Chapter 6

Deep Insights

Until now we have mostly studied the proposed approximate inference techniques empirically. In this chapter we turn to a more theoretical analysis, and concentrate mostly on the case of dropout, as it seems to be the most widely used among the various stochastic regularisation techniques. We begin by suggesting practical considerations for getting good uncertainty estimates, followed by a review of what affects predictive uncertainty characteristics. We then offer an analytical analysis in the linear case, answering many higher-level questions about the behaviour of the inference, and analyse dropout's evidence lower bound (ELBO) correlation with test log likelihood. We continue by discussing various alternative priors to the standard Gaussian prior: we discuss the properties of a discrete prior, and (approximately) derive the optimal variational posterior with a spike and slab prior. The latter, quite surprisingly, turns out to be closely related to the structure of the dropout approximating distribution. We finish the chapter with a more philosophical discussion, examining the different types of uncertainty available to us from the dropout neural networks, and suggest a new tool to optimise the dropout probabilities under the variational setting.

6.1 Practical considerations for getting good uncertainty estimates

I will suggest some "tricks of the trade" to get good predictive uncertainty estimates. First, it seems that "over-parametrised" models result in better uncertainty estimates than smaller models. Models with a large number of parameters can capture a larger class of functions, leading to more ways of explaining the data, and as a result larger uncertainty estimates further from the data. Similar behaviour was noted in [Neal, 1995] with regard

to model size. In the dropout case, this conforms with the observation that better RMSE can be obtained when a large number of parameters is used (larger than when dropout is not used as discussed in §4.5). The dropout probability is important as well, with larger models requiring a larger dropout probability: varying the dropout probability p (through either grid-search or Bayesian optimisation) we have that large models (large K) push p towards 0.5, since the weight of the entropy w.r.t. p (eq. (A.1)) is scaled by K. For a fixed model size K, smaller probabilities p result in decreasing predictive uncertainty. Further, short model length-scale results in more erratic functions drawn from the posterior hence higher uncertainty values. Intuitively, high model precision (large τ) and large amounts of data (large N) give the expected log likelihood a higher weight than the prior KL, resulting in models that can fit the data well but might overfit (this can be seen in eq. (3.14)). On the other hand, long prior length-scale (large l) gives the prior KL a higher weight than the expected log likelihood, resulting in heavily regularised models that might not fit the data as well (this can be seen in eq. (6.6) below). The prior length-scale, model precision, and dropout probability can be optimised using Bayesian optimisation and cross validation over test log likelihood (as was done in §4.3). Lastly, it seems that model structure affects predictive uncertainty considerably. Many existing models were designed and developed in order to obtain good RMSE, but the same model structure might not be ideal to get good uncertainty estimates. Adapting model structure to result in good uncertainty estimates as well as RMSE might be helpful in improving test log likelihood. The reason for this is discussed next.

6.2 What determines what our uncertainty looks like?

Predictive uncertainty is determined through a combination of model structure, model prior, and approximating distribution. This is similar to Gaussian processes (GPs), where the choice of covariance function is determined by our prior belief as to the properties of the predictive uncertainty we expect to observe. Choosing a squared exponential covariance function for example corresponds to a prior belief that predictive uncertainty increases far away from the data, and a choice of a "neural network" covariance function corresponds to a prior belief that the predictive mean does not collapse to zero far away from the data [Rasmussen and Williams, 2006]. Bayesian NNs can be seen as Gaussian process approximations [Gal and Turner, 2015; Neal, 1995; Williams, 1997], where the GP's covariance function is determined by the Bayesian NN non-linearity and prior.

Since we are approximating the Bayesian NN posterior with an approximating distribution $q(\cdot)$, this also affects our resulting predictive uncertainty. In section §4.2 we saw that the resulting predictive uncertainty depends heavily on the non-linearity and model prior (through the prior length-scale l). It seemed to be less affected by the approximating distribution $q(\cdot)$, with dropout, multiplicative Gaussian noise, and others resulting in similar predictive uncertainty.

6.3 Analytical analysis in Bayesian linear regression

We next study some of dropout's properties through its view as an approximating distribution in VI. Some interesting questions we answer include: 1) is dropout's regularisation data dependent? 2) does the dropout probability collapse to the MAP solution with finite data? and 3) does the approximate posterior collapse to a point mass in the limit of data? These questions can be answered through an analytical analysis in the special case of Bayesian linear regression. Even though we don't *need* VI in Bayesian linear regression, we can see what dropout VI looks like as we can solve everything analytically in this case. An empirical analysis of these questions for deep models is given later in §6.7.

We start with Bayesian linear regression with N data points, mapping Q-dimensional inputs $\mathbf{X} \in \mathbb{R}^{N \times Q}$ to D dimensional outputs $Y \in \mathbb{R}^{N \times D}$,

$$\mathbf{Y} = \mathbf{X}\mathbf{w} + \boldsymbol{\epsilon}.$$

Here we assume observation noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and place a standard normal prior over the weights $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. For simplicity we shall assume that $D = 1$, hence $\mathbf{w} \in \mathbb{R}^{Q \times 1}$ is a column vector.

We can find the posterior distribution $p(\mathbf{w}|\mathbf{X}, \mathbf{Y})$ analytically for this model:

$$\begin{aligned} p(\mathbf{w}|\mathbf{X}, \mathbf{Y}) &= p(\mathbf{Y}|\mathbf{X}, \mathbf{w}) p(\mathbf{w}) / p(\mathbf{Y}|\mathbf{X}) \\ &= \mathcal{N}\Big(\mathbf{w}; (\mathbf{X}^T\mathbf{X} + \mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y}, (\mathbf{X}^T\mathbf{X} + \mathbf{I})^{-1}\Big) \end{aligned} \tag{6.1}$$

The MAP estimate is thus

$$\mathbf{w} = (\mathbf{X}^T\mathbf{X} + \mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y}. \tag{6.2}$$

In the linear case, the dropout variational distribution can be derived analytically as well. This approximate posterior distribution can then be compared to the exact

posterior. Define $q_{\mathbf{m},p}(\mathbf{w}) = \prod_{i=1}^{Q} q_{m_q,p}(w_q)$ with

$$q_{m_q,p}(w_q) = p\delta(w_q - m_q) + (1-p)\delta(w_q - 0)$$

given some variational parameters $\mathbf{m}$ and retain probability[1] p. This can be rewritten as $w_q = m_q \boldsymbol{\epsilon}_q$ with $\boldsymbol{\epsilon}_q \sim \text{Bern}(p)$, and in vector form $\mathbf{w} = \text{diag}([\boldsymbol{\epsilon}_i]_{i=1}^{Q}) \cdot \mathbf{m}$. We shall write $q(\mathbf{w})$ for $q_{\mathbf{m},p}(\mathbf{w})$ for brevity.

Evaluating the log evidence lower bound (ELBO) in the variational case amounts to:

$$\begin{aligned}
\log p(\mathbf{Y}|\mathbf{X}) &= \log \int p(\mathbf{Y}|\mathbf{X},\mathbf{w})p(\mathbf{w})\mathrm{d}\mathbf{w} \\
&\geq \int q(\mathbf{w}) \log p(\mathbf{Y}|\mathbf{X},\mathbf{w})\mathrm{d}\mathbf{w} - \mathrm{KL}\Big(q(\mathbf{w})||p(\mathbf{w})\Big) \\
&\propto \sum_{n=1}^{N} \left(-\frac{1}{2}y_n^2 + y_n \mathbf{x}_n \mathbb{E}_q[\mathbf{w}] - \frac{1}{2}\mathbf{x}_n \mathbb{E}_q[\mathbf{w}\mathbf{w}^T]\mathbf{x}_n^T \right) - \mathrm{KL}\Big(q(\mathbf{w})||p(\mathbf{w})\Big).
\end{aligned}$$

Following the definition of $q(\mathbf{w})$, we know that $\mathbb{E}_q[\mathbf{w}] = p\mathbf{m}$. Further, $\mathbb{E}_q[\mathbf{w}\mathbf{w}^T] = \text{Cov}_q[\mathbf{w}] + \mathbb{E}_q[\mathbf{w}]\mathbb{E}_q[\mathbf{w}]^T = p(1-p)\text{diag}([m_i^2]_{i=1}^{Q}) + p^2\mathbf{m}\mathbf{m}^T$ since each element of $\mathbf{w} \sim q(\mathbf{w})$ is a product of a scalar and an i.i.d. Bernoulli (remember that $\mathbf{m}$ is a row vector, therefore $\mathbf{m}\mathbf{m}^T$ is a matrix).

The log evidence can thus be bounded by:

$$\begin{aligned}
\log p(\mathbf{Y}|\mathbf{X}) &\geq \sum_{n=1}^{N} \left(-\frac{1}{2}y_n^2 + p y_n \mathbf{x}_n \mathbf{m} - \frac{p^2}{2}\mathbf{x}_n \mathbf{m}\mathbf{m}^T \mathbf{x}_n^T - \frac{p(1-p)}{2}\mathbf{x}_n \text{diag}([m_i^2]_{i=1}^{Q})\mathbf{x}_n^T \right) \\
&\qquad - \mathrm{KL}\Big(q(\mathbf{w})||p(\mathbf{w})\Big) \\
&\propto \sum_{n=1}^{N} \left(-\frac{1}{2}(y_n - p\mathbf{x}_n\mathbf{m})^2 - \frac{p(1-p)}{2}\mathbf{x}_n \text{diag}([m_i^2]_{i=1}^{Q})\mathbf{x}_n^T \right) - \mathrm{KL}\Big(q(\mathbf{w})||p(\mathbf{w})\Big) \\
&= -\frac{1}{2}||\mathbf{Y} - p\mathbf{X}\mathbf{m}||^2 - \frac{p(1-p)}{2}\sum_n \mathbf{x}_n \text{diag}([m_i^2]_{i=1}^{Q})\mathbf{x}_n^T - \mathrm{KL}\Big(q(\mathbf{w})||p(\mathbf{w})\Big).
\end{aligned}$$

This last expression can be simplified by noting that $\text{diag}([m_i^2]_{i=1}^{Q}) = \text{diag}([m_i]_{i=1}^{Q})^2$ thus

$$\begin{aligned}
\mathbf{x}_n \text{diag}([m_i^2]_{i=1}^{Q})\mathbf{x}_n^T &= \mathrm{Tr}\Big(\mathbf{x}_n \text{diag}(\mathbf{m})^2 \mathbf{x}_n^T\Big) \\
&= \mathrm{Tr}\Big(\text{diag}(\mathbf{m})\mathbf{x}_n^T \mathbf{x}_n \text{diag}(\mathbf{m})\Big).
\end{aligned}$$

[1] Note that here we denote dropout probability as $1-p$ instead of p.

Summing these over n we obtain

$$\sum_n \mathbf{x}_n \text{diag}([m_i^2]_{i=1}^Q)\mathbf{x}_n^T = \text{Tr}\Big(\text{diag}(\mathbf{m})\mathbf{X}^T\mathbf{X}\text{diag}(\mathbf{m})\Big) = \mathbf{m}^T\text{diag}(\mathbf{X}^T\mathbf{X})\mathbf{m}.$$

Following appendix A, we approximate the KL between $q(\mathbf{w})$ and a standard Gaussian prior $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}; 0, I)$ as

$$\begin{aligned}\text{KL}\Big(q(\mathbf{w})||p(\mathbf{w})\Big) &= \sum_{i=1}^Q \text{KL}\Big(q(w_i)||p(w_i)\Big) \\ &\approx \frac{p}{2}\mathbf{m}^T\mathbf{m} - Q\mathcal{H}(p) + C,\end{aligned} \tag{6.3}$$

defining $\mathcal{H}(p) = -p\log p - (1-p)\log(1-p)$, and with a constant C.

Maximising the ELBO can be written as a minimisation objective:

$$\mathcal{L}(\mathbf{m}, p) = \underbrace{||\mathbf{Y} - p\mathbf{X}\mathbf{m}||^2 + p(1-p)\mathbf{m}^T\text{diag}(\mathbf{X}^T\mathbf{X})\mathbf{m}}_{\text{likelihood terms}} + \underbrace{p\mathbf{m}^T\mathbf{m} - 2Q\mathcal{H}(p)}_{\text{prior terms}}.$$

Remark (Is dropout's regularisation data dependent?)**.** It was suggested in [Srivastava et al., 2014, section 9.1] that the term $p(1-p)\mathbf{m}^T\text{diag}(\mathbf{X}^T\mathbf{X})\mathbf{m}$ in the equation above is a regularisation term dependent on the data. Following the interpretation of dropout as approximate inference with our specific distribution $q(\cdot)$ and a standard Gaussian prior we have that the term is derived from the likelihood contribution, i.e. the term is part of the generative model. But one could claim that this result follows from our ad hoc choice for $q(\cdot)$. Below we will see an alternative prior under which we can (approximately) derive the variational distribution *structure* optimally. For that alternative prior we recover an optimal $q(\cdot)$ with a distribution structure similar to the dropout variational distribution. Apart from suggesting why the dropout approximating distribution structure is sensible, the new objective will also possess similar properties to the ones studied here.

The solution to this last minimisation objective can be found analytically. We rewrite the objective as

$$\begin{aligned}\mathcal{L}(\mathbf{m}, p) = \sum_{n=1}^N (y_n^2 - 2py_n\mathbf{x}_n\mathbf{m} + p^2\mathbf{m}^T\mathbf{x}_n^T\mathbf{x}_n\mathbf{m}) + \mathbf{m}^T\Big(p(1-p)\text{diag}(\mathbf{X}^T\mathbf{X}) + p\mathbf{I}\Big)\mathbf{m} \\ - 2Q\mathcal{H}(p)\end{aligned} \tag{6.4}$$

$$= \sum_{n=1}^{N}(y_n^2 - 2py_n\mathbf{x}_n\mathbf{m}) + \mathbf{m}^T\Big(p^2\mathbf{X}^T\mathbf{X} + p(1-p)\text{diag}(\mathbf{X}^T\mathbf{X}) + p\mathbf{I}\Big)\mathbf{m} - 2Q\mathcal{H}(p)$$

Differentiating w.r.t. $\mathbf{m}$ we have

$$\frac{\partial\mathcal{L}(\mathbf{m},p)}{\partial\mathbf{m}} = -2p\mathbf{X}^T\mathbf{Y} + 2\Big(p^2\mathbf{X}^T\mathbf{X} + p(1-p)\text{diag}(\mathbf{X}^T\mathbf{X}) + p\mathbf{I}\Big)\mathbf{m}$$

Setting $\frac{\partial\mathcal{L}(\mathbf{m},p)}{\partial\mathbf{m}} = \mathbf{0}$ leads to optimal $\mathbf{m}^*$ (under the constraint $p > 0$):

$$\mathbf{m}^* = \Big(pN\boldsymbol{\Sigma} + (1-p)N\boldsymbol{\Lambda} + \mathbf{I}\Big)^{-1}\mathbf{X}^T\mathbf{Y} \tag{6.5}$$

with $\boldsymbol{\Sigma} = \mathbf{X}^T\mathbf{X}/N$ and $\boldsymbol{\Lambda} = \text{diag}(\mathbf{X}^T\mathbf{X})/N$. This effectively shrinks the off-diagonal covariance terms, reducing the sensitivity of linear regression to colinear inputs[2].

Remark (Does the dropout probability collapse to the MAP solution with finite data?)**.** Here we see that the optimal variational parameter $\mathbf{m}^*$ equals the MAP estimate given in eq. (6.2) only for $p = 1$ or for $\boldsymbol{\Lambda}$ constant. We will next see that in the limit of data, optimal p^* is indeed 1.

Plugging $\mathbf{m}^*$ into $\mathcal{L}(\mathbf{m}, p)$ we obtain:

$$\begin{aligned}\mathcal{L}(\mathbf{m}^*, p) &= \mathbf{Y}^T\mathbf{Y} - 2p\mathbf{Y}^T\mathbf{X}\mathbf{m}^* + p\mathbf{m}^{*T}\Big(pN\boldsymbol{\Sigma} + (1-p)N\boldsymbol{\Lambda} + \mathbf{I}\Big)\mathbf{m}^* - 2Q\mathcal{H}(p)\\ &= \mathbf{Y}^T\mathbf{Y} - p\mathbf{Y}^T\mathbf{X}\Big(pN\boldsymbol{\Sigma} + (1-p)N\boldsymbol{\Lambda} + \mathbf{I}\Big)^{-1}\mathbf{X}^T\mathbf{Y} - 2Q\mathcal{H}(p).\end{aligned}$$

In the limit of data we have (assuming that the limit exists)

$$\begin{aligned}\frac{\mathcal{L}(\mathbf{m}^*, p)}{N} &= \frac{\mathbf{Y}^T\mathbf{Y}}{N} - p\frac{\mathbf{Y}^T\mathbf{X}}{N}\Big(p\boldsymbol{\Sigma} + (1-p)\boldsymbol{\Lambda} + \frac{\mathbf{I}}{N}\Big)^{-1}\frac{\mathbf{X}^T\mathbf{Y}}{N} - \frac{2Q\mathcal{H}(p)}{N}\\ &\xrightarrow[N\to\infty]{} a - \mathbf{b}^T\Big(\boldsymbol{\Sigma} + (p^{-1}-1)\boldsymbol{\Lambda}\Big)^{-1}\mathbf{b}\\ &=: \bar{\mathcal{L}}(p),\end{aligned}$$

with $a = \lim_{N\to\infty}\mathbf{Y}^T\mathbf{Y}/N$ and $\mathbf{b} = \lim_{N\to\infty}\mathbf{X}^T\mathbf{Y}/N$.

[2]Equation (6.5) was previously presented in [Srivastava et al., 2014; Wager et al., 2013; Wang and Manning, 2013], where the dropout objective was interpreted as a form of ridge regression with the design matrix columns normalized. Dropout in linear networks was also studied in [Baldi and Sadowski, 2013].

Remark (Does the approximate posterior collapse to a point mass in the limit of data?)**.** We have that $p = 1$ (no dropout) is a minimiser of $\bar{\mathcal{L}}(p)$ (and the only one when $\mathbf{\Lambda}$ is positive definite): $\mathbf{\Sigma}$ and $\mathbf{\Lambda}$ are positive semi-definite (PSD). For $p < 1$, we have $p^{-1} - 1 > 0$, therefore $(p^{-1} - 1)\mathbf{\Lambda}$ is PSD, and so is $\mathbf{\Sigma} + (p^{-1} - 1)\mathbf{\Lambda}$. Since $\mathbf{\Sigma} \preceq \mathbf{\Sigma} + (p^{-1} - 1)\mathbf{\Lambda}$, we have $\left(\mathbf{\Sigma} + (p^{-1} - 1)\mathbf{\Lambda}\right)^{-1} \preceq \mathbf{\Sigma}^{-1}$, and from the definition of PSD: $\mathbf{b}^T\left(\mathbf{\Sigma} + (p^{-1} - 1)\mathbf{\Lambda}\right)^{-1}\mathbf{b} \leq \mathbf{b}^T\mathbf{\Sigma}^{-1}\mathbf{b}$. As a result,

$$a - \mathbf{b}^T\left(\mathbf{\Sigma} + (p^{-1} - 1)\mathbf{\Lambda}\right)^{-1}\mathbf{b} \geq a - \mathbf{b}^T\mathbf{\Sigma}^{-1}\mathbf{b},$$

resulting in $\bar{\mathcal{L}}(p) \geq \bar{\mathcal{L}}(1)$ for all $p \in (0, 1)$. For $\mathbf{\Lambda}$ positive definite the inequalities above are strict, leading to $\bar{\mathcal{L}}(p) > \bar{\mathcal{L}}(1)$ for all $p \in (0, 1)$, i.e. $p^* = 1$ (no dropout) is the unique minimiser of our optimisation objective, and the approximate posterior collapses to a point mass in the limit of data.

6.4 ELBO correlation with test log likelihood

This has been joint work with Mark van der Wilk.

In an attempt to maximise model performance, it might be tempting to optimise dropout's probability p as a variational parameter following its VI interpretation. Interestingly enough, optimising the dropout probability can give mixed results. In this section we analyse this behaviour. We plot model ELBO versus test log likelihood for different dropout probabilities, and assess the correlation between the ELBO and the test log likelihood.

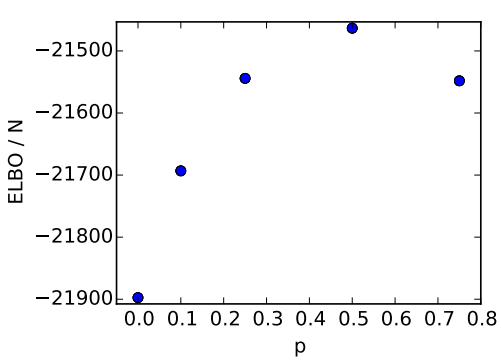
We repeat the experiment setup of §4.2 and use ReLU models with 4 hidden layers and 1024 units in each layer evaluated on Snelson and Ghahramani [2005]'s dataset with $N = 5000$ training points. We set model precision to $\tau = 50$ and assess model ELBO and test log likelihood at the end of optimisation for various probabilities[3] p of the weights to be set to zero. Each experiment was repeated three times. These can be seen in fig. 6.1. As can be seen, the ELBO correlates strongly with the test log likelihood—as the ELBO increases so does the test log likelihood. In fig. 6.2 we can see how models with dropout probability $p = 0$ (no dropout) cannot capture the full range of noise of the data, whereas by increasing the dropout probability we manage to model the noise better and better. For too large dropout probability ($p = 0.75$) the model starts underfitting.

[3]We performed a grid-search over p since we want to plot model performance for different p values.

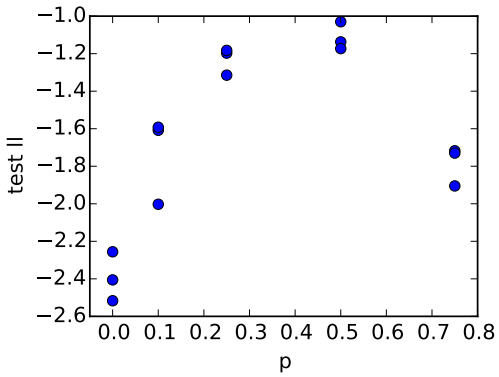
It is worth mentioning that for $p = 0.9$ the model does not converge at all. It is also interesting to note that test RMSE does not seem to correlate to the ELBO as is seen in fig. 6.3. The test log likelihood is composed of the test RMSE scaled by the uncertainty with an added uncertainty "penalty" term. This means that predictive uncertainty forms an important factor in the ELBO correlation with the test log likelihood. Repeating this experiment with $\tau = 20$ gives very similar results.

Repeating the same experiment with a different model precision value $\tau = 10$ we observe very different results though. Figure 6.4 shows that in this model when we set the dropout probability to $p = 0.5$ we obtain a worse test log likelihood than setting $p = 0$, even though the ELBO for the former is still higher than that of the latter. Note though that the model fits well with no dropout in this setting, and adding dropout with a large probability leads to deteriorated results (fig. 6.5). This might stem from the ELBO being too loose in some model setups, explaining why some attempts at optimising the dropout probability have failed in the past. However, for models in which the bound is tight enough, dropout optimisation can be done fairly well. This is explored in §6.7.

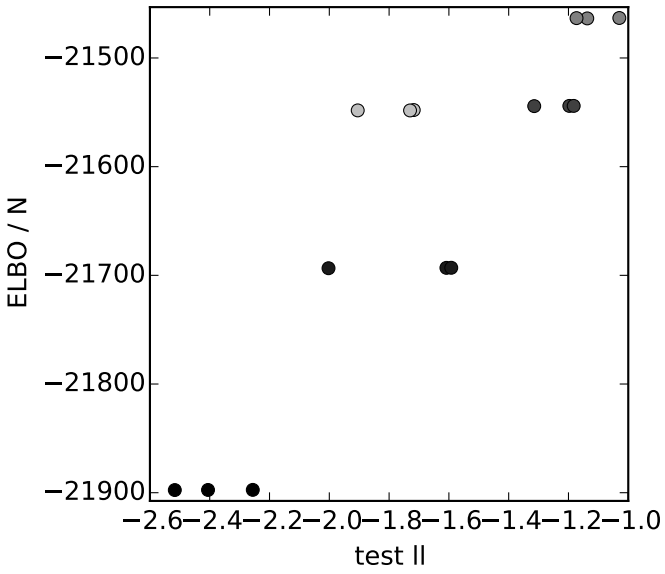
Remark (Model selection and bound tightness)**.** In a wider context, this last result above suggests some interesting questions. The assumption underlying variational inference is that models with higher ELBO correspond to "better" models (better according to what metric is a different question though). This is because the higher the ELBO, the lower the KL divergence between the approximate posterior and the true posterior would become. This fact is what drives us to look for the variational parameters that maximise the ELBO, and in turn minimise this notion of "distance" between our approximation and the true posterior. Since in our setting above the dropout probability p is a variational parameter, we would assume that maximising the ELBO w.r.t. p would lead to improved model performance. The fact that this is the case for some models ($\tau = 20, 50$) but not others ($\tau = 10$) is perplexing.
These results are related to the concept of *bound tightness*. MacKay [1992a] approximated the model evidence (which our ELBO is a lower bound to) and showed positive correlation between the evidence and test RMSE. He then performed model selection by choosing the model with the highest evidence. In VI the ELBO is often used as a proxy to the model evidence when the ELBO is tight enough [Rasmussen and Williams, 2006] (i.e. models with higher model evidence are assumed to have higher bound to that evidence). When changing model precision, different model precision values τ correspond to different models, and therefore each ELBO is a bound to a different (constant) model evidence. In Gaussian process approximations



(a) ELBO as a function of dropout p

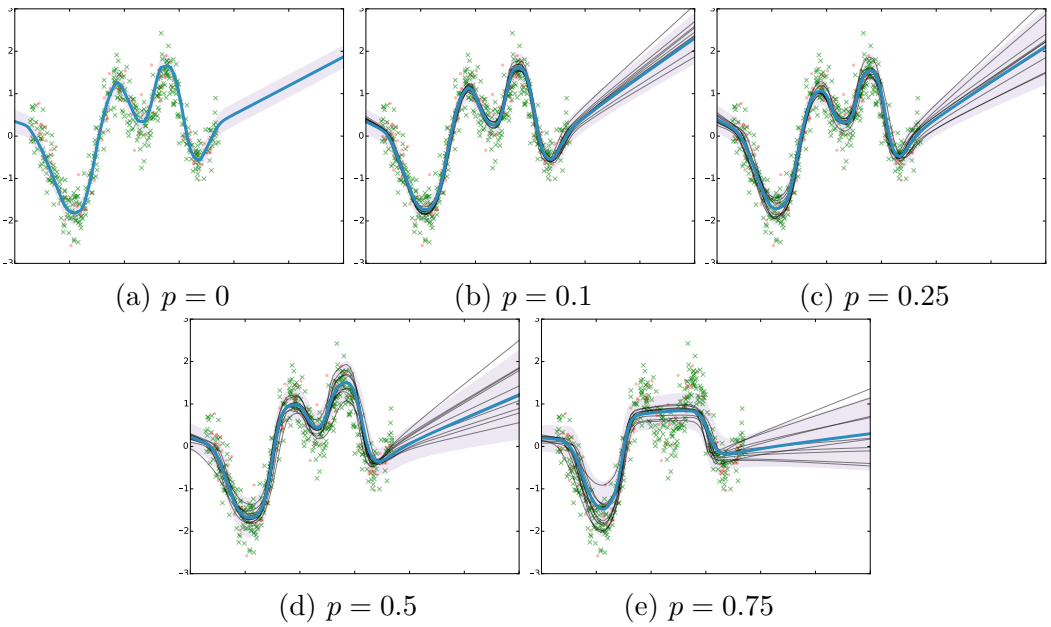


(b) Test log likelihood as a function of p



(c) ELBO as a function of test log likelihood for various p (darker = lower probability).

Fig. 6.1 ELBO (per training point) and test log likelihood (per test point) for various values of dropout probability for a model with 4 hidden layers, 1024 units, and model precision $\tau = 50$.



(a) $p = 0$ (b) $p = 0.1$ (c) $p = 0.25$

(d) $p = 0.5$ (e) $p = 0.75$

Fig. 6.2 Model fit for model precision $\tau = 50$ with various dropout probabilities

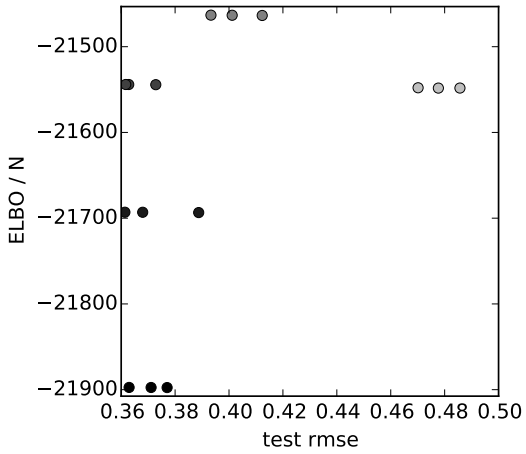
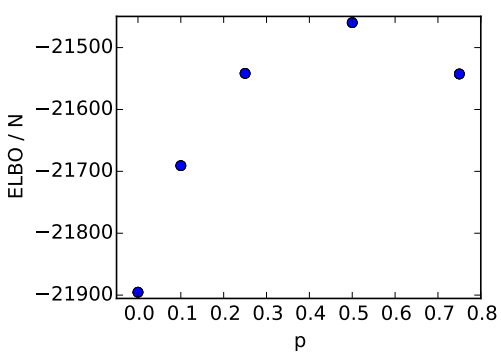
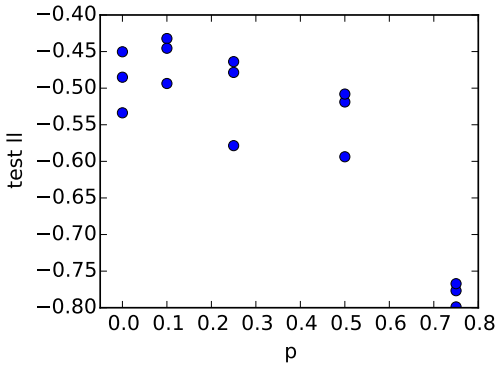


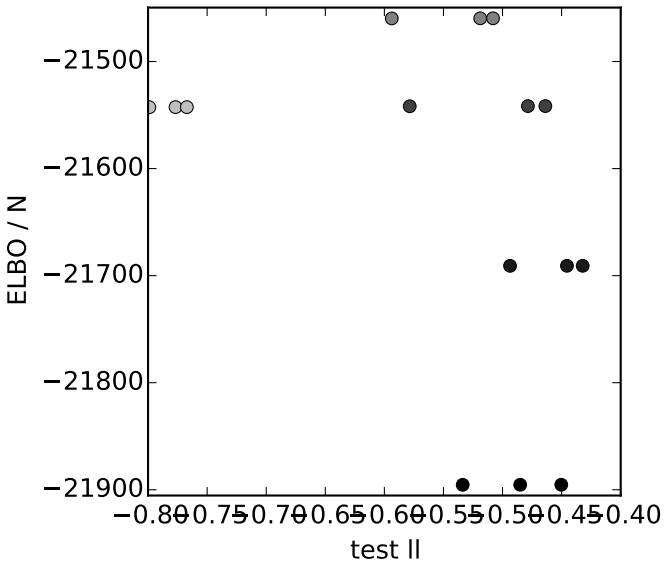
Fig. 6.3 ELBO as a function of test RMSE for $\tau = 50$



(a) ELBO as a function of dropout p

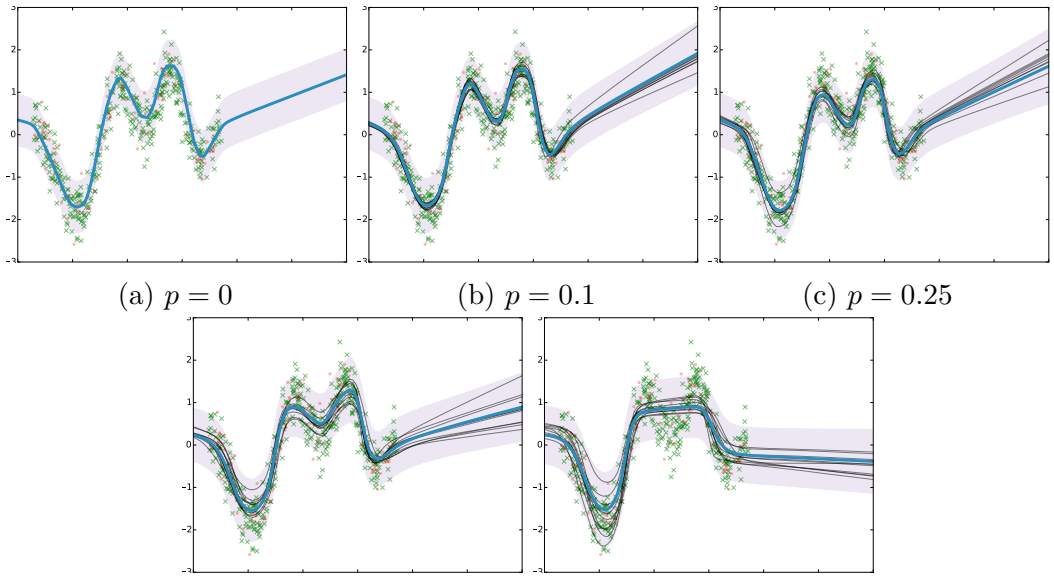


(b) Test log likelihood as a function of p



(c) ELBO as a function of test log likelihood for various p (darker = lower probability).

Fig. 6.4 ELBO (per training point) and test log likelihood (per test point) for various values of dropout probability for a model with 4 hidden layers, 1024 units, and model precision $\tau = 10$.



(a) $p = 0$ (b) $p = 0.1$ (c) $p = 0.25$

(d) $p = 0.5$ (e) $p = 0.75$

Fig. 6.5 Model fit for model precision $\tau = 10$ with various dropout probabilities

for example the bound is empirically observed to be tight enough to allow us to choose a model precision τ based on the ELBO. In our setting the ELBO seems to not be indicative of what model we should choose in some cases (fig. 6.6 shows the ELBO for all models above, with highest ELBO obtained at $p = 0.5$; further, fig. 6.7 shows that $\tau = 10$ has slightly higher ELBO than $\tau = 20$ and than $\tau = 50$). This suggests that the bound is not as tight as in the GP case.
Attempting to explain the above, there are several possible hypotheses one could propose:

1. The prior might dominate the optimisation objective in some models because there is not enough data compared to model size. The entropy in p stemming from the KL contribution (eq. (A.1)), which gives the ELBO its parabola-like shape, is scaled by the model size. In the limit of data this term would vanish. This interpretation is supported by the results in fig. 6.8 where the experiment with $\tau = 10$ above is repeated with $50 \cdot 10^6$ data points instead of $50 \cdot 10^3$, and resulting in positive correlation. Note that in this setting $\tau = 50$ has a "kink" in the ELBO-test log likelihood plot, but in both cases ELBO-test RMSE is highly correlated (not shown).

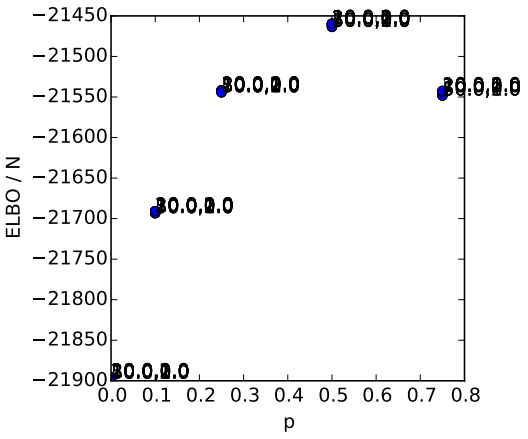


Fig. 6.6 ELBO as a function of dropout p for all model precision values $\tau = 10, 20, 50$

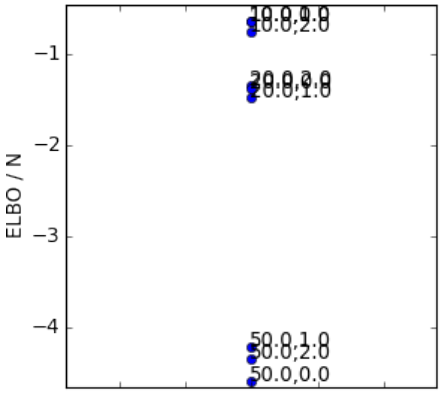


Fig. 6.7 ELBO as a function of dropout p for all model precision values $\tau = 10, 20, 50$, **zoomed in at** $p = 0.5$ (with labels of the form (τ, repetition number))

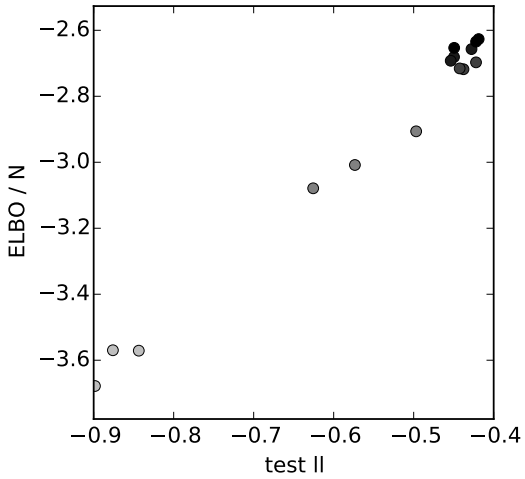


Fig. 6.8 ELBO as a function of test log likelihood for various p (darker = lower probability), with 4 hidden layers, 1024 units, and model precision $\tau = 10$ with $\mathbf{50 \cdot 10^6}$ **data points**.

2. The model might be misspecified (as discussed in [MacKay, 1992a, section 3.4]). MacKay [1992a] showed that when using inappropriate prior distributions a "kink" appeared in his evidence-test RMSE plots; he solved this by changing his prior. A lack of correlation between model evidence and test RMSE would lead to a lack of correlation between the ELBO and test RMSE even if the bound *is* tight.

3. The ELBO constant terms might not be specified correctly. The KL condition in appendix A only requires us to specify a $q(\boldsymbol{\omega})$ and $p(\boldsymbol{\omega})$ s.t. the derivatives of the ELBO and the dropout optimisation objective agree w.r.t. $\mathbf{W}$. The terms w.r.t. p however can change arbitrarily.

4. Lastly, the dropout probability might not be a variational parameter at all but rather a model parameter. Changing the dropout probability might change the model evidence—the quantity we bound.

Hypotheses 2 and 3 suggest that various terms apart from $\mathbf{W}$ in $p(\boldsymbol{\omega})$ might be affecting the optimisation objective. For example, specifying a different prior s.t. the KL condition is still satisfied w.r.t. the same approximating distribution $q(\boldsymbol{\omega})$ results in a different ELBO (as we will see next). Hypothesis 4 relates to prior selection as well. As we will see in a later section in this chapter, under some priors the dropout probability will be determined by our prior hyper-parameters, hence changing the dropout probability would change the evidence we bound (explaining the peculiar behaviour above).
In the final sections of this chapter we give some evidence in support of the various possible hypotheses suggested above. Further study of these hypotheses is left for future research.

6.5 Discrete prior models

In the previous chapters we attempted to use a discrete approximating distribution $q(\mathbf{w}_k) = p\delta(\mathbf{w}_k - \mathbf{0}) + (1-p)\delta(\mathbf{w}_k - \mathbf{m}_k)$ to recover dropout. But our prior had a continuous probability density function (a standard Gaussian distribution). We were therefore forced to approximate the KL between the approximating distribution and the prior by embedding the approximating distribution in a continuous space using a mixture of two Gaussians with small standard deviations σ (appendix A). But for a

fixed small σ the constant in the KL to the prior (eq. (A.1)) is rather large, making the lower bound quite loose. This bound becomes looser and looser for smaller and smaller σ, and diverges to negative infinity at $\sigma = 0$. This raises issues when we attempt to optimise the ELBO w.r.t. model hyper-parameters such as the length-scale for example[4]. More specifically, we will look at the bound when the prior above is set as $\mathcal{N}(0, l^{-2}I)$. For this prior the constant C contains the terms $-l^2\sigma^2 + \log l^2\sigma^2$. If we maximise the ELBO w.r.t. length-scale l, we would choose (with a very small fixed σ^2) a very large l. In fact, we can make the model completely ignore the likelihood term by setting σ^2 to be small enough: for every fixed number of points N there exists a σ^2 value such that $\log l^2$ dominates $\sum_n \frac{\tau}{2}(\mu - y_n)^2$ with μ being the mean of the data[5]. Because σ is held constant and identical in all models, if we perform model comparison based on the ELBO we would prefer the model with the longer length-scale l^2.

This suggests that the model might be misspecified w.r.t. our choice of approximating distribution (which we chose in order to recover dropout). Model misspecification is often discussed with respect to model evidence—when the evidence does not correlate to test error (see for example [MacKay, 1992a]). But in our case we compare log evidence *lower bounds*, and are interested to *define a model* in which a given approximating distribution specifies a tight bound.

A possible way to fix this issue is to specify an alternative prior, inducing a slightly altered model. Instead of using a continuous probability density function $p_l(\mathbf{w}) = \mathcal{N}(0, l^{-2}I)$, we will use a discrete probability *mass* function $p_l(\mathbf{w}) \propto e^{-\frac{l^2}{2}\mathbf{w}^T\mathbf{w}}$ defined over a *finite* space $\mathbf{w} \in X$ (a similar approach of quantising the space was used in [Hinton and Van Camp, 1993]). A continuous prior forced us to embed the discrete approximating distribution $q(\cdot)$ in a continuous space in order for the KL between the two to be properly defined. We did this using a mixture of Gaussians with small standard deviations σ. But this approach has led to our increasingly loose bound as the standard deviation σ decreased. The use of a discrete prior instead allows us to evaluate the KL divergence of both distributions over a finite space X.

In our case, since we optimise model parameters $\mathbf{w}$ on a computer, we define the space X to be a finite vector space defined over the finite field of numbers representable on a computer (for example numbers up to a certain precision). To ensure parameter gradients are properly defined, we relax the objective and embed the parameters in a continuous space *for optimisation*.

[4] This can be circumvented by grid-searching w.r.t. validation log likelihood instead of the ELBO.

[5] For large enough l^2 the optimal model parameters would be zero because of the $-\frac{l^2 p}{2}\mathbf{m}^T\mathbf{m}$ term in the prior; this leads to the model predicting the mean of the data denoted μ here.

Given the discrete approximating distribution $q(\mathbf{w}_k) = p\delta(\mathbf{w}_k - \mathbf{0}) + (1-p)\delta(\mathbf{w}_k - \mathbf{m}_k)$ the expected log likelihood terms stay the same as before, and the KL to the discrete prior distribution above is given by:

$$\begin{aligned}
\mathrm{KL}(q(\mathbf{w})||p_l(\mathbf{w})) &= \sum_{\mathbf{w}\in X} q(\mathbf{w}) \log \frac{q(\mathbf{w})}{p_l(\mathbf{w})} \\
&= \sum_{\mathbf{w}\in X} q(\mathbf{w}) \log \frac{q(\mathbf{w})}{e^{-\frac{l^2}{2}\mathbf{w}^T\mathbf{w}}} + \log Z_l \\
&= p \log \frac{p}{1} + (1-p) \log \frac{1-p}{e^{-\frac{l^2}{2}\mathbf{m}^T\mathbf{m}}} + \log Z_l \\
&\qquad + \sum_{\mathbf{w}\in X/\{0,\mathbf{m}\}} 0 \log \frac{0}{e^{-\frac{l^2}{2}\mathbf{w}^T\mathbf{w}}} \\
&= -\mathcal{H}(p) + \frac{l^2(1-p)}{2}\mathbf{m}^T\mathbf{m} + \log Z_l
\end{aligned}$$

with $\mathcal{H}(p) = -p\log p - (1-p)\log(1-p)$ and with the last transition following the identity $0 \log 0 = 0$. Since Z_l is the normaliser of $p_l(\mathbf{w})$:

$$Z_l = \sum_{\mathbf{w}\in X} e^{-\frac{l^2}{2}\mathbf{w}^T\mathbf{w}} \approx |\Delta\mathbf{w}|^{-1}(2\pi l^{-2})^{K/2}$$

with K being the dimensionality of the vector $\mathbf{w}$ and $|\Delta\mathbf{w}|$ the quantisation interval of the space X, we have:

$$\log Z_l \approx -K \log l - \log |\Delta\mathbf{w}| + \frac{K}{2} \log 2\pi.$$

This leads to

$$\mathrm{KL}(q(\mathbf{w})||p_l(\mathbf{w})) \approx \frac{l^2(1-p)}{2}\mathbf{m}^T\mathbf{m} - K \log l + \frac{K}{2} \log 2\pi - \mathcal{H}(p) - \log |\Delta\mathbf{w}|. \tag{6.6}$$

Surprisingly perhaps, the derivatives of this KL is identical to the one we used before (eq. (A.1)) for the terms $\mathbf{m}$ and p, but unlike before, there are no additional terms that diverge to infinity (terms dependent on σ). This means that approximate inference with this prior following the setting of §4.2 would give the same results observed in that chapter, where the only difference is the bound being more tight. An alternative interesting prior is discussed next, a prior that sheds light on dropout's peculiar *structure*.

6.6 Dropout as a proxy posterior in spike and slab prior models

I thank Nilesh Tripuraneni for suggesting the relation between dropout and spike and slab priors.

In his thesis from 1992, David MacKay discussed the possibility of placing a spike and slab prior over a neural network's weights [MacKay, 1992a, section 7.4]. MacKay recalled personal communication with Geoff Hinton, who suggested that an ideal prior in BNNs would set part of the weights to be exactly zero[6]. Many works at the time tried to approximate inference in similarly motivated models to varying degrees of success [Ji et al., 1990; Nowlan and Hinton, 1992; Weigend et al., 1991], none of which survived to modern day. In the following we will develop approximate inference in BNNs with spike and slab priors relying on recent VI advances, in effect formalising these ideas from 25 years ago. We will approximate the *optimal structure* of the approximating distribution, which as we will see turns out to be closely related to dropout's structure.

6.6.1 Historical context

MacKay [1992a]'s comments mentioned above were made with respect to an early draft of Nowlan and Hinton [1992], which extended on the work of Weigend et al. [1991].

Weigend et al. [1991] relied on MDL to motivate the addition of a regularisation term to the optimisation objective of a NN, in order to penalise complex models. They offered a Bayesian interpretation to their objective as maximising the log likelihood plus log prior (MAP) with a mixture prior over the weights, where the mixture was of a wide uniform and a Gaussian centred at zero, with the Gaussian's width being learnt.

Nowlan and Hinton [1992], following the Bayesian interpretation of the weight decay as a Gaussian prior, claimed that a Gaussian prior can be used to eliminate small weights. But at the same time it forces other weights to contract to the origin as well, weights that are needed to "explain the data as well" and thus should not be forced to the origin (the demand for a prior to contract "unneeded" weights to zero might have stemmed from the MDL interpretation of model complexity, but was not justified in the paper). Nowlan and Hinton [1992] proposed to use a mixture of a narrow Gaussian together with a wide Gaussian. They implemented this by placing a mixture of Gaussians prior and optimising the MAP over the means and standard deviations of the Gaussians.

[6]Note that the *prior* is constructed to be sparse here rather than the *posterior* as in dropout.

MacKay [1992a], repeating the desire to set part of the weights to be exactly zero, commented on the work of Nowlan and Hinton [1992]. MacKay [1992a] said that the non-zero width of the narrow Gaussian was a consequence of computational practicality, thus Nowlan and Hinton [1992]'s approach of inferring the width of the narrow Gaussian was not appropriate. MacKay [1992a] concluded that it will be interesting to see if a priori setting part of the weights to be exactly zero could be formalised, leading to a "well-founded" technique.

6.6.2 Spike and slab prior models

The desire to set a subset of weights to be exactly zero a priori can be materialised by placing a *spike and slab* prior over the weights. Spike and slab distributions return realisations which are identically zero with some probability (the spike) or sampled from a wide Gaussian otherwise (the slab), and have been used in the context of variable selection for example [Chipman; George and McCulloch, 1993, 1997; Ishwaran and Rao, 2005; Madigan and Raftery, 1994; Mitchell and Beauchamp, 1988]. We start by placing a spike and slab prior over each row of each weight matrix $\mathbf{w}_{ik}$, and assume that the rows are a priori independent of each other:

$$p(\mathbf{w}_{ik}) = f\delta(\mathbf{w}_{ik} - \mathbf{0}) + (1 - f)\mathcal{N}(\mathbf{w}_{ik}; \mathbf{0}, l^{-2}I) \tag{6.7}$$

with prior probability f and prior length-scale l. The first component is a point mass at a vector of zeros, setting an entire weight matrix row to zero with probability f (**a priori**). The second component draws a weight vector from a multivariate Gaussian distribution with probability $1 - f$. Draws from this distribution will give high frequencies for short length-scale l (resulting in erratic functions), and low frequencies for long length-scales (resulting in smooth functions). Placing the distribution over the rows of the matrix $\mathbf{W}$ captures correlations over the function's frequencies ($\mathbf{W}$'s columns, §3.2.3).

Given the prior above, we use a Gaussian likelihood for regression:

$$p(\mathbf{y}|\mathbf{x}, \boldsymbol{\omega}) = \mathcal{N}(\mathbf{h}_{L-1}(...\mathbf{h}_1(\mathbf{x})...)\mathbf{W}_L, \tau^{-1}I)$$

with $\mathbf{h}_i(\mathbf{x}) = \sigma(\mathbf{x}W_i)$, L layers, $\boldsymbol{\omega} = \{W_i\}_{i=1}^L$ a set of random matrices, and observation noise τ^{-1}.

6.6.3 Related work

Some of the techniques above have been revisited in other modern literature [Blundell et al., 2015; Goodfellow et al., 2012; Louizos, 2015; Titsias and Lázaro-Gredilla, 2011]. Blundell et al. [2015] for example use a BNN prior which resembles that of Nowlan and Hinton [1992]. Louizos [2015] looks at a spike and slab prior distribution in a VI setting, but makes use of a loose lower bound in order to evaluate the KL divergence between a spike and slab approximate posterior and the prior.

6.6.4 Approximate inference with free-form variational distributions

The posterior distribution over $\boldsymbol{\omega}$ given observed data $\mathbf{X}, \mathbf{Y}$ is difficult to evaluate. Instead, we use variational inference and approximate it with a variational distribution $q(\boldsymbol{\omega})$. We define the approximating distribution to factorise over the rows of the weight matrices $\mathbf{w}_{ik}$:

$$q(\boldsymbol{\omega}) = \prod_{i=1}^{L} \prod_{k=1}^{K} q(\mathbf{w}_{ik})$$

with $\mathbf{W}_i$ composed of k rows.

Unlike previous work, here we will not specify a structure for $q(\cdot)$, but rather find the optimal variational distribution structure using *calculus of variations*. This can be done using the following lemma and the corollary following it:

Lemma 1. Let $p(W)$ and $q(W)$ be two distributions defined over the same space $W \in \mathcal{W}$. Further, assume the distribution $q(W)$ factorises over $W = [\mathbf{w}_1, ..., \mathbf{w}_K]$: $q(W) = \prod_k q(\mathbf{w}_k)$.

The optimal distribution $q^*(\mathbf{w}_k)$ minimising $\mathrm{KL}(q(W)||p(W))$ is given by:

$$q^*(\mathbf{w}_k) \propto e^{E_{q(W)/q(\mathbf{w}_k)}[\log p(W)]}.$$

Proof. Using calculus of variations,

$$\begin{aligned}
&\frac{\partial}{\partial q(\mathbf{w}_k)} \left(\mathrm{KL}(q(W)||p(W)) + \lambda \left(\int q(\mathbf{w}_k) \mathrm{d}\mathbf{w}_k - 1 \right) \right) \\
&\quad = \int \frac{q(W)}{q(\mathbf{w}_k)} \log \frac{q(W)}{p(W)} \mathrm{d}(\mathbf{w}_i)_{\neq k} + \lambda \\
&\quad = -E_{q(W)/q(\mathbf{w}_k)}[\log p(W)] + \int \frac{q(W)}{q(\mathbf{w}_k)} \log \frac{q(W)}{q(\mathbf{w}_k)} \mathrm{d}(\mathbf{w}_i)_{\neq k} + \log q(\mathbf{w}_k) + \lambda
\end{aligned}$$

Setting this last quantity to zero we obtain

$$\log q(\mathbf{w}_k) = -\lambda - E_{q(W)/q(\mathbf{w}_k)}[\log q(W)/q(\mathbf{w}_k)] + E_{q(W)/q(\mathbf{w}_k)}[\log p(W)]$$

which leads to

$$q^*(\mathbf{w}_k) \propto e^{E_{q(W)/q(\mathbf{w}_k)}[\log p(W)]}.$$

□

This result is known in the literature. From this lemma we can derive the following corollary:

Corollary 1. Given a prior of the form $p(W) = \prod_k p(\mathbf{w}_k)$ and likelihood $p(Y|W)$, the posterior $p(W|Y)$ can be approximated with an optimal factorised distribution $q^*(W) = \prod_k q^*(\mathbf{w}_k)$ given by

$$q^*(\mathbf{w}_k) \propto e^{E_{q(W)/q(\mathbf{w}_k)}[\log p(Y|W)]} p(\mathbf{w}_k).$$

This last equation is fundamental in variational message passing for example [Winn and Bishop, 2005].

6.6.5 Proxy optimal approximating distribution

In our BNN case above, the optimal approximating distribution can be split into two terms:

$$q^*(\mathbf{w}_{ik}) \propto \underbrace{e^{E_{q(\boldsymbol{\omega})/q(\mathbf{w}_{ik})}[\log p(\mathbf{Y}|\mathbf{X},\boldsymbol{\omega})]}}_{\text{Nasty distribution}} \underbrace{p(\mathbf{w}_{ik})}_{\text{Nice dist.}}.$$

The nasty distribution term can be evaluated analytically for linear models. But for deep networks this distribution becomes rather complex. Instead, we will moment-match the nasty distribution. We fit it with a Gaussian parametrised with mean $\mathbf{m}_{ik}$ and diagonal variance $\sigma_{ik}^2 I$. This alters the optimal approximating distribution structure to

$$q_{\mathbf{m}_{ik},\sigma_{ik}}(\mathbf{w}_{ik}) \propto \mathcal{N}(\mathbf{w}_{ik}; \mathbf{m}_{ik}, \sigma_{ik}^2 I) p(\mathbf{w}_{ik}).$$

We then maximise the ELBO w.r.t. the variational parameters $\mathbf{m}_{ik}, \sigma_{ik}$ for each approximating distribution $q_{\mathbf{m}_{ik},\sigma_{ik}}(\mathbf{w}_{ik})$.

The structure of the approximating distribution for the spike and slab prior (eq. (6.7)) can be evaluated analytically now:

$$\begin{aligned} q_{\mathbf{m}_{ik},\sigma_{ik}}(\mathbf{w}_{ik}) &\propto \mathcal{N}(\mathbf{w}_{ik}; \mathbf{m}_{ik}, \sigma_{ik}^2 I)\Big(f\delta(\mathbf{w}_{ik} - \mathbf{0}) + (1-f)\mathcal{N}(\mathbf{w}_{ik}; \mathbf{0}, l^{-2} I)\Big) \\ &\propto fC_1\delta(\mathbf{w}_{ik} - \mathbf{0}) + (1-f)C_2\mathcal{N}\left(\mathbf{w}_{ik}; \frac{\mathbf{m}_{ik}}{1+l^2\sigma_{ik}^2}, \frac{\sigma_{ik}^2}{1+l^2\sigma_{ik}^2} I\right) \end{aligned}$$

with $C_1 = \mathcal{N}(\mathbf{0}; \mathbf{m}_{ik}, \sigma_{ik}^2 I)$ and $C_2 = \mathcal{N}(\mathbf{0}; \mathbf{m}_{ik}, (\sigma_{ik}^2 + l^{-2})I)$.

The normaliser for this distribution is given by

$$\begin{aligned} Z_q &= \int fC_1\delta(\mathbf{w}_{ik} - \mathbf{0}) + (1-f)C_2\mathcal{N}\left(\mathbf{w}_{ik}; \frac{\mathbf{m}_{ik}}{1+l^2\sigma_{ik}^2}, \frac{\sigma_{ik}^2}{1+l^2\sigma_{ik}^2} I\right) \mathrm{d}\mathbf{w}_{ik} \\ &= f\mathcal{N}(\mathbf{0}; \mathbf{m}_{ik}, \sigma_{ik}^2 I) + (1-f)\mathcal{N}(\mathbf{0}; \mathbf{m}_{ik}, (\sigma_{ik}^2 + l^{-2})I). \end{aligned}$$

Writing

$$\begin{aligned} \alpha(\mathbf{m}_{ik}, \sigma_{ik}, f) &= \frac{f\mathcal{N}(\mathbf{0}; \mathbf{m}_{ik}, \sigma_{ik}^2 I)}{(1-f)\mathcal{N}(\mathbf{0}; \mathbf{m}_{ik}, (\sigma_{ik}^2 + l^{-2})I)} \\ &= \frac{f}{(1-f)}(1 + l^{-2}\sigma_{ik}^{-2})^{K/2} e^{-\frac{1}{2}(l^2\sigma_{ik}^4 + \sigma_{ik}^2)^{-1}\mathbf{m}_{ik}^T\mathbf{m}_{ik}} \end{aligned} \tag{6.8}$$

we have

$$q_{\mathbf{m}_{ik},\sigma_{ik}}(\mathbf{w}_{ik}) = \frac{\alpha}{\alpha+1}\delta(\mathbf{w}_{ik} - \mathbf{0}) + \frac{1}{\alpha+1}\mathcal{N}\left(\mathbf{w}_{ik}; \frac{\mathbf{m}_{ik}}{1+l^2\sigma_{ik}^2}, \frac{\sigma_{ik}^2}{1+l^2\sigma_{ik}^2} I\right). \tag{6.9}$$

This approximating distribution sets each weight row to zero with probability $\alpha/(\alpha+1)$, which is determined by the magnitude of the variational parameters and the prior probability f. With probability $1/(\alpha+1)$, a row's weight is drawn from a normal distribution centred around mean $\mathbf{m}_{ik}$ scaled by $1/(1+l^2\sigma_{ik}^2)$, and with variance $\sigma_{ik}^2/(1+l^2\sigma_{ik}^2)$. For a large prior probability f tending towards 1, we have that α will tend towards infinity, and $q(\cdot)$ will tend towards a point estimate at 0. For small prior probability f, the distribution $q(\cdot)$ will tend towards a multivariate Gaussian distribution. Further, letting σ_{ik} tend towards 0 we recover dropout's "spike and spike" behaviour (but with data dependent dropout probability, which tends towards 1). On the other hand, by increasing σ_{ik}, the variance of the Gaussian component will tend towards l^{-2}.

6.6.6 Spike and slab and dropout

Evaluating the ELBO with our approximating distribution, we get the following objective:

$$\begin{aligned}\log p(\mathbf{Y}|\mathbf{X}) &\geq \sum_{i=1}^{N} -\frac{\tau}{2}\int q(\boldsymbol{\omega})||\mathbf{y}_i - \mu^{\boldsymbol{\omega}}(\mathbf{x}_i)||^2 \mathrm{d}\boldsymbol{\omega} \\ &\quad + \mathrm{KL}(q(\boldsymbol{\omega})||p(\boldsymbol{\omega})) + \frac{D}{2}\log\tau - \frac{D}{2}\log 2\pi \\ &=: \mathcal{L}\end{aligned}$$

defining $\mu^{\boldsymbol{\omega}}(\mathbf{x}_i) = \mathbf{h}_{L-1}(...\mathbf{h}_1(\mathbf{x})...)\mathbf{W}_L$ with D output dimensions. This last quantity can be approximated by Monte Carlo integration with a single draw $\widehat{\boldsymbol{\omega}}_i \sim q(\boldsymbol{\omega})$ for each integral in the summation:

$$\widehat{\mathcal{L}} := \sum_{i=1}^{N} -\frac{\tau}{2}||\mathbf{y}_i - \mu^{\widehat{\boldsymbol{\omega}}_i}(\mathbf{x}_i)||^2 + \mathrm{KL}(q(\boldsymbol{\omega})||p(\boldsymbol{\omega})) + \frac{D}{2}\log\tau - \frac{D}{2}\log 2\pi$$

with $\widehat{\mathcal{L}}$ forming an unbiased estimator to the exact ELBO, $E_{q(\boldsymbol{\omega})}[\widehat{\mathcal{L}}] = \mathcal{L}$. In appendix C we evaluate the KL divergence between the approximating distribution and the prior analytically, which results in a term similar to L_2 regularisation.

Dropout can be seen as a proxy to this last objective. Evaluating the last objective is identical to performing a stochastic forward pass through the deep model, where each weight row is dropped with probability determined by the magnitude of the row. Gaussian noise is added to rows which are not dropped. Apart from the added noise, this is similar to dropout, which sets each weight row to zero with a certain probability. But unlike dropout, the probability of a row being dropped is determined by the data[7]. Performing a grid-search over the dropout probabilities mimics this to a certain extent.

Remark (Dropout's probability as a model parameter)**.** Note that with this spike and slab prior, the dropout probability $p = \alpha(\mathbf{M}, \boldsymbol{\sigma}, f)/(\alpha(\mathbf{M}, \boldsymbol{\sigma}, f) + 1)$ is not a variational parameter and cannot be optimised directly, but only by changing $\mathbf{M}$ and f. However, changing the prior probability f changes the model and hence the model evidence. This means that for a fixed $\mathbf{M}$ if we were to change the posterior dropout probability p (which depends on the variational parameter $\mathbf{M}$ and model hyper parameter f) we would effectively be *changing the model evidence*—hence

[7]Note that $\frac{\alpha}{\alpha+1} \neq f$ due to the second term in eq. (6.8).

obtain bounds to different constant quantities! This can be seen as evidence towards hypothesis 4 in §6.4.

This approximate inference cannot be easily implemented because the parameter α depends on the exponent of $-\mathbf{m}^T\mathbf{m}$ which can be numerically unstable (as it decreases to zero very quickly). We leave further work on this to future research. Concentrating on the previous prior then, we next explore the dropout probability's effects on the model's epistemic uncertainty.

6.7 Epistemic, Aleatoric, and Predictive uncertainties

I thank Jiri Hron for contributing the code for the Concrete distribution used in this section.

We finish this chapter with a more philosophical discussion of the different types of uncertainty available to us, and ground our discussion in the development of new tools to better understand these. In section 1.2 we discussed the different types of uncertainty encountered in Bayesian modelling: epistemic uncertainty which captures our ignorance about the models most suitable to explain our data, aleatoric uncertainty which captures noise inherent in the environment, and predictive uncertainty which conveys the model's uncertainty in its output.

Epistemic uncertainty reduces as the amount of observed data increases—hence its alternative name "reducible uncertainty". When dealing with models over functions, this uncertainty can be captured through the range of possible functions and the probability given to each function. This uncertainty is often depicted by generating function realisations from our distribution and estimating the variance in the set of functions (for example over a finite input set $\mathbf{X}$). Aleatoric uncertainty captures noise sources such as measurement noise—noises which cannot be explained away even if more data were available (although this uncertainty *can* be reduced through the use of higher precision sensors for example). This uncertainty is often modelled as part of the likelihood, at the top of the model, where we place some noise corruption process on the model output. Gaussian corrupting noise is often assumed in regression, although other noise sources are popular as well such as Laplace noise. By inferring the Gaussian likelihood's precision parameter τ for example we can estimate the amount of aleatoric noise inherent in the data.

It can be difficult to distinguish different types of noise in a single model though, and in section 4.6 we proposed a possible model to do this. In that model we learnt both the aleatoric noise by optimising the (per point) model precision, and captured the epistemic uncertainty using dropout through a grid-search over the dropout probability (note that we could have searched over the model precision as well instead of optimising it, and indeed standard practice in the field of deep learning is to grid-search over it indirectly as part of the NN's weight decay).

Combining both types of uncertainty gives us the predictive uncertainty—the model's confidence in its prediction taking into account noise it can explain away and noise it cannot. This uncertainty is often obtained by generating multiple functions from our model and corrupting them with noise (with precision τ). Calculating the variance of these outputs on multiple inputs of interest we obtain the model's predictive uncertainty. This uncertainty has different properties for different inputs. Inputs near the training data will have a smaller epistemic uncertainty component, while inputs far away from the training data will have higher epistemic uncertainty. Similarly, some parts of the input space might have larger aleatoric uncertainty than others, with these inputs producing larger measurement error for example. These different types of uncertainty are of great importance in fields such as AI safety [Amodei et al., 2016] and autonomous decision making, where the model's epistemic uncertainty can be used to avoid making uninformed decisions with potentially life-threatening implications.

When using dropout NN (or any other SRT), we need to optimise over both the dropout probability p and the model weight decay parameters λ. This is in order to find the epistemic uncertainty and aleatoric uncertainty, respectively. This optimisation can be done by performing a grid-search over both quantities (which we performed in section 4.3 for example w.r.t. validation log-likelihood). But one of the difficulties with the approach above is that grid-searching over both parameters can be expensive and time consuming, especially when done with large models. Even worse, when operating in a continuous learning setting such as reinforcement learning, the model should collapse its epistemic uncertainty as it collects more data. This means that the data has to be set-aside such that a new model could be trained with a smaller dropout probability when the dataset is large enough. This is infeasible in many RL tasks. Instead, the model precision and dropout probability parameters can be optimised with gradient methods, where we seek to minimise some objective w.r.t. to these parameters.

In section 4.6 we specified a heteroscedastic loss which led to the optimisation objective

$$\widehat{\mathcal{L}}_{\text{dropout}}(\mathbf{M}_1, \mathbf{M}_2, \mathbf{b}, \tau) := \frac{1}{M} \sum_{i \in S} E^{\widehat{\mathbf{W}}_1^i, \widehat{\mathbf{W}}_2^i, \mathbf{b}}(\mathbf{x}_i, \mathbf{y}_i) + \lambda_1 ||\mathbf{M}_1||^2 + \lambda_2 ||\mathbf{M}_2||^2 + \lambda_3 ||\mathbf{b}||^2,$$

$$E^{\mathbf{W}_1,\mathbf{W}_2,\mathbf{b}}(\mathbf{x},\mathbf{y}) := \frac{\tau}{2}\left|\left|\mathbf{y} - \mathbf{f}^{\mathbf{W}_1,\mathbf{W}_2,\mathbf{b}}(\mathbf{x})\right|\right|_2^2 - \frac{D}{2}\log\tau$$
$$\lambda_i = \frac{l_i^2(1-p_i)}{2N}$$

with D the output dimensionality, and where we optimise the objective w.r.t. τ as well as the weights to find the model's aleatoric uncertainty. This objective was derived from eq. (3.7) with a Gaussian likelihood with precision τ, and written in the form of eq. (3.9). Compared to eq. (3.9), our optimisation objective is scaled by τ and has an added "regularisation" term $-\log\tau$, a term derived from the likelihood definition which would be omitted when the precision need not be optimised (since τ is constant w.r.t. model weights $\mathbf{W}$).

Similar to this, we can optimise our objective in eq. (3.7) w.r.t. the dropout probability to find the epistemic uncertainty as well. The KL contribution term between the approximate posterior and the prior depends on p through the entropy (eq. (6.6)):

$$\mathcal{H}(p) := -p\log p - (1-p)\log(1-p)$$

which results in an added *dropout regulariser* term to our objective in eq. (3.9). To obtain the KL contribution term between the approximate posterior and the prior *over all weights* we sum the KL per weight row in each weight matrix. The term is then divided by the number of training points in the objective (§3.2.3) to obtain the dropout regulariser in the form of eq. (3.9):

$$\frac{1}{N}\mathrm{KL}(q(\mathbf{W})||p(\mathbf{W})) \propto \frac{l^2(1-p)}{2N}||\mathbf{M}||^2 - \frac{K}{N}\mathcal{H}(p)$$

with K the input dimensionality of the layer and N the number of data points. This regularisation term depends on the dropout probability p alone, which means that the term is constant w.r.t. model weights (thus omitted when the dropout probability is not optimised, but crucial when it is optimised). The entropy of the Bernoulli random variable with probability p pushes the dropout probability towards 0.5—the highest it can attain. The scaling of the regularisation term means that large models will push the dropout probability towards 0.5 much more than smaller models, but as the amount of data increases the dropout probability will be pushed towards 0.

An issue arises when gradient methods are used for the dropout probability optimisation. The Bernoulli distribution which is used in dropout NNs in the expected log

likelihood term (first term in equation (3.7)):

$$\widehat{\mathcal{L}}_{\text{MC}}(\theta) = -\frac{N}{M}\sum_{i\in S}\log p(\mathbf{y}_i|\mathbf{f}^{g(\theta,\boldsymbol{\epsilon})}(\mathbf{x}_i)) + \text{KL}(q_\theta(\boldsymbol{\omega})||p(\boldsymbol{\omega}))$$

depends on p. The Bernoulli distribution is non-differentiable with respect to its parameter p, which means that the pathwise derivative estimator cannot be used with it (forcing us to use the high variance score function estimator). Instead we use the Concrete distribution relaxation [Jang et al., 2016; Maddison et al., 2016] to approximate the Bernoulli distribution when generating the dropout masks. Instead of sampling our dropout masks from the Bernoulli distribution (generating zeros and ones) we sample realisations from the Concrete distribution with temperature $t = 1/10$ which results in masks with values in the interval $[0, 1]$. This distribution concentrates most mass on the boundaries of the interval 0 and 1. In fact, for the one dimensional case here with the Bernoulli distribution, the Concrete distribution relaxation $\tilde{\mathbf{z}}$ of the Bernoulli random variable $\mathbf{z}$ reduces to a simple sigmoid distribution which has a convenient parametrisation:

$$\tilde{\mathbf{z}} = \text{sigmoid}\left(\frac{1}{t}\cdot\Big(\log p - \log(1-p) + \log u - \log(1-u)\Big)\right)$$

with uniform $u \sim \text{Unif}(0, 1)$.

These tools allow us to find both epistemic and aleatoric uncertainties with ease. To assess how different uncertainties behave with different amounts of data, we optimise both the dropout probability p as well as the (per point) model precision τ. We generated synthetic data from the function $y = 2x + 8 + \epsilon$ with $\epsilon \sim \mathcal{N}(0, 1)$ (i.e. corrupting the observations with noise with a fixed standard deviation 1), creating datasets increasing in size ranging from 10 data points (example in figure 6.9) up to $10,000$ data points (example in figure 6.10). We used models with three hidden layers of size 1024 and relu non-linearities, and repeated each experiment three times, averaging the experiments' results. Figure 6.11 shows the epistemic uncertainty (in standard deviation) decreasing as the amount of data increases. This uncertainty was computed by generating multiple function draws and evaluating the functions over a test set generated from the same data distribution. Figure 6.12 shows that the model obtains an increasingly improved estimate to the model precision (aleatoric uncertainty) as more data is given. Finally, figure 6.13 shows the predictive uncertainty obtained by combining the variances of both plots above. This uncertainty seems to converge towards a constant trend.

Lastly, the optimised dropout probabilities corresponding to the various dataset sizes are given in figure 6.14. As can be seen, the optimal dropout probability in each layer decreases as more data is observed, starting from the smallest dataset with near 0.5 probabilities in all layers but the first (input layer, #1), and converging to values ranging between 0.1 and 0.2 when 10,000 data points are given to the model. Further, the dropout probability of the first layer converges to a near-zero value for all data sizes, supporting our observations in §4.2.1. This empirical observation further confirms with our theoretical analysis in section 6.3.

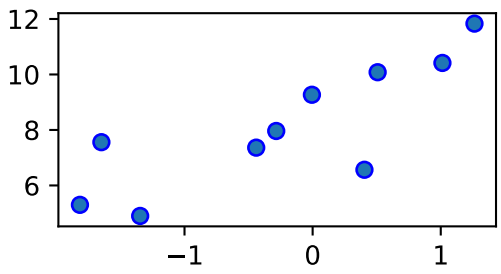


Fig. 6.9 Example dataset with 10 data points.

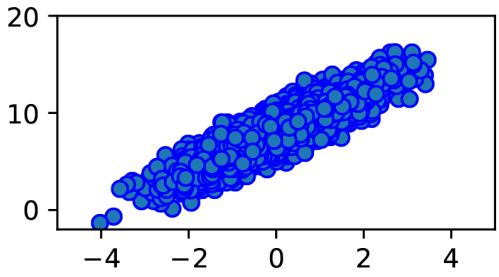


Fig. 6.10 Example dataset with $10,000$ data points.

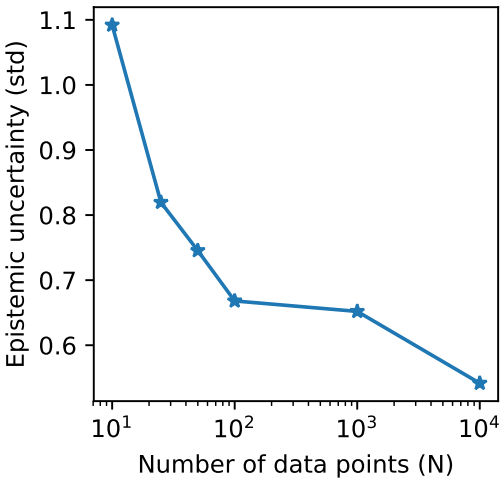


Fig. 6.11 Epistemic uncertainty (in std) as the number of data points increases.

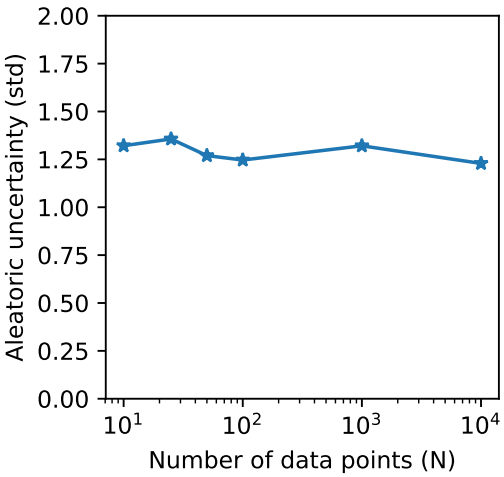


Fig. 6.12 Aleatoric uncertainty (in std) as the number of data points increases.

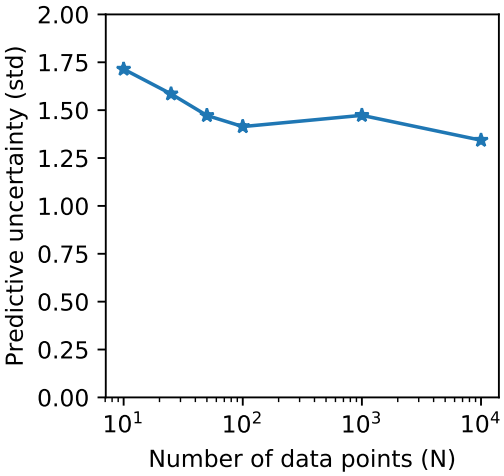


Fig. 6.13 Predictive uncertainty (in std) as the number of data points increases.

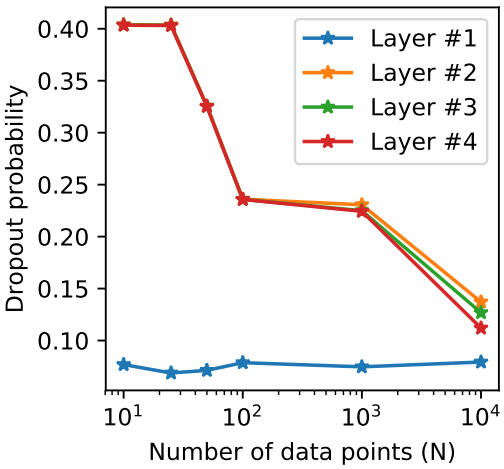


Fig. 6.14 Optimised dropout probability values (per layer) as the number of data points increases.