Evaluating Bayesian Theories of Generalization

Daniel J. Navarro
School of Psychology
University of Adelaide

January 29, 2008

In this note, I present more detailed derivations and technical remarks associated with the Bayesian generalization model outlined in

> Navarro, D. J., Lee, M. D., Dry, M. J. & Schultz, B. (submitted). Extending and testing the Bayesian theory of generalization. Submitted to the *Proceedings of the 30th Annual Conference of the Cognitive Science Society*

It is intended solely as a companion to the original manuscript, so the reader interested in the motivation behind the work, the relevant references or the experiment that tests the model is referred to the original paper.

Analytic Expressions for the Generalization Model

The Bayesian theory of generalization assumes, as per Shepard's original analysis, that there exists some unknown, connected region r, such that stimuli falling inside r will possess the property, and stimuli outside r will not. People are assumed to have some prior beliefs about r, and then report a value such as $p(y \in r|x_1, \ldots, x_n)$, the probability that novel item y lies inside the region, given that the previously observed items $x_1, \ldots, x_n$ are known to fall inside the region. In order to determine this probability, the learner relies on some assumptions about the manner in which the old items x_i were generated. Two proposals exist in the literature, known as strong and weak sampling respectively. Under *weak sampling* (Shepard's original proposal), the method by which x_i is generated is assumed not to depend on the region itself, so $p(x_i|r) = 1$ if $x_i \in r$, and $p(x_i|r) = 0$ otherwise. In contrast, Tenenbaum and Griffiths propose that the observations may be assumed to be explicitly sampled from the region (with uniform probability density on r), implying that $p(x_i|r) = 1/|r|$ if $x_i \in r$ (where $|r|$ denotes the size of the region), and $p(x_i|r) = 0$ otherwise. This latter proposal is referred to as *strong sampling*.

General Expressions

Irrespective of the sampling method, the Bayesian theory implies that the probability that $y \in r$ is

$$p(y \in r|x_1, \ldots, x_n) = \int_{\mathcal{R}} p(y \in r')p(r = r'|x_1, \ldots, x_n)\ dr' \tag{1}$$

where r' denotes one possibility as to the identity of the unknown region r, and the integration is taken over $\mathcal{R}$, the set of all such regions. Noting that $p(y \in r')$ is simple an indicator function that equals 1 if y falls inside r' and zero if it does not, we can expand the integral, and obtain:

$$\begin{aligned} p(y \in r|x_1, \ldots, x_n) &= \int_{\mathcal{R}} p(y \in r')p(r'|x_1, \ldots, x_n)\ dr' \\ &= \int_{\mathcal{R}_y} p(r'|x_1, \ldots, x_n)\ dr' \\ &= \int_{\mathcal{R}_y} \frac{p(x_1, \ldots, x_n|r')p(r')}{p(x_1, \ldots, x_n)}\ dr' \\ &= \frac{\int_{\mathcal{R}_y} p(x_1, \ldots, x_n|r')p(r')\ dr'}{\int_{\mathcal{R}} p(x_1, \ldots, x_n|r')p(r')\ dr'} \end{aligned} \tag{2}$$

where $\mathcal{R}_y \subset \mathcal{R}$ denotes the set of possible regions that include y. For the current purposes we assume a uniform prior over regions, $p(r') \propto 1$, and we restrict ourselves to the simple case in which stimuli vary only along a single dimension. Moreover, since the tasks that we consider explicitly provide upper and lower bounds for the possible regions, we may assume without loss of generality that $x_i \in [0, 1]$ for all i, that $y \in [0, 1]$, and that $\mathcal{R}$ is the set of all connected subsets of the interval $[0, 1]$ (i.e., all intervals $[a, b]$ such that $0 \leq a \leq b \leq 1$). Letting $z_l = \min(x_1, \ldots, x_n)$ and $z_u = \max(x_1, \ldots, x_n)$, it is straightforward to see that in general, the denominator in Eq. 2 is given by

$$\int_{\mathcal{R}} p(x_1, \ldots, x_n|r')p(r')\ dr' = \int_0^{z_l} \int_{z_u}^1 p(x_1, \ldots, x_n|r' = (l, u))\ du\ dl \tag{3}$$

Similarly, the numerator in Eq. 2 is given by

$$\int_{\mathcal{R}_y} p(x_1, \ldots, x_n|r')p(r')\ dr' = \int_0^{\min(y, z_l)} \int_{\max(y, z_u)}^1 p(x_1, \ldots, x_n|r' = (l, u))\ du\ dl \tag{4}$$

Since $\mathcal{R}$ consists only of intervals, when $z_l \leq y \leq z_u$, $p(x_1, \ldots, x_n|r') = 0$ for all $r' \in \mathcal{R} - \mathcal{R}_y$, and so the integrals in Eqs 3 and 4 are identical. Accordingly, when we substitute into Eq. 2, over this range $p(y \in r|x_1, \ldots, x_n) = 1$ irrespective of sampling scheme. However, for other cases (i.e., $y < z_l$ or $z_u < y$) the integrals must be explicitly evaluated. Since the integrals behave differently for the two cases (strong and weak sampling), we deal with them separately. We begin with the weak sampling case, since it is the simpler of the two.

The Weak Sampling Case

For weak sampling, the denominator in Eq. 2 is simply

$$\int_{\mathcal{R}} p(x_1, \ldots, x_n|r')\ dr' = \int_0^{z_l} \int_{z_u}^1 1\ du\ dl = (1 - z_u)z_l. \tag{5}$$

where u denotes the upper bound on r' and l is the lower bound. The numerator in those cases where $y < z_l$ or $z_u < y$ becomes

$$\int_{\mathcal{R}_y} p(x_1, \ldots, x_n|r')\ dr' = \int_0^{y} \int_{z_u}^1 1\ du\ dl = (1 - z_u)y \quad \text{if } y < z_l, \tag{6}$$

$$\int_{\mathcal{R}_y} p(x_1, \ldots, x_n|r')\ dr' = \int_0^{z_l} \int_{y}^1 1\ du\ dl = (1 - y)z_l \quad \text{if } y > z_u. \tag{7}$$

Thus, when we substitute into Eq. 2, under weak sampling the Bayesian theory predicts that the generalization gradients are linear:

$$p(y \in r|x_1, \ldots, x_n) = \begin{cases} y/z_l & \text{if} \quad y < z_l \\ 1 & \text{if} \quad z_l \leq y \leq z_u \\ (1 - y)/(1 - z_u) & \text{if} \quad z_u < y \end{cases}. \tag{8}$$

The Strong Sampling Case

Under strong sampling, the story is slightly more complex. For the moment, let $n > 2$ (the cases for $n = 1$ and $n = 2$ need to be dealt with separately). Under strong sampling, each observation is generated independently from a uniform distribution on r, so

$$\begin{aligned} \int_{\mathcal{R}} p(x_1, \ldots, x_n|r')\ dr' &= \int_0^{z_l} \int_{z_u}^1 p(x_1, \ldots, x_n|r' = (l, u))\ du\ dl \\ &= \int_0^{z_l} \int_{z_u}^1 (u - l)^{-n}\ du\ dl \\ &= \int_0^{z_l} \left[(1 - n)^{-1}(u - l)^{1-n}\right]_{z_u}^1\ dl \\ &= (1 - n)^{-1} \int_0^{z_l} (1 - l)^{1-n} - (z_u - l)^{1-n}\ dl \\ &= (1 - n)^{-1} \left[-(2 - n)^{-1}(1 - l)^{2-n} + (2 - n)^{-1}(z_u - l)^{2-n}\right]_0^{z_l} \\ &= (1 - n)^{-1}(2 - n)^{-1}(1 + (z_u - z_l)^{2-n} - (1 - z_l)^{2-n} - {z_u}^{2-n}) \end{aligned} \tag{9}$$

The integral over $\mathcal{R}_y$ has the same form, but with the outer integral taken from 0 to y when $y < z_l$, and the inner integral taken from y to 1 when $y > z_u$. Thus, by substitution into Eq. 2, it is straightforward to note that when $n > 2$, the Bayesian theory with strong

sampling predicts that

$$p(y \in r|x_1,\ldots,x_n) = \begin{cases} \dfrac{1+(z_u-y)^{2-n}-(1-y)^{2-n}-{z_u}^{2-n}}{1+(z_u-z_l)^{2-n}-(1-z_l)^{2-n}-{z_u}^{2-n}} & \text{if} \quad y < z_l \\ 1 & \text{if} \quad z_l \leq y \leq z_u \\ \dfrac{1+(y-z_l)^{2-n}-(1-z_l)^{2-n}-y^{2-n}}{1+(z_u-z_l)^{2-n}-(1-z_l)^{2-n}-{z_u}^{2-n}} & \text{if} \quad z_u < y \end{cases} \tag{10}$$

In the case where $n = 1$ we observe that,

$$\begin{aligned} \int_{\mathcal{R}} p(x_1|r')\ dr' &= \int_0^{z_l}\int_{z_u}^1 p(x_1|r'=(l,u))\ du\ dl \\ &= \int_0^{z_l}\int_{z_u}^1 (u-l)^{-1}\ du\ dl \\ &= \int_0^{z_l} [\ln(u-l)]_{z_u}^1\ dl \\ &= \int_0^{z_l} \ln(1-l)-\ln(z_u-l)\ dl \\ &= [(l-1)\ln(1-l)-l]_0^{z_l} - [(l-z_u)\ln(z_u-l)-l]_0^{z_l} \\ &= ((z_l-1)\ln(1-z_l)-z_l) - ((z_l-z_u)\ln(z_u-z_l)-z_l+z_u\ln z_u) \\ &= (z_u-z_l)\ln(z_u-z_l)-(1-z_l)\ln(1-z_l)-z_u\ln z_u \end{aligned} \tag{11}$$

Applying the same procedure as before yields the expression

$$p(y \in r|x_1) = \begin{cases} \dfrac{(z_u-y)\ln(z_u-y)-(1-y)\ln(1-y)-z_u\ln z_u}{(z_u-z_l)\ln(z_u-z_l)-(1-z_l)\ln(1-z_l)-z_u\ln z_u} & \text{if} \quad y < z_l \\ 1 & \text{if} \quad z_l \leq y \leq z_u \\ \dfrac{(y-z_l)\ln(y-z_l)-(1-z_l)\ln(1-z_l)-y\ln y}{(z_u-z_l)\ln(z_u-z_l)-(1-z_l)\ln(1-z_l)-z_u\ln z_u} & \text{if} \quad z_u < y \end{cases} \tag{12}$$

In this case, however, the expression can be further simplified since $z_l = z_u = x_1$:

$$p(y \in r|x_1) = \begin{cases} \dfrac{(1-y)\ln(1-y)+x_1\ln x_1-(x-y)\ln(x_1-y)}{(1-x_1)\ln(1-x_1)+x_1\ln x_1} & \text{if} \quad y < x_1 \\ 1 & \text{if} \quad y = x_1 \\ \dfrac{(1-x_1)\ln(1-x_1)+y\ln y-(y-x_1)\ln(y-x_1)}{(1-x_1)\ln(1-x_1)+x_1\ln x_1} & \text{if} \quad x_1 < y \end{cases} \tag{13}$$

(Obviously, this expression could be derived directly, rather than found as a limiting case of the "z_l, z_u" formulation, but the more general version is useful for other purposes, which is why we derive it this way). Turning to the case where $n = 2$,

$$\int_{\mathcal{R}} p(x_1,x_2|r')\ dr' = \int_0^{z_l}\int_{z_u}^1 p(x_1,x_2|r'=(l,u))\ du\ dl$$

$$
\begin{aligned}
&= \int_0^{z_l} \int_{z_u}^1 (u-l)^{-2}\, du\, dl \\
&= \int_0^{z_l} \left[-(u-l)^{-1}\right]_{z_u}^1 \, dl \\
&= -\int_0^{z_l} (1-l)^{-1} - (z_u - l)^{-1}\, dl \\
&= [\ln(1-l)]_0^{z_l} - [\ln(z_u - l)]_0^{z_l} \\
&= \ln(1-z_l) + \ln z_u - \ln(z_u - z_l) \qquad (14)
\end{aligned}
$$

Thus,

$$
p(y \in r | x_1, x_2) = \begin{cases} \dfrac{\ln(1-y) + \ln z_u - \ln(z_u - y)}{\ln(1-z_l) + \ln z_u - \ln(z_u - z_l)} & \text{if} \quad y < z_l \\ 1 & \text{if} \quad z_l \leq y \leq z_u \\ \dfrac{\ln(1-z_l) + \ln y - \ln(y - z_l)}{\ln(1-z_l) + \ln z_u - \ln(z_u - z_l)} & \text{if} \quad z_u < y \end{cases} \qquad (15)
$$

A More General Sampling Scheme

Arguably, strong and weak sampling are best viewed as two end points on a continuum: at one end the training items are sampled in a way that is completely dependent on the region itself, whereas at the other end observations are completely independent of the consequence at hand. However, in many realistic scenarios our observations arrive in a manner that is only partially correlated with the phenomenon in which we are interested. As a simple example, consider the sampling process involved when one is trying to guess whether a patient in a doctor's office is sick. Not everyone who enters the office is in fact sick, so strong sampling is impossible. However, people who are seeking treatment *are* more likely to be sick than randomly chosen people, so weak sampling seems inappropriate too. In short, a more general approach is necessary.

Perhaps the simplest scheme that satisfies this criterion is a *mixed sampling* approach; with probability θ, items are sampled from the region in question, but with probability $(1-\theta)$ they are generated independently of the region, coming from a uniform distribution on the whole space. Obviously, this is considerably simpler than the kind of thing that presumably underlies realistic sampling schemes such as the one that underlies the doctor-clinic problem. Specifically, the non-region samples are likely to be correlated with the region in the doctor's case: the proposed scheme is essentially a "sick person or random person" sampler, which is clearly unrealistic, but is at least considerably more plausible than purely strong or purely weak methods. In any case, the probability of sampling item x_i such that $x_i \in r$ is

$$
p(x_i, x_i \in r | r, \theta) = (1-\theta) + \theta |r|^{-1}. \qquad (16)
$$

However, if $x_i \notin r$, then

$$
p(x_i, x_i \notin r | r, \theta) = 1 - \theta. \qquad (17)
$$

So, letting $X = (x_1, \ldots, x_n)$, we obtain the expression

$$\begin{aligned} p(X, X \in r | r = (u,l), \theta) &= \prod_{k=1}^{n} (1-\theta) + \theta(u-l)^{-1} \\ &= \left((1-\theta) + \theta(u-l)^{-1}\right)^n \\ &= \sum_{k=0}^{n} \binom{n}{k} (1-\theta)^k \theta^{n-k} (u-l)^{n-k} \end{aligned} \quad (18)$$

Having written the posterior probability in this form, it is simple to note that

$$\int_{\mathcal{R}} p(X, X \in r | r = r', \theta)\ dr' = \sum_{k=0}^{n} \binom{n}{k} (1-\theta)^k \theta^{n-k} \int_0^{z_l} \int_{z_u}^{1} (u-l)^{n-k}\ du\ dl \quad (19)$$

and hence

$$p(y \in r | X, X \in r, \theta) = \frac{\sum_{k=0}^{n} \binom{n}{k} (1-\theta)^k \theta^{n-k} \int_0^{\min(y,z_l)} \int_{\max(y,z_u)}^{1} (u-l)^{n-k}\ du\ dl}{\sum_{k=0}^{n} \binom{n}{k} (1-\theta)^k \theta^{n-k} \int_0^{z_l} \int_{z_u}^{1} (u-l)^{n-k}\ du\ dl} \quad (20)$$

where the integrals in this expression are identical to those solved in the previous section. When interpreting this class of sampling models, it is important to recognize that the goal is to consider a class of sampling assumptions (indexed by θ) that allows the learner to vary the dependence of the sampling on r in a straightforward fashion. This by no means covers the full range of likelihood functions that one might consider, but it is sufficient for the present purposes. Having obtained these general expressions, it is straightforward to plot the effect of varying θ on the generalization gradient, as shown in Figure 1.

More General Priors

In the previous sections, we assumed a uniform prior distribution over regions, such that $p(r) = p(u,l) \propto 1$. However, since people are likely to have different intuitions *a priori* as to the nature of the region, it seems likely that a broader class of priors may be of some use. For simplicity, we will assume that the prior is *location invariant*[1]. Noting that we can reparameterize the region in terms of a centre $c = (u+l)/2$ and a size $s = u - l$, we write the location-invariant prior as $p(c,s) \propto p(s)$, and hence we may say that $p(r) \propto p(u-l)$. In reality, people may have location preferences, but it seems unlikely to matter a great deal in the current context, since the experimental design leaves little room for it to matter. Prior beliefs about size, however, can matter a great deal, since these will determine the steepness of the gradient (and can make the gradients concave if necessary). Noting that the experiment is such that the end points are explicitly made

[1]Note that a uniform prior over locations *does not* imply symmetric or location-invariant generalization gradients, since a shift in location alters the information provided by the known edge points. If, however, we took the limiting case where the edgepoints are arbitrarily distant, then symmetry and location-invariance would hold for the gradients.

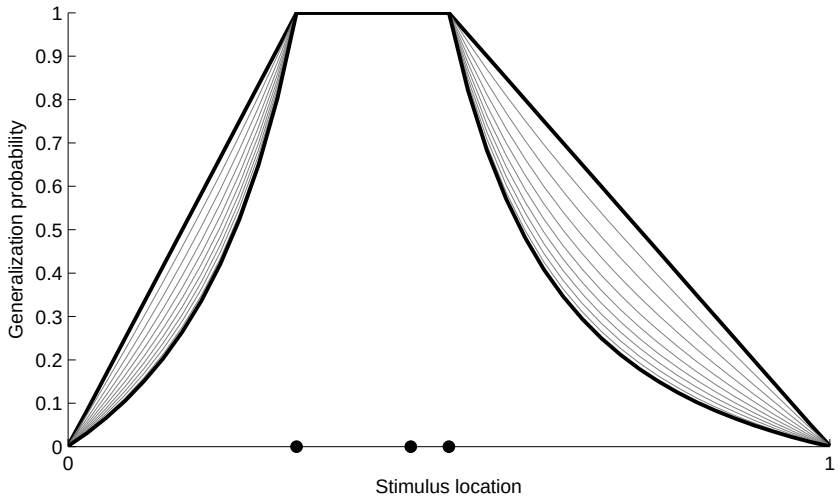

Figure 1: The effect of varying θ, for a case involving three training items (black dots). When $\theta = 0$ (weak sampling), we obtain a linear interpolation model (the uppermost black curve), whereas when $\theta = 1$ (strong sampling), we obtain the tightest generalization gradients (the lowest black curve). Varying θ in increments of 0.1 produces the various intermediate gradients shown with the grey curves.

clear to participants, we have a prior on sizes over the range $(0, 1)$. Again, a very broad range of priors are possible over this range, but for the current purposes we will restrict ourselves to the one-parameter Beta$(1, \phi)$ family, in which $p(u - l) = \phi(u - l)^{\phi-1}$, for $u - l \in (0, 1)$. This family has the nice property that it has much the same form as the likelihood (allowing ϕ to be interpreted as pseudo-data), and allows the integrals to be solved simply. In terms of the kinds of priors it encompasses, it allows people to have prior biases towards large regions ($\phi > 1$), prior biases towards small regions ($\phi < 1$), or no preference at all ($\phi = 1$). However, it does not allow a prior preference for "medium sized" regions. The full range of possible prior preferences allowed under this prior is shown in Figure 2. In any case, noting that constant of proportionality ϕ vanishes since it appears in every term in the numerator and denominator when we substitute Eq. 20 into Eq. 2. Accordingly, if we let

$$f(w, a, b) = \int_0^a \int_b^1 (u - l)^{-w} \, du \, dl \tag{21}$$

denote the basic integrals in the model for $w > 0$, $0 \leq a \leq b \leq 1$, and let

$$b(n, k, \theta) = \binom{n}{k} (1 - \theta)^k \theta^{n-k} \tag{22}$$

denote the binomial probability for k successes out of n trials with success rate parameter θ, then we may write the "complete" model predictions for the two-parameter generaliza-

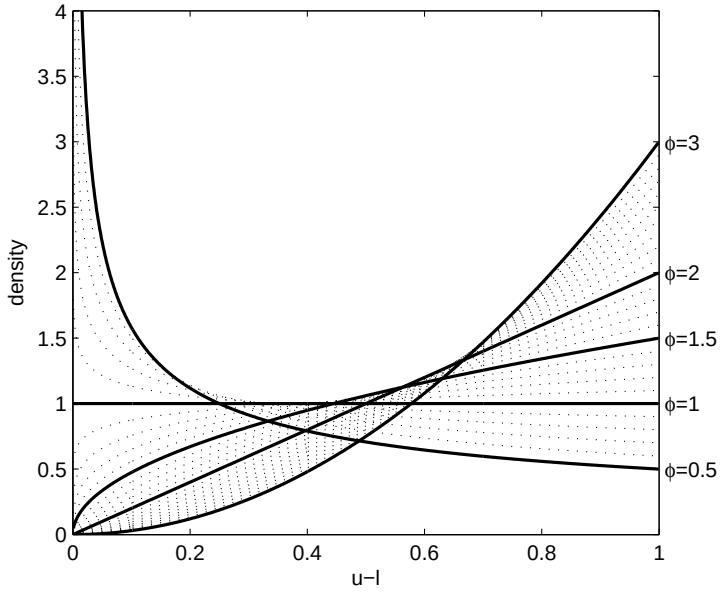

Figure 2: The Beta(1,ϕ) family of priors over possible region sizes, shown over the range $0.5 \leq \phi \leq 3$. Note that while the prior is flexible enough to capture some kinds of preferences that we might expect, it is still extremely constrained. In particular, preferences for medium sized regions, or bimodal preferences concentrated at 0 and 1 also make sense, suggesting that the full Beta family may be more plausible. However, since the full family is less tractable analytically, we restrict ourselves to the simpler case for now.

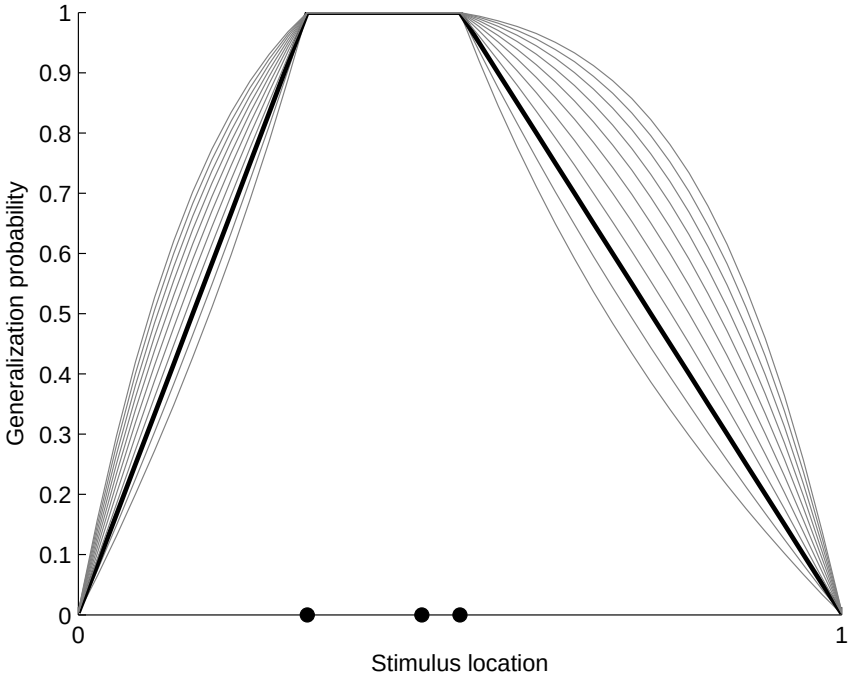

Figure 3: The effect of varying ϕ, when the sampling model is weak $\theta = 0$ for a case involving three training items (black dots). When $\phi = 1$ (the black curve), the standard linear interpolation function for weak sampling is obtained. When $\phi < 1$, the gradients become convex and dip below the linear one, whereas when $\phi > 1$ the gradients become concave.

tion model as follows:

$$p(y \in r|X, X \in r, \theta, \phi) \quad = \quad \frac{\sum_{k=0}^{n} b(n,k,\theta) f(n-k-\phi+1, \min(y, z_l), \max(y, z_u))}{\sum_{k=0}^{n} b(n,k,\theta) f(n-k-\phi+1, z_l, z_u)} \quad (23)$$

The effect of allowing a range of priors is illustrated in Figures 3 and 4. In both cases, ϕ varies from 0 to 5 in increments of 0.5. However, in Figure 3 a weak sampling model is used, and in Figure 4 a strong sampling model is used. Note that, as is always the case for Bayesian models, varying the prior has an effect not dissimilar to varying the likelihood (the sampling model). Accordingly, careful experimental design is required to discriminate between the two effects. Moreover (and again, obvious in hindsight), when there is a prior expectation that the region will be large (i.e., $\phi > 1$), the generalization gradients can in fact be *concave*.

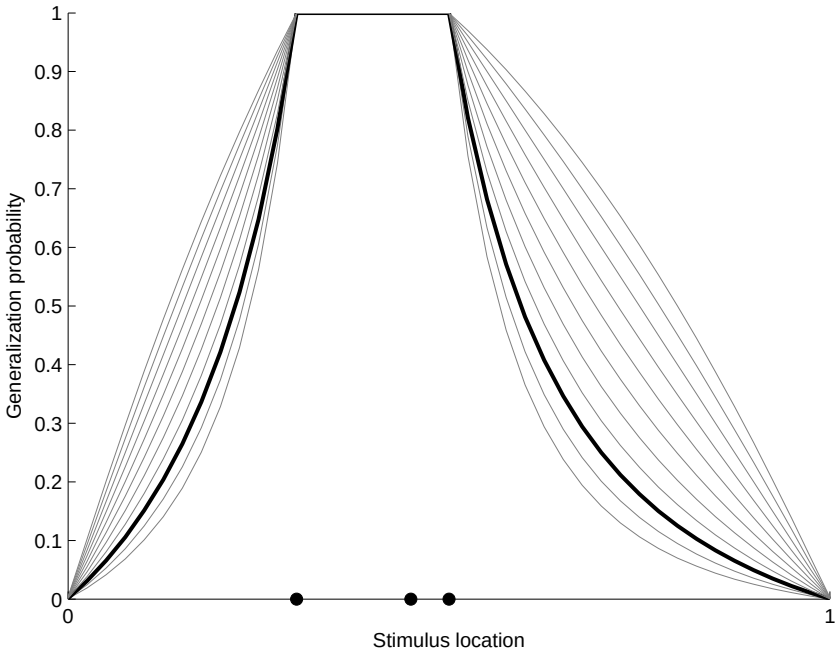

Figure 4: The effect of varying ϕ, when the sampling model is strong $\theta = 1$ for a case involving three training items (black dots). When $\phi = 1$ (the black curve), the tight convex gradients appear. When $\phi < 1$, the gradients are tighter, whereas when $\phi > 1$ they are less tight. In fact, when $\phi = n$ ($= 3$ in this case), the prior exactly cancels out the effect of sample size induced by the strong sampling model, yielding a linear model. When $\phi > n$, the gradients become concave.

Auxiliary Topics

In this section, I expand on some of the technical aspects to the model development in the original paper that were omitted, either due to space considerations or because they are of little interest except insofar as they are needed to fully specify the model.

Miscalibration in Probability Judgment

The main issue at hand is that participants' responses need not exactly correspond to the idealized probabilities *even if* people might generally endorse the kinds of generalization gradients that the Bayesian models imply (we are concerned here with the beliefs people hold about unknown concepts, not the manner in which people translate beliefs into probability statements). Indeed, the nature of understanding how people report degrees of endorsement or confidence that some proposition holds is itself a fairly complicated field of study. To a first approximation, however, we might assume that Bayesian generalization probabilities may be accorded a status not dissimilar to an "objective class-inclusion probability", and treat participants' responses as reflecting some "subjective confidence of class-inclusion"; implying that the relationship between the two should bear a strong similarity to confidence calibration curves, which appear to be approximately linear functions (see references in original paper). With this in mind, a natural way to link the Bayesian predictions $p(y \in r|x_1, \ldots, x_n)$ about y to typical subjective judgments $\tilde{p}(y)$ is to assume a linear relationship:

$$\tilde{p}(y) = j_l + (j_u - j_l)p(y \in r|x_1, \ldots, x_n) \tag{24}$$

where the function is parameterized by j_u and j_l, the upper and lower bounds on values that the participant is willing to report when making probability judgments.

Error Models and Contaminant Distributions

At this point, we have a fairly plausible method for describing the response that the model would predict a participant is most likely to provide. However, since judgments are noisy for no shortage of reasons, we require an explicit error theory in order to draw conclusions about the plausibility of any particular generalization function. While it would be possible to fit the model using standard OLS methods, which rely on the assumptions of normality and homoscedasticity of errors, it would be somewhat unwise to do so in this case since probability judgments are constrainted to lie on the interval [0,1], and the model predictions do reach the endpoints, which inevitably introduces skew and heteroscedasticity. In this case, the Beta distribution is likely to provide a better account of the errors. If we assume that the dependent variable d has probability density

$$d|\tilde{p} \sim \text{Beta}(1 + \tilde{p}\tau, 1 + (1 - \tilde{p})\tau)$$

then we obtain error models of the kind shown in Figure 5. It should be noted, however, that on occasions people provide somewhat arbitrary responses, either due to inattention,

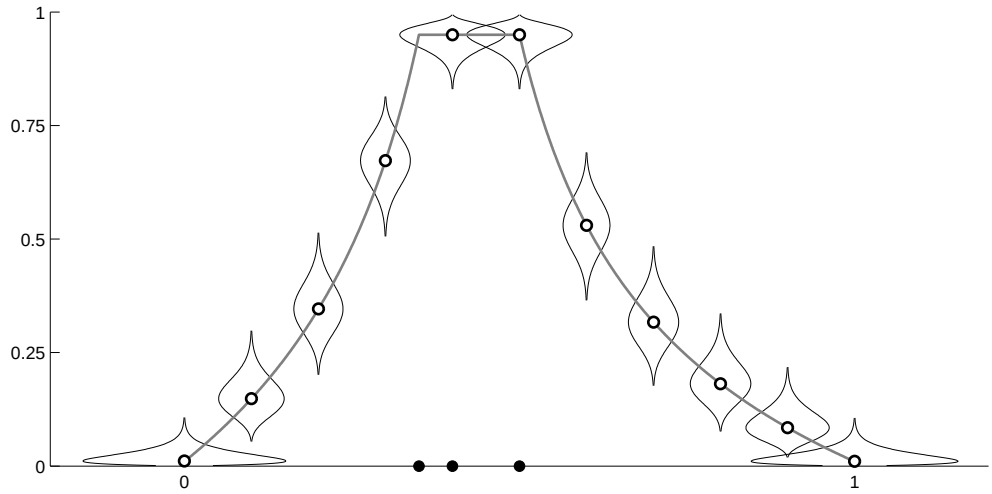

Figure 5: Distribution of errors for a generalization problem with training items at $x = (.35, .4, .5)$, and test items located from $y = (0, .1, .2, \ldots, .9, 1)$. In this example the cognitive parameters are $\theta = .5$ and $\phi = 1$, the response mapping parameters are $j_u = .95$ and $j_l = .01$, and the precision parameter is $\tau = 75$.

accidental responding, or any of a range of possibilities. Thus, we assume that with some probability ϵ, d is sampled from a uniform distribution on [0,1].

Priors

As noted previously, the most general formulation of the model requires the consideration six parameters; the two cognitive parameters θ and ϕ, the two calibration parameters j_u and j_l, and the two parameters for the error model τ and ϵ. In all six cases it is possible to specify reasonable priors that roughly approximate our expectations about the likely values these parameters could take.

Priors over θ and ϕ. We begin by specifying priors over the cognitive parameters. So, for instance, consider the simplest case, namely θ. If people are neither restricted to weak sampling $\theta = 0$ nor to strong sampling $\theta = 1$, it seems reasonable to specify a uniform prior density $p(\theta) = 1$, since this is a maximum entropy distribution conditional on the knowledge that both outcomes (sampling independent of the region and sampling from the region) are possible. Turning to our prior over possible participant priors, namely $p(\phi)$, an obvious consideration is that we would like the uniform case $\phi = 1$ to be the prior mode, since it corresponds to the most obvious constrained choice. Noting that $\phi \in [0, \infty)$, a fairly obvious class of priors is the Gamma family. The choice $\phi \sim \text{Gamma}(1, 2)$ yields the fairly simple prior density $p(\phi) = \phi \exp(-\phi)$, shown in Figure 6. An important point

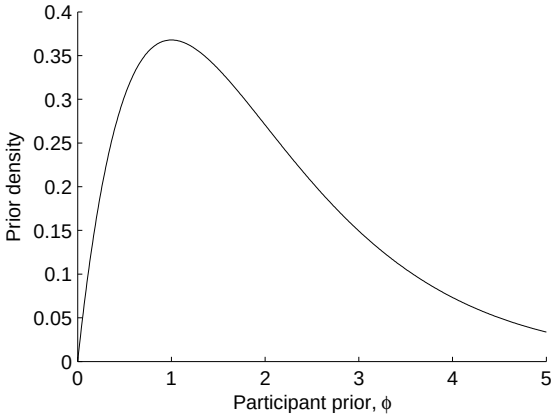

Figure 6: Prior density for ϕ.

is that the prior mode is $\phi = 1$, corresponding to the case where the *learner* has a uniform prior over regions.

Priors over j_u and j_l. In the case of j_u and j_l, even if these are not fixed at 1 and 0 respectively, we would not expect them to vary too far from these values. Moreover, we would want to preserve the constraint that $j_l \leq j_u$. If for the moment we were to imagine that this dependence need not hold, then a simple prior would be the triangular distributions $p(j_l) = (1 - j_l)/2$ and $p(j_u) = j_u/2$. Incorporating the constraint that the upper bound must exceed the lower bound gives the following prior

$$p(j_u, j_l) = \frac{1}{Z} j_u (1 - j_l) I(j_l \leq j_u) \tag{25}$$

where the constant of proportionality Z is

$$\begin{aligned} Z &= \int_0^1 \int_0^1 j_u(1 - j_l) I(j_l \leq j_u) dj_l dj_u \\ &= \int_0^1 \int_0^{j_u} j_u(1 - j_l) dj_l dj_u \\ &= \int_0^1 j_u \left[j_l - (1/2) {j_l}^2 \right]_0^{j_u} dj_u \\ &= \int_0^1 {j_u}^2 - (1/2) {j_u}^3 dj_u \\ &= \left[(1/3) {j_u}^3 - (1/8) {j_u}^4 \right]_0^1 \\ &= 5/24 \end{aligned} \tag{26}$$

Notice, however, that the inclusion of this constraint alters the marginal priors to some extent, since for example:

$$p(j_u) = \int_0^1 p(j_u, j_l) dj_l$$

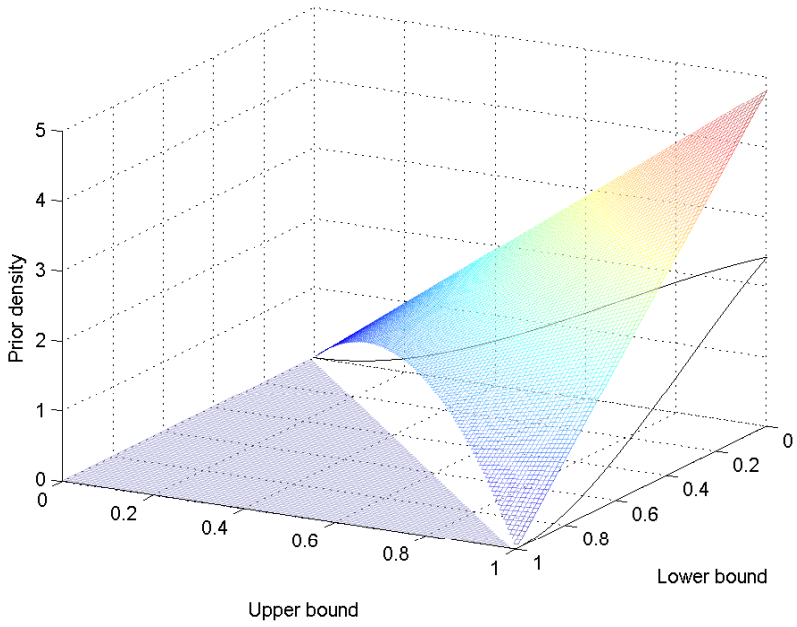

Figure 7: Joint and marginal priors over the parameters j_u and j_l governing potential calibration functions.

$$\begin{aligned} &= (24/5)\int_0^1 j_u(1-j_l)I(j_l \leq j_u)dj_l \\ &= (24/5)\int_0^{j_u} j_u(1-j_l)dj_l \\ &= (24/5)j_u\left[j_l - (1/2){j_l}^2\right]_0^{j_u} \\ &= (12/5){j_u}^2(2-j_u). \end{aligned} \tag{27}$$

The joint and marginal priors are shown in Figure 7: although much more constrained than the maximum entropy prior used for θ, this prior is deliberately chosen to be broader than a genuine prior belief distribution might be, for reasons of scientific caution.

Priors over τ and ϵ. The prior over the precision parameter τ is a little trickier. A conservative approach would be to propose that we do not know in advance the standard deviation associated with the error distribution, and so propose that *a priori* these should be uniformly distributed. However, since the variance of the Beta-error distribution depends on $\tilde{p}$ as well as τ, this is difficult to do precisely. To provide a simple approximation, we note that this variance is maximized when $\tilde{p} = .5$, and will use a prior over τ based on the notion that these worst-case variances are uniformly distributed. Now, when $\tilde{p} = .5$, the standard deviation of the error distribution for a given τ is $(\tau+3)^{-1/2}/2$. Accordingly, if this function is to be uniformly distributed on $[0, 12^{-1/2}]$ (noting that the uniform distribution has standard deviation $12^{-1/2}$ and is the highest variance distribution we wish

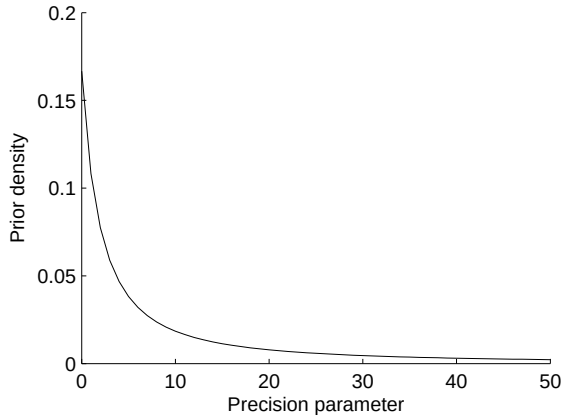

Figure 8: Prior over precision parameters, designed to ensure that the standard deviations of possible error densities are uniformly distributed a priori.

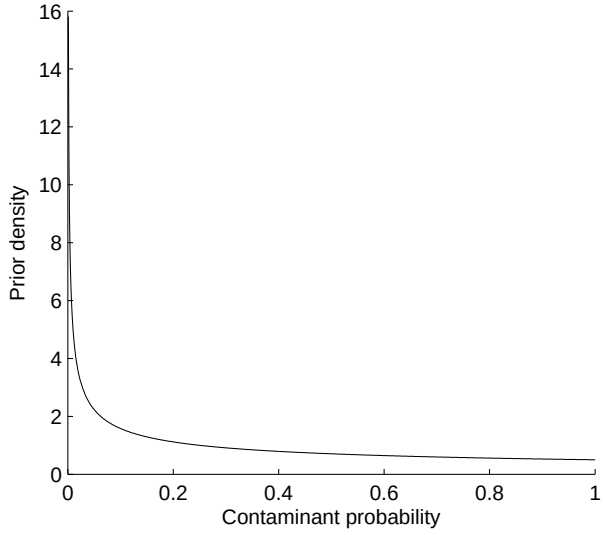

Figure 9: Prior over contaminant probabilities.

to include), then τ can be sampled by calculating the monotonically increasing function $\tau = 3((1-u)^{-2} - 1)$, where u is uniformly distributed on [0,1]. Noting that this function is therefore the inverse cumulative distribution function $p^{-1}(\tau \leq t)$ for τ, we obtain the cumulative distribution function,

$$p(\tau \leq t) = 1 - ((t/3) + 1)^{-1/2}. \tag{28}$$

Finally, differentiating gives the prior density for τ,

$$p(\tau) = (\sqrt{3}/2)(\tau + 3)^{-3/2} \tag{29}$$

This density is shown in Figure 8. Finally, we use a very tight prior $p(\epsilon) \propto \epsilon^{-1/2}$ over the contaminant probability so as not to encourage the model to "throw away" too many observations as outliers. In this case,

$$\int_0^1 \epsilon^{-1/2} d\epsilon = \left[2\epsilon^{1/2}\right]_0^1 = 2 \tag{30}$$

so $p(\epsilon) = (1/2)\epsilon^{-1/2}$. Note, however, that this density is not defined at $\epsilon = 0$.