B_3O: Bayes Empirical Bayes by Bayesian Optimization

James McInerney
contact@jamesmc.com

Abstract

There is rapidly growing interest in using Bayesian optimization to tune learning parameters and hyperparameters for approximate inference algorithms that take a long time to run. For example, Spearmint is a popular software package for selecting the optimal number of layers and learning rate in neural networks. But given that there is uncertainty about which hyperparameters give the best predictive performance, and given that fitting a model for each choice of hyperparameters is costly, it is arguably wasteful to "throw away" all but the best result, as per Bayesian optimization. In this paper, I consider an alternative approach that uses all the samples from the Bayesian optimization procedure to average over the uncertainty in model hyperparameters. The resulting approach, Bayes Empirical Bayes by Bayesian Optimization (B_3O), predicts held-out data better that Bayesian optimization in two experiments on logistic regression and latent Dirichlet allocation.

1 Introduction

There is rapidly growing interest in using Bayesian optimization (BO) to tune learning parameters and hyperparameters for approximate inference algorithms that take a long time to run (Snoek et al., 2012). Such tuning by hand was previously a time consuming task requiring trial, error, and expert knowledge of the model. To capture this knowledge, BO uses a meta-model (usually a Gaussian process) as a guide to regions of high performance to sequentially explore hyperparameters.

BO for approximate inference algorithms is a form of model selection in which some objective, such as predictive likelihood or root mean squared error, is optimized with respect to hyperparameters η. Thus, it is an empirical Bayesian procedure where the marginal likelihood is replaced by a proxy objective. Empirical Bayes optimizes the marginal likelihood of data set X,

$$\hat{\eta} \coloneqq \arg\max_{\eta} \int p(X \mid \theta) p(\theta \mid \eta) \mathrm{d}\theta, \tag{1}$$

then uses $p(\theta \mid X, \hat{\eta})$ as the posterior distribution over the unknown model parameters θ (Carlin and Louis, 2000). Empirical Bayes is applied in different ways, e.g., gradient-based optimization of Gaussian process kernel parameters, optimization of hyperparameters to conjugate priors in variational inference. What is special about BO is that it performs empirical Bayes in a way that requires calculating the posterior $p(\theta \mid X, \eta^{(s)})$ for *each* member in a sequence $1, \ldots, S$ of candidate hyperparameters $\eta^{(1)}, \eta^{(2)}, \ldots, \eta^{(S)}$. Often these posteriors will be approximate, such as a point estimate, a Monte Carlo estimate, or a variational approximation. Nonetheless, these operations are usually expensive to compute.

Therefore, what is surprising about BO for approximate inference is that it disregards most of the computed posteriors $J = \{p(\theta \mid X, \eta^{(s)}) \mid s \in [1, S], \eta^{(s)} \neq \hat{\eta}\}$ and keeps only the posterior $p(\theta \mid X, \hat{\eta})$ that optimizes the marginal likelihood. It is surprising because the intermediate posteriors J have something to say about the data, even if they condition on hyperparameter configurations

that do not maximize the marginal likelihood. In other words, when we harbour uncertainty about η, should we be more Bayesian? Yes, especially if one believes there is a danger of overfitting η on the validation set, which is the case as the dimensionality of the hyperparameters grows.

Bayes empirical Bayes (Carlin and Louis, 2000) extends the empirical Bayes paradigm by introducing a family of hyperpriors $p(\eta \mid \lambda)$ indexed by λ and calculates the posterior over the model parameters by integrating,

$$p(\theta \mid X, \lambda) = \int p(\theta \mid X, \eta) p(\eta \mid \lambda) \mathrm{d}\eta. \tag{2}$$

This leads naturally to what one might call *type II* empirical Bayes where where the nuisance hyper-hyper-parameter λ is maximized out,

$$\hat{\lambda} \coloneqq \arg\max_{\lambda} \int \int p(X \mid \theta) p(\theta \mid \eta) p(\eta \mid \lambda) \mathrm{d}\theta \mathrm{d}\eta, \tag{3}$$

and $p(\theta \mid X, \hat{\lambda})$ is used as the posterior. Comparing Eq. 3 to Eq. 1 highlights that type II empirical Bayes adds an extra layer of marginalization that can be exploited with set J at hand. Note the distinction between marginalizing the hyperparameters to the model vs. hyperparameters to the GP. Eq. 3 describes the former; the latter is already a staple of BO (Osborne, 2010).

In this paper, I present *Bayes empirical Bayes for Bayesian optimization* (or *B_3O*), an extension to BO that approximates Eq. 3 and makes use of all the posteriors computed during standard Bayesian optimization in the following Monte Carlo approximations to give an approximate posterior,

$$p(\theta \mid X, \hat{\lambda}) \approx \sum_{s=1}^{S} \bar{w}_{\hat{\lambda}}^{(s)} p(\theta \mid X, \eta^{(s)}), \quad \text{where} \;\; \eta^{(s)} \sim A_s(\cdot) \tag{4}$$

$$\hat{\lambda} \approx \arg\max_{\lambda} \sum_{s=1}^{S} \bar{w}_{\lambda}^{(s)} \int p(X \mid \theta) p(\theta \mid \eta^{(s)}) \mathrm{d}\theta, \quad \text{where} \;\; \eta^{(s)} \sim A_s(\cdot) \tag{5}$$

for the set of importance weights $W = \{\bar{w}_{\lambda}^{(s)} \mid s \in [1, S]\}$ that correct for the optimization procedure $A_s(\eta^{(s)})$ used to select candidate $\eta^{(s)}$ while accounting for the prior $p(\eta^{(s)} \mid \lambda)$. In the sequel, I expand on the definition of $\bar{w}^{(s)}$, present the $\mathrm{B}_3\mathrm{O}$ algorithm, test it on some concrete models, and briefly discuss related work.

2 Bayes Empirical Bayes by Bayesian Optimization ($\mathrm{B}_3\mathrm{O}$)

The approximation in Eq. 4 sums over the uncertainty of hyperparameter selection for learning from a dataset X. This sum expresses the posterior $p(\theta \mid X, \hat{\lambda})$ in terms of a weighted sum over every posterior candidate $p(\theta \mid X, \eta^{(s)})$ calculated during BO. The weights play the role of correcting for the bias introduced by the candidate selection procedure $A_s(\eta^{(s)})$ and account for the prior over hyperparameters, $\bar{w}_{\lambda}^{(s)} \propto \frac{p(\eta^{(s)} \mid \lambda)}{A_s(\eta^{(s)})}$ such that $\sum_s \bar{w}_{\lambda}^{(s)} = 1$ (normalized to reduce the variance of the overall estimate (Owen, 2014)). Both distributions are given next.

Acquisition Probability There are many choices of acquisition function in BO, such as the expected probability of improvement, probability of improvement, and upper confidence bound (Brochu et al., 2010). These are all deterministic policies meaning that they do not share the same non-zero support as the target distribution $p(\eta \mid \lambda)$, making them unsuitable for the importance sampling reweighting in Eq. 5 for $\mathrm{B}_3\mathrm{O}$.

Thompson sampling is a stochastic policy consisting of two steps: (1) sample $f^{(s)} \sim \mathcal{N}(y | m^{(s)}, \Sigma^{(s)})$, and, (2) choose $\eta^{(s)} = \arg\min_{\eta} f^{(s)}(\eta)$, where $(m^{(s)}, \Sigma^{(s)})$ is the posterior mean and covariance of the meta-model. This procedure has been shown to be competitive with other acquisition functions for a range of problems (Chapelle and Li, 2011). It balances exploration with exploitation because regions with small posterior means and regions with high variance are both more likely to appear as the minimum point in the sampled function $y^{(s)}$. For these reasons, Thompson sampling is the assumed acquisition method for the rest of the paper.

Table 1: Logistic Regression, MNIST dataset

Method	Log Lik (per image)	Accuracy (per image)
Bayes empirical Bayes by Bayesian opt. (B_3O)	**-0.3804**	**8.81%**
Bayesian optimization	-0.3839	8.94%

Table 2: LDA, 20 Newsgroups dataset

Method	Total Log Lik	Log Lik (per word)
Bayes empirical Bayes by Bayesian opt. (B_3O)	**-396747**	**-6.39**
Bayesian optimization	-474772	-7.65

After discretizing the input space to D possible candidates, the hyperparameter selection probability for Thompson sampling is,

$$A(\eta_i) = \mathbb{E}_{\mathcal{N}(f(\eta_{1:D}) \mid m, \Sigma)} \left[\prod_{j \neq i}^{D} \mathbb{I}[f(\eta_i) < f(\eta_j)] \right], \tag{6}$$

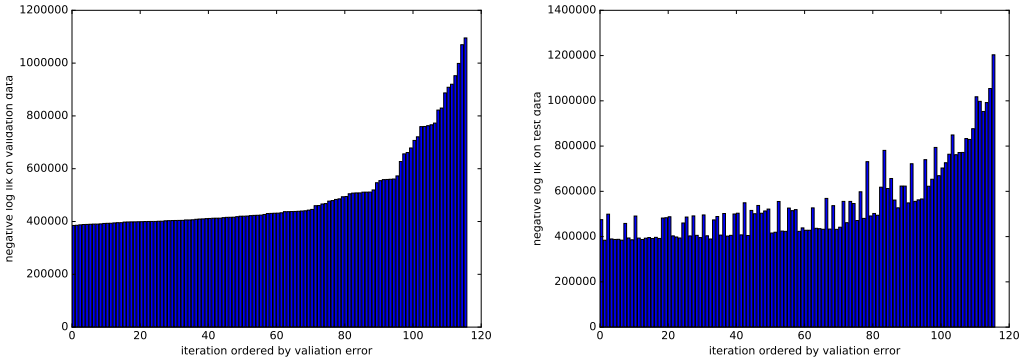
where I have suppressed the (s) index to simplify notation, and $\mathbb{I}[b]$ is the Iverson bracket evaluating to 1 when b is true and 0 otherwise. The Monte Carlo estimator of Eq. 6 has high variance so I discuss an approximation in Appendix B that assumes independence of the dimensions of f.

Hyperparameter Prior The other component for the weights in Eq. 4 is the prior over the hyperparameters $p(\eta \mid \lambda)$. In this work, I place a uniform mass over a subset $J' \subset J$ of the posteriors, where $|J'| = \frac{1}{10}|J|$. Therefore, after discretizating the input space, $p(\eta_i \mid \lambda) = \frac{\mathbb{I}[\lambda_i = 1]}{|J'|}$, where λ is a binary vector of length $|J|$ and $\sum_i \lambda_i = \frac{1}{10}|J|$. B_3O maximizes Eq. 5 with respect to λ.

In summary, the B_3O algorithm is given in Algorithm 1 in Appendix A. The computational complexity of B_3O is the same as for BO, which is dominated by the D posterior approximations and inverting the D-by-D matrix to calculate the GP posterior (an $\mathcal{O}(D^3)$ operation).

3 Experiments

I apply B_3O and BO to two approximate inference algorithms and two data sets. The first is maximum *a posteriori* logistic regression on the MNIST digit data set (LeCun et al., 1998). The second is stochastic variational inference on latent Dirichlet allocation (SVI-LDA) applied to the



(a) Negative log likelihood on validation data

(b) Negative log likelihood on test data

Figure 1: Performance in negative logarithm of the predictive likelihood for the validation data (left plot) and test data (right plot). Each iteration represents a different hyperparameter setting.

20 Newsgroups data. In both cases, I find that B_3O outperforms BO on predictive likelihood and accuracy. The performance improvement is larger in SVI-LDA, which has 4 hyperparameters to optimize, supporting the hypothesis stated in Section 1 that BO is vulnerable to overfitting on the validation data and that B_3O protects against this via approximate marginalization.

Throughout, I randomly split the data into into training, validation, and test sets. BO and B_3O evaluate performance on the validation set (after training on the training set) to determine the best hyperparameters. The test data is only used at the very end to report overall performance.

3.1 Logistic Regression

I apply logistic regression to the MNIST digit data set, comprising 70k images of handwritten digits 0-9, with supervision labels. The data set was binarized and split into 50k training, 10k validation, and 10k test images. The approximate inference algorithm used was the maximum *a posteriori* estimate of the coefficients θ in Eq. 5 as a delta function $p(\theta \mid X, \eta) \approx \delta_{X,\eta}(\hat{\theta})$ found through gradient ascent. This requires the selection of two hyperparameters: the strength of the prior, expressed on the scale $[0.0001, 1]$ as the relative proportion of the prior w.r.t. the data, and the constant learning rate, expressed on the $\log_e$ scale $[-10, 0]$. Table 1 compares the performance in log likelihood and accuracy on the test set for both approaches, showing that B_3O does slightly better than BO.

3.2 Latent Dirichlet Allocation

Latent Dirichlet allocation (LDA) finds topic structure in a set of text documents expressed as K word distributions (one per topic) and D topic distributions (one per document). I apply stochastic variational inference to LDA (Hoffman et al., 2013), a method that approximates the posterior over parameters $p(\theta \mid X, \eta)$ in Eq. 5 with variational distribution $q(\theta \mid v, \eta)$. The algorithm minimizes the KL divergence between q and p by adjusting the variational parameters v. SVI-LDA has several hyperparameters; four of them were varied during these experiments comprising three model-related hyperparameters and one learning rate hyperparameter. Full details of hyperparameters, document, and vocabulary selection are given in Appendix C.

Table 2 shows performance in log likelihood on the test data of the two approaches. B_3O performs significantly better than BO in this problem. To understand why, Figure 1 examines the error (negative log liklihood) on both the validation and test data for all the hyperparameters selected during BO. In the test scenario, BO chooses the hyperparameters corresponding to the left-most bar in Figure 1b because those hyperparameters minimized error on the validation set. However, Figure 1b clearly shows that other hyperparameter settings outperform this selection. For finite validation data, there is no way of knowing how the optimal hyperparameter will behave on test data before seeing it, motivating an averaging approach like B_3O.

4 Related Work

Empirical Bayes has a long history started by Robbins (1955) with a nonparametric approach, to parametric EB (Efron and Morris, 1972) and modern applications of EB (Snoek et al., 2012; Rasmussen and Williams, 2006). Another use of the GP as a meta-model arises in Bayesian quadrature, which uses GPs to approximately marginalize over model parameters θ (Osborne et al., 2012). However, quadrature is computationally infeasible for hyperparameter marginalization because $p(X \mid \eta)$ is expensive to compute. Finally, B_3O resembles ensemble methods, such as boosting and bagging, because it is a weighted sum over posteriors. Boosting trains models on data reweighted to emphasize errors from previous models (Freund et al., 1999) while bagging takes an average of models trained on bootstrapped data (Breiman, 1996).

5 Conclusions and Future Work

In this paper I introduced the B_3O, an extension to BO that approximately integrates over the unknown hyperparameters of arbitrary machine learning algorithms. I applied the method to two algorithms finding an improvement over BO. In future work, I will make comparisons on more algorithms and larger-scale data. I will also investigate new hyperpriors $p(\eta \mid \lambda)$ and other acquisition functions.

References

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2):123–140.

Brochu, E., Cora, V. M., and De Freitas, N. (2010). A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599*.

Carlin, B. P. and Louis, T. A. (2000). Empirical Bayes: Past, present and future. *Journal of the American Statistical Association*, 95(452):1286–1289.

Chapelle, O. and Li, L. (2011). An empirical evaluation of Thompson sampling. In *Advances in Neural Information Processing Systems*, pages 2249–2257.

Efron, B. and Morris, C. (1972). Limiting the risk of Bayes and empirical Bayes estimators—Part II: The empirical Bayes case. *Journal of the American Statistical Association*, 67(337):130–139.

Freund, Y., Schapire, R., and Abe, N. (1999). A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780):1612.

Girolami, M. and Rogers, S. (2006). Variational Bayesian multinomial probit regression with Gaussian process priors. *Neural Computation*, 18(8):1790–1817.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. W. (2013). Stochastic variational inference. *Journal of Machine Learning Research*, 14(1):1303–1347.

LeCun, Y., Cortes, C., and Burges, C. J. (1998). *The MNIST database of handwritten digits*.

Osborne, M. (2010). *Bayesian Gaussian Processes for Sequential Prediction, Optimisation and Quadrature*. PhD thesis, PhD thesis, University of Oxford.

Osborne, M., Garnett, R., Ghahramani, Z., Duvenaud, D. K., Roberts, S. J., and Rasmussen, C. E. (2012). Active Learning of Model Evidence Using Bayesian Quadrature. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 25*, pages 46–54. Curran Associates, Inc.

Owen, A. B. (2014). *Monte Carlo theory, methods and examples (book draft)*.

Rasmussen, C. E. and Williams, C. K. (2006). Gaussian processes for machine learning. *the MIT Press*, 2(3):4.

Robbins, H. (1955). The empirical Bayes approach to statistical decision problems. In *Herbert Robbins Selected Papers*, pages 49–68. Springer.

Snoek, J., Larochelle, H., and Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959.

A Algorithm for $\mathrm{B}_3\mathrm{O}$

Data: $X_{\mathrm{train}}, X_{\mathrm{validation}}$
Result: posterior $p(\theta \mid X_{\mathrm{train}}, X_{\mathrm{valid}}, \hat{\lambda})$ with optimal hyper-hyper-parameter $\hat{\lambda}$
initialize validation error history $V = \{\}$
while V *not converged* **do**
  approximate meta-model posterior $\mathrm{GP}(m, \Sigma \mid V)$
  sample error curve $f^{(s)} \sim \mathrm{GP}(m, \Sigma \mid V)$
  select hyperparameter $\eta^{(s)} := \arg\min_\eta f(\eta)$
  calculate/approximate model posterior $p(\theta \mid X_{\mathrm{train}}, \eta^{(s)})$
  evaluate model posterior $\mathrm{v}^{(\mathrm{s})} := \int p(X_{\mathrm{valid}} \mid \theta) p(\theta \mid X_{\mathrm{train}}, \eta^{(s)}) \mathrm{d}\theta$
  append performance to history $V := V \cup \{(\eta^{(s)}, -v^{(s)})\}$
end
find $\hat{\lambda}$ using Eq. 5
approximate $p(\theta \mid X_{\mathrm{train}}, X_{\mathrm{valid}}, \hat{\lambda})$ using Eq. 4

Algorithm 1: $\mathrm{B}_3\mathrm{O}$

B Approximation of Acquisition Probability

The expectation in Eq. 6 can be approximated by replacing the off-diagonal entries of the posterior covariance matrix Σ with zero, which makes $y_j \perp\!\!\!\perp y_i, \forall j \neq i$. The resulting expectation is identical to the label probability in the multinomial probit (Girolami and Rogers, 2006),

$$p(y = i) = \mathbb{E}_{\mathcal{N}(u \mid 0, \sigma_i^2)} \left[\prod_{j \neq i} \phi(u - m_i + m_j) \right], \tag{7}$$

for posterior mean and variance (m_j, σ_j) for dimension j.

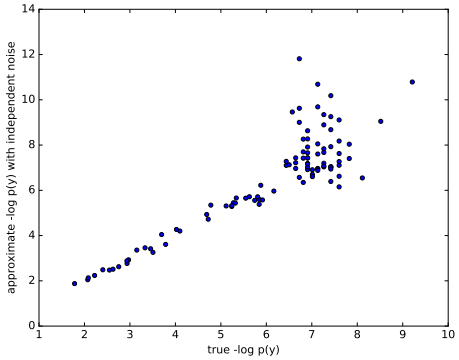


Figure 2: Scatter of approximate $-\log(y)$ against true $-\log(y)$ shows that the approximation works well except for extreme values of y.

To explore the effectiveness of the approximation I used the hyperparameter and performance results from the SVI-LDA detailed in Section 3.2. Figure 2 shows the approximated probability (Eq. 7) of selecting hyperparameter η in the Thompson sampling acquisition method against the true probability in Eq. 6 (evaluated using 100,000 samples per point). The approximation works well except for extreme values of y. Intuitively, when function draws at point η are far away from the mean and highly correlated with nearby points, they make the event of choosing η much more likely in the full covariance normal than the diagonalized covariance normal distribution.

C Hyperparameter Settings for SVI-LDA

SVI-LDA has four variable hyperparameters in the experiments for Section 3.2:

- K, range $[50, 200]$, the number of topics;
- α, $\log_e$ range $[-5, 0]$, the hyerparameter to the Dirichlet document-topic prior;
- η, $\log_e$ range $[-5, 0]$, the hyperparameter to the Dirichlet topic-word distribution prior;
- κ, range $[0.5, 0.9]$, the decay parameter to the learning rate $(t_0 + t)^{-\kappa}$, where t_0 was fixed at 10 for this experiment.

Several other hyperparameters are required and were kept fixed during the experiment. The minibatch size was fixed at 100 documents and the vocabulary was selected from the top 1,000 words, excluding stop words, words that appear in over 95% of documents, and words that appear in only one document. The 11,314 resulting documents were randomly split 80%-10%-10% into training, validation, and test sets.