp. 363–378, 1998.

Francesco Orabona and Dávid Pál. Coin betting and parameter-free online learning. In *Advances in Neural Information Processing Systems*, pp. 577–585, 2016.

Rajesh Ranganath, Sean Gerrish, and David Blei. Black Box Variational Inference. volume 33 of *Proceedings of Machine Learning Research*, pp. 814–822. PMLR, 2014.

Carl Edward Rasmussen. The infinite gaussian mixture model. In *Advances in neural information processing systems*, pp. 554–560, 2000.

Douglas A Reynolds, Thomas F Quatieri, and Robert B Dunn. Speaker verification using adapted gaussian mixture models. *Digital signal processing*, 10(1-3):19–41, 2000.

J. Rissanen. Modeling by the shortest data description. *Automatica*, 14:465–471, 1978a.

J. Rissanen. Minimax codes for finite alphabets. *EEE Trans. Inform. Theory*, IT-24(3):389–392, 1978b.

J. Rissanen. Universal coding, information, prediction, and estimation. *IEEE Trans. Inform. Theory*, IT-30(4):629–636, Jul. 1984.

J. Rissanen. Stochastic complexity and modeling. *The Annals of Statistics*, 14:1080–1100, 1986.

Mrinal K Sen and Paul L Stoffa. Bayesian inference, gibbs' sampler and uncertainty estimation in geophysical inversion 1. *Geophysical Prospecting*, 44(2):313–350, 1996.

Shai Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and trends in Machine Learning*, 4(2):107–194, 2011.

Gil I. Shamir. Minimum description length (MDL) regularization for online learning. In *Proceedings of the 1st International Workshop on Feature Extraction: Modern Questions and Challenges at NIPS 2015*, pp. 260–276, 2015.

Gil I Shamir. Logistic regression regret: What's the catch? In *COLT - Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 3296–3319, 2020.

Hsi Guang Sung. *Gaussian mixture regression and classification*. PhD thesis, 2004.

Andrew Gordon Wilson. The case for bayesian deep learning. *arXiv preprint arXiv:2001.10995*, 2020.

Q. Xie and A. Barron. Minimax redundancy for the class of memoryless sources. *IEEE Trans. Information Theory*, pp. 647–657, 1997.

Q. Xie and A. Barron. Asymptotic minimax regret for data compression, gambling, and prediction. *IEEE Trans. Information Theory*, 46:431–445, 2000.

M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *ICML*, pp. 928–936, 2003.

A Mean and Variance Updates

Eq. (15) gives an update for the mean $\mu_{i,t+1}$ that cannot be solved in closed form. This is beacuse $p_{i,t+}$ is a function of $\mu_{i,t+1}$. However, the update is easily solvable with a few iterations of Newton's method. We start by plugging $\mu_{i,t+1}^{(\ell=0)} = \mu_{i,t}$ for iteration $\ell = 0$. The solution for $\mu_{i,t+1}$ is the value of w_i that minimizes the negative logarithm of the r.h.s. of (12) where $\rho_{i,t}(w_i)$ is Gaussian with mean $\mu_{i,t}$ and variance $\sigma_{i,t}^2$. At iteration ℓ, we can compute $p_{i,t+}^{(\ell)}$ with (14), using $\mu_{i,t+1}^{(\ell-1)}$. The gradient of the negative logarithm of the posterior w.r.t. w_i at $w_i = \mu_{i,t+1}^{(\ell-1)}$ is given by

$$g_\ell = \frac{\mu_{i,t+1}^{(\ell-1)} - \mu_{i,t}}{\sigma_{i,t}^2} - \frac{y_t x_{i,t}}{\sqrt{1 + \frac{\pi}{8}\sigma_{-i,t}^2}} \cdot \left(1 - p_{i,t+}^{(\ell)}\right). \tag{20}$$

The second derivative w.r.t. w_i is given by

$$h_\ell = \frac{1}{\sigma_{i,t}^2} + \frac{y_t^2 x_{i,t}^2}{1 + \frac{\pi}{8}\sigma_{-i,t}^2} \cdot p_{i,t+}^{(\ell)} \cdot \left(1 - p_{i,t+}^{(\ell)}\right) \tag{21}$$

Then, $\mu_{i,t+1}$ is updated by

$$\mu_{i,t+1}^{(\ell)} = \mu_{i,t+1}^{(\ell-1)} - \frac{g_\ell}{h_\ell}. \tag{22}$$

The process terminates when the difference $\left|\mu_{i,t+1}^{(\ell)} - \mu_{i,t+1}^{(\ell-1)}\right|$ is smaller than some threshold, or a specified number of iterations had already executed.

Eq. (18) gives an update to the variance. It is interesting to interpret some observations from this update. For $x_{i,t} \geq 0$, $p_{i,t+} \geq p_t$. This leads to $y_t x_{i,t}(\mu_{i,t+1} - \mu_{i,t}) > 0$ because for every y_t, $\mu_{i,t+1}$ has to move away from $\mu_{i,t}$ with the sign of y_t. This implies that we add a positive term from the log-odds converted to p_t to those converted into $p_{i,t+}$ and also apply less shrinkage using $\sigma_{-i,t}^2$ instead of σ_t^2, yielding this claim. Hence, the coefficient of $\sigma_{i,t}$ outside the exponential term in (18) is upper bounded by 1. The argument of the exponential term is always nonnegative, which implies that the exponential term is lower bounded by 1. However, if the self-excluding variance $\sigma_{-i,t}^2$ is large, it makes the argument small. Similarly, if $p_{i,t+}$ is large relative to p_t it makes the whole expression smaller. These two observations imply that if the self excluding prior is less certain, or if the self excluding prior has a different belief about the label from the one observed overall, the uncertainty of the current feature reduces more, because it is deemed responsible for the observation y_t. On the other hand, if the opposite holds, i.e., either the uncertainty of the self excluding prior is low, or the self excluding prior agrees more with the observed label y_t, then, feature i matters less for the observation, and therefore, its uncertainty is not reduced as much.

B Summary of Other Methods

We briefly describe the updates we performed with other Gaussian approximation methods.

B.1 MULTI-DIMENSIONAL GAUSSIAN APPROXIMATION

Instead of marginalizing on all other features to update w_i, for which $x_{i,t} \neq 0$, we can apply multi-dimensional update on all features at round t, for which $x_{i,t} \neq 0$. For that we assume that the true posterior consists of a product between a diagonal covariance prior and a Sigmoid, and we apply Lapace approximation to obtain a new mean vector and covariance. We define

$$p_{t+} \triangleq \text{Sigma}\left(y_t \mathbf{x}_t^{\mathcal{T}} \mathbf{u}_{t+1}\right) \tag{23}$$

as the probability of y_t with the value of weights after they have been updated (with no shrinkage), where $\mathbf{u}_{t+1}$ is the updated vector of means. Let Σ_t be the diagonal covariance matrix, taking the values of $\sigma_{i,t}^2$. Then, applying Laplace approximation gives

$$\mathbf{g} = \Sigma_t^{-1}(\mathbf{w} - \mathbf{u}_t) - y_t \mathbf{x}_t \cdot (1 - p_{t+}) \tag{24}$$

and

$$H = \Sigma_t^{-1} + y_t^2 \mathbf{x}_t \mathbf{x}_t^{\mathcal{T}} p_{t+}(1 - p_{t+}). \tag{25}$$

These equations can, again, be used to apply Newton's method to find the solution $\mathbf{u}_{t+1}$ for $\mathbf{w}$ that minimizes the negative logarithm of the posterior. Inverting H gives the updated covariance, whose diagonal elements can be now used to update $\sigma_{i,t+1}^2$.

In the sparse problem, however, this approach may utilize correlations that are not there, and as empirical results suggest may not be as good as the marginalization approach.

B.2 EXPECTATION PROPAGATION - ASSUMED DENSITY FILTERING

Instead of minimizing the divergence between the approximate Q and the true posterior, we can use the expectation propagation approach, as proposed in Minka (2001), which essentially minimizes the opposite KL divergence, and attempts to match the first two moments. More details can be found in Minka (2001).

B.3 MARGINALIZED VARIATIONAL BAYES

Instead of using Laplace approximation or matching the location of the peak of the true posterior and the estimated one together with either its curvature or its value, we can apply a full VB, by matching the approximate posterior Q with the true one through minimizing the KL divergence between Q and the true posterior. This can be done on the posterior we obtained in (12) on feature i after we marginalized on all other features. We need not perform full VB on all dimensions, instead we can apply it on each i individually, directly on this posterior. This KL divergence is decomposed into $KL(Q_{i,t}(w_i)||p(w_i|\mathbf{x}_t, y_t)) = KL(Q_{i,t}||\rho_{i,t}) + E_{Q_{i,t}}\left[\log\left\{1 + \exp\left(-\frac{y_t(\mu_{-i,t} + x_{i,t}w_i)}{\sqrt{1+\frac{\pi}{8}\sigma_{-i,t}^2}}\right)\right\}\right]$. The KL term can be computed in closed form. For the second term, w_i is replaced by $\mu_{i,t+1} + S\sigma_{i,t+1}$, where S is randomly drawn form $\mathcal{N}(0,1)$, to obtain a Monte Carlo estimate of the expected approximation. Then, Newton method can be used to solve resulting equations for both $\mu_{i,t+1}$ and $\sigma_{i,t+1}^2$. Note that without marginalization, a similar VB approach would require expensive Monte Carlo sampling or iterative mean field approximation EM to converge on all components.

C ADDITIONAL EMPIRICAL RESULTS

In this appendix, we bring a collection of simulation results demonstrating the performance of Algorithm 1 and other methods in different settings. In all the simulations we performed we observed that Algorithm 1, with a proper prior for the setting, consistently gives regret rates that are at the rate of the lower bound (e.g., $0.5 \log T$ cost for each unknown parameter). The other methods, while in some cases exhibit performance close to that of Algorithm 1, fail to do so consistently in all conditions. The marginalized VB approach requires very large complexity as shown in Fig 1 to approach the performance of Algorithm 1.

Fig. 2 gives a two dimensional grid of varying d_t and varying "true" feature weights. In all these simulations, Algorithm 1 with prior matched to the "true" one, gives minimal regret curves. While SGD

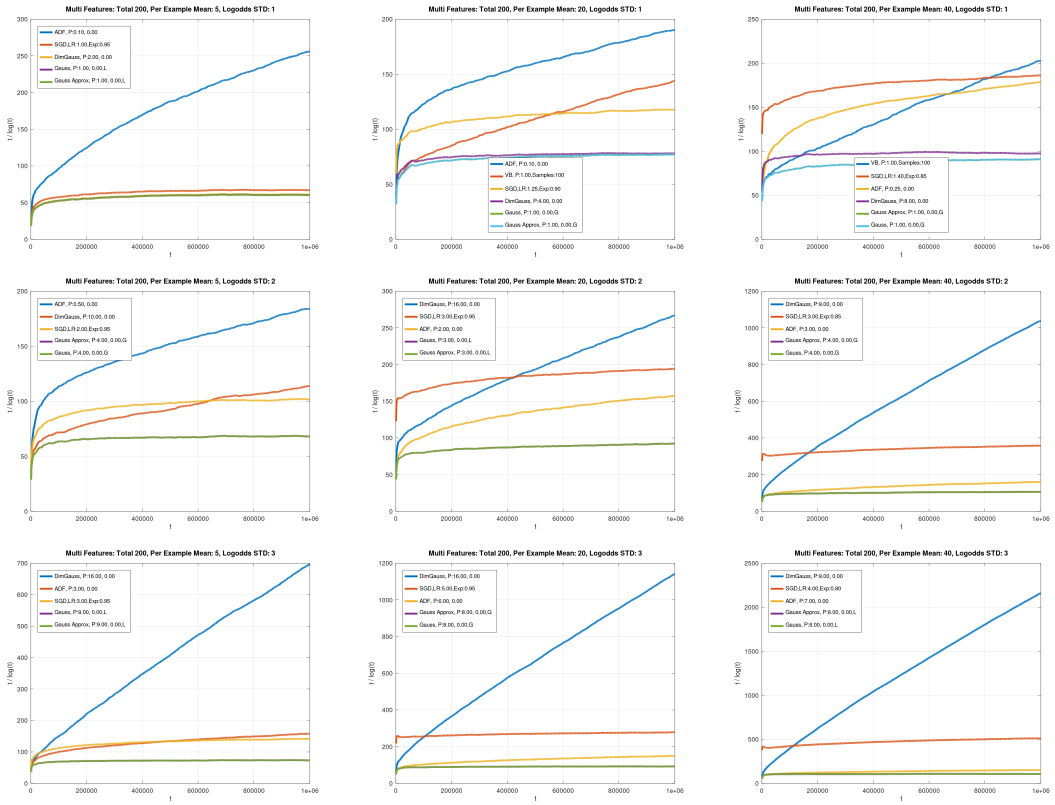

Figure 2: Regret normalized by $\log t$ vs. round t for various methods with randomly drawn binary features, with d, expected d_t/d, and standard deviation of "true" log-odds noted in each graph. Graphs shown for $d = 200$, $E[d_t] \in \{5, 20, 40\}$ and "true" log-odds std in $\{1, 2, 3\}$.

seems, with the right choices of parameters, to approach logarithmic regret, its regret is larger, and increases with the "true" parameters' variance and number of active features d_t. Multi-dimensional Gaussian approximation appears to be mis-calibrated on the prior, and depending on the feature density d_t tends to become better only with much larger priors. EP ADF gives larger regrets and while appearing reasonable with higher feature density, seems to be inferior with lower feature densities. Similar results are obtained when replacing the data generation model with models with uniform priors on the feature weights with various ranges.

Fig. 3 shows curves for simulations with different models. On the left, features are nonbinary, and on the right, the model consists of order of magnitude more features, where in average an order of magnitude more features occur in each example. In both cases, Algorithm 1 persists with similar regret rates, whereas other algorithms exhibit larger regrets. With a large number of features, both the multi dimensional Gaussian approximation and SGD have larger regrets, although with even lower prior / learning rates slightly better regret may be possible.

Fig 4 demonstrates curves of the various algorithms for categorical features, where in each example a fixed set of features are selected from each category. On the right, the selection gives exponentially decaying distribution over the features, so that some features are selected very often while others rarely occur. Again, regret rates behave similarly for Algorithm 1 and SGD. The multi-dimensional Gaussian approximation completely breaks in this setting. This is because it hypothesizes correlations in its updates with features that rarely reoccur.

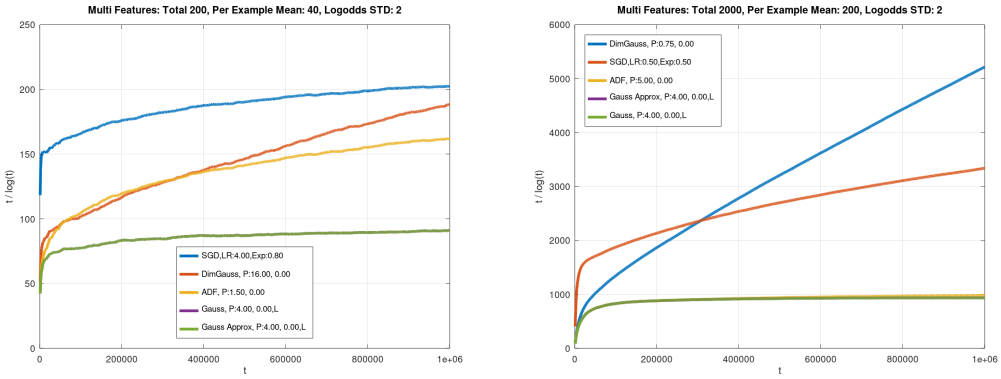

Figure 3: Regret normalized by $\log t$ vs. round t for different algorithms and different data generation models. On the left, a model with $d = 200$ features, out of which in average $d_t = 40$ occur in an example, and weight standard deviation is 2, with nonbinary feature values uniform in $[0, 1]$. Right: models with $d = 2000$, average $d_t = 200$, data generation standard deviation of 2, with binary features.

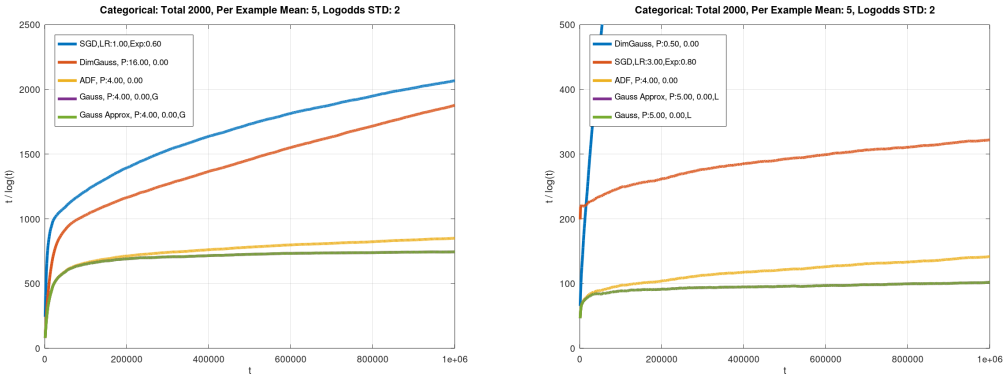

Figure 4: Regret normalized by $\log t$ vs. round t for categorical models and different algorithms. Left: 2000 features in 10 categories of 200 features each, where in each example 5 features from each category are present with "true" weight standard deviation of 2. On the right: A similar setting, except that in each category features are selected with a long tail exponential distribution (prioritizing a few features and rarely selecting others).