On Sequential Regret Bounds for Bayesian Model Averaging

Abstract

We consider the problem of online density estimation using Bayesian model averaging. We begin with a general setting, without any assumptions about the prior distribution or about the point of comparison. We do, however, require a positive definite upper bound on the family of negative log-densities. This allows us to get finite-sequence regret bounds using the Laplace approximation. In the general case, the resulting regret bounds are difficult to use because they involve the score function, which we may not have control over. We show how variational techniques can be used to alleviate this problem. In the case of Gaussian priors, the regret bound simplifies substantially, and the resulting bounds improves earlier results for generalized linear models and Gaussian process regression.

1. Bayesian Model Averaging

Many prediction problems, such as predicting tomorrow's high temperature, can be thought of as sequential prediction games that proceed in distinct rounds. At the t'th round, we make a prediction p_t about an outcome (e.g. the next day's maximum temperature), and then we observe the actual outcome y_t. Here we consider the "log-loss" setting, in which the predictions $p_1, p_2, \dots$ are density functions on the outcome space $\mathcal{Y}$, and the loss at the t'th round is $-\log p_t(y_t)$, which is the negative log-likelihood of the observation under the predicted density. The general goal is to find prediction strategies that perform well with respect to "cumulative loss," which is the sum of these losses over the course of the sequence. However, a level of cumulative loss that is quite good for a difficult sequence may be unacceptably high for an easy sequence. This motivates the introduction of a "target loss" $T(y^n)$, which defines the level of cumulative loss that's "good" for each sequence $y^n = (y_1, \dots, y_n)$. The gap between the cumulative loss and the target loss is known as the *regret*. In this paper, we present several regret bounds for a particular family of prediction strategies, known as *Bayesian model averaging*.

In Bayesian model averaging (BMA), we begin with a probability model $\{p(y \mid \theta) \mid \theta \in \Theta\}$ of probability densities on $\mathcal{Y}$ with index set Θ, together with a prior distribution π_0 on Θ. In the each round, we get an observation and use Bayes rule to get an updated posterior distribution on Θ. At the t'th round, the distribution on Θ is $d\pi_t(\theta) = d\pi_0(\theta \mid y_1, \dots, y_{t-1})$, and we play the predictive density

$$p_t(y) = p(y \mid y_1, \dots, y_{t-1}) = \int p(y \mid \theta)\, d\pi_t(\theta).$$

The cumulative loss for BMA with prior π_0 on the sequence $y^n = (y_1, \dots, y_n)$ is given by

$$\begin{aligned} L(\pi_0, y^n) &= -\sum_{t=1}^{n} \log p_t(y_t) \\ &= -\log\left(\prod_{t=1}^{n} p(y_t \mid y_1, \dots, y_{t-1})\right) \\ &= -\log p(y^n) = -\log \mathbb{E}_{\theta\sim\pi_0} p(y^n \mid \theta). \end{aligned}$$

As the target loss for this setup, it is common to use $-\log p(y^n \mid \theta_0)$ for some $\theta_0 \in \Theta$, where θ_0 *may depend on* y^n. For example, a common comparison point is the maximum likelihood estimator (MLE) $\theta_0 = \arg\max_{\theta\in\Theta} p(y^n \mid \theta)$. Most of our results below allow for any $\theta_0 \in \Theta$, though some take much simpler forms for special choices of θ_0. The target loss $-\log p(y^n \mid \theta_0)$ corresponds to the cumulative loss attained by playing $p_t(y) = p(y \mid \theta_0)$ for each of the n rounds, since

$$\begin{aligned} L(\theta_0, y^n) &= -\sum_{t=1}^{n} \log p_t(y_t) \\ &= -\sum_{t=1}^{n} \log p(y_t \mid \theta_0) \\ &= -\log p(y^n \mid \theta_0) \end{aligned}$$

Thus for a fixed probability model indexed by Θ, the regret for using BMA with prior π_0 compared to using

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

the fixed strategy $\theta_0 \in \Theta$ is, for any sequence $y^n \in \mathcal{Y}^n$, given by

$$\begin{aligned} r(\pi_0, \theta_0) &:= L(\pi_0, y^n) - L(\theta_0, y^n) \\ &= -\log \int p(y^n \mid \theta)\, d\pi_0(\theta) - (-\log p(y^n \mid \theta_0)) \\ &= -\log \mathbb{E}_{\theta\sim\pi_0} e^{\ell(\theta)} + \ell(\theta_0), \end{aligned}$$

where we have defined the sequence log-likelihood function $\ell(\theta) = \log p(y^n \mid \theta)$.

2. Overview

We will now take as our starting point the expression for regret found at the end of Section 1. Although we are still motivated by the sequential BMA setting, we can abstract away from that setting by taking $\ell(\theta)$ and a distribution π on Θ as our primary ingredients. We define the regret for BMA with prior π and loss $-\ell(\theta)$ as

$$r(\pi, \theta_0) := \ell(\theta_0) - \log \mathbb{E}_{\theta\sim\pi} e^{\ell(\theta)}. \tag{1}$$

Note that there is no explicit reference to a sequence $y^n = (y_1, \dots, y_n)$. In the BMA setting, the dependence on the sequence is built into the function $\ell(\theta)$.

Upper bound on $r(\pi, \theta_0)$, known as *regret bounds,* have previously been derived for many specific cases of ℓ, θ_0, and π. In this paper, we consider the cases in which $\Theta \subset \mathbf{R}^d$ and is open and convex (or star-shaped about θ_0). We generally do not place any restriction on $\theta_0 \in \Theta$, though we do make mention of some standard special cases. Finally, we take $\ell : \Theta \to \mathbf{R}$ to be any twice-differentiable function with a lower bound on the Hessian. That is, we assume there exists a symmetric positive definite (spd) matrix J for which

$$-\nabla^2_\theta \ell(\theta) \preceq J, \quad \forall \theta \in \Theta,$$

where $A \preceq B$ means that the matrix $B - A$ is spd. Certain more general settings, such as Gaussian process regression in which Θ is an infinite-dimensional function space, can be reduced to this setting in a straightforward manner (See Section 5).

Our most general bound are given in Section 3. The bound is tight, in the sense that it becomes an equality when $\ell(\theta)$ is quadratic. We give two short proofs of this theorem. The most direct is the classic "Laplace approximation," which follows from a single application of Taylor's theorem. The variational approach is a generalization of the work in (Kakade & Ng, 2004; Kakade et al., 2005; Seeger et al., 2008), which assumed that π was a Gaussian distribution or a Gaussian process. Besides generalizing to arbitrary distributions, we tighten the variational analysis and get an improved bound. Next we look in more detail at the proof techniques of (Kakade & Ng, 2004; Kakade et al., 2005; Seeger et al., 2008) and compare them to the our variational approach. While the approaches taken there do not give the tightest bounds in the specific case of Gaussisan priors, their techniques apply nicely to situations with non-Gaussian priors. In Theorem 7, we specialize our bounds to the case of a Gaussian prior, and in Theorem 10 we apply it to the case of a Gaussian process prior.

3. General Regret Bounds

We first state our general regret bound. Below, we write $\mathcal{N}(\theta \mid \mu, \Sigma)$ to denote the multivariate Gaussian density function with mean μ and covariance $\Sigma \succ 0$, evaluated at θ. For notational ease, we define the "score" vector as

$$s = \nabla_\theta \ell(\theta_0).$$

We only use s in contexts for which the point $\theta_0 \in \Theta$ has already been fixed.

Theorem 1. *Suppose $\ell(\theta)$ is twice differentiable on an open, convex*[1] *set $\Theta \subset \mathbf{R}^d$, and $-\nabla^2_\theta \ell(\theta) \preceq J, \quad \forall \theta \in \Theta$, and $J \succ 0$. Let π be any probability distribution on Θ. Then for any $\theta_0 \in \Theta$, we have*

$$\begin{aligned} r(\pi, \theta_0) \le &-\frac{1}{2} s^T J^{-1} s + \frac{d}{2}\log\frac{1}{2\pi} + \frac{1}{2}\log\det(J) \\ &- \log \mathbb{E}_{\theta\sim\pi} \mathcal{N}(\theta \mid \theta_0 + J^{-1}s, J^{-1}). \end{aligned} \tag{2}$$

We have equality when $\ell(\theta)$ is quadratic. In particular, if $\ell(\theta) = -\frac{1}{2}(y-\theta)'\Sigma^{-1}(y-\theta) + c$, for some $y \in \mathbf{R}^d, c \in \mathbf{R}$, we have

$$\begin{aligned} r(\pi, \theta_0) = &-\frac{1}{2}(y-\theta_0)^T \Sigma^{-1}(y-\theta_0) + \frac{d}{2}\log\frac{1}{2\pi} \\ &+ \frac{1}{2}\log\det\left(\Sigma^{-1}\right) - \log \mathbb{E}_{\theta\sim\pi}\mathcal{N}(\theta \mid y, \Sigma). \end{aligned} \tag{3}$$

A comparison point of particular interest is $\theta_{\text{MLE}} \in \arg\max_{\theta\in\Theta} \ell(\theta)$, where the notation is motivated by the case when $\ell(\theta)$ is a log-likelihood. In this case, $s = \nabla_\theta \ell(\theta_{\text{MLE}}) = 0$, and the bound simplifies. We record this in

Corollary 2. *Under the conditions of Theorem 1,*

[1] Here, and elsewhere, we may relax the requirement of convexity to being star-convex with respect to θ_0. A set $\Theta \subset \mathbf{R}^d$ is star-convex with respect to $\theta_0 \in \Theta$, if $[\theta, \theta_0] \subset \Theta$, for all $\theta \in \Theta$, where $[\theta, \theta_0]$ denotes the line segment connecting θ and θ_0.

$$r(\pi, \theta_{MLE}) \leq \frac{d}{2} \log \frac{1}{2\pi} + \frac{1}{2} \log \det (J) - \log \mathbb{E}_{\theta \sim \pi} \mathcal{N}(\theta \mid \theta_{MLE}, J^{-1}). \tag{4}$$

In Equation 2, we are taking the expectation of a Gaussian density centered at $\hat{\theta} = \theta_0 + J^{-1}s$. If we had $J^{-1} = \nabla_\theta^2 \ell(\theta_0)$, then $\hat{\theta}$ would exactly be the result of a Newton step from θ_0 towards the θ_{MLE}. So even when we're not comparing to θ_{MLE}, the Gaussian density we are integrating will be centered near θ_{MLE}, at least when $\nabla_\theta^2 \ell(\theta_0) \approx J^{-1}$, and $\ell(\theta)$ is closely approximated by a quadratic. Indeed, for the quadratic case in Equation (3), the Gaussian density is centered at y, which is the maximizer of $\ell(\theta)$.

3.1. Proofs

Below we give two proofs of Theorem 1. A key step in each proof is to use Taylor's theorem to establish a quadratic lower bound on $\ell(\theta)$. We have the following:

Lemma 3. *Let $\Theta \subset \mathbf{R}^d$ be a convex, open set. Suppose that $\ell : \Theta \to \mathbf{R}$ is twice differentiable on Θ, with $-\nabla_\theta^2 \ell(\theta) \preceq J$, $\forall \theta \in \Theta$, and $J \succeq 0$. Then for any $\theta \in \Theta$,*

$$\ell(\theta) - \ell(\theta_0) \geq -\frac{1}{2}(\theta - \theta_0)^T J (\theta - \theta_0) + (\theta - \theta_0)^T s. \tag{5}$$

When $J \succ 0$, we can write this as

$$\ell(\theta) - \ell(\theta_0) \geq -\frac{1}{2}\left[(\theta - \hat{\theta})^T J(\theta - \hat{\theta}) - s' J^{-1} s\right], \tag{6}$$

where $\hat{\theta} := \theta_0 + J^{-1}s$. When $\ell(\theta)$ is quadratic with $-\nabla_\theta^2 \ell(\theta) \equiv J$ for all $\theta \in \Theta$, these bounds are equalities.

Proof. By Taylor's theorem, for each $\theta \in \Theta$, there exists $\tilde{\theta} \in [\theta, \theta_0]$ (the line segment connecting θ and θ_0) for which

$$\ell(\theta) - \ell(\theta_0) = (\theta - \theta_0)^T s + \frac{1}{2}(\theta - \theta_0)^T \nabla_\theta^2 \ell(\tilde{\theta})(\theta - \theta_0).$$

Applying $\nabla_\theta^2 \ell(\tilde{\theta}) \succeq -J$, we get (5). We get (6) by completing the quadratic form, which may be checked directly. □

Proof by Laplace Approximation

In the inequality step below, we use Lemma 3 to bound $\ell(\theta)$ in terms of a quadratic function (the "Laplace approximation"):

$$\begin{aligned} r(\pi, \theta_0) =& \ell(\theta_0) - \log \mathbb{E}_\pi e^{\ell(\theta)} \text{ (by definition)} \\ \leq& \ell(\theta_0) - \log \mathbb{E}_\pi e^{\ell(\theta_0) - \frac{1}{2}(\theta - \hat{\theta})^T J (\theta - \hat{\theta}) + \frac{1}{2} s' J^{-1} s} \\ =& -\frac{1}{2} s' J^{-1} s - \log \mathbb{E}_\pi e^{-\frac{1}{2}(\theta - \hat{\theta})^T J (\theta - \hat{\theta})} \end{aligned} \tag{7}$$

From here, we simply note that the integrand is proportional to a Gaussian density with covariance J^{-1}. Including the appropriate scaling for the Gaussian density, we get Equation 2. Since Lemma 3 gives equality when $\ell(\theta)$ is quadratic, we get equality here as well. For $\ell(\theta) = -\frac{1}{2}(y - \theta)' \Sigma^{-1} (y - \theta) + c$, we have $-\nabla_\theta^2 \ell(\theta) \equiv \Sigma^{-1} =: J$ and $s = \nabla_\theta \ell(\theta_0) = \Sigma^{-1}(y - \theta_0)$. Plugging these values into Equation 2, we get Equation 3 □

Proof by Variational Methods

The key to getting tight bounds with the variational approach is the following lemma, which is a consequence of Fenchel-Legendre duality. We cite Lemma 1 of (Banerjee, 2006) for a short proof, though we modify the statement slightly to emphasize the necessary technical conditions.

Lemma 4. *Let Q and P be any probability measures on $\mathcal{H}$, with $Q \ll P$. Let $\phi : \mathcal{H} \to \mathbf{R}$ be in $L^1(Q)$ or nonnegative. Then*

$$\mathbb{E}_Q [\phi(h)] - \log \left[\mathbb{E}_P e^{\phi(h)}\right] \leq KL(Q, P),$$

where $KL(Q, P) = \mathbb{E}_P \log \left(\frac{dQ}{dP}(h)\right)$. If $dQ = \frac{1}{Z} e^{\phi(h)} dP$, where $Z = \int e^{\phi(h)}\, dP(h) < \infty$, then we get equality.

Below in Equation (8), we use Lemma 4 to get a variational upper bound. We relax the bound in (9) by replacing $-\ell(\theta)$ with a quadratic upper bound using Lemma 3. Finally, we use the equality part of Lemma 4 to get (10):

$$
\begin{aligned}
r(\pi,\theta_0) &= -\log \mathbb{E}_\pi e^{\ell(\theta)} + \ell(\theta_0) \text{ (by definition)} \\
&\le \inf_{Q:Q\ll\pi} \left[-\mathbb{E}_Q \ell(\theta) + \mathrm{KL}(Q,\pi)\right] + \ell(\theta_0) \qquad (8) \\
&\le \inf_{Q:Q\ll\pi} \Big\{ \mathbb{E}_Q^\theta \big[-\ell(\theta_0) + \tfrac{1}{2}(\theta-\hat\theta)^T J(\theta-\hat\theta) \qquad (9) \\
&\qquad - \tfrac{1}{2} s' J^{-1} s \big] + \mathrm{KL}(Q,\pi) \Big\} + \ell(\theta_0) \\
&= \inf_{Q:Q\ll\pi} \Big\{ \mathbb{E}_Q^\theta \tfrac{1}{2}(\theta-\hat\theta)^T J(\theta-\hat\theta) + \mathrm{KL}(Q,\pi) \Big\} \\
&\qquad - \tfrac{1}{2} s' J^{-1} s \\
&= -\log \mathbb{E}_\pi^\theta e^{-\frac{1}{2}(\theta-\hat\theta)^T J(\theta-\hat\theta)} - \frac{1}{2} s' J^{-1} s \qquad (10)
\end{aligned}
$$

The rest of the proof is the same as in the Laplace approximation approach. □

Discussion

This variational proof was based on the proof of Theorem 2.2 in (Kakade & Ng, 2004) and the proof of Theorem 1 in (Seeger et al., 2008). The goal was to generalize the approach from their more specific setting (Gaussian priors for regression coefficients) to the case of arbitrary priors π in a generic density estimation setting. Beyond their restriction to a Gaussian prior, the essential difference in the proof of (Seeger et al., 2008) is that they restrict Q to have mean θ_0[2], while we leave the mean free during the optimization of Q. If we directly generalize their proof technique to the case of arbitrary π, we get Theorem 6 below. The advantage of their approach is that it eliminates the dependence on the score s, which may be difficult to control. However, we first separate out an intermediate result that combines the quadratic bound on $\ell(\theta)$ with a variational bound restricting Q to have mean θ_0. The result may be seen as a generalization of Theorem 11.10 in (Cesa-Bianchi & Lugosi, 2006), which performed a similar separation on the method in (Kakade & Ng, 2004). We get the following

Lemma 5. *Under the conditions of Theorem 1,*

$$
\begin{aligned}
r(\pi,\theta_0) &\le \inf_{\substack{Q:Q\ll\pi \\ \mathbb{E}_Q\theta=\theta_0}} \Big\{ \mathbb{E}_Q \tfrac{1}{2}(\theta-\theta_0)^T J(\theta-\theta_0) + KL(Q,\pi) \Big\} \\
&= \inf_{\Sigma\succeq 0} \inf_{\substack{Q:Q\ll\pi \\ \mathbb{E}_Q\theta=\theta_0 \\ Cov(\theta)=\Sigma}} \Big\{ \frac{1}{2}\operatorname{tr}(J\Sigma) + KL(Q,\pi) \Big\}
\end{aligned}
$$

Proof. In (11) we use a weakened form of Lemma 4, in which we restrict Q to have expectation θ_0. The bound in (12) follows from inequality (5) in Lemma 3. To get (13) we apply the condition that $\mathbb{E}_Q\theta = \theta_0$. In the last step, we apply basic probability theory and linear algebra to compute the expectation with respect to Q.

$$
\begin{aligned}
r(\pi,\theta_0) &\le \inf_{\substack{Q:Q\ll\pi \\ \mathbb{E}_Q\theta=\theta_0}} \left[-\mathbb{E}_Q \ell(\theta) + \mathrm{KL}(Q,\pi)\right] + \ell(\theta_0) \qquad (11) \\
&\le \inf_{\substack{Q:Q\ll\pi \\ \mathbb{E}_Q\theta=\theta_0}} \Big\{ \mathbb{E}_Q^\theta \big[-\ell(\theta_0) + \tfrac{1}{2}(\theta-\theta_0)^T J(\theta-\theta_0) \qquad (12) \\
&\qquad - (\theta-\theta_0)^T s \big] + \mathrm{KL}(Q,\pi) \Big\} + \ell(\theta_0) \\
&= \inf_{\substack{Q:Q\ll\pi \\ \mathbb{E}_Q\theta=\theta_0}} \Big\{ \mathbb{E}_Q \tfrac{1}{2}(\theta-\theta_0)^T J(\theta-\theta_0) + \mathrm{KL}(Q,\pi) \Big\} \qquad (13) \\
&= \inf_{\Sigma\succeq 0} \inf_{\substack{Q:Q\ll\pi \\ \mathbb{E}_Q\theta=\theta_0 \\ \mathrm{Cov}(\theta)=\Sigma}} \Big\{ \frac{1}{2}\operatorname{tr}(J\Sigma) + \mathrm{KL}(Q,\pi) \Big\} \qquad (14)
\end{aligned}
$$

□

If we make the optimal choice for Q in Lemma 5, we get the following theorem:

Theorem 6. *Under the conditions of Theorem 1,*

$$
r(\pi,\theta_0) \le -\log \sqrt{(2\pi)^d |J|} - \log \mathbb{E}_\pi \left[\mathcal{N}(\theta \mid \theta_0, J^{-1})\right]
$$

Proof. Starting with the bound in Lemma 5, we follow Lemma 4 and choose $dQ \propto e^{-\frac{1}{2}(\theta-\theta_0)^T J(\theta-\theta_0)} d\pi$. Since this Q has mean θ_0, we the equality in (15). To finish the proof, we include the appropriate normalization constant to rewrite the expression in terms of a normal density.

$$
\begin{aligned}
r(\pi,\theta_0) &\le \inf_{\substack{Q:Q\ll\pi \\ \mathbb{E}_Q\theta=\theta_0}} \Big\{ \mathbb{E}_Q \tfrac{1}{2}(\theta-\theta_0)^T J(\theta-\theta_0) + \mathrm{KL}(Q,\pi) \Big\} \\
&= -\log \mathbb{E}_\pi \exp\left(-\frac{1}{2}(\theta-\theta_0)^T J(\theta-\theta_0)\right) \qquad (15) \\
&= -\log \mathbb{E}_\pi \left[\sqrt{(2\pi)^d |J|}\, \mathcal{N}(\theta \mid \theta_0, J^{-1})\right] \qquad (16)
\end{aligned}
$$

□

Discussion

While the bound in Theorem 1 is our tightest and most general bound, it can present challenges for practical use. First, it has a term involving the "score" vector $s = \nabla_\theta \ell(\theta_0)$, which may not be easy to control in most natural settings. As we note in Corollary 2, the

[2] They also *a priori* restrict Q to be Gaussian with a particular parameterized covariance, but these choices turn out to be optimal once π is Gaussian and Q has mean θ_0.

dependence on s vanishes when we take θ_{MLE} as our point of comparison. We can also use the variational technique where we restrict Q to have mean θ_0. This gave us Theorem 6, which is much simpler and makes no reference to s. Our discussion in Section 6.1, for the specific case of generalized linear models, will illustrate the penalty we take for this simplification.

The second major challenge in the bound of Theorem 1 is the term involving the integral of a Gaussian density with respect to the prior distribution. While still present in Theorem 6, Lemma 5 preceding it reveals a way around this integral: we can restrict the choice of Q to a family of distributions for which $\mathrm{KL}(Q, \pi)$ has a nice form. One such case is suggested in Exercise 11.20 of (Cesa-Bianchi & Lugosi, 2006), where the prior is uniform on a cube.

4. With Gaussian Prior

While the integral of the Gaussian density in Theorem 1and Theorem 6 present a challenge in general, they can be evaluated in closed form when π is itself Gaussian. In fact, using a matrix identity we can isolate the appearance of the score s in the regret bound to a single term that is always negative. Thus with a Gaussian prior, we can deal effectively with both the s and the integral in the general regret bound.

In Theorem 7 below, we consider the case that π is Gaussian with covariance Σ_0. When θ_0 is in the column space of Σ_0, we get a particularly nice form for the regret, which is exactly what we need when we consider Gaussian process regression in Section 5. When the covariance matrix is nonsingular, the bound simplifies further.

Theorem 7. *Under the conditions of Theorem 1, suppose $\pi = \mathcal{N}(0, \Sigma_0)$, for $\Sigma_0 \succeq 0$. Then for any $\theta_0 \in \Theta$ we have*

$$r(\pi, \theta_0) \le -\frac{1}{2} s^T J^{-1} s + \frac{1}{2} \log\left[\det\left(I + J\Sigma_0\right)\right] + \frac{1}{2}\left(\theta_0 + J^{-1}s\right)^T \left(\Sigma_0 + J^{-1}\right)^{-1}\left(\theta_0 + J^{-1}s\right). \tag{17}$$

If $\theta_0 = \Sigma_0\alpha$ for some $\alpha \in \mathbf{R}^d$, then

$$r(\pi, \theta_0) \le \frac{1}{2}\log\left[\det\left(I + J\Sigma_0\right)\right] + \frac{1}{2}\alpha^T\Sigma_0\alpha - \frac{1}{2}(s-\alpha)^T M (s-\alpha) \tag{18}$$

or equivalently,

$$r(\pi, \theta_0) \le \frac{1}{2}\log\left[\det\left(I + J\Sigma_0\right)\right] + \frac{1}{2}\theta_0^T\Sigma_0^-\theta_0 - \frac{1}{2}(s - \Sigma_0^-\theta_0)^T M (s - \Sigma_0^-\theta_0) \tag{19}$$

where $M = \Sigma_0 - \Sigma_0\left(J^{-1} + \Sigma_0\right)^{-1}\Sigma_0 \succeq 0$, and Σ_0^- is any generalized inverse[3] of Σ_0. If $\Sigma_0 \succ 0$, then $M = \left(J + \Sigma_0^{-1}\right)^{-1}$.

In each of these cases, we have equality when $\ell(\theta)$ is quadratic. In particular, if $\ell(\theta) = -\frac{1}{2}(y-\theta)'\Sigma^{-1}(y-\theta) + c$, for some $\Sigma \succ 0$, $y \in \mathbf{R}^d$, and *$c \in \mathbf{R}$, we get equalities by replacing J with Σ^{-1} and s with $\Sigma^{-1}(y-\theta_0)$.*

We will use two lemmas to prove this theorem. The first is a standard property of multivariate normal distributions, whose proof we include for completeness. The second lemma is a matrix identity that we use to transform (17) into (18) and (19) in the theorem statement. These latter bounds are crucial for exposing the improvement in Theorem 7 over earlier approaches.

Lemma 8. *Let $\pi = \mathcal{N}(\mu, \Sigma)$. Then for any $\theta \in \mathbf{R}^d$ and any $d \times d$ matrix $\Sigma_0 \succ 0$, we have*

$$\mathbb{E}_{y\sim\pi}\mathcal{N}(y\,;\,\theta, \Sigma_0) = \mathcal{N}(0;\,\mu - \theta, \Sigma + \Sigma_0). \tag{20}$$

Proof. Let $X \sim \mathcal{N}(\theta, \Sigma_0)$ and $Y \sim \pi$, with X and Y independent. Then

$$Y - X \sim \mathcal{N}(\mu - \theta, \Sigma + \Sigma_0).$$

Since $\Sigma + \Sigma_0 \succ 0$, this distribution has a density, which we will denote by $p_{Y-X}(z)$. Note that the RHS of Equation 20 is exactly $p_{Y-X}(0)$. We can also write $p_{Y-X}(z)$ using a convolution formula in terms of the density $p_X(x)$ for X (which exists, since $\Sigma_0 \succ 0$):

$$p_{Y-X}(z) = \mathbb{E}p_X(Y - z) = \mathbb{E}_{y\sim\pi}\mathcal{N}\left(y - z \mid \theta, \Sigma_0\right).$$

The claim now follows by taking $z = 0$ and equating the two expressions for $p_{Y-X}(0)$. □

Lemma 9. *For any $d \times d$ matrices $J \succ 0$ and $\Sigma \succeq 0$, for any $s, \alpha \in \mathbf{R}^d$, we have*

$$-s'J^{-1}s + \left(\Sigma\alpha + J^{-1}s\right)'\left(\Sigma + J^{-1}\right)^{-1}\left(\Sigma\alpha + J^{-1}s\right) = \alpha'\Sigma\alpha - (s-\alpha)'M(s-\alpha),$$

where $M = \Sigma - \Sigma\left(\Sigma + J^{-1}\right)^{-1}\Sigma \succeq 0$. The identity also holds if we replace α by $\Sigma^-\Sigma\alpha$, where Σ^- is any generalized inverse of Σ. If $\Sigma \succ 0$, then $M = \left(J + \Sigma^{-1}\right)^{-1}$.

We defer the proof of this lemma to Appendix A.

[3] Σ^{-1} is a generalized inverse of Σ iff $\Sigma\Sigma^-\Sigma = \Sigma$. A particular example is the Moore-Penrose pseudoinverse.

Proof of Theorem 7.

By Lemma 8, we have

$$\begin{aligned}&-\log \mathbb{E}_{\pi}^{\theta}\mathcal{N}(\theta \mid \hat{\theta}, J^{-1}) = -\log \mathcal{N}(-\hat{\theta} \mid 0, \Sigma_0 J^{-1})\\ =&\frac{d}{2}\log(2\pi) + \frac{1}{2}\log\det(\Sigma_0 + J^{-1}) + \frac{1}{2}\hat{\theta}^T(\Sigma_0 + J^{-1})^{-1}\hat{\theta}.\end{aligned}$$

Plugging this result into Theorem 1 and combining terms, we get (17). If we then replace θ_0 by $\Sigma_0\alpha$, the rest of the claims follow from the matrix identities in Lemma 9.

5. Gaussian Process Regression

Here we get an improvement of a result of (Seeger et al., 2008) as a special case of our Gaussian density estimation framework. The conditions of this theorem are exactly those of Theorem 1 of (Seeger et al., 2008), but the bound here is tightened by an additional term that is never positive.

Theorem 10. *Let π be a zero-mean Gaussian Process on Θ, a space of functions from $\mathcal{X}$ to $\mathbf{R}$, with covariance function k. Let $\mathcal{H}$ be the RKHS[4] with kernel k. For any points $x_1, \ldots, x_n \in \mathcal{X}$, and for any $f \in \mathcal{H}$, define*

$$\ell(f) = \sum_{t=1}^{n} g_t(f(x_t)),$$

such that there exists $c > 0$ for which $g_t : \mathbf{R} \to \mathbf{R}$ satisfies[5]

$$-g_t''(a) \le c \;\forall t \in \{1, \ldots, n\}, \forall a \in \{f(x) \mid f \in \mathcal{H}, x \in \mathcal{X}\}.$$

Then for any $f \in \mathcal{H}$, we have

$$\begin{aligned}r(\pi, f) \le&\frac{1}{2}\|f\|_{\mathcal{H}}^2 + \frac{1}{2}\log|I + cK|\\ &-\frac{1}{2}(\boldsymbol{g}'(\boldsymbol{f}) - K^{-}\boldsymbol{f})^T M(\boldsymbol{g}'(\boldsymbol{f}) - K^{-}\boldsymbol{f}),\end{aligned}$$

where $K = K(x_i, x_j) \in \mathbf{R}^{n\times n}$ is the data kernel matrix, $M = \left[K - K\left(K + c^{-1}I\right)^{-1}K\right] \succeq 0$, $\boldsymbol{g}'(\boldsymbol{f}) = (g_1'(f(x_1)), \ldots, g_n'(f(x_n)))^T$, $\boldsymbol{f} = (f(x_1), \ldots, f(x_n))^T$, and K^{-} is any generalized inverse of K. If $K \succ 0$, then $M = \left(cI + K^{-1}\right)^{-1}$.

Proof. The main step in the proof is to reduce the parameter space from an infinite dimensional RKHS to $\mathbf{R}^n$, and the GP prior to a Gaussian prior on $\mathbf{R}^n$. Notice that $\ell(f) = \sum_{t=1}^{n} g_t(f(x_t))$ depends on f only via its evaluations at $x_1, \ldots, x_n$. Thus with a slight abuse of notation, we will write ℓ for the function $\ell(\boldsymbol{f}) = \sum_{t=1}^{n} g_t(\boldsymbol{f}_t)$. In particular, we can write $\ell(f) = \ell(\boldsymbol{f})$, where on the LHS $\ell : \mathcal{H} \to \mathbf{R}$ and on the RHS, $\ell : \mathbf{R}^n \to \mathbf{R}$.

Let $\mathcal{L} = \operatorname{span}\{k(x_1, \cdot), , \ldots, , k(x_n, \cdot)\}$ be the "span of the data" in the RKHS $\mathcal{H}$. Let $f_{\|} \in \mathcal{L}$ be the projection of f onto $\mathcal{L}$. Then for any $x \in \{x_1, \ldots, x_n\}$, we have

$$\begin{aligned}f(x) &= \langle f, k(x, \cdot)\rangle_{\mathcal{H}} = \left\langle f_{\|} + f - f_{\|}, k(x, \cdot)\right\rangle_{\mathcal{H}}\\ &= \left\langle f_{\|}, k(x, \cdot)\right\rangle_{\mathcal{H}} + \underbrace{\left\langle f - f_{\|}, k(x, \cdot)\right\rangle_{\mathcal{H}}}_{=0} = f_{\|}(x).\end{aligned}$$

Therefore, for any $f \in \mathcal{H}$, we have $\ell(f) = \ell(f_{\|})$, which implies $r(\pi, f) = r(\pi, f_{\|})$. For any $f_{\|} \in \mathcal{L}$, we can write $f_{\|}(\cdot) = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot)$, for some $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^T \in \mathbf{R}^n$, and we have

$$\boldsymbol{f} = \left(\sum_{i=1}^{n} \alpha_i k(x_i, x_j)\right)_{j=1}^{n} = K\boldsymbol{\alpha}.$$

By definition of a GP, π induces a marginal distribution $\pi_n = \mathcal{N}(0, K)$ on the elements $\boldsymbol{f} \in \mathbf{R}^n$. We conclude that $r(\pi, f) = r(\pi_n, \boldsymbol{f})$, and we can apply Theorem 7. We take $J = cI$, since $-\left[\nabla_{\boldsymbol{f}}^2 \ell(\boldsymbol{f})\right] = -\operatorname{diag}(g_t''(\boldsymbol{f}_t))_{t=1}^{n} \preceq cI$. Since $\boldsymbol{f} = K\boldsymbol{\alpha}$, Equation 19 applies and we get

$$\begin{aligned}r(\pi_n, \boldsymbol{f}) \le&\frac{1}{2}\log|I + cK| + \frac{1}{2}\boldsymbol{f}^T K^{-}\boldsymbol{f}\\ &-\frac{1}{2}\left(\boldsymbol{g}'(\boldsymbol{f}) - K^{-}\boldsymbol{f}\right)^T M\left(\boldsymbol{g}'(\boldsymbol{f}) - K^{-}\boldsymbol{f}\right).\end{aligned} \tag{21}$$

Note that

$$\begin{aligned}\|f\|_{\mathcal{H}}^2 \ge&\|f_{\|}\|_{\mathcal{H}}^2 = \left\langle \sum_{i=1}^{n} \alpha_i k(x_i, \cdot), \sum_{i=1}^{n} \alpha_i k(x_i, \cdot)\right\rangle_{\mathcal{H}}\\ =&\sum_{i,j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) = \boldsymbol{\alpha}^T K\boldsymbol{\alpha} = \boldsymbol{\alpha}^T K K^{-} K\boldsymbol{\alpha}\\ =&\boldsymbol{f}^T K^{-}\boldsymbol{f}.\end{aligned}$$

Replacing $\boldsymbol{f}^T K^{-}\boldsymbol{f}$ with $\|f\|_{\mathcal{H}}^2$ and $r(\pi_n, \boldsymbol{f})$ with $r(\pi, f)$ in (21) completes the proof. □

[4]A reproducing kernel Hilbert space (RKHS) of functions from $\mathcal{X}$ to $\mathbf{R}$ is a Hilbert Space $\mathcal{H}$ that possesses a reproducing kernel, *i.e.*, a function $k : \mathcal{X} \times \mathcal{X} \to \mathbf{R}$ for which the following properties hold:

1. $k(x, .) \in \mathcal{H}$ for all $x \in \mathcal{X}$, and
2. $\langle f, k(x, .)\rangle_{\mathcal{H}} = f(x)$, for all $x \in \mathcal{X}$ and $f \in \mathcal{H}$, where $\langle\cdot, \cdot\rangle_{\mathcal{H}}$ denotes the inner product in $\mathcal{H}$.

[5]We typically have $g_t : a \mapsto \log p(y_t \mid a)$, so that $\ell(f) = -L(f, y^n, x^n)$, and $r(\pi_0, f)$ corresponds to the regret discussed above. However, we prefer to state the theorem without any direct reference to a probability model.

We note that the term $-\frac{1}{2}\left(\boldsymbol{g}'(\boldsymbol{f}) - K^{-}\boldsymbol{f}\right)^T M \left(\boldsymbol{g}'(\boldsymbol{f}) - K^{-}\boldsymbol{f}\right)$ constitutes the improvement in our bound over the bound in Theorem 1 of (Seeger et al., 2008).

6. Generalized Linear Models

A generalized linear model (GLM) maps from $x \in \mathbf{R}^d$ to a distribution on $\mathcal{Y}$ defined by the probability density $dP_{\theta,x}(y) = \exp\left[g(y, \theta' x)\right] d\mu$, with respect to some base measure μ on $\mathcal{Y}$. Now, for any $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbf{R}^d \times \mathcal{Y}$, define

$$\ell(\theta) = \sum_{t=1}^{n} g_t(\theta^T x_t).$$

When $g_t(\mu) = g(y_t, \mu)$ defined above, $\ell(\theta)$ is the conditional log-likelihood of the observed values $y_1, \ldots, y_n$ under the generalized linear model defined above. For the results below, we only require a uniform lower bound on the second derivative of g_t:

$$-g_t''(a) \leq c \; \forall t \in \{1, \ldots, n\}, \forall a \in \mathbf{R}.$$

Theorem 11. *Let* $\ell(\theta) = \sum_{t=1}^{n} g_t(\theta^T x_t)$ *with* $-g_t''(a) \leq c$ *for* $t = 1, \ldots, n$, *and* $a \in \mathbf{R}$. *Let* $\pi = \mathcal{N}(0, \Sigma_0)$ *with* $\Sigma_0 \succ 0$. *Then*

$$\begin{aligned} r(\pi, \theta_0) &\leq \tfrac{1}{2}\log\left|I + cX^T X \Sigma_0\right| + \tfrac{1}{2}\theta_0^T \Sigma_0^{-1}\theta_0 \\ &-\tfrac{1}{2}(s - \Sigma_0^{-1}\theta_0)^T \left(cX^T X + \Sigma_0^{-1}\right)^{-1}(s - \Sigma_0^{-1}\theta_0), \end{aligned} \tag{22}$$

where X is the matrix with x_t in the t'th row.

Proof. We have

$$\begin{aligned} -\nabla_\theta^2 \ell(\theta) &= -\sum_{t=1}^{n} \nabla_\theta^2 \left[g_t(\theta^T x_t)\right] \\ &= -\sum_{t=1}^{n} g_t''(\theta^T x_t) x_t x_t' \preceq c \sum_{t=1}^{n} x_t x_t^T = cX^T X, . \end{aligned}$$

where X is a matrix whose t'th row is x_t. We also have $s = \nabla_\theta \ell(\theta_0) = \sum_{t=1}^{n} g_t'(\theta_0' x_t)\, x_t$. Let $J = \varepsilon I + cX^T X$, for any $\varepsilon > 0$. Then by Theorem 7,

$$\begin{aligned} r(\pi, \theta_0) &\leq \inf_{\varepsilon > 0} \tfrac{1}{2}\log\left[\det\left(I + \varepsilon\Sigma_0 + cX^T X \Sigma_0\right)\right] + \tfrac{1}{2}\theta_0^T \Sigma_0^{-}\theta_0 \\ &- \frac{1}{2}(s - \Sigma_0^{-1}\theta_0)^T \left(\varepsilon I + cX^T X + \Sigma_0^{-1}\right)^{-1}(s - \Sigma_0^{-1}\theta_0) \end{aligned}$$

By continuity, we can take $\varepsilon = 0$ to complete the proof. □

6.1. Discussion

To get a sense for the asymptotic behavior of these bounds, let us now index everything by n. At time n the cumulative log-likelihood is $\ell_n(\theta) = \sum_{t=1}^{n} g_t(\theta^T x_t)$, the regret is given by $r_n(\pi, \theta_n) = \ell_n(\theta_n) - \log \mathbb{E}_\pi e^{\ell_n(\theta)}$, and $s_n = \nabla_\theta \ell_n(\theta_n)$. For simplicity, take the prior covariance to be $\Sigma_0 = aI$, and suppose that $\|x_t\| \leq 1$ for all t. Then $\log\left|I + cX^T X \Sigma_0\right| = O(\log n)$ and $\frac{1}{2}\theta_n^T \Sigma_0^{-1}\theta_n = O(1)$, if we take $\theta_n \to \theta_0$. Suppose $\frac{1}{n}s_n \to s_0$, as would typically be the case under iid sampling of the x_i's. Then asymptotically the three terms in the bound behave as follows:

$$\begin{aligned} r_n(\pi, \theta_n) &\leq O(\log(n)) + O(1) \\ -\frac{1}{2}(ns_0 - \Sigma_0^{-1}\theta_0)^T &\left(cX^T X + \Sigma_0^{-1}\right)^{-1}(ns_0 - \Sigma_0^{-1}\theta_0). \end{aligned}$$

In the last term, note that $X^T X = O(n)$. If $s_0 \neq 0$, the last term decreases linearly with n, which dominates the bound. On the other hand, if $s_0 = 0$, we see the last term decreases like $\frac{1}{n}$, and thus is just a small order correction to the $O(\log n)$ upper bound. These drastically different behaviors relate back to the choices of θ_n and the limit θ_0. Recall that when $\theta_0 = \theta_{\text{MLE}}$, we have $s_0 = 0$. If θ_n is bounded away from θ_{MLE}, then BMA with any nondegenerate prior will eventually perform better than θ_n, thus the regret will go negative and decrease with n. This behavior is captured by the last term, which isn't present in the bounds in (Kakade & Ng, 2004).

7. Conclusions

We have started with a very general BMA setting, allowing arbitrary priors and requiring only that the Hessian of the log-loss function have a positive definite upper bound — that is, the log-loss cannot be "too convex." This assumption seems reasonable in several practical settings (two of which we have examined), and it allows us to get non-asymptotic regret bounds for all sequences using the Laplace approximation as the first ingredient. While the bound in Theorem 1 is our tightest and most general bound, it presents challenges to being used in practice. First, it has a term involving the "score" vector $s = \nabla_\theta \ell(\theta_0)$, which may not be easy to control in most natural settings. As we note in Corollary 2, the dependence on s vanishes when we take θ_{MLE} as our point of comparison. We can also use the variational technique from (Kakade et al., 2005; Kakade & Ng, 2004) to eliminate the dependence on s. Theorem 6 illustrates their technique in our more general setting. Although the resulting bound is simpler, our discussion in Section 6.1 shows that the behavior illustrated by the bound is misleading when the comparison point θ_0 is bounded away from θ_{MLE}.

The second major challenge in the bound of Theorem 1 is the term involving the integral of a Gaussian density with respect to the prior distribution. We mention an approach to this problem, but we still consider it an open problem. For the specific case of a Gaussian prior, our results improve on earlier results, and thanks to a matrix identity, we can write the difference in the bounds with a single quadratic form.

References

Banerjee, Arindam. On Bayesian bounds. In *Proceedings of the 23rd international conference on Machine learning*, ICML '06, pp. 81–88, New York, NY, USA, 2006. ACM. ISBN 1-59593-383-2. doi: 10.1145/1143844.1143855. URL http://dx.doi.org/10.1145/1143844.1143855.

Cesa-Bianchi, Nicolo and Lugosi, Gabor. *Prediction, Learning, and Games*. Cambridge University Press, March 2006. ISBN 0521841089. URL http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0521841089.

Kakade, S., Seeger, M., and Foster, D. Worst-Case Bounds for Gaussian Process Models. In *Neural Information Processing Systems*, December 2005. URL http://www.kyb.tuebingen.mpg.de/bs/people/seeger/.

Kakade, Sham M. and Ng, Andrew Y. Online Bounds for Bayesian Algorithms. In *NIPS*, 2004.

Petersen, K. B. and Pedersen, M. S. The Matrix Cookbook. Technical University of Denmark, October 2008. URL http://www2.imm.dtu.dk/pubdb/p.php?3274.

Seeger, Matthias W., Kakade, Sham M., and Foster, Dean P. Information Consistency of Nonparametric Gaussian Process Methods. *IEEE Transactions on Information Theory*, 54(5):2376–2382, May 2008. ISSN 0018-9448. doi: 10.1109/TIT.2007.915707. URL http://dx.doi.org/10.1109/TIT.2007.915707.

A. Proof of Lemma 9 (Matrix Identity)

Proof. Let $F = \left(\Sigma + J^{-1}\right)^{-1}$. On the right hand side, we have

$$\begin{aligned}
&\alpha'\Sigma\alpha - (s-\alpha)'\left[\Sigma - \Sigma F\Sigma\right](s-\alpha)\\
=&\alpha'\Sigma\alpha - s'\Sigma s + 2\alpha'\Sigma s - \alpha'\Sigma\alpha\\
&+ s'\Sigma F\Sigma s + \alpha'\Sigma F\Sigma\alpha - 2\alpha'\Sigma F\Sigma s\\
=&\alpha'\Sigma F\Sigma\alpha + 2\alpha'(\underbrace{\Sigma - \Sigma F\Sigma}_{M})s + s'(\underbrace{\Sigma F\Sigma - \Sigma}_{N})s
\end{aligned}$$

On the left hand side, we have

$$\begin{aligned}
&-s'J^{-1}s + \left(\Sigma\alpha + J^{-1}s\right)' F \left(\Sigma\alpha + J^{-1}s\right)\\
=&-s'J^{-1}s + \alpha'\Sigma F\Sigma\alpha + 2\alpha'\Sigma FJ^{-1}s + 2\alpha'\Sigma FJ^{-1}s\\
=&\alpha'\Sigma F\Sigma\alpha + 2\alpha'\underbrace{\Sigma FJ^{-1}}_{M_0}s + s'(\underbrace{J^{-1}FJ^{-1} - J^{-1}}_{N_0})s
\end{aligned}$$

From the following standard matrix identity[6]

$$A - A(A+B)^{-1}A = B - B(A+B)^{-1}B,$$

it follows that $N = N_0$. We also have

$$\begin{aligned}
M - M_0 =&\Sigma - \Sigma F\Sigma - \Sigma FJ^{-1}\\
=&\Sigma\left(I - F\left(\Sigma + J^{-1}\right)\right)\Sigma(I - I) = 0.
\end{aligned}$$

Thus the first claim is true. Now note that if we replace α by $\Sigma^-\Sigma\alpha$ on the left hand side, nothing changes, since $\Sigma\Sigma^-\Sigma\alpha = \Sigma\alpha$, by definition of Σ^-. Thus we can make the replacement on both sides. Finally, we show that $M \succeq 0$: Define

$$A = \begin{pmatrix} \Sigma + J^{-1} & \Sigma \\ \Sigma & \Sigma \end{pmatrix}.$$

Then $M = \Sigma - \Sigma\left(\Sigma + J^{-1}\right)^{-1}\Sigma$ is the Schur complement of $\Sigma + J^{-1}$ in A. $A \succeq 0$ implies $M \succeq 0$. We know $A \succeq 0$ since for any $a, b \in \mathbf{R}^n$,

$$\begin{aligned}
&(a'\ b')\begin{pmatrix} \Sigma + J^{-1} & \Sigma \\ \Sigma & \Sigma \end{pmatrix}\begin{pmatrix} a \\ b \end{pmatrix}\\
&= (a'\ b')\left[\begin{pmatrix} I \\ I \end{pmatrix}\Sigma\begin{pmatrix} I & I \end{pmatrix} + \begin{pmatrix} J^{-1} & 0 \\ 0 & 0 \end{pmatrix}\right]\begin{pmatrix} a \\ b \end{pmatrix}\\
&= (a+b)^T\Sigma(a+b) + a'J^{-1}a \geq 0,
\end{aligned}$$

The last expression is nonnegative since $\Sigma \succeq 0$ and $J^{-1} \succ 0$. Finally, if $\Sigma \succ 0$, by the Woodbury Identity[7], we have $M = \left(J + \Sigma^{-1}\right)^{-1}$. □

[6] This is easy to prove if we use the substitution $B = C - A$. See Section 3.2.4 in (Petersen & Pedersen, 2008) for further references.

[7] When the inverses exist, $(A+BC)^{-1} = A^{-1} - A^{-1}B(I+CA^{-1}B)^{-1}CA^{-1}$. See Equation 148 in (Petersen & Pedersen, 2008).