Online Bounds for Bayesian Algorithms

Sham M. Kakade
Computer and Information Science Department
University of Pennsylvania

Andrew Y. Ng
Computer Science Department
Stanford University

Abstract

We provide competitive online bounds for a number of Bayesian algorithms, without making any assumptions on the data generation process. These bounds show that some of the most popular Bayesian algorithms (such as least squares regression and Bayesian logistic regression) perform favorably when compared, in retrospect, to the single best model in the model class. Interestingly, our lower bound also shows that the performance on the worst case sequence of examples is sometimes not much different from the performance on the worst i.i.d. sequence of examples. We also provide bounds for some widely used maximum a posteriori (MAP) estimation algorithms, including regularized logistic regression.

1 Introduction

The philosophy taken in the Bayesian methodology is often at odds with that in the online learning community. In the Bayesian methodology, one often makes rather stringent assumptions that the data are being generated from some known model. In contrast, the online learning setting makes rather minimal assumptions on the conditions under which the data are being presented to the learner — usually, Nature could provide examples in an adversarial manner.

We study the performance of Bayesian algorithms in a more adversarial setting, where the process generating the data are not restricted to come from the prior and may be from an adversary. Our motivation is party due to the fact that, in practice, models are often chosen to balance realism with computational tractability and often assumptions made by the Bayesian are not believed to hold (*eg* i.i.d. assumptions). Furthermore, due to the close relationship between Bayes rule and the multiplicative weights algorithms (see Freund and Schapire [1999], Cesa-Bianchi et al. [1997]) which have strong guarantees in a more adversarial setting, we would in fact expect that Bayesian algorithms also perform favorably in a more adversarial setting. Although it is widely known in the online learning community that when the loss function is the log loss, then the multiplicative weights algorithms are sometimes identical to Bayes rule (with particular choices of the parameters, see Freund and Schapire [1999], Vovk [2001]), there have only been a few studies of Bayesian algorithms themselves in an online learning setting. The sleeping experts setting (of Freund and Schapire [1997]) is one setting where Bayes rule is applied to obtain competitive bounds in an expert setting where some experts may be "awake" (actively predicting) or "asleep" (not predicting).

We consider the widely used class of generalized linear models — focusing on least squares and logistic regression. We provide competitive bounds when the cost function is the log loss, and we compare our performance to the best model in our model class (as in the experts setting). Though the regression problem (with square loss) has been studied in a competitive framework (see Foster [1991], Vovk [2001]), our focus here is to examine how the more widely used Bayesian algorithms perform in this setting [1]. Furthermore, again as with many multiplicative weights algorithms, we show that their worst case performance is not that different from an average case performance (when Nature is restricted to follow an i.i.d. strategy).

We also examine Maximum Aposteriori (MAP) algorithms for both least squares (*ie* ridge regression) and (regularized) logistic regression. These algorithms are often used in practice — particularly for the case of logistic regression (where Bayesian model averaging is computationally intractable, while the MAP algorithm involves solving a convex program). As expected, the MAP algorithms are somewhat less competitive than full Bayesian model averaging — though not unreasonably so.

2 Bayesian Model Averaging

We now consider the Bayesian model averaging (BMA) algorithm and give a bound on its worst-case online loss.

Let $x \in \mathbb{R}^n$ denote the inputs of a learning problem and $y \in \mathbb{R}$ the outputs. Consider a *generalized linear model* that can be written $p(y|x,\theta) = p(y|\theta^T x)$, where $\theta \in \mathbb{R}^n$ are the parameters of our model (θ^T denotes the transpose of θ)). Note that the predicted distribution on y depends only on $\theta^T x$, which is linear in θ. For example, in the case of linear least squares regression, we have

$$p(y|x,\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(\theta^T x - y)^2}{2\sigma^2}\right), \tag{1}$$

where σ^2 is a fixed, *known* constant that is not a parameter of our model. In logistic regression, we would have

$$\log p(y|x,\theta) = y \log \frac{1}{1+\exp(-\theta^T x)} + (1-y)\log\left(1 - \frac{1}{1+\exp(-\theta^T x)}\right). \tag{2}$$

where we assume $y \in \{0,1\}$.

Let $S = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(T)}, y^{(T)})\}$ be an arbitrary sequence of examples, possibly chosen by an adversary. We also use S_t to denote the subsequence consisting of only the first t examples. Unless otherwise stated, we assume throughout this paper that $||x^{(i)}|| \leq 1$ (where $||\cdot||$ denotes the ℓ_2 norm) and $|y^{(i)}| \leq 1$ for all i. (For some models, additional restrictions would make sense, such as $y \in \{0,1\}$ in the case of logistic regression.)

Assume that we are going to use a Bayesian algorithm to make our online predictions. Specifically, assume that we have a Gaussian prior on the parameters:

$$p(\theta) = \mathcal{N}(\theta; \vec{0}, \nu^2 I_n),$$

where I_n is the n-by-n identity matrix, $\mathcal{N}(\cdot;\mu,\Sigma)$ is the density of a Gaussian with mean μ and covariance Σ, and $\nu^2 > 0$ is some fixed constant governing the variance in our prior.

[1] The work of Vovk [2001] presents many competitive algorithms for statistics, and, as with multiplicative weights, his aggregating algorithm also has close connections to Bayes rule. However, the cases Vovk analyses typically do not use Bayes rule. Foster [1991] analyses a variant of the MAP algorithm and does not consider Bayesian model averaging.

Also, let

$$p_t(\theta) = p(\theta|S_t) = \frac{\left(\prod_{i=1}^{t} p(y^{(i)}|x^{(i)},\theta)\right) p(\theta)}{\int_\theta \left(\prod_{i=1}^{t} p(y^{(i)}|x^{(i)},\theta)\right) p(\theta) d\theta}$$

be the posterior distribution over θ given the first t training examples. We also have that $p_0(\theta) = p(\theta)$ is just the prior distribution.

On iteration t, we are given the input $x^{(t)}$, and our algorithm makes a prediction using the posterior distribution over the outputs:

$$p(y|x^{(t)}, S_{t-1}) = \int_\theta p(y|x^{(t)}, \theta) p(\theta|S_{t-1}) d\theta.$$

We are then given the true label $y^{(t)}$, and we suffer logloss $-\log p(y^{(t)}|x^{(t)}, S_{t-1})$. We define the cumulative loss of the BMA algorithm after T rounds to be

$$L_{\mathrm{BMA}}(S) = \sum_{t=1}^{T} -\log p(y^{(t)}|x^{(t)}, S_{t-1}).$$

Importantly, note that even though the *algorithm* we consider is a Bayesian one, our theoretical results do *not* assume that the data comes from any particular probabilistic model. In particular, the data may be chosen by an adversary.

We will be interested in comparing against the loss of any expert that uses some fixed parameters $\theta \in \mathbb{R}^n$. Define $\ell_\theta(t) = -\log p(y^{(t)}|x^{(t)}, \theta)$, and let

$$L_\theta(S) = \sum_{t=1}^{T} \ell_\theta(t) = \sum_{t=1}^{T} -\log p(y^{(t)}|x^{(t)}, \theta).$$

Sometimes, we also wish to compare against distributions over experts. Given a distribution Q over θ, define $\ell_Q(t) = \int_\theta -Q(\theta) \log p(y^{(t)}|x^{(t)}, \theta) d\theta$, and

$$L_Q(S) = \sum_{t=1}^{T} \ell_Q(t) = \int_\theta Q(\theta) L_\theta(S) d\theta.$$

This is the average logloss incurred by a procedure that first samples some $\theta \sim Q$, and then uses the parameter θ for all of its predictions. (Here, the expectation is over the random θ, not over the sequence of examples.) Note that this is the expectation of the logloss, which is a different type of averaging than we saw in BMA, which had the expectation and the log in the reverse order.

2.1 A General Online Bound

We now provide a worst case bound of the loss incurred by Bayesian algorithms. A result very similar to this (for finite model classes) is given by Freund and Schapire [1997]. For completeness, we prove the result here in its full generality, though our proof is similar to theirs. As usual, define $KL(q||p) = \int_\theta q(\theta) \log \frac{q(\theta)}{p(\theta)}$.

Lemma 2.1: *Let Q be any distribution over θ. Then for all sequences S*

$$L_{\mathrm{BMA}}(S) \le L_Q(S) + KL(Q||p_0).$$

Proof. Let $Y = \{y^{(1)}, y^{(2)}, \ldots, y^{(T)}\}$ and $X = \{x^{(1)}, x^{(2)}, \ldots, x^{(T)})\}$. By the chain rule of conditional probabilities, it follows that $L_{\mathrm{BMA}}(S) = -\log p(Y|X)$ and that $L_\theta(S) =$

$-\log p(Y|X,\theta)$. By Bayes rule, we have that $p_T(\theta) = \frac{p(Y|X,\theta)p_0(\theta)}{p(Y|X)}$, and so

$$\begin{aligned} L_{\mathrm{BMA}}(S) - L_Q(S) &= -\log p(Y|X) + \int_\theta Q(\theta)\log p(Y|X,\theta)d\theta \\ &= \int_\theta Q(\theta)\log\frac{p(Y|X,\theta)}{p(Y|X)}d\theta \\ &= \int_\theta Q(\theta)\log\frac{p_T(\theta)}{p_0(\theta)}d\theta \\ &= \int_\theta Q(\theta)\log\frac{Q(\theta)}{p_0(\theta)}d\theta - \int_\theta Q(\theta)\log\frac{Q(\theta)}{p_T(\theta)}d\theta \\ &= KL(Q||p_0) - KL(Q||p_T). \end{aligned}$$

Along with the fact that $KL(Q||p_T) \geq 0$, this proves the lemma. □

2.2 An Upper Bound for Generalized Linear Models

For the theorem that we shortly present, we need one new definition. Let $f_y(z) = -\log p(y|w^t\theta = z)$. Thus, $f_{y^{(t)}}(\theta^T x^{(t)}) = \ell_\theta(t)$. Note that for linear regression (as defined in Equation 1), we have

$$|f_y''(z)| \leq \frac{1}{\sigma^2} \tag{3}$$

and that for the logistic regression model (as defined in Equation 2), we have

$$|f_y''(z)| \leq 1\,.$$

Theorem 2.2: *Suppose $f_y(z)$ is continuously differentiable, and that $|f_y''(z)| \leq c$ for some constant c. For all T length sequences S and for all θ^**

$$L_{\mathrm{BMA}}(S) \leq L_{\theta^*}(S) + \frac{1}{2\nu^2}||\theta^*||^2 + \frac{n}{2}\log\left(1 + \frac{Tc\nu^2}{n}\right) \tag{4}$$

This bound is similar to ones provided by Vovk [2001] and Foster [1991] for the square loss case (though they use different algorithms). The $\frac{1}{2\nu^2}||\theta^*||^2$ is a penalty term from our prior. The $\log$ term is how fast our loss could grow in comparison to the best θ^*. Importantly, this extra loss is only logarithmic in this adversarial setting. However, in the next section, we show that if S is restricted to come from an (adversarial) distribution D where $D(y|x)$ is i.i.d., then our bound becomes no better. In this sense, this worst case bound behaves much like an average case one.

Proof: We use Lemma 2.1 with $Q(\theta) = \mathcal{N}(\theta;\theta^*,\epsilon^2 I_n)$ being a normal distribution with mean θ^* and covariance $\epsilon^2 I_n$. Here, ϵ^2 is a parameter that we tune later to get the tightest possible bound. Letting $\mathcal{H}(Q) = \frac{n}{2}\log 2\pi e\epsilon^2$ be the entropy of Q, we have that

$$\begin{aligned} KL(Q||p_0) &= \int_\theta Q(\theta)\log\left[\frac{1}{(2\pi)^{n/2}|\nu^2 I_n|^{1/2}}\exp\left(-\frac{1}{2\nu^2}\theta^T\theta\right)\right]^{-1}d\theta - \mathcal{H}(Q) \\ &= n\log\nu + \frac{1}{2\nu^2}\int_\theta Q(\theta)\theta^T\theta d\theta - \frac{n}{2} - n\log\epsilon \\ &= n\log\nu + \frac{1}{2\nu^2}\left(||\theta^*||^2 + n\epsilon^2\right) - \frac{n}{2} - n\log\epsilon. \end{aligned} \tag{5}$$

To prove the result, we also need to relate the error of L_Q to that of L_{θ^*}. By taking a Taylor expansion of f_y, we have that

$$f_y(z) = f_y(z^*) + f_y'(z^*)(z-z^*) + f_y''(\xi(z))\frac{(z-z^*)^2}{2},$$

for some appropriate function ξ. Thus, if z is a random variable with mean z^*, we have

$$\begin{aligned} E_z[f_y(z)] &= f_y(z^*) + f'_y(z^*) \cdot 0 + E_z\left[f''_y(\xi(z))\frac{(z-z^*)^2}{2}\right] \\ &\leq f_y(z^*) + cE_z\left[\frac{(z-z^*)^2}{2}\right] \\ &= f_y(z^*) + \frac{c}{2}\mathrm{Var}(z) \end{aligned}$$

Consider a single example (x, y). We can apply the argument above with $z^* = {\theta^*}^T x$, and $z = \theta^T x$, where $\theta \sim Q$. Note that $E[z] = z^*$ since Q has mean θ^*. Also, $\mathrm{Var}(\theta^T x) = x^T(\epsilon^2 I_n)x = ||x||^2\epsilon^2 \leq \epsilon^2$ (because we previously assumed that $||x|| \leq 1$). Thus, we have

$$E_{\theta\sim Q}[f_y(\theta^T x)] \leq f_y({\theta^*}^T x) + \frac{c\epsilon^2}{2}$$

Since $\ell_Q(t) = E_{\theta\sim Q}[f_{y^{(t)}}(\theta^T x^{(t)})]$ and $\ell_{\theta^*}(t) = f_{y^{(t)}}({\theta^*}^T x^{(t)})$, we can sum both sides from $t = 1$ to T to obtain

$$L_Q(S) \leq L_{\theta^*}(S) + \frac{Tc}{2}\epsilon^2$$

Putting this together with Lemma 2.1 and Equation 5, we find that

$$L_{\mathrm{BMA}}(S) \leq L_{\theta^*}(S) + \frac{Tc}{2}\epsilon^2 + n\log\nu + \frac{1}{2\nu^2}\left(||\theta^*||^2 + n\epsilon^2\right) - \frac{n}{2} - n\log\epsilon.$$

Finally, by choosing $\epsilon^2 = \frac{n\nu^2}{n+Tc\nu^2}$ and simplifying, Theorem 2.2 follows. $\square$

2.3 A Lower Bound

We now provide a lower bound for Bayesian linear regression, which demonstrates that our upper bound is tight. In the proof, Nature can satisfy this lower bound with a generating distribution of S, such that $D(y^{(i)}|x^{(i)})$ is i.i.d.

Theorem 2.3: *If $\nu^2 = \sigma^2 = 1$, then there exists a sequence S such that*

$$L_{\mathrm{BMA}}(S) \geq \inf_{\theta^*}(L_{\theta^*}(S) + \frac{1}{2}||\theta^*||^2) + \frac{n}{2}\log\left(1 + \frac{T}{n}\right) + O(1)$$

Before we provide the proof, let us first present the form of the posteriors and the predictions. Define $A_t = \frac{1}{\nu^2}I_n + \frac{1}{\sigma^2}\sum_{i=1}^t x^{(i)}{x^{(i)}}^T$, and $b_t = \sum_{i=1}^t x^{(i)}y^{(i)}$. We now have that

$$p_t(\theta) = p(\theta|S_t) = \mathcal{N}\left(\theta; \hat{\theta}_t, \hat{\Sigma}_t\right), \tag{6}$$

where $\hat{\theta}_t = A_t^{-1}b_t$, and $\hat{\Sigma}_t = A_t^{-1}$. Also, the predictions at time $t+1$ are given by

$$p(y^{(t+1)}|x^{(t+1)}, S_t) = \mathcal{N}\left(y^{(t+1)}; \hat{y}_{t+1}, s_{t+1}^2\right) \tag{7}$$

where $\hat{y}_{t+1} = \hat{\theta}_t^T x^{(t+1)}$, $s_{t+1}^2 = {x^{(t+1)}}^T\hat{\Sigma}_t x^{(t+1)} + \sigma^2$. In contrast, the prediction of a fixed expert using parameter θ^* would be

$$p(y^{(t)}|x^{(t)}, \theta^*) = \mathcal{N}\left(y^{(t)}; y_t^*, \sigma^2\right), \tag{8}$$

where $y_t^* = {\theta^*}^T x^{(t)}$. Now the BMA loss is:

$$L_{\mathrm{BMA}}(S) = \sum_{t=1}^T \frac{1}{2s_t^2}(y^{(t)} - \hat{\theta}_{t-1}^T x^{(t)})^2 + \log\sqrt{2\pi s_t^2} \tag{9}$$

Proof: We start with the case where $n = 1$ and choose an i.i.d. distribution D which generates the sequence S such that $x^{(i)} = 1$ and $y^{(i)}$ has mean 0 and unit variance. We then have $A_t = t+1$, and $b_t = \sum_{i=1}^t y^{(i)}$. This implies that for the posteriors, $\hat{\theta}_t = \frac{1}{t+1}\sum_{i=1}^t y^{(i)}$ and $\hat{\Sigma}_t = \frac{1}{t+1}$, and for the BMA prediction $\hat{y}_{t+1} = \frac{1}{t+1}\sum_{i=1}^t y^{(i)}$ and $s_{t+1}^2 = \frac{t+2}{t+1}$. The prediction of a fixed expert uses $y_t^* = \theta^*$.

Ignoring the irrelevant $\log(\sqrt{2\pi})$ term, the expected BMA loss, $\mathrm{E}_D[L_{\mathrm{BMA}}(S)]$, is

$$
\begin{aligned}
&= \mathrm{E}_D[\frac{1}{2}\sum_{t=0}^{T-1}\left(\frac{t+1}{t+2}\left(y^{(t+1)} - \frac{1}{t+1}\sum_{i=1}^{t} y^{(i)}\right)^2 + \log\frac{t+2}{t+1}\right)] \\
&= \mathrm{E}_D\left[\frac{1}{2}\sum_{t=0}^{T-1}\frac{t+1}{t+2}\left((y^{(t+1)})^2 + \frac{1}{(t+1)^2}\sum_{i=1}^{t}(y^{(i)})^2\right)\right] + \frac{1}{2}\log(1+T) \\
&= \frac{1}{2}\sum_{t=0}^{T-1}\frac{t+1}{t+2}\left(1 + \frac{t}{(t+1)^2}\right) + \frac{1}{2}\log(1+T) \\
&= \frac{T}{2} - \frac{T}{2(T+1)} + \frac{1}{2}\log(1+T)
\end{aligned}
$$

where the second step follows using $\mathrm{E}_D[y^{(i)}y^{(j)}] = 0$ for $i \neq j$ since D is i.i.d., the third from the fact that $y^{(t)}$ is zero mean, unit variance, and the last step follows from algebra.

We now sketch the argument for the θ^* term. Now it is straightforward to show that $L_{\theta^*}(S) + \frac{1}{2}||\theta^*||^2$ is minimized at $\theta^* = \frac{1}{T+1}\sum_{i=1}^T y^{(i)}$. Again ignoring the irrelevant $\log(\sqrt{2\pi})$ term, we have that $\mathrm{E}_D[\inf_{\theta^*}(L_{\theta^*}(S) + ||\theta^*||^2)]$ is equal to

$$
\mathrm{E}_D[\frac{1}{2}\sum_{t=0}^{T-1}(y^{(t+1)} - \frac{1}{T+1}\sum_{i=1}^{T} y^{(i)})^2 + \frac{1}{2}||\frac{1}{T+1}\sum_{i=1}^{T} y^{(i)}||^2] = \frac{T}{2} - \frac{T}{(T+1)^2}
$$

where one can use a similar argument to the previous one to see this.

Comparing these two terms proves the result for $n = 1$. The modification for n dimensions follows: S is broken into T/n subsequences where only one dimension $x^{(i)}{}_k = 1$ (and the other dimensions are set to 0) in any subsequence. The result follows due to the additivity of the losses on the subsequences. $\square$

3 MAP Estimation

We now present bounds for MAP algorithms for both linear least squares (ridge regression in this case) and logistic regression. These algorithms use the maximum $\hat{\theta}_{t-1}$ of $p_{t-1}(\theta)$ to form their prediction $p(y^{(t)}|x^{(t)}, \hat{\theta}_{t-1})$ at time t (as opposed to BMA's prediction of $p(y^{(t)}|x^{(t)}, S_{t-1})$). As expected these bounds are weaker than BMA, though perhaps not unreasonably so.

3.1 Ridge Regression

The log loss for the MAP algorithm is essentially the square loss. We now analyze the squared loss of using the MAP estimates (which is referred to as ridge regression). Define

$$
\widetilde{L}_{\mathrm{MAP}}(S) = \frac{1}{2}\sum_{t=1}^{T}(y^{(t)} + \hat{\theta}_{t-1}^T x^{(t)})^2\,, \quad \widetilde{L}_{\theta^*}(S) = \frac{1}{2}\sum_{t=1}^{T}(y^{(t)} + \theta^{*T} x^{(t)})^2 \qquad (10)
$$

where $\hat{\theta}_t$ is the MAP estimate (see Equation 6). Note these square losses are closely related to the log losses of using constant predictions, which are:

$$L_{\theta^*}(S) = \sum_{t=1}^{T} \frac{1}{2\sigma^2}(y^{(t)} - {\theta^*}^T x^{(t)})^2 + \log\sqrt{2\pi\sigma^2} \tag{11}$$

Contrast these with the BMA loss in Equation 9, which is adaptively weighting the squared term with the inverse variances.

Corollary 3.1: *Let $\gamma^2 = \sigma^2 + \nu^2$. For all S and for all θ^*, we have that*

$$\widetilde{L}_{\mathrm{MAP}}(S) \leq \frac{\gamma^2}{\sigma^2}\widetilde{L}_{\theta^*}(S) + \frac{\gamma^2}{2\nu^2}||\theta^*||^2 + \gamma^2\frac{n}{2}\log\left(1 + \frac{Tc\nu^2}{n}\right)$$

Proof: First, note that from the definitions in Equations (6, 7), we can show that $\sigma^2 \leq s_t^2 \leq \sigma^2 + \nu^2$. Using Equations(9, 11) and Theorem 2.2, we have

$$\begin{aligned}\sum_{t=1}^{T} \frac{1}{2s_t^2}(y^{(t)} - \hat{\theta}_{t-1}^T x^{(t)})^2 \quad &\leq \quad \sum_{t=1}^{T} \frac{1}{2\sigma^2}(y^{(t)} - {\theta^*}^T x^{(t)})^2 + \frac{1}{2\nu^2}||\theta^*||^2 \\ &\quad + \frac{n}{2}\log\left(1 + \frac{Tc\nu^2}{n}\right) + \sum_{t=1}^{T}\log\frac{\sqrt{2\pi\sigma^2}}{\sqrt{2\pi s_t^2}}\end{aligned}$$

Now the result follows by noting that the last term is negative (since $s_t^2 \geq \sigma^2$) and by multiplying both sides of the equation by $\sigma^2 + \nu^2$. $\square$

This bound is not unreasonable, since when there exists a θ^* with loss near to 0 then we are competitive. However, we might hope to be more competitive by having the coefficient in front of the $\widetilde{L}_{\theta^*}(S)$ term be 1 (as in Theorem 2.2) rather than $\frac{\gamma^2}{\sigma^2}$. Crudely, the reason that MAP is not as effective as BMA is that MAP does not taken into account *uncertainty* — the squared terms cannot be reweighted to take variance into account (compare Equations 9 and 10).

The algorithm of Vovk [2001] has a similar bound with a coefficient of 1 — though this algorithm is specialized to the square loss (and does not use Bayes rule). Foster [1991] provides a bound with the coefficient 1 with more restrictive assumptions. As the following lower bound shows, it is not possible for ridge regression to have a coefficient of 1 for $\widetilde{L}_{\theta^*}(S)$, with a reasonable regret bound. A similar lower bound is in Vovk [2001], which doesn't apply to our setting where we have the constraint $||x^{(i)}|| \leq 1$.

Theorem 3.2: *Let $\gamma^2 = \sigma^2 + \nu^2$. There exists an S such that*

$$\widetilde{L}_{\mathrm{MAP}}(S) \geq \inf_{\theta^*}(\widetilde{L}_{\theta^*}(S) + \frac{1}{2}||\theta^*||^2) + \Omega(T)$$

Proof: (sketch) Let S be a $T+1$ lengths sequence, with $n = 1$, where for the first T steps, $x^{(i)} = 1/\sqrt{T}$ and $y^{(i)} = 1$, and at $T+1$, $x^{(T+1)} = 1$ and $y^{(T+1)} = 0$. Here, one can show that $\inf_{\theta^*}(\widetilde{L}_{\theta^*}(S) + \frac{1}{2}||\theta^*||^2) = T/4$ and $\widetilde{L}_{\mathrm{MAP}}(S) \leq 3T/8$, and the result follows. $\square$

3.2 Logistic Regression

MAP estimation is widely used for logistic regression (and is essentially regularized logistic regression). The reason being that the MAP estimates $\hat{\theta}_{t-1}$ can be found by solving a convex program (while BMA has to deal with intractable integrals). Letting $\hat{\theta}_{t-1}$ be maximum of the posterior $p_t(\theta)$, define $L_{\mathrm{MAP}} = \sum_{t=1}^{T} p(y^{(t)}|x^{(t)}, \hat{\theta}_{t-1})$. As with square loss case, the bound we present is multiplicatively worse (by a factor of 4).

Theorem 3.3: *If $\nu \leq 0.5$, then for all S and for all θ^**

$$L_{\mathrm{MAP}}(S) \leq 4L_{\theta^*}(S) + \frac{2}{\nu^2}||\theta^*||^2 + 2n\log\left(1 + \frac{Tc\nu^2}{n}\right)$$

Proof: (sketch) Assume $n = 1$ (the general case is analogous). We show that $\ell_{\hat{\theta}_{t-1}}(t) \leq 4\ell_{\mathrm{BMA}}(t)$. Without loss of generality, assume $y^{(t)} = 1$ and $x^{(t)} = x \geq 0$, so the BMA prediction is $\int_\theta p(1|\theta, x)p_t(\theta)d\theta$, and $\ell_{\mathrm{BMA}}(t)$ is the negative log of this. Note $\theta = \infty$ gives probability 1 for $y^{(t)} = 1$ (and this setting of θ minimizes the loss at time t).

Since we do not have a closed form solution of the prior p_t, let us use another distribution q in lieu of p_t. Now we want to find the 'worst case' posterior distribution $q(\theta)$ such that $p_q = \int_\theta p(1|\theta, x)q(\theta)d\theta$ is greater than the BMA prediction and q obeys some constraints. First, let $q(\theta)$ have its maximum at the MAP estimator $\hat{\theta}_{t-1}$. Second, it suffices to set $q(\theta) = 0$ if $\theta < \hat{\theta}_{t-1}$, since mass to the left of $\hat{\theta}_{t-1}$ is just making the p_q smaller. Lastly, due to the Gaussian prior p_0, it is straightforward to show that $(\log p_t)''(\theta) \geq \frac{1}{\nu^2}$ (the prior imposes some minimum curvature), so we impose the constraint that $q''(\theta) \geq 1/\nu^2$. With these constraints, one can show the maximal q is the rectification of the Gaussian $\mathcal{N}(\theta; \hat{\theta}_{t-1}, \nu^2 I_n)$, such that there is positive probability only for $\theta > \hat{\theta}_{t-1}$ (and the distribution is renormalized). From these constraints, it follows that q has loss, $-\log p_q$, that is less than or equal to $\ell_{\mathrm{BMA}}(t)$.

Now we show that $\ell_{\hat{\theta}_{t-1}}(t) = -\log p(1|x, \hat{\theta}_{t-1}) \leq -4\log p_q$, which completes the proof since $-\log p_q \leq \ell_{\mathrm{BMA}}(t)$. We consider two cases, $\theta x < 0$ and $\theta x \geq 0$. We start with the case $\theta x < 0$. Using the boundedness of the derivative $|(\log(p(1|x,\theta))'| < 1$, we have

$$\begin{aligned} p_q &= \int_\theta \exp(\log(p(1|x,\theta))q(\theta)d\theta \\ &\leq \int_\theta \exp\left(\log(p(1|x,\hat{\theta}_{t-1}) + \theta - \hat{\theta}_{t-1}\right) q(\theta)d\theta \leq 1.6p(1|x,\hat{\theta}_{t-1}) \end{aligned}$$

where we have used $\int_\theta \exp(\theta - \hat{\theta}_{t-1})q(\theta)d\theta < 1.6$ (using the definition of q with $\nu \leq 0.5$). The claim follows using the lower bound $-\log p(1|x, \hat{\theta}_{t-1}) \geq \log 2$ for $\theta x \leq 0$, since this implies $-\log p_q \geq -\log p(1|x, \hat{\theta}_{t-1})(1 - \log 1.6/\log 2) \geq 0.3\ell_{\hat{\theta}_{t-1}}(t)$.

For $\theta x \geq 0$, let σ be the sigmoid function, so $p(1|x,\theta) = \sigma(\theta x)$ and $p_q = \int_\theta \sigma(x\theta)q(\theta)$. For $\theta x \geq 0$, σ is concave, so $p_q \leq \sigma\left(x\int_\theta \theta q(\theta)d\theta\right)$. Using the definition of q (and $||x|| < 1$), we then have that $p_q \leq \sigma(\hat{\theta}_{t-1} + \nu)$. Again using properties of σ, we have $|(\log\sigma)'(\theta)| < -\log\sigma(\theta)$. Furthermore, for all $\theta > \hat{\theta}_{t-1}$, $|(\log\sigma)'(\theta)| < -\log\sigma(\hat{\theta}_{t-1})$. Using this derivative condition, we have that $-\log p_q \geq -\log\sigma(\hat{\theta}_{t-1})(1-\nu) = \ell_{\hat{\theta}_{t-1}}(t)(1-\nu)$, which proves the claim for the $\theta x \geq 0$ condition. □

References

N. Cesa-Bianchi, Y. Freund, D. Haussler, D. Helmbold, R.. Schapire, and M. Warmuth. How to use expert advice. *J. ACM*, 44, 1997.

D. P. Foster. Prediction in the worst case. *Annals of Statistics*, 19, 1991.

Y. Freund and R. Schapire. Using and combining predictors that specialize. *In STOC*, 1997.

Y. Freund and R. Schapire. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 29:79–103, 1999.

V. Vovk. Competitive on-line statistics. *International Statistical Review*, 69, 2001.