0.1 Deviance of an oracle

To make deviance values more interpretable, it is helpful to normalise them with respect to the lowest deviance achievable by any model. The observer's responses are determined partly by the stimulus, and partly by internal noise. The problem with raw deviance values is that they depend on the influence of internal noise: the higher the internal noise, the larger the deviance of the best model will be, simply because the observer behaves unpredictably. The lowest deviance is achieved by the model which knows everything except the value of the internal noise on every trial. We call this model the "oracle". Our derivation follows on related derivations of human-model consistency bounds in classification image experiments in [**?**] and derivations of accountable variance in [**?**].

Assume that the observer maps the input to an internal variable $\mu = \mu(\mathbf{x})$, where $\mathbf{x}$ is the stimulus and μ is an arbitrarily complex function of the stimulus, with $\mu \in [0, 1]$, which is known to the oracle. The key insight here is that $\mathbf{x}$, the stimulus, is chosen at random on every trial from an arbitrarily large pool. It follows that $\mathbf{x}$ has a probability distribution $p(\mathbf{x})$, and this induces a probability distribution $p(\mu)$ on μ. If $p(\mu)$, $p(y|\mu)$ and $D(y, \mu)$ are known, then the expected deviance on a given trial may be computed.

Assume that the observer's response y is given by 1 with probability μ and 0 with probability $1 - \mu$, and that y is unknown to the oracle. Note that these labels are different from $y = \pm 1$ used in section **??** for reasons that will soon become clear. The likelihood of an observation y is given by:

$$p(y|\mu) = \mu^y (1 - \mu)^{1-y}$$

We assume that the oracle produces the prediction $\mu = \mu(\mathbf{x}) = \mathbb{E}(y)$ on any given trial. The deviance between the prediction and y is given by:

$$D(y, \mu) = -2 \log p(y|\mu) = -2(y \log \mu + (1 - y) \log(1 - \mu))$$

The expected deviance on N independent trials is then:

$$\begin{aligned} \mathbb{E}(D(\mathbf{y}, \mu)) &= N \sum_{i=0}^{1} \int D(y = i, \mu) p(y = i|\mu) p(\mu) d\mu \qquad (1) \\ &= -2N \int (\mu log \mu + (1 - \mu) \log(1 - \mu)) p(\mu) d\mu \end{aligned}$$

Our primary goal here will be to estimate $p(\mu)$. Suppose we have two pools of trials, T_1 and T_2. T_1 are one-pass trials, while T_2 are M-pass trials. The first pool give us information about the mean of $p(\mu)$, while the second pool gives us information both about its mean and variance. Now we let $p(\mu)$ have a parametric form, with parameters ξ_j. We obtain $p(\mu)$ by integrating over the parameters:

$$p(\mu) = \int p(\mu|\xi_j)p(\xi_j|T_1)p(\xi_j|T_2)p(\xi_j)d\xi_j$$

Here $p(\xi_j)$ is the prior probability of the parameters. We first deal with the repeated trials. By Bayes' theorem:

$$p(\xi_j|T_2) \propto p(T_2|\xi_j)$$

For trials in the pool T_2, we collect a vector $\mathbf{z}$, where each element is equal to the number of positive responses to a repeated presentation of a stimulus. z_i is the number of successes in M independent binary experiments which individually have a probability of success μ_i. Thus z_i follows a binomial distribution:

$$p(z_i|\mu_i) = \binom{M}{z_i}\mu_i^{z_i}(1-\mu_i)^{M-z_i}$$

We now integrate over the hidden variable μ:

$$\begin{aligned} p(T_2|\xi_j) &= \prod_i \int p(z_i|\mu_i)p(\mu_i|\xi_j)d\mu_i \\ &= \prod_i \binom{M}{z^i} \int \mu_i^{z_i}(1-\mu_i)^{M-z_i}p(\mu_i|\xi_j)d\mu_i \end{aligned} \tag{2}$$

$p(T_1|\xi_j)$ has the same form for the special case $M = 1$. The rest of the derivation depends on the choice of $p(\mu_i|\xi_j)$.

Beta distribution We first assume that μ follows a smooth, flexible probability distribution on [0,1] for which the integral 2 is analytically tractable. An obvious choice comes from noticing that 2 has the same form as a binomial likelihood multiplied by an unspecified prior. The conjugate prior for the binomial is the beta distribution. Thus, we let μ follow:

$$p(\mu|a,b) = Beta(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}$$

The integration is then easily done, as $p(z_i|\mu_i)p(\mu_i|\xi_j)$ itself follows a distribution $Beta(a+z_i, M+b-z_i)$. A further simplification comes from the fact that z_i may only take integer values $z_i = 0 \ldots M$. Denoting C_k as the number of elements of $\mathbf{z}$ taking the value k, we have:

$$p(T_2|a,b) = \prod_{k=0}^{M}\left(\binom{M}{k}\frac{\Gamma(a+b)\Gamma(a+k)\Gamma(M+b-k)}{\Gamma(a)\Gamma(b)\Gamma(a+b+M)}\right)^{C_k}$$

$p(T_1|\xi_j)$ is derived by simplifying the above formula for the case $M = 1$ suing the properties of the gamma function:

$$p(T_1|a,b) = \left(\frac{b}{a+b}\right)^{D_0} \left(\frac{a}{a+b}\right)^{D_1}$$

Here D_0 and D_1 are the number of trials in the first set where the outcome was 0 and 1, respectively. Finally, we impose a vague improper prior $p(a,b)$ of the form:

$$p(a,b) \propto \frac{1}{ab}$$

Employing a prior of this form removes degeneracies in the posterior distribution of a and b when there is little data; its influence vanishes in the limit of large quantities of data. The integral needed to obtain $p(\mu)$ is analytically intractable. However, we may write:

$$p(\mu) \approx \frac{\sum_{i,j} p(\mu|a_i,b_j)p(a_i,b_j|T_1,T_2)}{\sum_{i,j} p(a_i,b_j|T_1,T_2)}$$

Here $p(a_i,b_j|T_1,T_2) \equiv p(a_i,b_j|T_1)p(a_i,b_j|T_2)p(a_i,b_j)$. The minimum of $-\log p(a_i,b_j|T_1,T_2)$ may be found readily using nonlinear optimization software, e.g. `fmincon` in Matlab. From the Hessian of $-\log p(a_i,b_j|T_1,T_2)$, we can obtain a Gaussian approximation for the probability distribution; we sample a_i, b_j in a grid aligned with the principal axes of this Gaussian such that at the extremes of the grid the Gaussian probability falls below 10^{-4}. This yields the full probability distribution $p(\mu)$. The expected deviance under the model is then obtained by numerical integration. The function `evalexpectedd` included in our software package performs these numerical operations.

Conjectured lower bound distribution Here, as in the next choice of distribution, our goal is to get explicit formulas for the expected deviance under extreme circumstances. We make three simplifying assumptions. First, we assume that the observer is unbiased. Second, we assume that trials in the pool T_2 are double-pass (e.g. $M = 2$). Third, we assume that all trials are in the second pool.

Given the form of 1, it seems reasonable to conjecture that the distribution which induces the lowest expected deviance consistent with a certain level of human-human consistency should have as much mass around the values 0 and 1 possible; all other mass should go towards reducing human-human consistency as fast as possible. We thus assume that:

$$p(\mu|c) = \begin{cases} 0 & \textit{with probability } (1-c)/2 \\ 1/2 & \textit{with probability } c \\ 1 & \textit{with probability } (1-c)/2 \end{cases}$$

Here $\mathbb{E}(\mu) = 1/2$, as desired for an unbiased observer. Now:

$$
\begin{aligned}
p(T_2|c) &= \prod_i \binom{M}{z^i} (p(z_i|\mu_i = 0)p(\mu_i = 0|c) + \\
&\quad p(z_i|\mu_i = 1/2)p(\mu_i = 1/2|c) + \\
&\quad p(z_i|\mu_i = 1)p(\mu_i = 1|c))
\end{aligned}
$$

Considering cases $z_i = 0, 1, 2$, we obtain:

$$
p(T_2|c) = \left(\frac{1}{2} - \frac{1}{4}c\right)^{C_0+C_2} \left(\frac{1}{4}c\right)^{C_1}
$$

We assume no prior on c. We may differentiate this estimate with regards to c to find its maximum likelihood estimate:

$$
c_{ML} = \frac{2C_1}{C_0 + C_1 + C_2} = 2(1 - C_{hh})
$$

Where C_{hh} is human-human consistency, the proportion of repeat trials where the response is the same. Now:

$$
\begin{aligned}
p(\mu) &\propto \int p(\mu|c)p(T_2|c)dc \\
&\approx \int p(\mu|c)\delta(c_{ML})dc \\
&= p(\mu|c_{ML})
\end{aligned}
$$

Consider an optimal, non-probabilistic model [**?**] which predicts the observer's response to be "1" when $\mu \geq 1/2$ and "0" when $\mu < 1/2$, . What is the expected percentage of correct responses, e.g. the model-human consistency, under this optimal model? The model gets one half of the trials with $\mu = 1/2$ wrong, hence:

$$
C_{mh} = 1 - 1/2c_{ML} = C_{hh}
$$

We have thus given a parametric form for a model which achieves $C_{mh} = C_{hh}$, which is the lowest achievable value for C_{mh} [NeriLevi]. The expected deviance for this model is then:

$$
\mathbb{E}(D(y, \mu)) = 4N(1 - C_{hh}) \log 2 \tag{3}
$$

Conjectured upper bound distribution Given 1, it seems reasonable to conjecture that the distribution which induces the *highest* expected deviance consistent with a certain level of human-human consistency should have mass as far away from 0 and 1 as possible. Under the constraint that $\mathbb{E}(\mu) = 1/2$, we choose the following form for the distribution:

$$p(\mu|c) = \begin{cases} c & with\, probability\, 1/2 \\ 1-c & with\, probability\, 1/2 \end{cases}$$

This yields:

$$\begin{aligned} p(T_2|c) &= \prod_i \int p(z_i|\mu_i)p(\mu_i|c)d\mu_i \\ &= \prod_i \frac{1}{2}\binom{M}{z^i}(p(z_i|\mu_i = c) + p(z_i|\mu_i = 1-c)) \end{aligned}$$

Again considering cases $z_i = 0, 1, 2$, we obtain:

$$p(T_2|c) = \left(\frac{1}{2}((1-c)^2 + c^2)\right)^{C_0+C_2} (2c(1-c))^{C_1}$$

We differentiate and solve a quadratic equation to find:

$$c_{ML} = \frac{1 + \sqrt{2C_{hh} - 1}}{2}$$

The model-human consistency for an optimal model under this distribution is given by:

$$C_{mh} = \frac{1 + \sqrt{2C_{hh} - 1}}{2}$$

In this case, the model attains the upper bound for model-human consistency, derived in [NeriLevi]. The expected deviance for the oracle in this case is given by:

$$\mathbb{E}(D(y,\mu)) = -2N(c_{ML}\log c_{ML} + (1-c_{ML})\log(1-c_{ML})) \qquad (4)$$

Here, as with 3, the expression goes to 0 when $C_{hh} = 1$, and to $2N\log 2$ when $C_{hh} = 1/2$. Interestingly,

$$-2N(c_{ML}\log c_{ML} + (1-c_{ML})\log(1-c_{ML})) \geq 4N(1-C_{hh})\log 2$$

We conjecture that the preceeding two models yield lower and upper bounds for expected deviance given a certain level of human-human consistency, a conjecture which we now back up with simulation data.

Simulations We performed several simulations on the assumption of a variable v with a mixture distribution:

$$v \sim \frac{1}{2}\mathcal{N}_0(-a, \sigma^2) + \frac{1}{2}\mathcal{N}_1(a, \sigma^2)$$

It should be noted that the internal variable in the linear model [Ref] has exactly this distribution assuming the noise patterns are generated from Gaussian distributions. We define a "correct" trial as one where the observer responds "1" when v is taken from the second distribution or "0" when v is the taken from the first. We considered the following models for μ:

1. $\mu = \Phi_l(v)$, where Φ_l is the logistic function
2. $\mu = \Phi_l(v^3)$
3. $\mu = 0.01 + \Phi_l(v) \cdot 0.98$, equivalent to (1) but where one response out of 50 is stimulus-independent, corresponding to a lapse rate of 0.02 [**?**].
4. $\mu = \Phi_g(v)$, where Φ_g is the Gaussian cdf
5. $\mu = \Phi_l(v)$, adjusted so that the observer is correct on 69% of trials ($d' \approx 1$)
6. $\mu = \Phi_l(v + b)$, where b is adjusted so that the observer is slightly biased (55% of "1" responses)

In all cases, a and σ where adjusted manually so that observers roughly showed a human-human consistency of 0.65, 0.75 and 0.90 increments, and so that they performed at 75% (except for generative model 5). We performed 50 simulations for each case, with 2000 trials in each simulation, where 1200 trials where one-pass and 400 were repeated twice. The average computed deviances, and the average predicted deviances based on the beta, lower bound, and upper bound distributions are plotted in Figure 1.

The beta distribution estimate does not appear to fit to the exact distribution induced by the used generative model. For example, the predicted mean deviance under the beta distribution model does not differ appreciably in cases (2) and (3), although the computed deviances are quite different across these cases. Rather, the beta estimate takes a grand average over many beta distributions which are consistent with the data. It appears as though the average predicted deviance hugs the upper bound more closely when the observer is very noisy and the lower bound when the observer is less noisy. Interestingly, this trend appears to hold with the computed deviances as well.

We note that in all cases, the mean computed deviance is indeed bounded above and below by what we conjecture are bounds for it. Furthermore, the beta distribution estimate is always smaller than the upper bound and larger than the lower bound, not just on average but on any particular simulation. Given these facts, we recommend the use of the beta distribution estimate, which appears less biased in these simulations than, say, the midline between the upper and lower bounds, and yet as roughly the same variance.

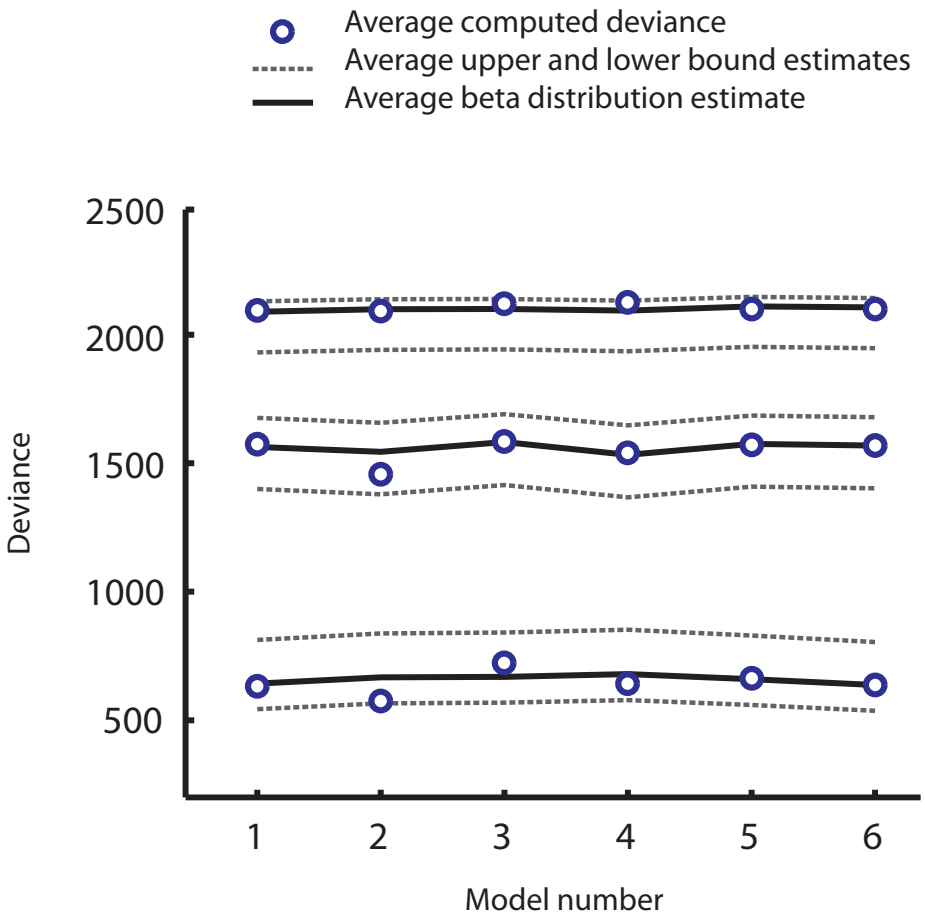

Figure 1: Average computed and predicted deviance values under different distributional assumptions for μ. From top to bottom: $C_{hh=}0.65$, $C_{hh} = 0.75$, $C_{hh} = 0.90$.

The beta distribution estimates are closely parallel to the upper and lower bound estimates; these would be straight lines if human-human consistency was adjusted with perfect accuracy across conditions. The beta distribution estimates, like the upper and lower bound estimates, are only as good as the accuracy of the human-human consistency estimate $\hat{C}_{hh}$. As such, we do not recommend directly quoting cross-validated scores as $D_{CV} - D_{oracle}$, as these can be deceiving. For example, under the distribution $p(\mu)$ assumed for the lower bound, it is easy to verify by simulation that with 400 trials repeated twice, and a nominal C_{hh} of 0.75, the 95% confidence range for $\hat{C}_{hh}$ is [0.71,0.79], which implies a confidence range for the lower bound of D for 2000 trials based on this estimate as $[1164, 1608]$. Rather, we suggest simply quoting D_{oracle} with the understanding that this gives at rough, but useful order of magnitude for the baseline.