Data augmentation in Bayesian neural networks and the cold posterior effect (supplementary material)

Seth Nabarro[*1] **Stoil Ganev**[*2] **Adrià Garriga-Alonso**[3]
Vincent Fortuin[3,4] **Mark van der Wilk**[†1] **Laurence Aitchison**[†2]

[1]Department of Computing, Imperial College London
[2]Department of Computer Science, University of Bristol
[3]Department of Engineering, University of Cambridge
[4]Department of Computer Science, ETH Zürich

A AVERAGING LOSSES EMERGES WHEN USING DA IN VI AND SGLD

There are two particularly important algorithms for doing Bayesian inference in neural networks: stochastic gradient Langevin dynamics [SGLD; Welling and Teh, 2011] and variational inference [VI; Blundell et al., 2015]. In SGLD without DA, we draw samples from the posterior over weights by following gradient of the log-probability with added noise,

$$(\Delta \mathbf{w})_{\text{noaug}} = \frac{\epsilon}{2} \nabla_{\mathbf{w}} \left[\log \mathrm{P}\left(\mathbf{w}\right) + \sum_{i=1}^{N} \log \mathrm{P}_{\text{noaug}}\left(y_i | \mathbf{x}_i, \mathbf{w}\right) \right] + \sqrt{\epsilon}\, \boldsymbol{\eta} \tag{22}$$

where $\boldsymbol{\eta}$ is standard Gaussian IID noise, and for simplicity we give the expression for full-batch Langevin dynamics rather than minibatched SGLD (they do not differ for the purposes of reasoning about DA). Likewise the variational inference objective is,

$$\text{ELBO}_{\text{noaug}} = \mathbb{E}_{\mathrm{Q}(\mathbf{w})} \left[\log \mathrm{P}\left(\mathbf{w}\right) + \sum_{i=1}^{N} \log \mathrm{P}_{\text{noaug}}\left(y_i | \mathbf{x}_i, \mathbf{w}\right) - \log \mathrm{Q}\left(\mathbf{w}\right) \right] \tag{23}$$

where $\mathrm{Q}\left(\mathbf{w}\right)$ is the variational approximate posterior learned by optimizing this objective. To understand the effect of the standard approach to DA, we replace $\log \mathrm{P}_{\text{noaug}}\left(y_i | \mathbf{x}_i, \mathbf{w}\right)$ with $\mathcal{L}^i_{\text{loss}}(y_i; \mathbf{w})$. Then, we consider the expected update to the weights, averaging over the augmented images, $\mathbf{x}'_i$ conditioned on the underlying unaugmented images, $\mathbf{x}_i$,

$$\mathbb{E}\left[(\Delta \mathbf{w})_{\text{aug}}\right] = \frac{\epsilon}{2} \nabla_{\mathbf{w}} \left[\log \mathrm{P}\left(\mathbf{w}\right) + \sum_{i=1}^{N} \mathcal{L}^i_{\text{loss}}\left(y_i; \mathbf{w}\right) \right] + \sqrt{\epsilon}\, \boldsymbol{\eta}, \tag{24}$$

$$\text{ELBO}_{\text{aug}} = \mathbb{E}_{\mathrm{Q}(\mathbf{w})} \left[\log \mathrm{P}\left(\mathbf{w}\right) + \sum_{i=1}^{N} \mathcal{L}^i_{\text{loss}}\left(y_i; \mathbf{w}\right) - \log \mathrm{Q}\left(\mathbf{w}\right) \right]. \tag{25}$$

In both cases, this ultimately replaces $\log \mathrm{P}_{\text{noaug}}\left(y_i | \mathbf{x}_i, \mathbf{w}\right)$ with $\mathcal{L}^i_{\text{loss}}\left(y_i; \mathbf{w}\right)$, which as discussed in Sec. 3 is not a valid log-likelihood.

B THE APPROXIMATE POSTERIOR IN VI REDUCES VARIANCE

Here, we derive the ELBO using Jensen's inequality; we take x to be the data and z to be a latent variable. Our goal is to compute the model evidence, $\mathrm{P}\left(x\right)$, by integrating out z,

$$\mathrm{P}\left(x\right) = \int dz\, \mathrm{P}\left(x|z\right) \mathrm{P}\left(z\right) = \int dz\, \mathrm{P}\left(x, z\right) \tag{26}$$

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

where $\mathrm{P}(z)$ is the prior, $\mathrm{P}(x|z)$ is the likelihood and $\mathrm{P}(x,z)$ is the joint. We introduce an approximate posterior, $\mathrm{Q}(z)$, and rewrite the integral as an expectation over that approximate posterior and apply Jensen's inequality,

$$\log \mathrm{P}(x) = \log \int dz\, \mathrm{Q}(z)\, \frac{\mathrm{P}(x,z)}{\mathrm{Q}(z)} \tag{27}$$

$$= \log \mathbb{E}_{\mathrm{Q}(z)} \left[\frac{\mathrm{P}(x,z)}{\mathrm{Q}(z)} \right] \geq \mathbb{E}_{\mathrm{Q}(z)} \left[\log \frac{\mathrm{P}(x,z)}{\mathrm{Q}(z)} \right]. \tag{28}$$

Now it is evident that the tightness of the bound is controlled by the variance of $\mathrm{P}(x,z) \,/\, \mathrm{Q}(z)$. Critically, if $\mathrm{Q}(z)$ matches the true posterior,

$$\mathrm{Q}(z) = \mathrm{P}(z|x) \propto \mathrm{P}(x,z) \tag{29}$$

then $\mathrm{P}(x,z) \,/\, \mathrm{Q}(z)$ is constant (zero variance) and the bound is tight.

C KINETIC DIAGNOSTIC RESULTS

The values of kinetic temperature during inference are plotted in Fig. 5.

D GENERALIZATION OUTSIDE OF CLASSIFICATION

We may be interested in generalizing the averaging logits and averaging probabilities ideas outside classification. For averaging probabilities, we use,

$$\mathrm{P}(y_i|\mathbf{x}_i) = \int d\mathbf{x}'_i\, \mathrm{P}(y_i|\mathbf{x}'_i)\, \mathrm{P}(\mathbf{x}'_i|\mathbf{x}_i) = \mathbb{E}\left[\mathrm{P}(y_i|\mathbf{x}'_i)\right] \tag{30}$$

Intuitively, each augmentation forms one component of a (potentially infinite) mixture model over the outputs, y_i. Importantly, this expression makes no assumption about the support of distributions over y_i, so y_i could be a finite set (classification), real-valued (regression), or anything else (a string, a graph, etc.) Note that directly applying a multi-sample estimator to (the logarithm of) (Eq. 30) gives us a log-likelihood lower bound as in (Eq. 14).

To generalize averaging logits, consider a situation where a distribution over an arbitrary y_i is parameterized by a vector, $\mathbf{f}_i$ output by a neural network,

$$\mathrm{P}(y_i|\mathbf{x}) = \pi(y_i; \mathbf{f}_i). \tag{31}$$

In the standard case with no augmentation, we would take $\mathbf{f}_i = \mathbf{f}(\mathbf{x}_i; \mathbf{w})$, (where we take $\mathbf{f}_i$ as the specific vector for input i, and $\mathbf{f}(\cdot;\cdot)$ as a function represented by a neural network, that takes an image and weights and returns a vector). In the case with augmentation, we can average neural network outputs across different augmentations,

$$\mathbf{f}_i = \int d\mathbf{x}'_i\, \mathrm{P}(\mathbf{x}'_i|\mathbf{x}_i)\, \mathbf{f}(\mathbf{x}'_i; \mathbf{w}) = \mathbb{E}\left[\mathbf{f}(\mathbf{x}'_i; \mathbf{w})\right]. \tag{32}$$

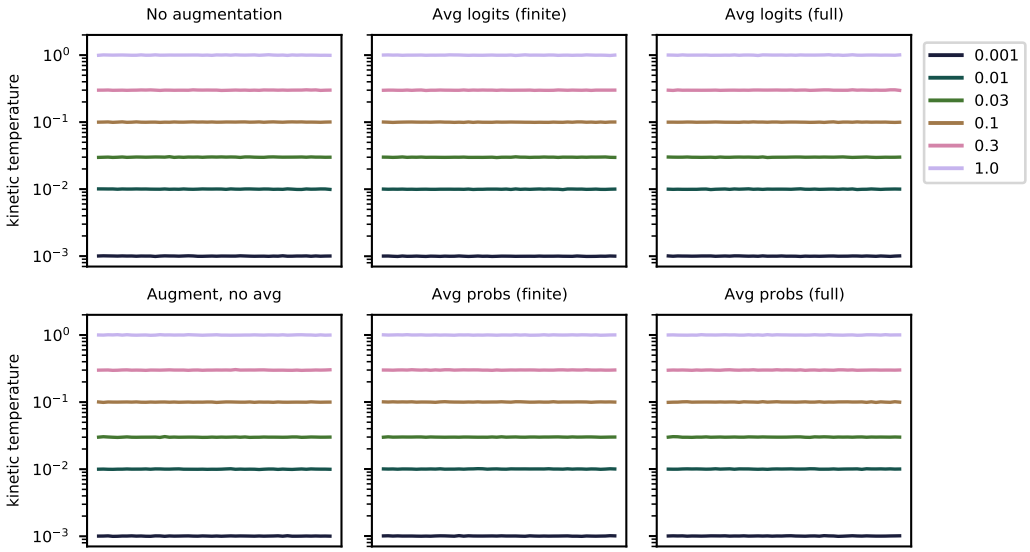
Note that in this case we need additional conditions for the multi-sample estimator to form a lower bound. In particular, we need $\log \pi(y_i; \mathbf{f}_i)$ to be concave when treated as a function of $\mathbf{f}_i$ for a fixed y_i.

E PERSPECTIVES ON PROBABILISTIC DATA AUGMENTATION

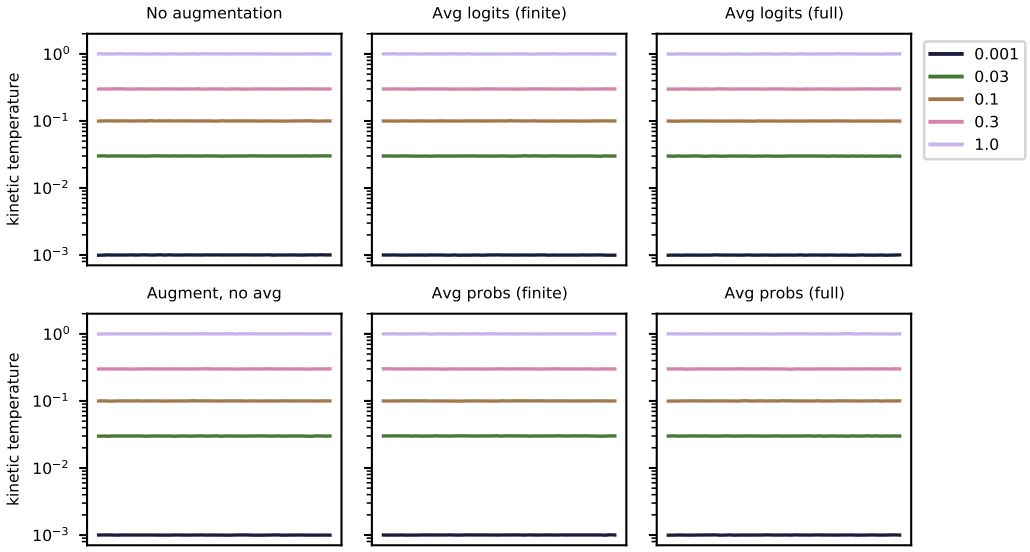
Here we explore in depth the two general approaches to probabilistic data augmentation (Eq. 30) and (Eq. 32). We discuss their justifications in Sec. E.1 and E.2, and compare their properties in Sec. E.3.

E.1 INVARIANCE CONSTRUCTION

In the main text, we suggest two ways of incorporating data augmentation: 1) by averaging logits output by the neural network, and 2) by averaging the predicted probabilities. In classification, both of these methods can be understood as justified by attempting to create a prediction that is more invariant to the transformations in the data augmentation.



(a) MNIST, FCNN.



(b) CIFAR-10, ResNet20.

Figure 5: The evolution of the kinetic temperature diagnostic [Leimkuhler and Matthews, 2015] during inference. Good agreement between the diagnostic temperature and intended temperature (in legend) suggests accurate inference.

When averaging logits, we aim to make the neural network mapping $\mathbf{f} : \mathbb{R}^D \to \mathbb{R}^C$ more invariant by averaging the outputs as in (Eq. 32). This construction influences only the regression function, and so has a similar effect to changing the neural network architecture or changing the prior on the functions $\mathbf{f}(\cdot)$ in the Bayesian case [van der Wilk et al., 2018]. Since only the outputs are affected, this can be directly applied to any likelihood that depends only on an evaluation of the function, i.e. any likelihood which can be written as $\mathrm{P}(y_i|\mathbf{f}_i)$.

In the case of averaging the probabilities, we can consider the model to be learning a mapping from image inputs to probability vectors $\mathbf{p} : \mathbb{R}^D \to \mathbb{P}^{\mathbb{C}}$. We can make this mapping more invariant in the same way:

$$\begin{aligned}\mathbf{p}_{\text{inv}}(\mathbf{x}_i; \mathbf{w}) &= \mathbb{E}\left[\operatorname{softmax}\left(\mathbf{f}(\mathbf{x}'_i; \mathbf{w})\right)\right] \\ &= \int \operatorname{softmax}\left(\mathbf{f}(\mathbf{x}'_i; \mathbf{w})\right) \mathrm{P}(\mathbf{x}'_i|\mathbf{x}_i)\mathrm{d}\mathbf{x}'_i \,. \end{aligned} \tag{33}$$

The straightforward generalization of this construction would be to replace the softmax with the appropriate likelihood (see general case in (Eq. 30)). When considering likelihoods other than softmax classification (e.g. Gaussian likelihoods for regression), stronger differences between these constructions emerge in both behaviour and justification. We investigate further in Appendix E.3.

E.2 NOISY-INPUT MODEL

As stated above, we can generalize averaging the classification probabilities by replacing the softmax with the appropriate likelihood as in (Eq. 30). This modified likelihood, which incorporates data augmentation, was also discussed in Wenzel et al. [2020, Appendix K] and is a (potentially continuous) mixture model on the observation y_n, where each augmentation introduces a mixture component. This is as a *noisy-input* model [Girard and Murray-Smith, 2003, McHutchon and Rasmussen, 2011, Damianou et al., 2016] where the input $\mathbf{x}_i$ is corrupted via the augmentation distribution.

E.3 MODEL COMPARISON

The forms of the invariance construction (Eq. 32) and the noisy-input model (Eq. 30) imply a difference of purpose. In using the invariance construction, we seek a regression function with the specified symmetry, which is consistent with the data according to the likelihood function $\mathrm{P}\left(y_i|\mathbf{f}_i\right)$. Conversely, with the noisy-input model (Eq. 30) we aim to find a function which gives rise to an invariant likelihood, consistent with the observed outputs for inputs randomly perturbed by $\mathrm{P}(\mathbf{x}'|\mathbf{x})$. The role of $\mathbf{x}'$ is different in each case. In the noisy-input model, $\mathbf{x}'$ is a latent variable on which we could, in principle, do inference (with e.g. an amortized variational approach). While in the invariance construction, we integrate over $\mathbf{x}'$ to parameterize $\mathbf{f}(\mathbf{x}; \mathbf{w})$.

We now compare the behaviours of the invariance and noisy-input constructions. We will see that they result in quite different posteriors.

In the main text, we compared the empirical performance of averaging probabilities and averaging logits for BNN classification (see Figs. 2 and 4). However, as the invariance perspective justifies both averaging logits and probabilities, this comparison does not clearly distinguish between the noisy-input and invariance viewpoints. Further, we are interested not only in predictive performance but also in understanding how each construction behaves. With this in mind, we investigate the models with an illustrative example, where we can both integrate over the orbit and do inference in closed form.

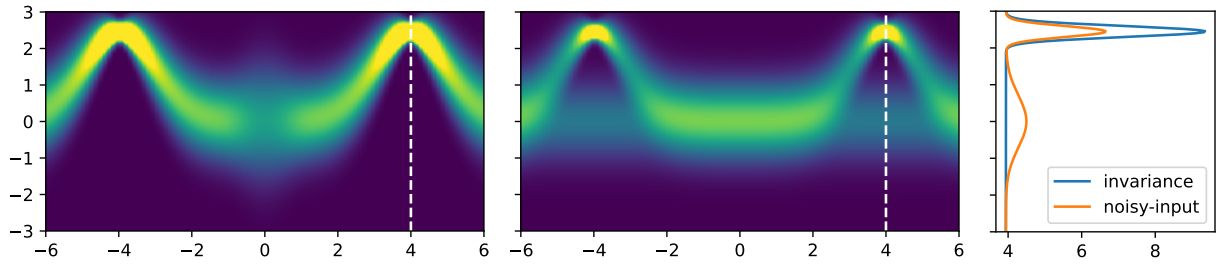
We consider Gaussian process (GP) regression with a one-dimensional input and data augmentation which enforces symmetry about $x = 0$, i.e. $\mathrm{P}(x'|x) = \frac{1}{2}\left(\delta(x' - x) + \delta(x' + x)\right)$. From Van der Wilk et al. [2018], the invariance view can be expressed in the kernel of the GP:

$$g \sim \mathcal{GP}(\mathbf{0}, k_{\text{base}}) \tag{34}$$

$$f(x) = g(x) + g(-x) \tag{35}$$

$$\implies f \sim \mathcal{GP}(\mathbf{0}, k_{\text{inv}}), \tag{36}$$

$$\text{where} \quad k_{\text{inv}}(x_i, x_j) = \sum_{c_i \in \{-1,1\}} \sum_{c_j \in \{-1,1\}} k_{\text{base}}(c_i x_i, c_j x_j). \tag{37}$$



(a) Invariant GP (b) Noisy-input GP (c) $p(y|x = 4, \mathcal{D})$

Figure 6: Posterior densities for the model constructions for a single observation at $x_1 = -4, y_1 = 2.5$.

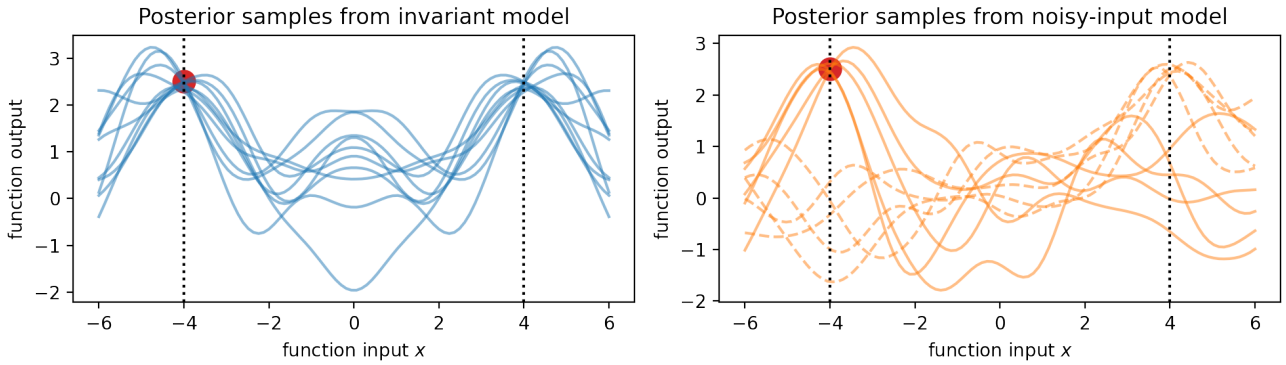


Figure 7: Samples from the model posteriors. The red dot marks the location of the observation $(x_1 = -4, y_1 = 2.5)$. The noisy input posterior comprises two components: one conditioned on (x_1, y_1) (dashed lines), the other on $(-x_1, y_1)$ (solid lines).

We then follow standard GP inference to find the posterior over invariant functions. Note that unlike Van der Wilk et al. [2018], we are not concerned with learning invariances here.

The noisy-input model for this case is

$$\mathrm{P}(\mathbf{x}, \mathbf{y}, \mathbf{f}) = \mathrm{P}(\mathbf{f}) \prod_{i=1}^{N} \int \mathrm{P}(y_i|f(x_i')) \, \mathrm{P}(x_i'|x_i) dx_i' \tag{38}$$

$$\mathrm{P}(y_i|f(x_i')) = \mathcal{N}\left(y_i; f(x_i'), \sigma^2\right) \tag{39}$$

$$f \sim \mathcal{GP}(0, k). \tag{40}$$

Given a single observation (x_1, y_1), the noisy-input posterior is

$$\mathrm{P}(f|x_1, y_1) = \frac{1}{Z} \mathrm{P}(f(x_1), x_1, y_1) \tag{41}$$

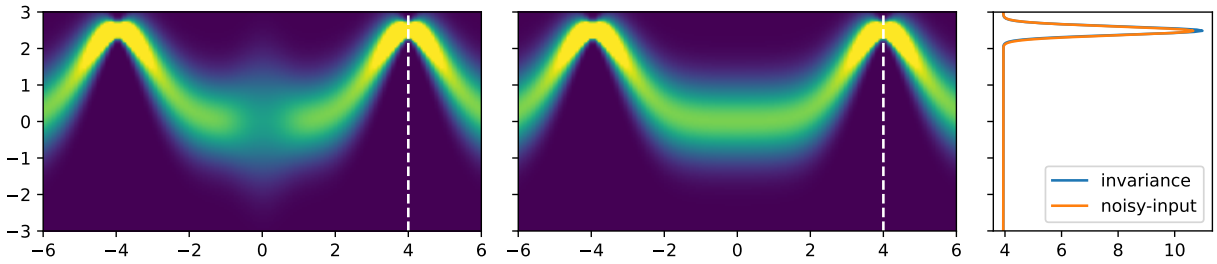
$$= \frac{1}{2Z} \mathrm{P}\left(f(x_1)\right) \left[\mathrm{P}(y_1|f(x_1)) + \mathrm{P}(y_1|f(-x_1))\right] \tag{42}$$

$$= \frac{1}{2} \left[\mathrm{P}(f|x_1, y_1) + \mathrm{P}(f|-x_1, y_1)\right], \tag{43}$$

a mixture of GP posteriors, with two components (one for each point in the orbit).

How do these posteriors compare? For an observation at $(x_1 = -4, y_1 = 2.5)$ we plot the posterior predictive densities in Fig. 6. Both posteriors are symmetric around $x = 0$ as we expect, however the noisy-input model is bimodal in the regions surrounding $x = 4$ and $x = -4$, where the invariance posterior has unimodal density concentrated around the observed y value of 2.5. The difference is clear in Fig. 6c, which shows the marginal predictive densities at $x = 4$.

In the noisy-input case, our observation is (x_1, y_1), but x is uncertain, so the observation could have been generated by $(-x_1, y_1)$ with equal probability. This results in a mixture posterior with two components: one component has "seen"



(a) Invariant GP (b) Noisy-input GP (c) $p(y|x = 4, \mathcal{D})$

Figure 8: Posterior densities for the model constructions for ten observations at $\{(x_i = -4, y_i = 2.5)\}_{i=1}^{10}$.

(x_1, y_1), while the other "saw" $(-x_1, y_1)$. The first component's prediction at $-x_1$ remains uninformed by its "observation" and the same is true for the second component's prediction at x_1. Thus, the predictions made by these components at these locations revert to the zero-mean prior.

From the invariance perspective, we condition on the point (x_1, y_1) but the double-sum kernel forces the function to be the same at $(-x_1, y_1)$. As the posterior is a single GP, it has unimodal marginals with high density around both points.

We can gain further intuition by looking at samples from both posteriors (Fig. 7). We can see that the *every* sample from the invariance posterior is symmetric about $x = 0$, where the functions drawn from the noisy input posterior are not symmetric in general.

The samples illustrate the key difference between the models. For the noisy input model, we can see the two components of the mixture posterior arise from conditioning on different locations in the orbit of x_1 as described above. The component going through $(x = 4, y_1)$ (samples drawn with dashed lines) is close to the prior at $(x = -4, y_1)$, the other (solid lines) goes through $(x = -4, y_1)$ and is close to the prior at $(x = 4, y_1)$. However, under the invariance model, inference on the observation concentrates all model density around y_1 for both points in the orbit of x_1.

We now consider how this comparison changes as we observe more data. The noisy-input model (Eq. 38) requires integration over $\mathrm{P}(x'|x)$ to compute the likelihood of each datapoint, all of which are multiplied together to calculate their combined likelihood. Thus, the number of posterior components grows exponentially with the number of observations: A^N (for orbit size A). Suppose all observations are at the same location (x_1, y_1). In this case, the posterior density due to prior reversion at $\{x_1, -x_1\}$ decreases exponentially with N. This is because the fraction of mixture components conditioned on all input observations being at the same point in the orbit of x_1, i.e. all at x_1 or $-x_1$, is given by A^{1-N}. The predictive posteriors for ten observations, each at $(x = -4, y = 2.5)$ is shown in Fig. 8. Contrasting this noisy-input posterior (Fig. 8b) to that for one observation (Fig. 6b), we can see the reduction in density around to the prior mean for points around the orbit of x_1. In summary, the noisy-input and invariance posteriors become more alike as we observe more data in the same orbit.