Are you using test log-likelihood correctly?

Anonymous Author(s)
Affiliation
Address
email

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[1] 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 commonly implied claims about test log-likelihood:

- 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 3)
- The higher the test log-likelihood, the more accurately an approximate inference algorithm recovers the Bayesian posterior distribution of the latent model parameters (Section 4).

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

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 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 were 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 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 comparably simpler 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.

[1] 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

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

2 Background

We typically model training data $\mathcal{D} = \{x_n\}_{n=1}^N$, which we assume are 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, we often compute 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 $\pi(\theta|\mathcal{D})$, we 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 predictive density is given by

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

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

Typically, we assess how well our 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, we 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 we have made explicit the dependence on the testing data $\mathcal{D}^\star$ and the chosen model Π. Note that we have chosen the abbreviation lpd as a reminder that test log-likelihood depends on evaluating the log-predictive density.

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, we introduce 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 1: the higher the lpd, the more accurate the predictive mean

We will show that although lpd roughly measures closeness in a KL sense, a comparison of lpd can be misleading when it comes to comparing predictive means. 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 root mean square error.

Specifically, suppose that we observe $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^{10,000}$ generated according to

$$x_n \sim \mathcal{U}(0, 25), \quad y_n | x_n \sim \mathcal{N}(5 + x_n, 1), \tag{2}$$

which we model using one of the following misspecified conditional linear models:

$$\Pi : y_n | x_n \sim \mathcal{N}(w x_n, \sigma^2) \qquad \text{or} \qquad \tilde{\Pi} : y_n | x_n \sim \mathcal{N}(10.5 + w x_n, \sigma^2). \tag{3}$$

Both Π and $\tilde{\Pi}$ depend on two unknown parameters – a slope w and a residual variance σ^2 – but differ in their fixed intercepts, with Π fixing the intercept at 0 and $\tilde{\Pi}$ fixing the intercept at 10.5.

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

$$\Pi(y^\star | x^\star, \mathcal{D}) : y^\star | x^\star \sim \mathcal{N}(\hat{w}_\Pi x^\star, \hat{\sigma}_\Pi^2) \qquad \text{and} \qquad \tilde{\Pi}(y^\star | x^\star, \mathcal{D}) : y^\star | x^\star \sim \mathcal{N}(10.5 + \hat{w}_{\tilde{\Pi}} x^\star, \hat{\sigma}_{\tilde{\Pi}}^2).$$

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 $N^\star = 39{,}500$, we observed $\text{lpd}(\mathcal{D}^\star; \Pi) = -2.59 < -2.49 = \text{lpd}(\mathcal{D}^\star; \tilde{\Pi})$. 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):

$$\left(\frac{1}{N^\star} \sum_{n=1}^{N^\star} (y_n^\star - \hat{w}_\Pi x_n^\star)^2 \right)^{\frac{1}{2}} = 2.68 < 2.9 = \left(\frac{1}{N^\star} \sum_{n=1}^{N^\star} (y_n^\star - 10.5 - \hat{w}_{\tilde{\Pi}} x_n^\star)^2 \right)^{\frac{1}{2}}.$$

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 model to achieve both largest lpd and smallest RMSE. To explain our intuition, had we not fixed the intercept of model Π, 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, it is not difficult to verify 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 $\tilde{\Pi}(y^\star | x^\star, \mathcal{D})$.

4 Claim 2: lpd accurately assesses posterior approximation quality

In most Bayesian analyses, the posterior density $\pi(\theta | \mathcal{D})$ is analytically intractable and we 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.

Formally, suppose that we have two approximate inference algorithms that produce two approximate posteriors $\hat{\Pi}_1(\theta|\mathcal{D})$ and $\hat{\Pi}_2(\theta|\mathcal{D})$, which respectively induce predictive distributions $\hat{\Pi}_1(x^\star|\mathcal{D})$ 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 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 (in a 2-Wasserstein sense) 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 A.

Well-specified case. Consider the following Bayesian linear model

$$\theta \sim \mathcal{N}([0,0]^\top, [1, 0.9; 0.9, 1]), \quad y_n \mid \theta, \phi_n \sim \mathcal{N}(\theta^\top \phi_n, 0.25^2), \tag{4}$$

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

$$\theta_* = [-2, -1]^T, \quad x_n \sim \mathcal{N}(0, 1), \quad y_n \mid \theta_*, \phi_n \sim \mathcal{N}(\theta_*^T \phi_n, 0.25^2). \tag{5}$$

The left panel of Figure 1 plots the contours of the true posterior distribution $\Pi(\phi|\mathcal{D}_{10})$ (A); a mean field variational approximation constrained to the isotropic Gaussian family (B); and modified variational approximations with re-scaled marginal variances (C–F). 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 (5).

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 1 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 1 (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. In Figure 3 in Appendix A, we show directly that an approximate posterior with larger lpd can provide worse estimates of the standard deviation than an approximation with smaller lpd.

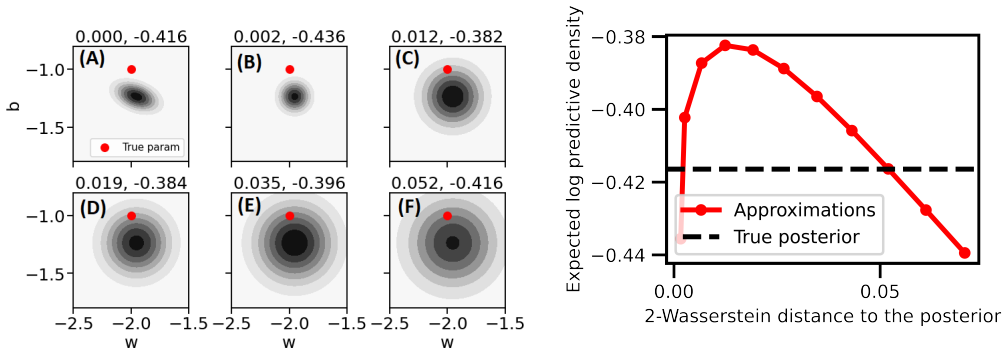

Figure 1: *(Left)*. Contours of the (A) true posterior, (B) the mean field variational approximation, and (C)–(F) re-scaled mean field approximations. The two numbers in the title of each plot are the 2-Wasserstein distance to the true posterior and test log-likelihoods computed on 10^4 validation instances sampled from the true data generating process. *(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.

Mis-specified case. The disagreement between lpd and posterior approximation error is more pronounced when the Bayesian model is mis-specified, which is arguably always the case. To illustrate, suppose we observe $\mathcal{D}_{500} = \{(x_n, y_n)\}_{n=1}^{500}$ drawn from the non-linear model

$$\theta_* = [-2, -1]^T, \quad x_n \sim \mathcal{N}(0, 1), \quad y_n \mid \theta_*, \phi_n \sim \mathcal{N}(\theta_*^T \phi_n + x_n^2, 0.5), \tag{6}$$

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

$$\theta \sim \mathcal{N}([0,0]^T, [1, 0; 0, 1]), \quad y_n \mid \theta, \phi_n \sim \mathcal{N}(\theta^T \phi_n, 0.5). \tag{7}$$

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 (6). In this case, there is little uncertainty in the Bayesian posterior and it is well approximated by the variational posterior. The Wasserstein distance between the two is 10^{-4}. As we scale up the variance of this approximation, we move away from the Bayesian posterior but cover more of the data, leading to higher lpd but poorer approximations to the posterior. Eventually, when the approximation is diffuse enough, lpd decreases, again highlighting its non-monotonic relationship with posterior approximation quality. In Figure 4 in Appendix A, we similarly see that elpd can be higher even when the difference in standard deviations is larger.

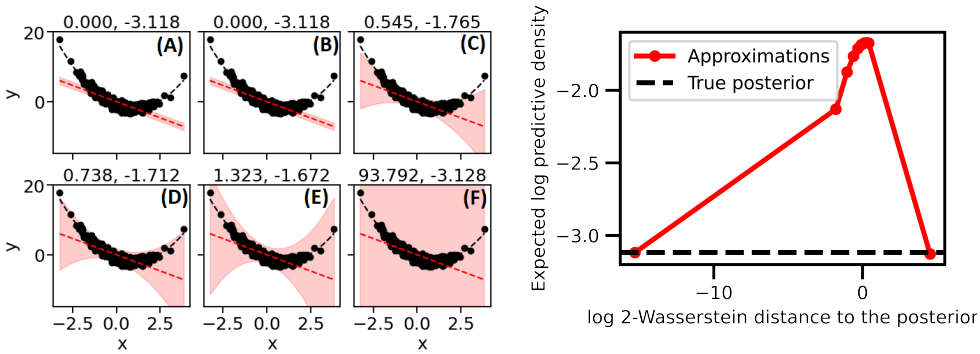

Figure 2: *(Left)*. Predictive distributions under the Bayesian posterior and mean field variational approximations. The two numbers in the title of each plot are the 2-Wasserstein distance to the true posterior and test log-likelihoods computed on 10^4 validation instances sampled from the true data generating process. *(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.

In our two examples, the non-monotonic relationship between lpd and 2-Wasserstein distance means that lpd is, at best, a poor proxy of posterior approximation quality. One could argue that our example is 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 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.

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 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.

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.

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

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

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.

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.

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

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

A Additional Plots on Non-Monotonicity

Well-specified case.

Figure 3 displays a similar kind of non-monotonicity as the right panel of Figure 1. The experimental setup is identical to that of Figure 1: we have only changed what is plotted on the x-axis.

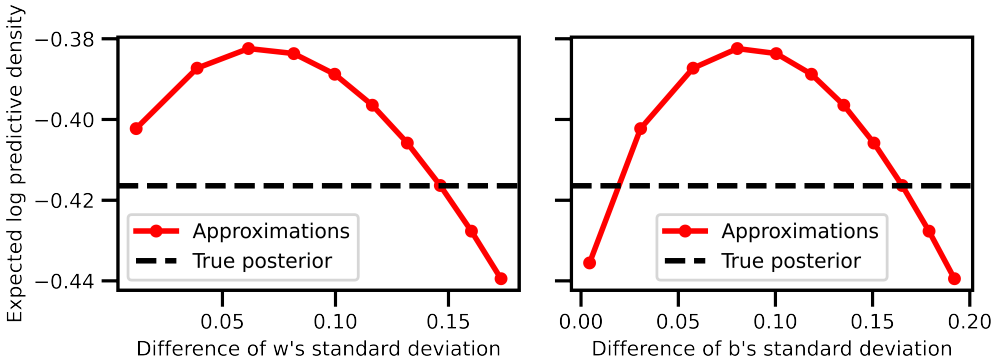

Figure 3: The non-monotonic relationship between difference in marginal standard deviations and test log-predictive density in a well-specified case. *(Left)* The x-axis reports the difference in the standard deviation of the weight w between an approximation and the posterior. *(Right)* The x-axis reports the difference in the standard deviation of the bias b between an approximation and the posterior.

Mis-specified case.

Figure 4 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.

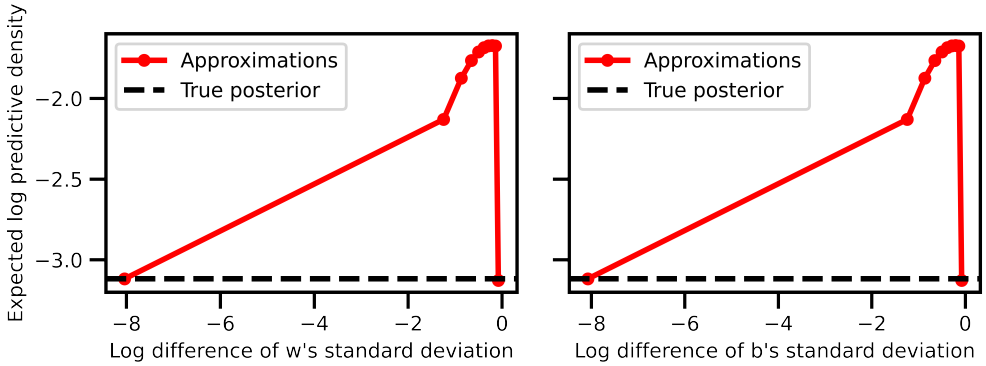

Figure 4: The non-monotonic relationship between difference in marginal standard deviations and test log-predictive density in a mis-specified case. The meaning of x-axis is similar to that of Figure 3.