Are you using test log-likelihood correctly?

Sameer K. Deshpande[*]
University of Wisconsin–Madison
Madison, WI 53706
`sameer.deshpande@wisc.edu`

Soumya Ghosh[*]
MIT-IBM Watson AI Lab & IBM Research
Cambridge, MA 02142
`ghoshso@us.ibm.com`

Tin D. Nguyen[*]
MIT-IBM Watson AI Lab & Massachusetts Institute of Technology
Cambridge, MA 02139
`tdn@mit.edu`

Tamara Broderick
MIT-IBM Watson AI Lab & Massachusetts Institute of Technology
Cambridge, MA 02139
`tbroderick@mit.edu`

Abstract

Test log-likelihood is commonly used to compare different models of the same data and different approximate inference algorithms for fitting the same probabilistic model. We present simple examples demonstrating how comparisons based on test log-likelihood can contradict comparisons according to other objectives. Specifically, our examples show that (i) conclusions about forecast accuracy based on test log-likelihood comparisons may not agree with conclusions based on other distributional quantities like means; and (ii) that approximate Bayesian inference algorithms that attain higher test log-likelihoods need not also yield more accurate posterior approximations.

1 Introduction

Test log-likelihood[2] is often used to compare different models of the same data or to compare different algorithms used to fit the same probabilistic model. Although there are compelling reasons for this practice (Section 2.1), we provide counter-examples to the following, usually implicit, claims:

- **Claim**: The higher the test log-likelihood, the more accurately an approximate inference algorithm recovers the Bayesian posterior distribution of latent model parameters (Section 3).
- **Claim**: The higher the test log-likelihood, the more accurately the model can predict the mean, variance, or quantiles of the true data generating process (Section 4).

Our counter-examples demonstrate that test log-likelihood is not always a good proxy for posterior approximation error. They further demonstrate that forecast evaluations based on test log-likelihood may not agree with forecast evaluations based on other important distributional quantities.

We are not the first to highlight discrepancies between test log-likelihood and other analysis objectives. For instance, Quinonero-Candela et al. (2005) and Kohonen and Suomela (2005) showed that when

---

[*] Contributed equally

[2] Also known as predictive log-likelihood or test log-predictive. It is computed as the log-predictive density averaged over a set of held-out data

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

predicting discrete data with continuous distributions, test log-likelihood can be made arbitrarily large by concentrating probability into vanishingly small intervals. Chang et al. (2009) observed that topic models with larger test log-predictive densities can be less interpretable. Yao et al. (2019) highlighted the disconnect between test log-likelihood and posterior approximation error in the context of Bayesian neural networks. Our examples, however, reveal more fundamental issues with test log-likelihood. In particular, we show how comparisons based on test log-likelihood can contradict comparisons based on other objectives even in simple models like linear regression.

After introducing our notation, we precisely define test log-likelihood and review arguments for its use in Section 2. In Sections 3 and 4, we present counter-examples highlighting some limitations of using test log-likelihood as a default predictive loss function or model selection criterion. We conclude in Section 5 with a reflection on when we should use test log-likelihood in practice.

2 Background

Practitioners often model training data $\mathcal{D} = \{x_n\}_{n=1}^N$, which are assumed to be distributed according to an unknown probability distribution $\mathcal{P}$ with density $p(x)$, by introducing a parameter θ and specifying a conditional distribution $\Pi(X|\theta)$ with density $\pi(x|\theta)$. In a non-Bayesian analysis, one usually computes a point estimate $\hat{\theta}$ of the unknown parameter (e.g. by maximum likelihood). A Bayesian analysis elaborates the conditional model by specifying a prior distribution $\Pi(\theta)$ and formally computes the density $\pi(\theta|\mathcal{D})$ of the posterior distribution $\Pi(\theta|\mathcal{D})$ from the assumed joint distribution $\Pi(X, \theta)$.

Upon computing $\hat{\theta}$ or the posterior density $\pi(\theta|\mathcal{D})$, one can ask how well the fitted model predicts new data generated from $\mathcal{P}$. Given a point estimate $\hat{\theta}$, the predictive density evaluated at $x^\star$ is just $\pi(x^\star|\hat{\theta})$. The Bayesian posterior predictive density is given by

$$\pi(x^\star|\mathcal{D}) = \int \pi(x^\star|\theta)\pi(\theta|\mathcal{D})d\theta.$$

Observe that $\pi(x^\star|\hat{\theta})$ is numerically equal to $\pi(x^\star|\mathcal{D})$ when the prior $\Pi(\theta)$ is a point-mass at $\hat{\theta}$.

Practitioners commonly assess how well their fitted model predicts out-of-sample using a held-out set of testing data $\mathcal{D}^\star = \{x_n^\star\}_{n=1}^{N^\star}$, which was not used to train the model. To compute test log-likelihood, they average evaluations of the log-predictive density function over the testing set.

$$\text{lpd}(\mathcal{D}^\star; \Pi) := \frac{1}{N^\star} \sum_{n=1}^{N^\star} \log \pi(x_n^\star|\mathcal{D}), \tag{1}$$

where our notation makes explicit the dependence of lpd on testing data $\mathcal{D}^\star$ and the chosen model Π. The abbreviation lpd is a reminder that test log-likelihood involves log-predictive density evaluations.

2.1 The case for test log-likelihood

Researchers commonly use lpd to select between two models of the data, say Π and $\tilde{\Pi}$; that is, they select model Π over $\tilde{\Pi}$ whenever $\text{lpd}(\mathcal{D}^\star; \Pi) > \text{lpd}(\mathcal{D}^\star; \tilde{\Pi})$. To understand this criterion, consider the *expected log-predictive density*,

$$\text{elpd}(\Pi) := \int \log \pi(x^\star|\mathcal{D})p(x^\star)dx^\star.$$

Our use of the abbreviation elpd follows the example of Gelman et al. (2014, Equation 1). Under mild assumptions about $\mathcal{P}$ and Π, $\text{lpd}(\mathcal{D}^\star; \Pi) \to \text{elpd}(\Pi)$ as the number of testing points $N^\star$ diverges. Expected log-predictive density is closely related to the Kullback–Leibler divergence:

$$\text{KL}\left(\mathcal{P}(x^\star) \parallel \Pi(x^\star|\mathcal{D})\right) = \int p(x^\star) \log p(x^\star)dx^\star - \text{elpd}(\Pi).$$

Thus, assuming that test set $\mathcal{D}^\star$ is sufficiently large, if $\text{lpd}(\mathcal{D}^\star; \Pi) > \text{lpd}(\mathcal{D}^\star; \tilde{\Pi})$ we can reasonably conclude that $\text{elpd}(\Pi) > \text{elpd}(\tilde{\Pi})$, which in turn implies that $\Pi(x^\star|\mathcal{D})$ is closer to $\mathcal{P}(x^\star)$ than

$\tilde{\Pi}(x^\star|\mathcal{D})$ in a KL sense. In other words, we would expect predictions made using the fitted model with larger lpd to be closer (in a KL sense) to realizations from the true data generating process.

In addition to being essentially the only strictly proper local scoring rule (Bernardo and Smith, 2000, Proposition 3.13), in the absence of application-specified predictive loss, lpd may be seen as a "non-informative" choice (Robert, 1996; Gelman et al., 2014). When $\Pi(x^\star|\mathcal{D})$ is assumed to be Gaussian, elpd is intimately related to another proper scoring rule: the Dawid–Sebastiani score (Dawid and Sebastiani, 1999). Namely, elpd is equal to the Dawid–Sebastiani score plus a constant that does not depend on the model or the data-generating process. Further, for Gaussian predictive distributions, the highest possible elpd is obtained whenever the means and variances of $\Pi(x^\star|\mathcal{D})$ and $\mathcal{P}(x^\star)$ are identical. By contrast, minimizing mean square error is equivalent to only matching the means of $\Pi(x^\star|\mathcal{D})$ and $\mathcal{P}(x^\star)$.

Model comparison with lpd makes two implicit assumptions: (i) that $\text{lpd}(\mathcal{D}^\star;\cdot)$ is a close approximation to $\text{elpd}(\cdot)$ and (ii) that closeness between $\Pi(x^\star|\mathcal{D})$ and $\mathcal{P}$ in a KL sense is desirable. As we will see shortly, however, KL closeness to $\mathcal{P}$ does not necessarily imply closeness of other distributional quantities or of posterior approximation quality.

3 Claim: lpd accurately assesses posterior approximation quality

In most Bayesian analyses, the posterior density $\pi(\theta|\mathcal{D})$ is analytically intractable and practitioners must instead rely on approximate posterior computations. There are myriad approximate inference algorithms (e.g. Laplace approximation, Hamiltonian Monte Carlo, coordinate ascent mean-field variational inference, to name just a few). All of these algorithms aim to approximate the same posterior $\Pi(\theta|\mathcal{D})$. Log predictive-density is often used to compare the quality of different approximations, with higher lpd values assumed to reflect more accurate approximations, in the context of variational inference (see, e.g., Hoffman et al., 2013; Ranganath et al., 2014; Hernández-Lobato et al., 2016; Liu and Wang, 2016; Shi et al., 2018) and Bayesian deep learning (see, e.g., Hernández-Lobato and Adams, 2015; Gan et al., 2016; Li et al., 2016; Louizos and Welling, 2016; Sun et al., 2017; Ghosh et al., 2018; Mishkin et al., 2018; Wu et al., 2019; Izmailov et al., 2020, 2021; Ober and Aitchison, 2021).

Formally, suppose that our exact posterior is $\Pi(\theta|\mathcal{D})$ and that we have two approximate inference algorithms that produce two approximate posteriors $\hat{\Pi}_1(\theta|\mathcal{D})$ and $\hat{\Pi}_2(\theta|\mathcal{D})$. The exact posterior and its approximations respectively induce predictive distributions $\Pi(x^\star|\mathcal{D}), \hat{\Pi}_1(x^\star|\mathcal{D})$, and $\hat{\Pi}_2(x^\star|\mathcal{D})$. For instance, $\hat{\Pi}_1(\theta|\mathcal{D})$ could be the empirical distribution of samples drawn using HMC and $\hat{\Pi}_2(\theta|\mathcal{D})$ could be a mean-field variational approximation. Our first example demonstrates that it is possible that (i) $\text{lpd}(\mathcal{D}^\star;\hat{\Pi}_1) > \text{lpd}(\mathcal{D}^\star;\Pi)$ but (ii) using $\hat{\Pi}_1$ could lead to different inference about model parameters than using the exact posterior Π. Our second example demonstrates that it is possible that (i) $\text{lpd}(\mathcal{D}^\star;\hat{\Pi}_1) > \text{lpd}(\mathcal{D}^\star;\hat{\Pi}_2)$ but (ii) $\hat{\Pi}_1(\theta|\mathcal{D})$ is a worse approximation to the true posterior $\Pi(\theta|\mathcal{D})$ than $\hat{\Pi}_2(\theta|\mathcal{D})$. We use the 2-Wasserstein distance to assess posterior approximation quality because 2-Wasserstein can be used to bound differences in means and variances (Huggins et al., 2020, Theorem 3.4). We also assess differences in standard deviations directly in Appendix B.

lpd and downstream posterior inference. Relying on lpd for model selection can lead to different inferences than we would find by using the exact posterior. To illustrate, suppose we observe $\mathcal{D}_{100} = \{(x_n, y_n)\}_{n=1}^{100}$ drawn from the following heteroscedastic model,

$$x_n \sim \mathcal{N}(0,1), \quad y_n \mid x_n \sim \mathcal{N}(x_n, 1 + \log(1 + \exp(x_n))). \tag{2}$$

Further suppose we modeled these data with a mis-specified homoscedastic model,

$$\theta \sim \mathcal{N}([0,0]^T, [1,0;0,1]), \quad y_n \mid \theta, \phi_n \sim \mathcal{N}(\theta^T \phi_n, 1), \tag{3}$$

where $\phi_n = [x_n, 1]^T$, and $\theta = [\theta_1, \theta_2]$.

Figure 1 shows the posterior mean and the 95% predictive interval of the mis-specified regression line $\theta^\top \phi$ from (A) the Bayesian posterior; (B) the mean field variational approximation restricted to isotropic Gaussians; and (C)–(F) variational approximations with re-scaled marginal variances. In each plot, we overlaid the observed data $\mathcal{D}_{100}$, the true data generating function in dashed black, and also report the 2-Wasserstein distance between the true posterior and each approximation and the lpd averaged over $N^* = 10^4$ test data points drawn from Equation (2). Note that the variational approximation (panel (B) of Figure 1) is quite accurate: the 2-Wasserstein distance between the approximation and the exact posterior is $\sim 10^{-4}$. As we scale up the variance of this approximation, we move away from the exact posterior over the parameters but the posterior predictive distribution covers more of the data, yielding higher lpd. Figure 4 in Appendix B shows the contours of these approximate posterior distributions.

Now suppose we are interested in understanding whether there is a relationship between the covariates and the responses, i.e., whether $\theta_1 = 0$. The actual 95% posterior credible interval is $[0.63, 1.07]$. Since the interval does not contain zero, we could correctly infer that $\theta_1 \neq 0$. On the other hand, the approximation with highest lpd (panel (F) in Figure 1) yields an approximate 95% credible interval of [-0.29,1.99], which does cover zero. Had we used the approximate interval, we might have failed to conclude $\theta_1 \neq 0$.

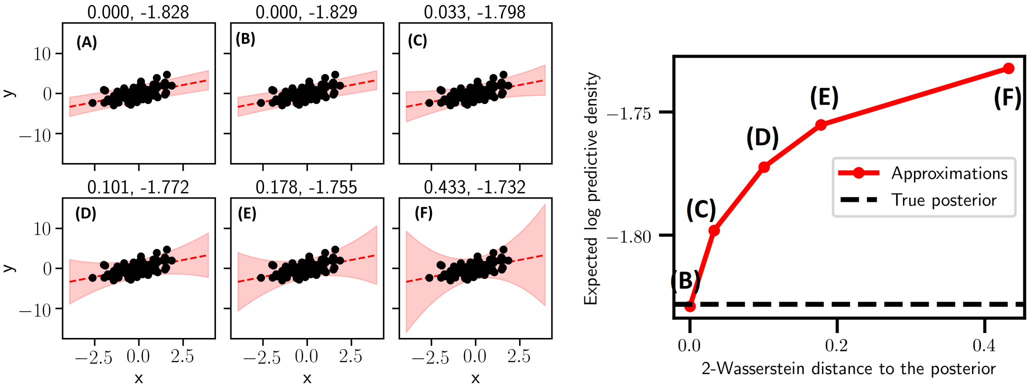

Figure 1: *(Left).* Predictive distributions under the Bayesian posterior and mean field variational approximations. The two numbers in the title of each plot are the 2-Wasserstein distance to the true posterior and test log-likelihood computed on 10^4 test set observations. *(Right).* The relationship between distance to posterior and test log-predictive density.

Another mis-specified case. As another example illustration that a posterior approximation with higher lpd is not necessarily in more agreement with the exact posterior, suppose we observe $\mathcal{D}_{500} = \{(x_n, y_n)\}_{n=1}^{500}$ drawn from the non-linear model

$$\theta_* = [-2, -1]^T, \quad x_n \sim \mathcal{N}(0, 1), \quad y_n \mid \theta_*, \phi_n \sim \mathcal{N}(\theta_*^T \phi_n + x_n^2, 0.5), \tag{4}$$

where $\phi_n = [x_n, 1]^T$. Further suppose we modeled these data with a mis-specified linear model

$$\theta \sim \mathcal{N}([0, 0]^T, [1, 0; 0, 1]), \quad y_n \mid \theta, \phi_n \sim \mathcal{N}(\theta^T \phi_n, 0.5). \tag{5}$$

Figure 2 shows the posterior mean and the 95% predictive interval of the mis-specified regression line $\theta^\top \phi$ from (A) the Bayesian posterior; (B) the mean field variational approximation restricted to isotropic Gaussians; and (C)–(F) several re-scaled variational approximations. In each plot, we overlaid the observed data $\mathcal{D}_{500}$, the true data generating function in dashed black, and also report the 2-Wasserstein distance between the true posterior and each approximation and the lpd averaged over $N^* = 10^4$ test data points drawn from Equation (4). Like in our previous example, the mean field approximation (panel (B) of Figure 2) is very close to the exact posterior. Further, as we scale up the marginal variance of the approximate posteriors, the posterior predictive distributions cover more data, yielding higher lpd, while simultaneously moving away from the exact posterior over the model parameters in a 2-Wasserstein sense. Interestingly, when the approximation is diffuse enough, lpd decreases, again highlighting its non-monotonic relationship with posterior approximation quality. Figure 6 in Appendix B shows a similar phenomenon with posterior standard deviations.

In this example of a mis-specified model, the non-monotonic relationship between lpd and 2-Wasserstein distance means that lpd is, at best, a poor proxy of posterior approximation quality. In fact, we can observe a similar non-monotonic relationship even when the likelihood model is correctly specified (see Appendix A). One could argue that these examples are not cause for worry in situations where prediction is of primary interest and inference about latent parameters is of secondary concern. We note, however, that Proposition 3.6 of Huggins et al. (2020) shows that approximation

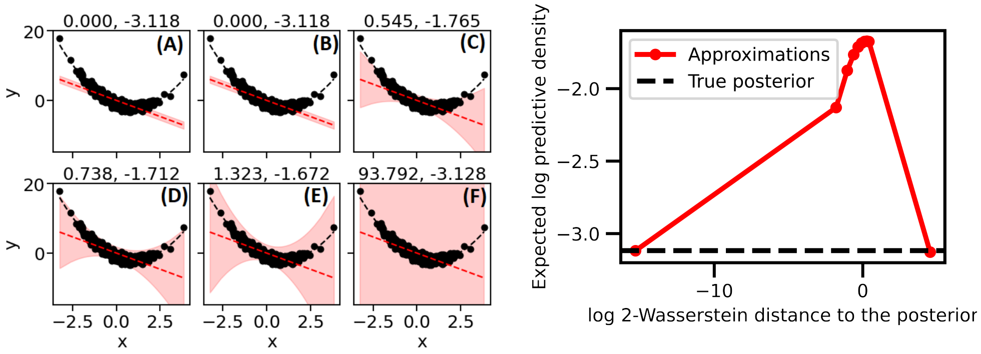

Figure 2: *(Left).* Predictive distributions under the Bayesian posterior and mean field variational approximations. The two numbers in the title of each plot are the 2-Wasserstein distance to the true posterior and test log-likelihoods computed on 10^4 test set observations. *(Right).* The relationship between distance to posterior and test log-predictive density. Observe the log scale of the x-axis and the non-monotonic relationship between test log-predictive density and 2-Wasserstein distance to the Bayesian posterior.

error for predictive distributions closely tracks approximation errors of posterior distributions of latent parameters. That is, even when one only cares about prediction, an approximate inference algorithm with large lpd may still produce large predictive errors *because* it has not approximated the posterior $\Pi(\theta|\mathcal{D})$ well.

4 Claim: the higher the lpd, the more accurate the predictive mean

We next show that although lpd roughly measures closeness in a KL sense, a comparison of lpd can disagree with a comparison based on root mean squared error (RMSE). To this end, we construct two models Π and $\tilde{\Pi}$ such that $\text{lpd}(\mathcal{D}^\star; \Pi) < \text{lpd}(\mathcal{D}^\star; \tilde{\Pi})$ but $\tilde{\Pi}$ yields larger predictive RMSE.

Specifically, suppose that we observe $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^{100{,}000}$ generated according to

$$x_n \sim \mathcal{U}(0, 25), \quad y_n | x_n \sim \text{Laplace}(x_n, 1/\sqrt{2}), \tag{6}$$

which we model using one of the following mis-specified conditional linear models:

$$\Pi : y_n | x_n \sim \mathcal{N}(w x_n, \sigma^2) \qquad \text{or} \qquad \tilde{\Pi} : y_n | x_n \sim \text{Laplace}(0.45 + w x_n, \lambda). \tag{7}$$

Both Π and $\tilde{\Pi}$ depend on two unknown parameters. Π depends on a slope w and a residual variance σ^2 and $\tilde{\Pi}$ depends on a slope w and a residual scale λ. The kind of mis-specification is different across models: while Π has the correct mean specification but incorrect noise specification, $\tilde{\Pi}$ has incorrect mean specification but correct noise specification.

We computed the maximum likelihood estimates (MLEs) $(\hat{w}_\Pi, \hat{\sigma}_\Pi)$ and $(\hat{w}_{\tilde{\Pi}}, \hat{\lambda}_{\tilde{\Pi}})$ for both models. The two fitted models induce the following predictive distributions of $y^\star | x^\star$:

$$\Pi(y^\star | x^\star, \mathcal{D}) : y^\star | x^\star \sim \mathcal{N}(\hat{w}_\Pi x^\star, \hat{\sigma}^2_\Pi) \qquad \text{and} \qquad \tilde{\Pi}(y^\star | x^\star, \mathcal{D}) : y^\star | x^\star \sim \text{Laplace}(0.45 + \hat{w}_{\tilde{\Pi}} x^\star, \hat{\lambda}_{\tilde{\Pi}}).$$

The means of these predictive distributions are natural point estimates of the output $y^\star$ at input $x^\star$.

Using a test set of size $N^\star = 395{,}000$, we observed $\text{lpd}(\mathcal{D}^\star; \Pi) = -1.42 < -1.39 = \text{lpd}(\mathcal{D}^\star; \tilde{\Pi})$. These values suggest that on average over inputs $x^\star$, $\tilde{\Pi}(y^\star | x^\star, \mathcal{D})$ is closer to $\mathcal{P}(y^\star | x^\star)$ than $\Pi(y^\star | x^\star, \mathcal{D})$ in a KL sense. However, using the same test set, we found that Π yielded more accurate point forecasts, as measured by root mean square error (RMSE):

$$\left(\frac{1}{N^\star} \sum_{n=1}^{N^\star} (y_n^\star - \hat{w}_\Pi x_n^\star)^2 \right)^{1/2} = 1.00 < 1.03 = \left(\frac{1}{N^\star} \sum_{n=1}^{N^\star} (y_n^\star - 0.45 - \hat{w}_{\tilde{\Pi}} x_n^\star)^2 \right)^{1/2}.$$

The comparison of RMSEs suggest that on average over inputs $x^\star$, the predictive mean of $\Pi(y^\star | x^\star, \mathcal{D})$ is closer to the mean of $\mathcal{P}(y^\star | x^\star)$ than the predictive mean of $\tilde{\Pi}(y^\star | x^\star, \mathcal{D})$. In other words, the model with larger lpd – whose predictive distribution is ostensibly closer to $\mathcal{P}$ – makes worse point predictions than the model with smaller lpd.

At first glance, our deliberate use of mis-specified models may appear contrived. We note, however, that when a model is correctly specified and fitted using maximum likelihood, we would intuitively expect the correct model to achieve both largest lpd and smallest RMSE. To explain our intuition, had we not fixed the intercept of model $\tilde{\Pi}$, the MLE would converge to the true parameter values. Accordingly the predictive distribution at each $x^\star$ would converge to the true data generating process $\mathcal{P}(y^\star|x^\star)$. The strict propriety of the log-score implies that $\mathcal{P}(y^\star|x^\star)$ achieves the smallest possible elpd among all possible models of $y^\star|x^\star$. Further, one can show that $\mathcal{P}(y^\star|x^\star)$ also achieves the smallest possible RMSE among all possible models. Thus, so long as $\mathcal{D}$ and $\mathcal{D}^\star$ are large enough, we might reasonably expect a correctly specified model to achieve both largest lpd and smallest RMSE.

Of course, virtually all models are mis-specified in practice. As our example illustrates, we might expect to see "rank reversals" where one model might achieve a higher lpd but larger RMSE than another in the mis-specified regime. We conjecture that there are similar examples of mis-specified regression models Π and $\tilde{\Pi}$ for which $\text{lpd}(\mathcal{D}^\star;\tilde{\Pi}) > \text{lpd}(\mathcal{D}^\star;\Pi)$ but other moments or quantiles of $\Pi(y^\star|x^\star,\mathcal{D})$ are closer to the true moments and quantiles of $\mathcal{P}(y^\star|x^\star)$ than $\tilde{\Pi}(y^\star|x^\star,\mathcal{D})$.

5 Discussion

So when should one use test log-likelihood? We argue that comparing probabilistic forecasts using test log-likelihood is reasonable when either (a) the predictive distribution is Gaussian (so maximizing lpd ensures correct mean and variance forecasts) or when (b) being close to the true data generating distribution in a Kullback-Leibler sense is substantively important. It is important to note, however, that just because two distributions are close in KL, their means and variances need not be close; in fact, Propositions 3.1 & 3.2 of Huggins et al. (2020) show that the means and variances of distributions that are close in KL can be arbitrarily far apart.

We further argue that it is inappropriate to compare different Bayesian posterior approximations solely on the basis of the implied test log-predictive densities. Bluntly, just because an approximate inference algorithm produces a larger lpd, it does not necessarily follow that the algorithm has produced a more accurate posterior approximation. Even in cases where one only cares about the implied predictive distribution rather than the underlying posterior distribution simply examining lpd can obfuscate the interplay between modeling choices and posterior approximation algorithms.

Acknowledgments and Disclosure of Funding

We are grateful to Will Stephenson for helping us find examples of discrepancies between posterior approximation quality and lpd.

This work was supported in part by the MIT-IBM Watson AI Lab, an NSF Career Award, an ONR Early Career Grant, the DARPA I2O LwLL program, an ARPA-E project with program director David Tew, and the Wisconsin Alumni Research Foundation.

References

Bernardo, J. M. and Smith, A. F. (2000). *Bayesian Theory*. Wiley.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J., and Blei, D. (2009). Reading tea leaves: How humans interpret topic models. *Advances in Neural Information Processing Systems*, 22.

Dawid, A. P. and Sebastiani, P. (1999). Coherent dispersion criteria for optimal experimental design. *Annals of Statistics*, 27(1):65–81.

Gan, Z., Li, C., Chen, C., Pu, Y., Su, Q., and Carin, L. (2016). Scalable Bayesian learning of recurrent neural networks for language modeling. *arXiv pre-print arXiv:1611.08034*.

Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24:997–1016.

Ghosh, S., Yao, J., and Doshi-Velez, F. (2018). Structured variational learning of Bayesian neural networks with horseshoe priors. In *Proceedings of the* 35th *International Conference on Machine Learning*.

Hernández-Lobato, J. M. and Adams, R. (2015). Probabilistic backpropagation for scalable learning of Bayesian neural networks. In *Proceedings of the* 23*rd* *International Conference on Machine Learning*.

Hernández-Lobato, J. M., Li, Y., Rowland, M., Hernández-Lobato, D., and Turner, R. (2016). Black-box α-divergence minimization. In *Proceedings of the* 33rd *International Conference on Machine Learning*.

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

Huggins, J. H., Kasprzak, M., Campbell, T., and Broderick, T. (2020). Validated variational inference via practical posterior error bounds. In *Proceedings of the* 23*rd* *International Conference on Artificial Intelligence and Statistics*.

Izmailov, P., Maddox, W. J., Kirichenko, P., Garipov, T., Vetrov, D., and Wilson, A. G. (2020). Subspace inference for Bayesian deep learning. In *Uncertainty in Artificial Intelligence*.

Izmailov, P., Vikram, S., Hoffman, M. D., and Wilson, A. G. (2021). What are Bayesian neural network posteriors really like? In *Proceedings of the* 38th *International Conference on Machine Learning*.

Kohonen, J. and Suomela, J. (2005). Lessons learned in the challenge: making predictions and scoring them. In *Machine Learning Challenges Workshop*, pages 95–116. Springer.

Li, C., Chen, C., Fan, K., and Carin, L. (2016). High-order stochastic gradient thermostates for Bayesian learning of deep models. In *Proceedings of the Thirtieth AAAI Conference on Artifical Intelligence*.

Liu, Q. and Wang, D. (2016). Stein variaitonal gradient descent: A general purpose Bayesian inference algorithm. In *Advances in Neural Informational Processing Systems*.

Louizos, C. and Welling, M. (2016). Structured and efficient variational deep learning with matrix Gaussian posteriors. In *Proceedings of the* 33rd *International Conference on Machine Learning*.

Mishkin, A., Kunstner, F., Nielsen, D., Schmidt, M., and Khan, M. E. (2018). SLANG: Fast structured covariance approximations for Bayesian deep learning with natural gradient. In *Advances in Neural Informational Processing Systems*.

Ober, S. W. and Aitchison, L. (2021). Global inducing point variational posteriors for Bayesian neural networks and deep Gaussian processes. In *Proceedings of the* 38th *International Conference on Machine Learning*.

Quinonero-Candela, J., Rasmussen, C. E., Sinz, F., Bousquet, O., and Schölkopf, B. (2005). Evaluating predictive uncertainty challenge. In *Machine Learning Challenges Workshop*, pages 1–27. Springer.

Ranganath, R., Gerrish, S., and Blei, D. M. (2014). Black box variational inference. In *Proceedings of the* 17*th* *International Conference on Artificial Intelligence and Statistics*.

Robert, C. P. (1996). Intrinsic losses. *Theory and Decision*, 40:191–214.

Shi, J., Sun, S., and Zhu, J. (2018). Kernel implicit variational inference. In *International Conference on Learning Representations*.

Sun, S., Chen, C., and Carin, L. (2017). Learning structured weight uncertaitny in Bayesian neural networks. In *Proceedings of the* 20th *International Conference on Artificial Intelligence and Statistics*.

Wu, A., Nowozin, S., Meeds, E., Turner, R. E., Hernández-Lobato, J. M., and Gaunt, A. L. (2019). Deterministic variational inference for robust Bayesian neural networks. In *International Conference on Learning Representations*.

Yao, J., Pan, W., Ghosh, S., and Doshi-Velez, F. (2019). Quality of uncertainty quantification for Bayesian neural network inference. arXiv:1906.09686.

A Posterior approximation quality, lpd, and well-specified models

Our second example in Section 3 demonstrates that lpd may not be a good proxy of posterior approximation quality. Just like most applied situation, the Bayesian model in that example was mis-specified. We now demonstrate that an approximate posterior distribution with higher lpd may not be more accurate even when the model is correctly specified. To this end, consider the following Bayesian linear model

$$\theta \sim \mathcal{N}([0,0]^\top, [1, 0.9; 0.9, 1]), \quad y_n \mid \theta, \phi_n \sim \mathcal{N}(\theta^\top \phi_n, 0.25^2), \tag{8}$$

where $\phi_n = [x_n, 1]^\top$. Now, suppose we observe ten data points $\mathcal{D}_{10} = \{(x_n, y_n)\}_{n=1}^{10}$ sampled as

$$\theta_* = [-2, -1]^T, \quad x_n \sim \mathcal{N}(0,1), \quad y_n \mid \theta_*, \phi_n \sim \mathcal{N}(\theta_*^T \phi_n, 0.25^2). \tag{9}$$

The left panel of Figure 3 plots the contours of (A) the true posterior distribution $\Pi(\phi|\mathcal{D}_{10})$; (B) a mean field variational approximation constrained to the isotropic Gaussian family; and (C)–(F) variational approximations with re-scaled marginal variances. In each panel, we report the 2-Wasserstein distance between the approximate and true posterior and the test log-predictive averaged over $N^\star = 10^4$ test data points drawn from (9).

Interestingly, although we have correctly specified the conditional model of $y|(\theta, \phi)$, the true posterior has a smaller lpd than some of the approximate posteriors. The left panel of Figure 3 suggests that the more probability mass an approximate posterior places around the true data-generating parameter, the higher the lpd. Further, as the approximation becomes more diffuse, lpd begins to decrease (Figure 3 (right)). The non-monotonicity demonstrates that an approximate posterior with larger implied lpd can in fact be further away from the true posterior in a 2-Wasserstein sense than an approximate posterior with smaller implied lpd. Figure 5 in Appendix B shows that an approximate posterior with larger lpd can provide worse estimated standard deviation than an approximation with smaller lpd.

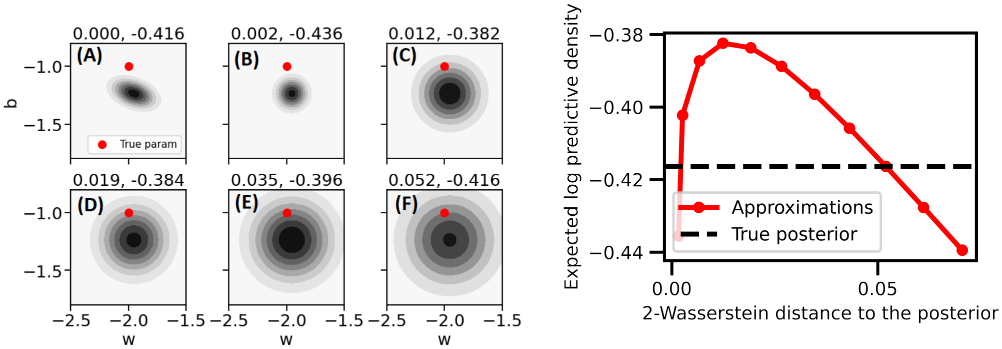

Figure 3: *(Left)*. Contours of the (A) true posterior, (B) the mean field variational approximation, and (C)–(F) re-scaled mean field approximations. The two numbers in the title of each plot are the 2-Wasserstein distance to the true posterior and test log-likelihoods computed on 10^4 test set observations. *(Right)*. The non-monotonic relationship between distance to posterior and test log-predictive density. Observe that the true posterior does not achieve highest test log-predictive density.

B Additional Plots

Misleading inference. Figure 4 shows the contours of each posterior approximation from our first example in Section 3. Notice that the actual posterior distribution (panel (A)) is concentrated on positive θ_1 values. Although the lpd increases as the approximations become more diffuse (panels (B)–(F)), the approximations begin to place non-negligible probability mass on negative θ_1 values.

Well-specified case.

Figure 5 displays a similar kind of non-monotonicity as the right panel of Figure 3. The experimental setup is identical to that of Figure 3: we have only changed what is plotted on the x-axis.

Mis-specified case.

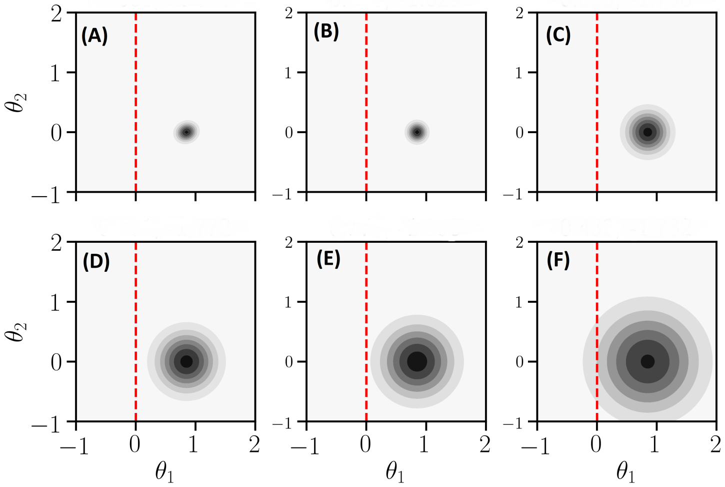

Figure 4: Contours of the (A) true posterior, (B) the mean field variational approximation, and (C)–(F) re-scaled mean field approximations.

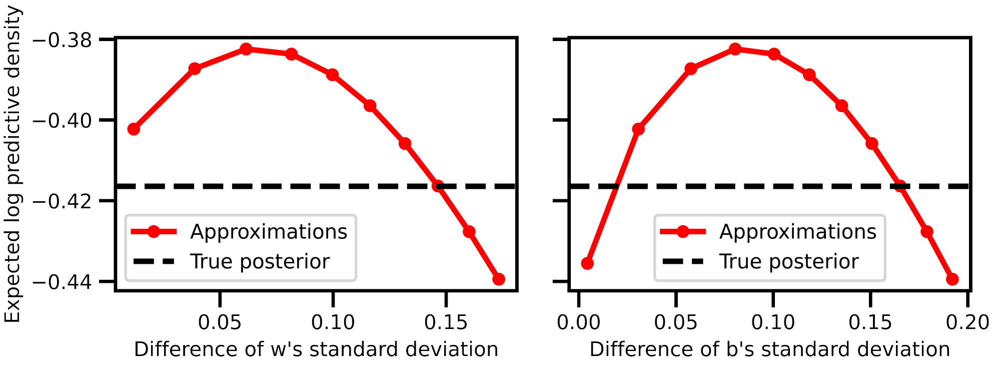

Figure 5: The non-monotonic relationship between difference in marginal standard deviations and test log-predictive density in a well-specified case. *(Left)* The x-axis reports the difference in the standard deviation of the weight w between an approximation and the posterior. *(Right)* The x-axis reports the difference in the standard deviation of the bias b between an approximation and the posterior.

Figure 6 displays a similar kind of non-monotonicity as the right panel of Figure 2. The experimental setup is identical to Figure 2: we have only changed what is plotted on the x-axis.

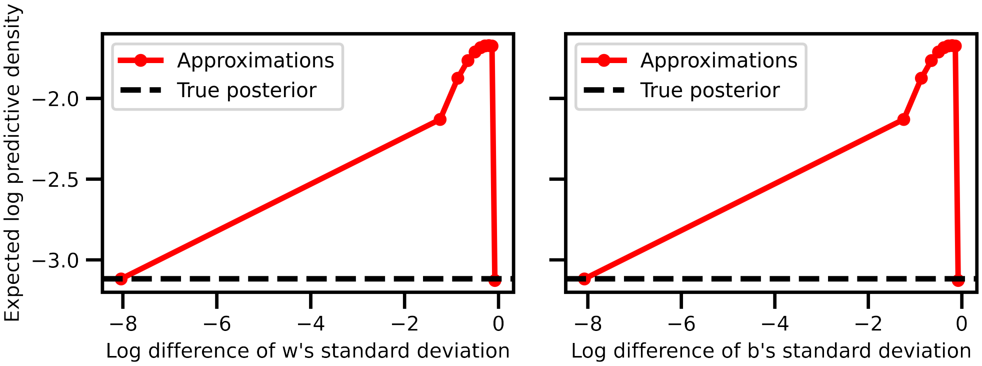

Figure 6: The non-monotonic relationship between difference in marginal standard deviations and test log-predictive density in a mis-specified case. The meaning of x-axis is similar to that of Figure 5.