Are you using test log-likelihood correctly?

Anonymous authors
Paper under double-blind review

Abstract

Test log-likelihood is commonly used to compare different models of the same data or 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) approximate Bayesian inference algorithms that attain higher test log-likelihoods need not also yield more accurate posterior approximations and (ii) conclusions about forecast accuracy based on test log-likelihood comparisons may not agree with conclusions based on root mean squared error.

1 Introduction

Test log-likelihood, also known as predictive log-likelihood or test log-predictive, is computed as the log-predictive density averaged over a set of held-out data. It 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 examples that falsify 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 better the predictive performance on held-out data according to other measurements, like root mean squared error (Section 4).

Our 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 root mean squared error.

We are not the first to highlight discrepancies between test log-likelihood and other analysis objectives. For instance, Quiñonero-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 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 discrepancies between test log-likelihood and other evaluation metrics. 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 Section 3, we show that over a range of posterior approximations provided by a recent method, those with higher test log-likelihood provide worse posterior approximation quality; in additional examples, we recover similar results even when using different approximations and even when there is little or no model misspecification. In Section 4, we show examples in both complex and simple models where test log-likelihood is higher but root mean squared error on held-out data is worse. Our examples in Section 4 do depend on model misspecification, but we note that model misspecification is unavoidable in practice. 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 $\mathcal{D} = \{y_n\}_{n=1}^{N}$, which are assumed to be distributed according to an unknown probability distribution $\mathcal{P}$, by introducing a parameter θ and specifying a conditional distribution $\Pi(Y|\theta)$ with density $\pi(y|\theta)$. 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 $\Pi(\mathcal{D}, \theta)$.

Upon computing a point estimate $\hat{\theta}$ or 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 $y^\star$ is just $\pi(y^\star|\hat{\theta})$. The Bayesian posterior predictive density is given by

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

Observe that $\pi(y^\star|\hat{\theta})$ is numerically equal to $\pi(y^\star|\mathcal{D})$ when the prior or posterior of θ 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 $\mathcal{D}^\star = \{y_n^\star\}_{n=1}^{N^\star}$, 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:

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

where our notation makes explicit the dependence of the test log-likelihood (TLL) on testing data $\mathcal{D}^\star$ and the chosen model Π.

2.1 The case for test log-likelihood

Researchers commonly use test log-likelihood to select between two models of the data, say Π and $\tilde{\Pi}$; that is, they select model Π over $\tilde{\Pi}$ whenever $\text{TLL}(\mathcal{D}^\star; \Pi) > \text{TLL}(\mathcal{D}^\star; \tilde{\Pi})$. Often practitioners will further observe that TLL may exhibit variability across draws of testing data and express confidence in choosing Π when the lower bound of a confidence interval around $\text{TLL}(\mathcal{D}^\star; \Pi)$ exceeds the upper bound of a confidence interval around $\text{TLL}(\mathcal{D}^\star; \tilde{\Pi})$.[1]

To understand these comparisons, consider the *expected log-predictive density*,

$$\text{elpd}(\Pi) := \int \log \pi(y^\star|\mathcal{D})d\mathcal{P}(y^\star).$$

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

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

Thus, assuming that the test set $\mathcal{D}^\star$ is sufficiently large, if the lower bound of a confidence interval around $\text{TLL}(\mathcal{D}^\star; \Pi)$ exceeds the upper bound of a confidence interval around $\text{TLL}(\mathcal{D}^\star; \tilde{\Pi})$, we can reasonably conclude that $\text{elpd}(\Pi) > \text{elpd}(\tilde{\Pi})$, which in turn implies that $\Pi(y^\star|\mathcal{D})$ is closer to $\mathcal{P}(y^\star)$ than $\tilde{\Pi}(y^\star|\mathcal{D})$ in a KL sense.

---

[1] In our experiments, the number of data points is typically high, so we assume that the sampling distribution of TLL is well-approximated by a normal distribution, and we report 95% confidence intervals as the mean plus or minus two standard errors. Since TLL takes the form of a mean, its standard error can be calculated using the usual formula for standard error of the mean, and that is what we use in our experiments below. Also, note that technically the estimates of the standard error of $\text{TLL}(\mathcal{D}^\star; \Pi)$ and of $\text{TLL}(\mathcal{D}^\star; \tilde{\Pi})$ may be correlated when computed in this way; we do not expect that more careful treatment of this correlation would change our substantive conclusions below.

In other words, we would expect predictions made using the fitted model with larger TLL 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, TLL may be seen as a "non-informative" choice (Robert, 1996; Gelman et al., 2014). When $\Pi(y^\star|\mathcal{D})$ 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 $\Pi(y^\star|\mathcal{D})$ and $\mathcal{P}(y^\star)$ are identical. By contrast, minimizing mean square error is equivalent to only matching the means of $\Pi(y^\star|\mathcal{D})$ and $\mathcal{P}(y^\star)$.

Model comparison with TLL makes two (often implicit) assumptions: (i) that $\mathrm{TLL}(\mathcal{D}^\star;\cdot)$ is a close approximation to $\mathrm{elpd}(\cdot)$ and (ii) that closeness between $\Pi(y^\star|\mathcal{D})$ 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: TLL accurately assesses posterior approximation quality

In this section, we give examples where test log likelihood is higher though the quality of an approximate posterior mean, variance, or other common summary is lower. We start with examples in mis-specified models and then give a correctly specified example.

Practitioners often use posterior expectations to summarize the relationship between a covariate and a response. For instance, the posterior mean serves as a point estimate, and the posterior standard deviation quantifies uncertainty. However, as the posterior density $\pi(\theta|\mathcal{D})$ is analytically intractable, practitioners must instead rely on approximate posterior computations. There are myriad approximate inference algorithms (e.g. Laplace approximation, Hamiltonian Monte Carlo, mean-field variational inference, to name just a few). All 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 TLL values assumed to reflect more accurate approximations, e.g. 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) or 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, respectively $\hat{\Pi}_1(\theta|\mathcal{D})$ and $\hat{\Pi}_2(\theta|\mathcal{D})$. The exact posterior and its approximations respectively induce predictive distributions $\Pi(y^\star|\mathcal{D}), \hat{\Pi}_1(y^\star|\mathcal{D})$, and $\hat{\Pi}_2(y^\star|\mathcal{D})$. 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[2] (i) $\mathrm{TLL}(\mathcal{D}^\star;\hat{\Pi}_1) > \mathrm{TLL}(\mathcal{D}^\star;\Pi)$ 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) $\mathrm{TLL}(\mathcal{D}^\star;\hat{\Pi}_1) > \mathrm{TLL}(\mathcal{D}^\star;\hat{\Pi}_2)$ but (ii) $\hat{\Pi}_1(\theta|\mathcal{D})$ is a worse approximation to the exact posterior $\Pi(\theta|\mathcal{D})$ than $\hat{\Pi}_2(\theta|\mathcal{D})$.

3.1 TLL and downstream posterior inference

Relying on TLL 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 model these data with a mis-specified homoscedastic model:

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

---

[2] In fact, we compare confidence intervals around the test log-likelihood values, but we write "$\mathrm{TLL}(\mathcal{D}^\star;\hat{\Pi}_1) > \mathrm{TLL}(\mathcal{D}^\star;\Pi)$" and below for brevity. In our experiments, we generally find that the confidence intervals are small on the scale of the comparison.

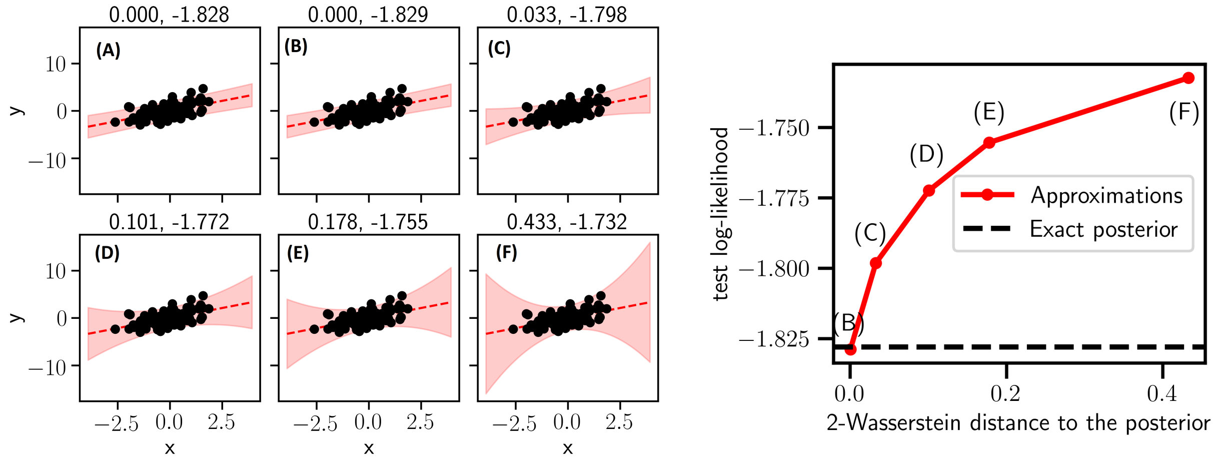

Figure 1: *(Left).* Predictive distributions under the Bayesian posterior and mean field variational approximations. The two numbers in the title of each plot are the 2-Wasserstein distance to the exact posterior and test log-likelihood computed on 10^4 test set observations. Two standard errors in the test log-likelihood estimate are (A) 0.03, (B) 0.03, (C) 0.02, (D) 0.02, (E) 0.02, (F) 0.02. *(Right).* The relationship between 2-Wasserstein distance to the posterior and test log-likelihood.

where $\phi_n = [x_n, 1]^\top$, 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 exact Bayesian posterior; (B) the mean field variational approximation restricted to isotropic Gaussians; and (C)–(F) variational approximations with re-scaled marginal variances. Each panel includes a scatter plot of the observed data, $\mathcal{D}_{100}$. We also report the 2-Wasserstein distance between the exact posterior and each approximation and the TLL averaged over $N^* = 10^4$ test data points drawn from Equation (2); note that the 2-Wasserstein distance can be used to bound differences in means and variances (Huggins et al., 2020). The variational approximation (panel (B) of Figure 1) is quite accurate: the 2-Wasserstein distance between the approximation and the exact posterior is $\sim 10^{-4}$. See also Figure 2, which shows the contours of the exact and approximate posterior distributions. 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 data, yielding higher TLL.

TLL and a discrepancy in inferences. Researchers are often interested in understanding whether there is a relationship between a covariate and response; a Bayesian analysis will often conclude that there is no relationship if the posterior on the corresponding effect-size parameter places substantial probability on an interval not containing zero. In our example, we wish to check whether $\theta_1 = 0$. Notice that the exact posterior distribution (panel (A) in Figures 1 and 2) is concentrated on positive θ_1 values. The 95% credible interval of the exact posterior[3] is $[0.63, 1.07]$. Since the interval does not contain zero, we would infer that $\theta_1 \neq 0$. On the other hand, as the approximations become more diffuse (panels (B)–(F)), TLL increases and the approximations begin to place non-negligible probability mass on negative θ_1 values. In fact, the approximation with highest TLL (panel (F) in Figures 1 and 2) yields an approximate 95% credible interval of [-0.29,1.99], which covers zero. Had we used this approximate interval, we would have failed to conclude $\theta_1 \neq 0$. That is, in this case, we would reach a different substantive conclusion about the effect θ_1 if we (i) use the exact posterior or (ii) use the approximation selected by highest TLL.

3.2 TLL in the wild

Next, we examine a more realistic scenario in which the difference between the quality of the posterior approximation and the exact posterior distribution TLL arises naturally, without the need to artificially

---

[3] Throughout we used symmetric credible intervals formed by computing quantiles: the 95% interval is equal to the 2.5%–97.5% interquantile range.

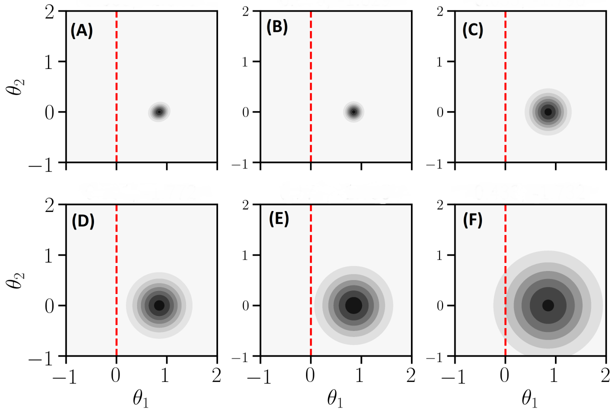

Figure 2: Contours of (A) the exact posterior, (B) the mean field variational approximation restricted to isotropic Gaussians, and (C)–(F) re-scaled mean field approximations. The line $\theta_1 = 0$ is highlighted in red.

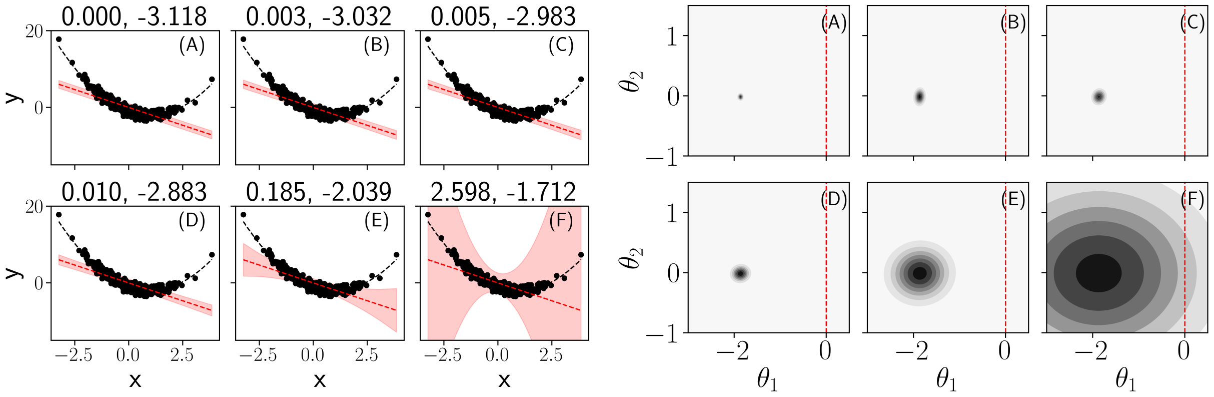
0.000, -3.118
0.003, -3.032
0.005, -2.983
0.010, -2.883
0.185, -2.039
2.598, -1.712

Figure 3: *(Left)*. Predictive distributions under the Bayesian posterior (A) and the SWAG posterior with SWAG learning rate of (B) 10^{-3}, (C) 10^{-2}, (D) 10^{-1}, (E) 1, and (F) 10. The two numbers in the title of each plot are the 2-Wasserstein distance to the exact posterior and test log-likelihood computed on 10^4 test set observations. Two standard errors in the test log-likelihood estimates are (A) 0.16, (B) 0.15, (C) 0.14, (D) 0.13, (E) 0.05, (F) 0.01. *(Right)*. Contours of the (A) exact posterior, and (B)–(F) SWAG approximations with different learning rates. The line $\theta_1 = 0$ is highlighted in red.

increase the marginal variance of the variational approximations. To explore this situation, we will first introduce another example of mis-specification and repeat the type of analysis described in Section 3.1.

Consider the following case: we observe 500 observations $\mathcal{D}_{500} = \{(x_n, y_n)\}_{n=1}^{500}$ drawn from a non-linear model:

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

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

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

While the misspecification here might appear egregious, linear models are widely used in practice for modeling non-linear phenomena when one is primarily interested in inferring whether the covariates are positively correlated, negatively correlated, or are uncorrelated with the responses (Berk et al., 2014; 2018; Blanca et al., 2018; Vowels, 2023). Next, we use SWAG (Maddox et al., 2019), an off-the-shelf approximate inference algorithm, to approximate the posterior $\Pi(\theta|\mathcal{D}_{500})$.[4] SWAG uses a gradient-based optimizer with a learning rate schedule that encourages the optimizer to oscillate around the optimal solution instead of converging to it. Then, a Gaussian distribution is fit to the set of solutions explored by the optimizer around the optimum using moment matching. In general, one must select the learning rate schedule in a heuristic fashion. One might be tempted to use TLL to tune the learning rate schedule. We use this heuristic[5] and run SWAG for a thousand epochs, annealing the learning rate down to a different constant value after 750 epochs. We vary this constant value over the set $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$. In Figure 3, we show the resulting posterior mean and the 95% predictive interval of the misspecified regression line $\theta^\top \phi$ from (A) the Bayesian posterior; (B)–(F) the SWAG posteriors using different learning rate schedules. In each plot, we overlay the observed data $\mathcal{D}_{500}$ (black dots) with the true data generating function in dashed black. We also report the 2-Wasserstein distance between the exact posterior and each approximation and the TLL averaged over $N^* = 10^4$ test data points drawn from Equation (4). In all cases, SWAG overestimates the posterior variance, with predictive distributions that better cover the data and consequently lead to a higher TLL. However, these SWAG posterior approximations are *farther* from the exact posterior. In fact, we found that a learning rate of 10 (Figure 3, *Left*, panel (F)) maximized TLL but led to the worst approximation of the exact posterior.

As in the previous section, next suppose we fit this misspecified linear model to understand whether there is a relationship between the covariates and the responses, i.e., whether $\theta_1 = 0$. Notice that the exact posterior distribution (Figure 3, *Right*, panel (A)) is concentrated on negative θ_1 values, with the 95% posterior credible interval being $[-1.96, -1.79]$. Since the interval is to the left of zero, we would infer that $\theta_1 < 0$ and that the covariate and the response are negatively correlated. In contrast, if we select the SWAG approximation with the highest TLL, we select the posterior approximation in panel (F) on the right side of Figure 3. The corresponding 95% posterior credible interval is $[-4.46, 0.74]$, which places non-negligible probability mass on $\theta_1 > 0$. In this case, we would not conclude that the response and the covariate are negatively correlated – by contrast to the conclusion using the exact posterior.

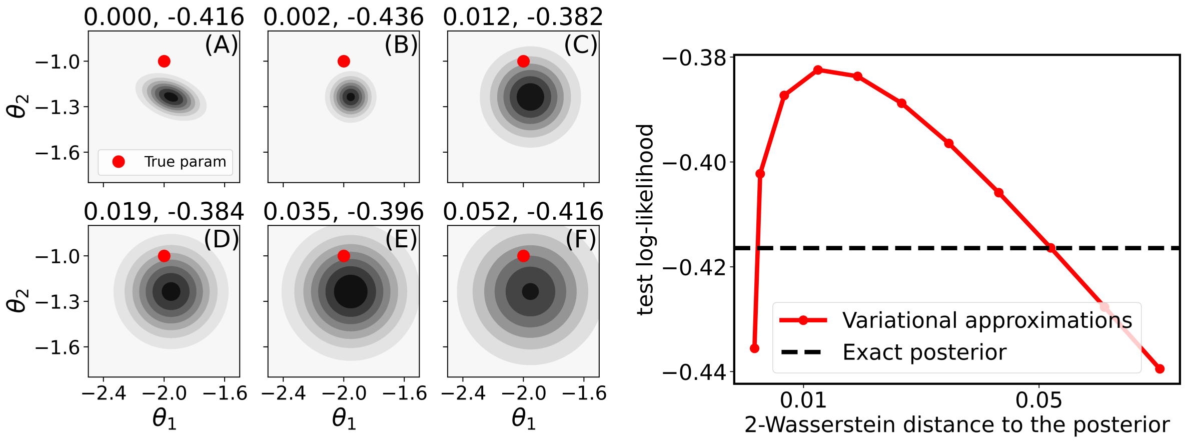

Figure 4: *(Left).* Contours of (A) the exact posterior, (B) the mean field variational approximation restricted to isotropic Gaussians, and (C)–(F) re-scaled mean field approximations. The two numbers in the title of each plot are the 2-Wasserstein distance to the exact posterior and test log-likelihoods computed on 10^4 test set observations. Two standard errors in the test log-likelihood estimates are (A) 0.019, (B) 0.020, (C) 0.014, (D) 0.013, (E) 0.011, (F) 0.009. *(Right).* The non-monotonic relationship between distance to posterior and test log-likelihood. Observe that the exact posterior does not achieve highest test log-likelihood.

---

[4] We also repeat the re-scaled variational inference experiment from Section 3.1 with this set of data and models (Equations (4) and (5)). See Appendix A.

[5] Although used pedagogically here, similar heuristics have been used in practice (di Langosco et al., 2022), where the learning rate is tuned based on the accuracy achieved on held-out data.

3.3 TLL and well-specified models

The examples above demonstrated that TLL is not a reliable proxy to posterior approximation quality when the model is mis-specified. Though mis-specified models are the norm in practice, we now demonstrate that a distribution with higher TLL may not provide a more accurate posterior approximation even when the model is correctly specified.

To this end, consider the following Bayesian linear model:

$$\theta \sim \mathcal{N}([0,0]^\top, [1, 0.9; 0.9, 1]), \quad y_n \mid \theta, \phi_n \sim \mathcal{N}(\theta^\top \phi_n, 0.25^2), \tag{6}$$

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

$$\theta_* = [-2, -1]^\top, \quad x_n \sim \mathcal{N}(0,1), \quad y_n \mid \theta_*, \phi_n \sim \mathcal{N}(\theta_*^\top \phi_n, 0.25^2). \tag{7}$$

The left panel of Figure 4 plots the contours of (A) the exact posterior distribution $\Pi(\phi|\mathcal{D}_{10})$; (B) the 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 exact posterior and the test log-predictive averaged over $N^\star = 10^4$ test data points drawn from Equation (7).

Although we have correctly specified the conditional model of $y|(\theta, \phi)$, the exact posterior has a lower TLL than some of the approximate posteriors; in particular, the 95% confidence intervals for (C) and (D) are disjoint from the 95% confidence interval for the exact posterior, shown in (A). The left panel of Figure 4 suggests that the more probability mass an approximate posterior places around the true data-generating parameter, the higher the TLL. Eventually, as the approximation becomes more diffuse, TLL begins to decrease (Figure 4 (right)). The non-monotonicity demonstrates that an approximate posterior with larger implied TLL can in fact be further away from the exact posterior in a 2-Wasserstein sense than an approximate posterior with smaller implied TLL. Figure 8 in Appendix A shows that, in the well-specified case, a distribution with larger TLL can provide a worse approximation of the posterior standard deviation than a distribution with smaller TLL.

4 Claim 2: the higher the TLL, the more accurate the predictive mean

We now show that although TLL roughly measures closeness in a KL sense, a comparison based on TLL 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 $\mathrm{TLL}(\mathcal{D}^\star; \Pi) < \mathrm{TLL}(\mathcal{D}^\star; \tilde{\Pi})$ but $\tilde{\Pi}$ yields larger predictive RMSE.

Misspecified Gaussian process regression. Suppose we observe $\mathcal{D}_{100} = \{(x_n, y_n)\}_{n=1}^{100}$ from the following data generating process:

$$x_n \sim \mathcal{U}(-5, +5) \quad y_n | x_n \sim \mathcal{N}(\sin(2x_n), 0.1). \tag{8}$$

Further suppose we model this data using a zero-mean Gaussian process (GP) with Gaussian noise,

$$f \sim \mathrm{GP}(\mathbf{0}, k(x, x')), \quad y_n | f_n \sim \mathcal{N}(f_n, \sigma^2), \tag{9}$$

where f_n is shorthand for $f(x_n)$. First consider the case where we employ a periodic kernel,[6] constrain the noise nugget σ^2 to 1.6, and fit all other hyper-parameters by maximizing the marginal likelihood. The resulting fit is shown in Figure 5 (A). Next, consider an alternate model where we use a squared-exponential kernel and fit all hyper-parameters including the noise nugget via maximum marginal likelihood. The resulting fit is displayed in Figure 5 (B). The squared exponential model fails to recover the predictive mean and reverts back to the prior mean (RMSE = 0.737, 95% confidence interval $[0.729, 0.745]$),[7] while the periodic model recovers

---

[6] PeriodicMatern32 in `https://github.com/SheffieldML/GPy`

[7] To compute the RMSE confidence interval, we first compute the mean of the squared errors (MSE, m) and its associated standard error of the mean (s). Since we have a large number of data points and the MSE takes the form of a mean, we assume the sampling distribution of the MSE is well-approximated by a normal distribution. We use $[m - 2s, m + 2s]$ as the 95% confidence interval for the MSE. We use $[\sqrt{m - 2s}, \sqrt{m + 2s}]$ as the 95% confidence interval for the RMSE. Note that the resulting RMSE confidence interval will generally not be symmetric.

the predictive mean accurately, as measured by RMSE $= 0.355$ (95% confidence interval $[0.351, 0.360]$).Despite the poor mean estimate provided by the squared exponential model, it scores a substantially higher TLL.

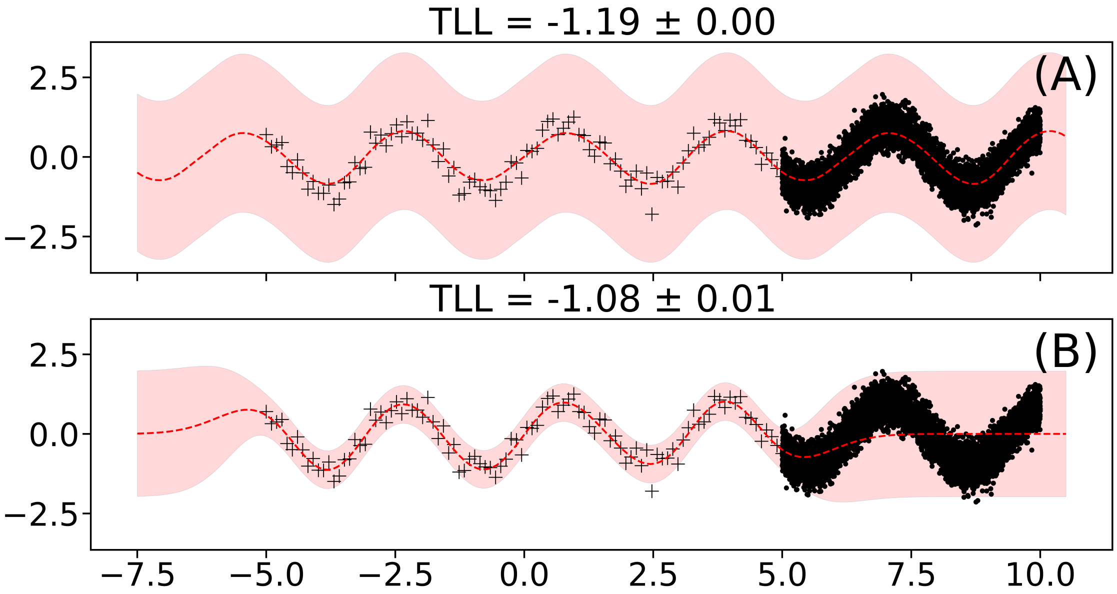

Figure 5: The plots display two Gaussian processes trained on the same set of data (represented by black plus symbols). The dashed red line shows the mean of the posterior Gaussian process, while the red highlighted region represents the 95% predictive interval. The subplot titles display the TLL (±2 standard error) attained by each Gaussian process. Although the Gaussian process in panel (A) achieves a better mean fit compared to panel (B), it has a worse TLL when evaluated on 10^4 test instances (represented by black dots).

In this example, we see that, even with an accurate point estimate, TLL can be reduced by, for instance, inflating the predictive uncertainty. And this discrepancy between TLL and RMSE is not necessarily removed by optimizing the parameters of a model.

Misspecified linear regression. Our next example illustrates that even when all parameters in a model are fit with maximum likelihood, a comparison based on TLL may still disagree with a comparison based on RMSE. It also illustrates that the discrepancy between TLL and RMSE can arise even in very simple and low-dimensional models and even when data is sufficiently large that we expect that TLL matches elpd well.

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

$$x_n \sim \mathcal{U}(0, 25), \quad y_n | x_n \sim \text{Laplace}(x_n, 1/\sqrt{2}), \tag{10}$$

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

$$\begin{gathered} \Pi : y_n | x_n \sim \mathcal{N}(\theta x_n, \sigma^2) \\ \text{or} \\ \tilde{\Pi} : y_n | x_n \sim \text{Laplace}(0.45 + \theta x_n, \lambda). \end{gathered} \tag{11}$$

Both Π and $\tilde{\Pi}$ depend on two unknown parameters. Π depends on a slope θ and a residual variance σ^2 and $\tilde{\Pi}$ depends on a slope θ 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) $(\hat{\theta}_\Pi, \hat{\sigma}_\Pi)$ and $(\hat{\theta}_{\tilde{\Pi}}, \hat{\lambda}_{\tilde{\Pi}})$ for both models. The two fitted models induce the following predictive distributions of $y^\star | x^\star$:

$$\begin{gathered} \Pi(y^\star | x^\star, \mathcal{D}) : y^\star | x^\star \sim \mathcal{N}(\hat{\theta}_\Pi x^\star, \hat{\sigma}^2_\Pi) \\ \text{and} \\ \tilde{\Pi}(y^\star | x^\star, \mathcal{D}) : y^\star | x^\star \sim \text{Laplace}(0.45 + \hat{\theta}_{\tilde{\Pi}} x^\star, \hat{\lambda}_{\tilde{\Pi}}). \end{gathered} \tag{12}$$

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 $\text{TLL}(\mathcal{D}^\star; \Pi) = -1.420 < -1.389 = \text{TLL}(\mathcal{D}^\star; \tilde{\Pi})$. The standard error of either TLL estimate is only 0.002. Hence, based on sample mean and standard error, we conclude that $\tilde{\Pi}$ has better elpd than Π. 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):

$$\left(\frac{1}{N^\star} \sum_{n=1}^{N^\star} (y^\star_n - \hat{\theta}_\Pi x^\star_n)^2 \right)^{1/2} = 1.000 < 1.025 = \left(\frac{1}{N^\star} \sum_{n=1}^{N^\star} (y^\star_n - 0.45 - \hat{\theta}_{\tilde{\Pi}} x^\star_n)^2 \right)^{1/2}. \tag{13}$$

In addition, the 95% confidence intervals for the RMSE do not overlap: the interval for Π's RMSE is $[0.997, 1.005]$ and that for $\tilde{\Pi}$'s RMSE is $[1.022, 1.029]$. The comparison of RMSEs suggests 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 TLL – whose predictive distribution is ostensibly closer to $\mathcal{P}$ – makes worse point predictions than the model with smaller TLL.

5 Discussion

Our paper is neither a blanket indictment nor recommendation of test log-likelihood. Rather, we hope to encourage researchers to explicitly state and commit to a particular data-analysis goal – and recognize that different methods may perform better under different goals. For instance, if the goal of a method is to approximate the exact Bayesian posterior, we would argue that it is inappropriate to use test log-likelihood as the principal metric. We have produced examples where a distribution can provide a better test log-likelihood but yield a (much) poorer approximation to the Bayesian posterior – in particular, leading to fundamentally different inferences and decisions.

Conversely, it may very reasonably be the case that a particular method is *not* designed to approximate the exact Bayesian posterior; indeed, many of the arguments for using the exact Bayesian posterior in decision making rely on correct model specification, which we cannot rely upon in practice. But then the choice of evaluation metrics would ideally be made plain. Test log-likelihood might be a good choice of evaluation metric when the goal is being close to the true data generating distribution in a Kullback–Leibler sense. 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. If there is a quantity of particular interest in the data-generating process, such as a moment or a quantile, a good choice of evaluation metric may be an appropriate scoring rule. Namely, one might choose a scoring rule whose associated divergence function is known to quantify the distance between the forecast's quantity of interest and that of the data-generating process. For instance, when comparing the quality of mean estimates, one option is using the squared-error scoring rule, whose divergence function is the integrated squared difference between the forecast's mean estimate and the mean of the data-generating process. See Gneiting and Raftery (2007) for a list of commonly used scoring rules and their associated divergences.

References

Berk, R., Brown, L., Buja, A., George, E., Pitkin, E., Zhang, K., and Zhao, L. (2014). Misspecified mean function regression: Making good use of regression models that are wrong. *Sociological Methods & Research*, 43(3):422–451.

Berk, R., Brown, L., Buja, A., George, E., and Zhao, L. (2018). Working with misspecified regression models. *Journal of Quantitative Criminology*, 34:633–655.

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

Blanca, M. J., Alarcón, R., and Bono, R. (2018). Current practices in data analysis procedures in psychology: What has changed? *Frontiers in Psychology*, 9:2558.

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.

di Langosco, L. L., Fortuin, V., and Strathmann, H. (2022). Neural variational gradient descent. In *Fourth Symposium on Advances in Approximate Bayesian Inference*.

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* 35^{th} *International Conference on Machine Learning*.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102:359–378.

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* 33^{rd} *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* 38^{th} *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* 33^{rd} *International Conference on Machine Learning*.

Maddox, W. J., Izmailov, P., Garipov, T., Vetrov, D. P., and Wilson, A. G. (2019). A simple baseline for Bayesian uncertainty in deep learning. *Advances in Neural Information Processing Systems*, 32.

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* 38^{th} *International Conference on Machine Learning*.

Quiñonero-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* 20^{th} *International Conference on Artificial Intelligence and Statistics*.

Vowels, M. J. (2023). Misspecification and unreliable interpretations in psychology and social science. *Psychological Methods*, 28(3):507.

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 Additional Experiments and Plots

A.1 Additional TLL in the wild experiments

SWAG with higher learning rates. In Figure 6 we continue the experiment described in Section 3.2 but using higher learning rates of 12, 15, and 20. Despite moving further from the exact posterior the test log-likelihood remains higher than those achieved by SWAG approximations with lower learning rates (panels (B) through (E) of Figure 3).

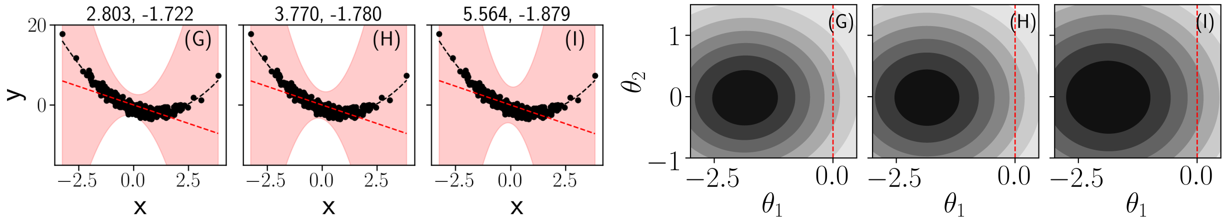

Figure 6: *(Left)*. Predictive distributions under the SWAG posterior with SWAG learning rate of (G) 12, (H) 15, (I) 20. The two numbers in the title of each plot are the 2-Wasserstein distance to the exact posterior and test log-likelihood computed on 10^4 test set observations. Two standard errors in the test log-likelihood estimates are (G) 0.01, (H) 0.009, (I) 0.08. *(Right)*. Contours of the SWAG approximations with different learning rates. The line $\theta_1 = 0$ is highlighted in red.

Mean field variational inference. Next, we reproduce the experimental setup described in Section 3.2, but instead of using SWAG to approximate the posterior, we use mean field variational inference and examine the relationship between TLL and posterior approximation quality under different re-scalings of the marginal variance of the optimal variational approximation. Figure 7 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 TLL averaged over $N^* = 10^4$ test data points drawn from Equation (2) Like in our previous example, the mean field approximation (panel (B) of Figure 7) 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 TLL, while simultaneously moving away from the exact posterior over the model parameters in a 2-Wasserstein sense. Interestingly, when the approximation is diffuse enough, TLL decreases, again highlighting its non-monotonic relationship with posterior approximation quality. In this example of a mis-specified model, the non-monotonic relationship between TLL and 2-Wasserstein distance means that TLL is, at best, a poor proxy of posterior approximation quality.

A.2 Non-monotonicity of TLL beyond the Wasserstein distance

We reproduce the experimental setup that produced Figure 4 but plot TLL against the difference of marginal standard deviations of the parameters of interest between an approximation and the exact posterior in Figure 8. We observe a similar kind of non-monotonicity as the right panel of Figure 4.

Figure 9 shows a similar phenomenon with posterior standard deviations. Figure 9 displays a similar kind of non-monotonicity as the right panel of Figure 7. The experimental setup is identical to Figure 7: we have only changed what is plotted on the x-axis.

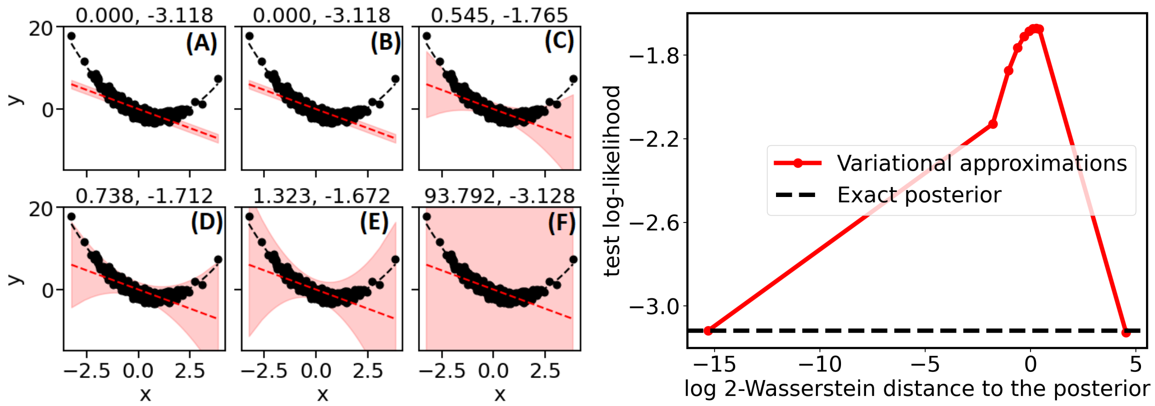

Figure 7: *(Left).* Predictive distributions under the Bayesian posterior and mean field variational approximations. The two numbers in the title of each plot are the 2-Wasserstein distance to the true posterior and test log-likelihoods computed on 10^4 test set observations. Two standard errors in the test log-likelihood estimates are (A) 0.16, (B) 0.16, (C) 0.03, (D) 0.02, (E) 0.02, (F) 0.01. *(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.

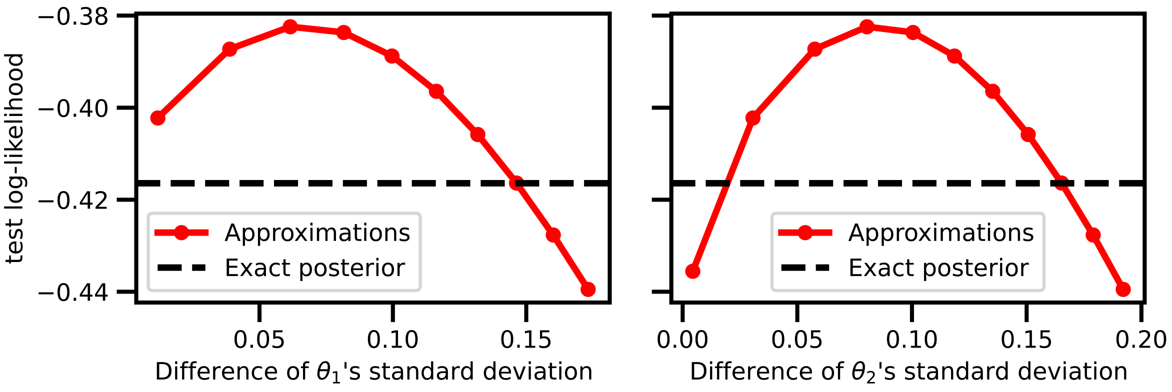

Figure 8: The non-monotonic relationship between difference in marginal standard deviations and test log-predictive density in a well-specified case. *(Left)* The x-axis reports the difference in the standard deviation of the weight θ_1 between an approximation and the posterior. *(Right)* The x-axis reports the difference in the standard deviation of the bias θ_2 between an approximation and the posterior.

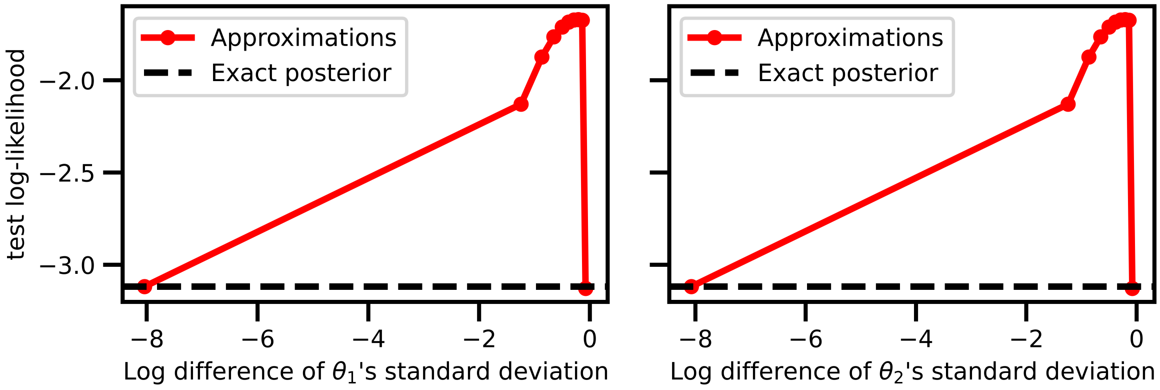

Figure 9: The non-monotonic relationship between difference in marginal standard deviations and test log-predictive density in a mis-specified case. The meaning of x-axis is similar to that of Figure 8.