Rate Distortion Theoretic and Bayesian Models for Speed Perception

Jacob Glenn

Advisor: Alan Stocker, Ph.D.

Advisor: Lingqi Zhang

University of Pennsylvania

Abstract

Rate Distortion Theory (R) and Bayesian (B) frameworks have both been used to describe optimal information processing. One such application of this is in the realm of visual perception. In general, B frameworks attribute errors in perception to sources of noise, while RDT frameworks attribute errors to a combination of noise and capacity constraints on communication channels. While it is clear that sometimes these models produce similar results, it is not clear in which ways they are fundamentally similar or different. In this paper, we take a look at two existing RDT and B models and interpret them in a general context. We discuss their mathematical definitions, address the question of whether they can be mathematically equivalent under certain circumstances, and apply them to a speed discrimination data set for which the B model has already been demonstrated to explain the data well. We discover a simple case in which the two models are mathematically equivalent. We also find that they can both explain the speed discrimination data adequately well.

Rate Distortion Theoretic and Bayesian Models for Speed Perception

Introduction

Both Rate Distortion Theoretic (RDT) and Bayesian (B) approaches have been used to model optimal information processing. From the standpoint of modeling human behavior, these approaches attempt to explain why human information processing, for example during perception, can be corrupted in predictable ways. While it is clear that the two frameworks can yield similar results under certain circumstances, it is unclear to what extent they are inherently similar or different.

In this report we will fix our discussion to two specific examples of existing Rate Distortion Theoretic (RDT) and Bayesian (B) models respectively, defined to produce a single conditional probability distribution (channel) which models a certain real world probabilistic relationship. First we take a look at the mathematical definitions of our particular models and ask whether they can ever be equivalent, that is, equal as conditional probability distributions. Next, we apply these models to a speed discrimination task in the context of variable contrast levels and compare the efficacy of the two models in describing human behavior.

When viewed in the context of Marr's categories of psychological models, these RDT and B approaches constitute computation level theories of perception (Marr, 1982). They both propose that the goal of the perceptual act is to minimize the expected cost of making perceptual errors but achieve this goal with respect to with different mathematical constraints. Because at the most basic level, both models simply represent conditional probabilities, comparing them seems like a simple task. However, both RDT and B have inherent flexibility which makes comparing them on an equal playing field challenging, requiring imposing meaningful constraints to the parameters of both models.

The Models

High Level Description

A convenient way to talk about RDT and B models representing a given channel is in terms of classes of channels. We say that a channel (conditional probability distribution) is modeled by RDT or B if it is in the class of RDT-optimal or B-optimal channels, respectively. We say that a channel is RDT-optimal or B-optimal if it attainable by an RDT-optimal or B-optimal observer.

An RDT-optimal observer starts with several intrinsic properties. Firstly, it has a prior distribution over the inputs, which encapsulates the observer's working knowledge of the source statistics of the inputs of the model. The observer also has a notion of the cost of reconstructing input element x as output element y, which it acts to minimize. Moreover, the communication channel between input and output (the entire process) is constrained by capacity. Rate Distortion Theory formalizes and integrates these components. The end result is that if the prior, cost, and capacity are fixed, then the observers' channel minimizes average cost with respect to the information capacity constraint.

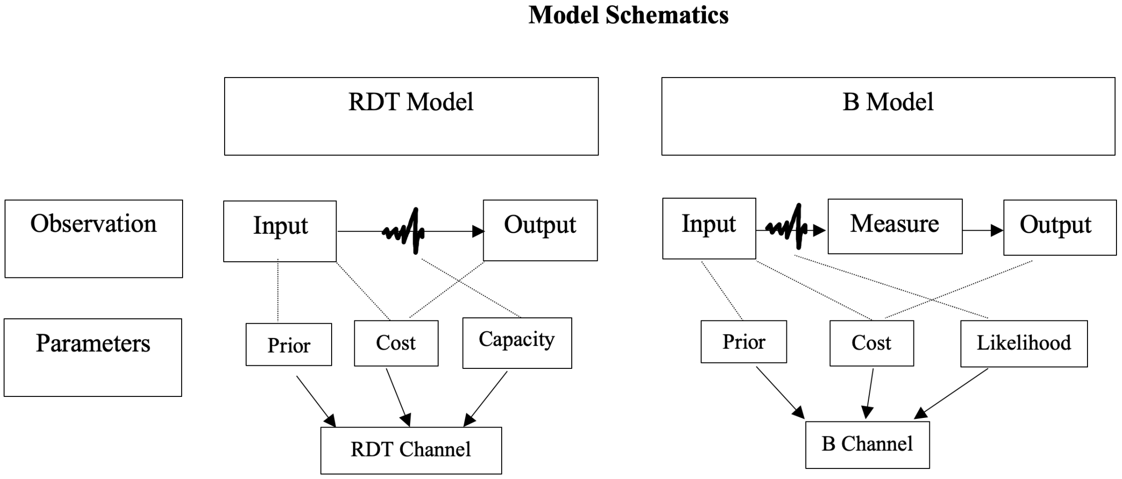


Figure 1. RDT and B model schematics: The RDT observer is described by a single noisy channel from inputs to outputs. The B observer is described by a two-step process, consisting of a noisy encoding stage and a noiseless decoding stage. For RDT, the prior represents a knowledge of source statistics. The cost function represents the cost of reconstructing input elements as output elements, for every combination of the two. The capacity encapsulates an information bottleneck in the channel, preventing perfect reconstruction. The RDT channel is considered optimal if it optimally integrates the

prior, capacity, and cost, minimizing average cost with respect to the capacity constraint. For B, the prior and cost represent the same things. The likelihood characterizes the relationship between incoming inputs and noisy measurements taken of the inputs. The B channel is considered optimal if it optimally integrates the prior, cost, and likelihood, representing the final input-output probabilistic relationship produced by choosing the decoding formula which minimizes average cost.

On the other hand, a B-optimal observer also starts with a prior over inputs, and cost over input and output, which it acts to minimize. However, in contrast to the RDT observer, the observation act is explicitly broken up into a measurement and estimation stage. In the first stage, inputs are encoded to form noisy sensory representations i.e., measurements. In the second stage, measurements are decoded deterministically. The decoding scheme is chosen using prior beliefs to minimize the end-to-end average cost between inputs and outputs. This joint distribution between inputs and outputs is implicitly given by first combining an internal likelihood of measurements given inputs and the input prior to form a posterior distribution of measurements given inputs, then taking the resulting distribution of the decoded measurements given the inputs.

Formulation

To formalize RDT-optimality and B-optimality, we must first define the underlying alphabets upon which the probability distributions in view are defined. For this section, we will show what the models look like for discrete input and output alphabets, and note that continuous analogs also exist. Therefore, we will assume we have a discrete input alphabet X and output or reconstruction alphabet Y. From this point, we will say that $p : X \to \mathbb{R}_{\geq 0}$ is distribution over X and that $d : X \times Y \to \mathbb{R}_{\geq 0}$ is a nonnegative cost function defined between the input and reconstruction alphabets with no limitations other than $\forall\ x\ \exists\ y \in Y$ st. $d(x, y) = 0$. In both models, $p(x)$ represents the prior probability of seeing input x and $d(x, y)$ represents the cost of reconstructing input x as output y. Furthermore, let $\mathcal{P}$ be the set of all probability distributions over X, $\mathcal{D}$ the set of such cost functions over $X \times Y$, and $\mathcal{Q}$ the set of all conditional probability distributions over $Y \times X$.

RDT. In this context, we use the standard formulation of the Rate Distortion function to define a class of channels which are RDT-optimal for some prior p, cost d, and capacity constraint $r \in \mathbb{R}_{\geq 0}$ (see Bates and Jacobs (2020), e.g. for an interpretation of the components). We write this class as

$$Q_{RDT}(p,d,r) = \underset{q \in \mathcal{Q};\, \sum_{x,y} p(x)q(y|x) \log \frac{q(y|x)}{\sum_x p(x)q(y|x)} \leq r}{\operatorname{argmin}} \sum_{x,y} p(x)q(y|x)d(x,y)$$

or for short, $Q_{RDT}(p,d,r) = \underset{q \in \mathcal{Q};\, I_q \leq r}{\operatorname{argmin}}\, D_q$ where I_q is the mutual information between X and Y and $D_q = \mathbb{E}[d(x,y)]$, both being determined by the joint probability defined by combining p and q.

B. Because the B model is a two-step process, in addition to input and output alphabets, we need to define a measurement alphabet M, denoting the alphabet of potential measurements. We will say that $l : M \times X \to [0,1]$ is some likelihood function such that $l(m,x)$ is the probability of attaining measurement m from input x. We use a previously tested model for a Bayesian estimation task to define a class of channels which are B-optimal for some prior p, cost d, and likelihood function l (see (Stocker & Simoncelli, 2006) for the experimental model and Wei and Stocker (2015) for this general model. We write this class as

$$Q_B(p,d,l) = \Big\{ q \in \mathcal{Q};\; q(y|x) = \sum_{m \in M;\, f(m)=y} l(m|x) \Big\}$$

where

$$f(m) = \underset{y \in Y}{\operatorname{argmin}} \sum_x g(x|m)d(x,y) \quad \text{and} \quad g(x|m) = \frac{p(x)l(m|x)}{\sum_x p(x)l(m|x)}$$

We can see the two-step process unfold in the formula, where the prior and the likelihood are combined to obtain a posterior g. This posterior is used to choose reconstructions for measurements that minimize average error from input to reconstruction. Then the distribution l of measurements given inputs is transformed to obtain the desired distribution q of reconstruction given input.

Qualitative Similarities

We now see the inherent flexibility of both models when considered for arbitrary priors and cost functions. It also may now be clear why we are interested in comparing these two models. The prominent reason is that they are both mathematically optimal in different ways, and they both incorporate both a prior and a cost function to characterize a channel as optimal. The difference is that RDT takes in yet another parameter, the capacity constraint r, and B takes in another parameter, the likelihood l.

Another way of looking at it is to say that given a prior p and cost d, there is already a class of channels which are candidates for RDT-optimality, that is

$$Q_{RDT}(p, d) := \{q \in \mathcal{Q};\ \ \exists\ r \in \mathbb{R}_{\geq 0} \text{ st. } q \in Q_{RDT}(p, d, r)\}$$

and that there is a class of channels which are candidates for B-optimality, that is

$$Q_B(p, d) := \{q \in \mathcal{Q};\ \ \exists\ l \in \mathcal{L} \text{ st. } q \in Q_B(p, d, l)\}$$

An additional similarity we can see at the formulaic level relates to the cost function parameter. It is clear from both formulas that if d and d' are such that $d(x, y) = kd'(x, y)$ where $k \in \mathbb{R}_{\geq 0}$, then $Q_{RDT}(p, d, r) = Q_{RDT}(p, d', r)$ for all p, r and $Q_B(p, d, l) = Q_B(p, d', l)$ for all p, l. In other words, it doesn't affect either model if we scale the cost function.

Formulation of Equivalency

The previous considerations point us to another question, namely, if there is a prior-cost pair for which $Q_{RDT}(p, d) \subset Q_B(p, d)$. If we fix prior p and distortion d, then is there a class of likelihoods $L \subset \mathcal{L}$ such that $Q_B(p, d, L) = Q_{RDT}(p, d, \mathbb{R}_{\geq 0})$? We are unaware of previous attempts to establish the mathematical equivalency of specific Rate Distortion Theoretic and Bayesian models as a strategy to understand how the RDT and B approaches, in general, are fundamentally similar or different (but see Bates and

Jacobs (2020), regarding how the approaches are different at a high level).

Perhaps it appears that this relationship between our specific RDT and B models is unlikely to be true for any nontrivial p and d, given that the definitions for Q_{RDT} and Q_B have no glaring connection. On the other hand, one might suspect that it is possible because not constraining the possible likelihoods for B in any way provides significant flexibility on the part of B. We looked into this question for one of the simplest cases and found and prior-cost pair for which the property holds. Specifically, for $p(x) = \mathcal{N}(x; \mu, \sigma)$ and $d(x, y) = (x - y)^2$, we demonstrated that $Q_{RDT}(p, d) \subset Q_B(p, d)$. Moreover, we found a simple parameterization for the channels in both cases and a formula for translating between these parameters (see Appendix A).

Results

We now consider the application of these models to a perceptual experiment. Both Rate Distortion Theoretic and Bayesian models, in general, have been applied to the subject of visual perception (Sims, 2016; Stocker & Simoncelli, 2006). A common theme is that both varieties of models have so many parameters that it is vital to apply them in such a way as to reduce the number of parameters to avoid over-fitting. As a foothold, we used the previous speed discrimination experiment (from which we obtained our B model), where it was implemented successfully to establish reasonably restrictive constraints (Stocker & Simoncelli, 2006). We used the data from this experiment for a representative subject for the following analysis.

Observer Model for Speed Discrimination

Our application of the RDT and B models to the speed discrimination task involves modeling the speed estimation step by RDT and B channels. This means that our input is stimulus speed, and our output is estimated speed (percept). The entire probabilistic relationship between stimuli and percepts is described by the RDT or B channel.

We then embed the RDT and B estimators in an observer model that includes an optimal decision stage (see Stocker and Simoncelli (2006)). Then we fit the models to trials of responses in a two-alternative forced choice (2AFC) speed discrimination experiment. Using this setup, we were able to recover the prior distribution and capacity constraints associated with the best-fitting RDT model, as well as the prior distribution and likelihood associated with the best-fitting B model.

In the 2AFC experimental protocol, subjects were presented with two stimuli and tasked with selecting the faster moving stimulus. Our optimal decision stage assumption assumes that subjects perform an independent estimate of the speeds of both stimuli and then select the one with the higher estimate. This allows us to generate response data from our models and compare that to the real subject response data (psychometric curves). Then, we can fit the parameters of our models using the response data via maximum likelihood estimation.

Constraints on RDT and B

Prior. For both RDT and B, we decided to solve for a nonparametric description of the prior, using the value of the prior at several logarithmically spaced sample points as parameters in our models.

Cost Function. For the B model, we assumed a difference-squared cost function $d(x, y) = (x - y)^2$ in order to simplify the implementation of our model. We attempted to use the same cost function for the RDT model but found that it resulted in bad fit quality (see Effect of Cost Function on RDT Model). We then experimented with other cost functions and found that a difference-squared cost function in log-space $d(x, y) = (\log(x) - \log(y))^2$ performed significantly better (in terms of log-likelihood).

Dependence of RDT and B Parameters on Contrast. In the task, subjects chose from two stimuli which stimulus was moving faster. The stimuli consisted of moving gratings with variable speed and contrast. The models needed to be flexible enough to account for different behaviors across these different speeds and contrasts.

For our RDT model, we assumed that the capacity constraint for the channel was purely a function of contrast and solved for a nonparametric description of this capacity function. For the B model, we assumed (as in Stocker, 2006) that likelihood width was a separable function of speed and contrast and solved for a nonparametric description of both the speed and contrast factors.

Estimating Priors, Capacity Function, and Likelihood

For the RDT model, we assumed that the prior was a monotonically decreasing function of speed. The resulting shape of the prior distribution recovered was similar in shape to the prior recovered by B, which we did not assume to be monotonically decreasing. However, the B prior decreased more precipitously than the RDT prior. Both priors appeared to be linear in the log-log domain for a substantial portion of speeds, i.e, well approximated by a power law function of speed for this range.

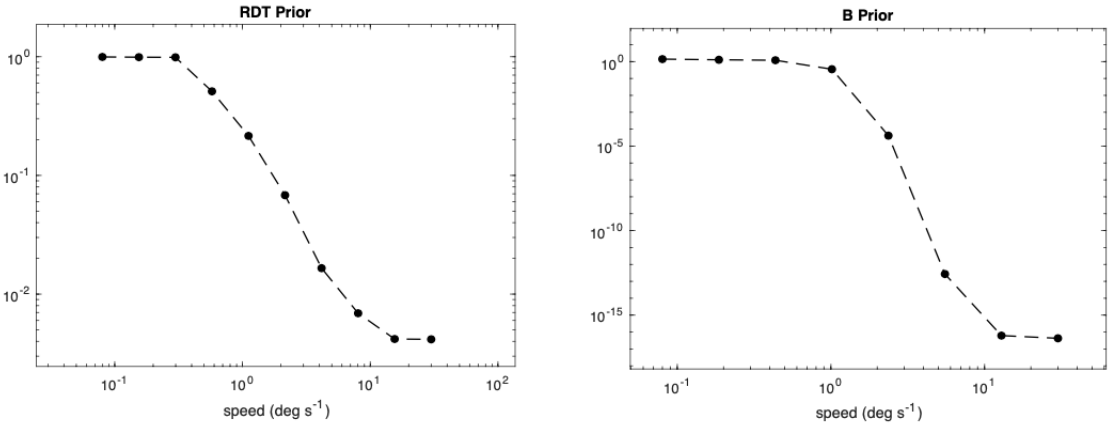


Figure 2. Recovered priors for RTD (left) and B models (right) for the best log-likelihood fits of the models to the data of the subject. Both priors were represented nonparametrically using logarithimically spaced sample points. There were 10 sample points for the RDT prior and 8 for the B prior. They have similar shapes, appearing to approximately follow a power law for a range of intermediate speeds. They both break that pattern for extremely slow and fast speeds.

For RDT, we assumed that the capacity function was a monotonically increasing function of contrast. The recovered capacity function for the RDT model took on a range of values, indicating that there may be utility in viewing contrast as a proxy for capacity in this case.

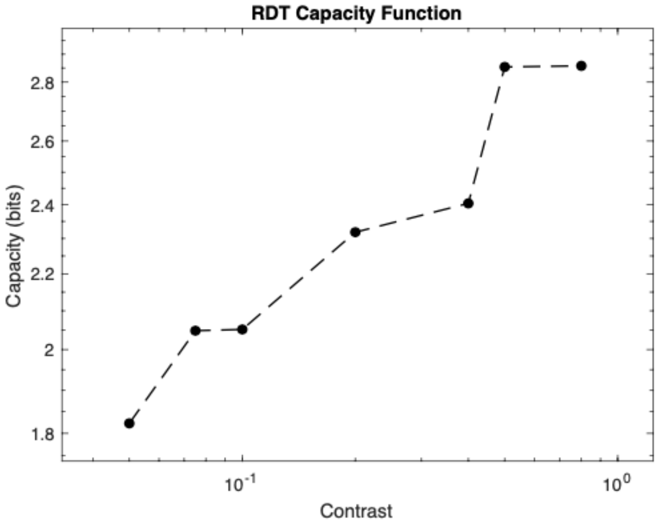


Figure 3. Recovered capacity function for the RDT model for the best log-likelihood fit to the experimental data of the subject. We assumed the capacity function was monotonically increasing. The effect of the capacity constraint on the RDT channel is shown in Figure 7.

For B, we found the recovered dependence of the width of the likelihood on speed and contrast. We did not identify a clear relationship between the speed and the likelihood width given our current assumptions. We observed a clearer dependence on contrast, where higher contrast corresponded to smaller width. Intuitively this makes sense because we would assume that the certainty of the speed given the measurement would be higher for higher contrast stimuli than for lower contrast stimuli.

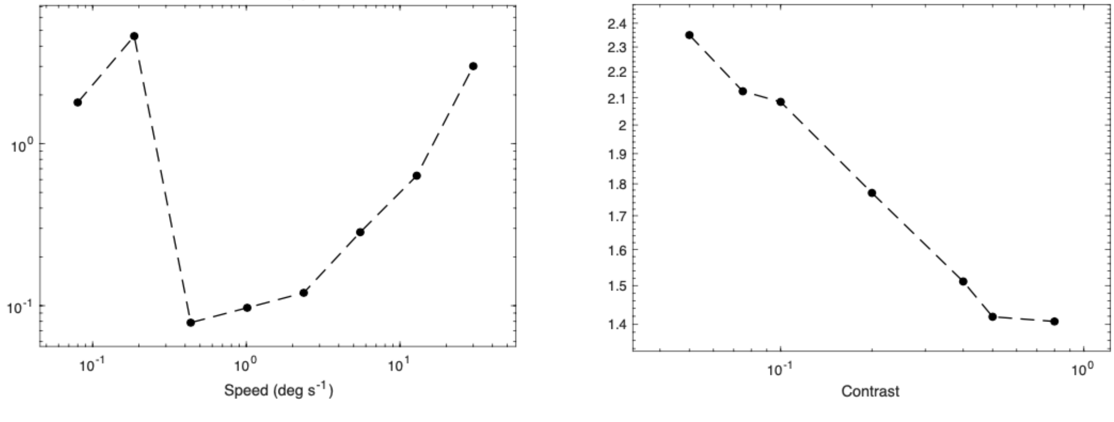


Figure 4. Width of likelihood for B as a function of speed and contrast, fitted to the subject data. We assumed that the likelihood was separable in terms of speed and contrast so that $\sigma(v, c) = g(v)h(c)$.

Model Comparison

To quantify how well the RDT and B models fit the data, we compared the log-likelihood of the data for our fitted RDT and B models to that of a model obtained by fitting a Weibull function to the psychometric curve associated with each experimental condition. The number of unique experimental conditions was 72 (Stocker & Simoncelli, 2006). We fit the models to a representative subject and compared the log-probability of all the data for the subject across all experimental conditions. We plotted the values for this subject on an axis where the minimum value is determined by the log-probability of a coin-flipping model, and the maximum value is determined by the log-probability of a model using the fitted Weibull functions.

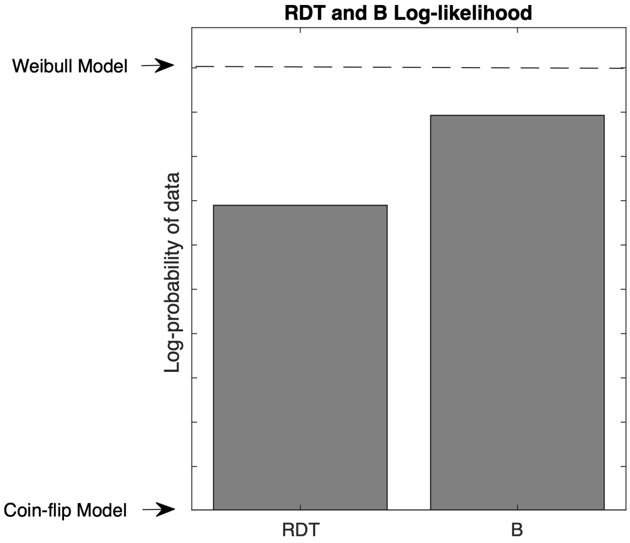


Figure 5. Log-likelihood of the data for the highest log-likelihood RDT and B fits respectively. The log-likelihoods are plotted on an axis that ranges from the value of a coin-flipping model to a model constituted by fitting a Weibull function to each stimulus condition. The comparison of the psychometric curves produced by RDT and B is shown in Appendix C.

For the subject, both RDT and B perform relatively well compared to the model obtained from Weibull fits, while the B model performs better. The degrees of freedom between the models are also relevant: the RDT model has 17 parameters; the B model has 23 parameters. The Weibull function consists of 72 fits each with two parameters, so 144 parameters in total.

Effect of Cost Function on RDT Channel

When we fitted the RDT model to the speed perception data, we noticed that our choice of cost function influenced the quality of the fit. Specifically, the fits using the cost function $d(x, y) = (\log(x) - \log(y))^2$ were much better than the fits using the cost function $d(x, y) = (x - y)^2$ using the log-likelihood performance metric.

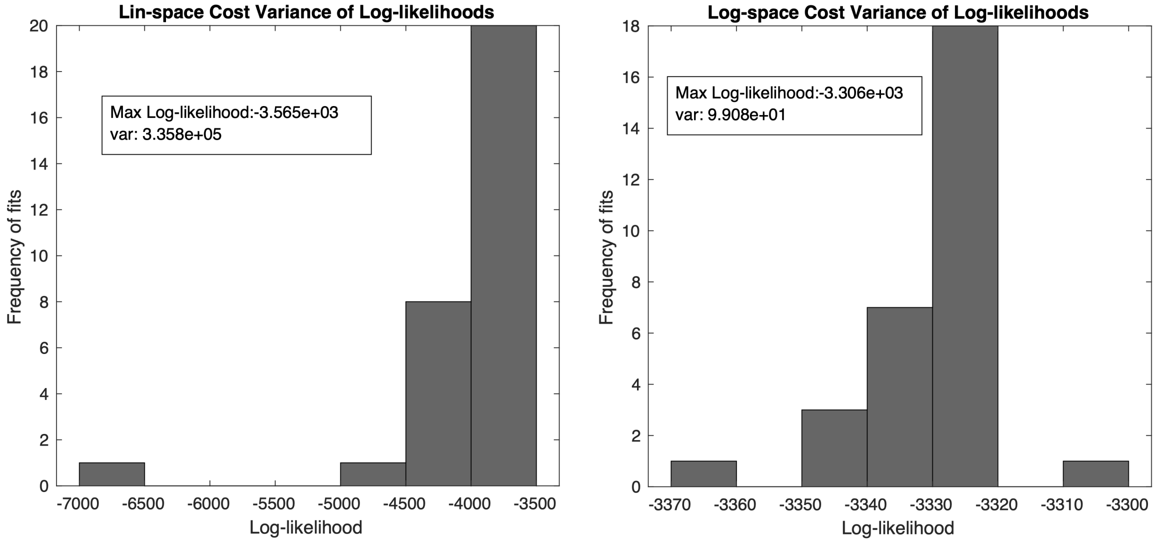


Figure 6. Histogram of log-probability values for 30 fits using the RDT model with $d(x, y) = (x - y)^2$ versus $d(x, y) = (\log(x) - \log(y))^2$. Each fit was obtained by doing a grid search over model parameters to minimize log-likelihood for about 5000 iterations with a random initial set of parameters. Histogram bars represent the number of fits whose log-likelihood falls within each the range of the bottom of the bar. On average, the log-likelihood was greater for the log-space cost. The maximum likelihood fit for the log-space cost was also greater than the maximum likelihood fit for the linear-space cost.

In addition to comparing the log-likelihood of the fits for the two cost functions, we decided to investigate what the effect on the channel was between the cost functions. To this end, we fixed the prior to the slow speed prior fit by the log-space cost and plot the linear and log-space channels produced by our RDT-optimal model for a small number of contrast levels.

The log-space cost channels had a higher variance for higher stimulus speeds for all rates.

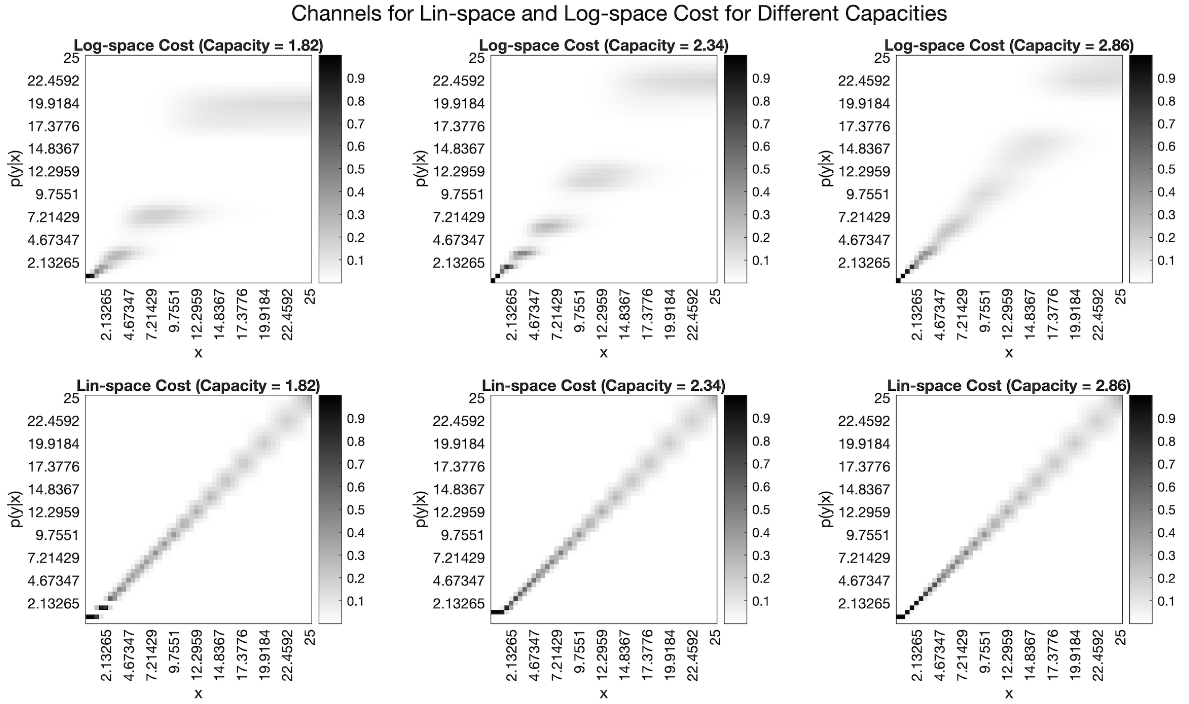


Figure 7. Channels for log-space (top) versus linear-space (bottom) cost functions ($d(x, y) = (\log(x) - \log(y))^2$ and $d(x, y) = (x - y)^2$ respectively). Prior was fixed to the prior fitted to by the best log-likelihood fit for the log-space cost RDT model. Channels were plotted for three different capacities, chosen from the capacity function for the best log-likelihood fit for the log-space cost model. The capacities chosen were the minimum of the capacity function, the maximum, and the average of the minimum and the maximum. The linear-space cost channels are less spread out for higher speeds. We also find that the channels for both costs have higher variance for lower capacities, which makes sense because intuitively, if there is a communication bottleneck between the input and output, there should be more noise in the channel. It also appears that the effect of capacity on the channel is more pronounced for the log-space cost model.

One explanation for why the log-space cost produces better fits is that the resulting channel captures the Weber–Fechner law to a greater extent. This indicates that there may be quantitative validity in describing human sensitivity via an RDT cost function. Other attempts to use RDT to model processes contingent on human visual sensitivity have seen log-space cost functions work better than linear-space cost functions (Frei & Baxter, 1977).

Discussion

We have formalized the RDT and B models, defined equivalence of the two, and shown an example. We have also demonstrated that our RDT and B models can provide an adequate description of speed perception given appropriate constraints. For

RDT, viewing contrast manipulation as an indirect form of capacity manipulation allowed us to reconstruct the shape of the prior and capacity function. For B, viewing contrast manipulation as an indirect form of likelihood width manipulation allowed us to reconstruct the shape of the prior and likelihood. For RDT specifically, we found that we achieved better results using a log-space cost function compared to a linear-space cost function. We explained this by observing that the log-space cost function channel reflects the Weber-Fechner law while the linear-space cost function does not.

Previous experiments have been done applying Rate Distortion Theoretic and Bayesian approaches to visual perceptual tasks by fitting priors, cost functions, or both. Usually, when these functions are fit, they are constrained to a particular parametric form (Sims, 2016). Generally speaking, this is because RDT and B are inherently flexible models that need to be constrained to prevent overfitting.

There are some limitations to the current RDT and B models. In terms of RDT, treating the entire process from stimulus to percept as a single channel may leave out important details of intermediary processes. Moreover, at the current level of analysis, RDT doesn't explain how a mechanism in the brain could be formed to produce behavior that was near RDT-optimal. Furthermore, some have remarked that it is unclear how the brain would act to minimize average cost between stimulus to percept given that the brain does not have direct access to the stimulus (Jozefowiez, 2019).

In the future, it would be beneficial to examine what RDT and B channels look like for more general distributions than what we considered in this paper. If the nature and patterns of these channels were better understood, then it would be possible to tell whether one could create interesting channels by using novel priors and cost functions. It also may be discovered that these optimal channels could be used as modules in larger models if it is the case that some component of the model is well described by an RDT or B channel.

Methods

Experiment. Subjects in the experiment were shown pairs of gratings moving horizontally. Each grating (v, c) had a speed and a contrast. Each pair of gratings consisted of a reference grating (v_1, c_1) and a test grating (v_2, c_2) assigned to the left and to the right at random from the point of view of the subject. The reference grating had one of two contrast values (0.075, 0.5) and one of five speeds (0.5, 1, 2, 4, 8, 12 deg s^{-1}). The test grading had one of seven contrast values (0.05, 0.075, 0.1, 0.2, 0.4, 0.5, 0.8) and a speed which varied. For each reference speed, reference contrast, and test contrast triplet (v_1, c_1, c_2), 80 trials were presented with varying test speeds from which the psychometric curves were determined. Speed was defined in terms of deg s^{-1} and contrast was defined in terms of the ratio between maximal intensity in the grating and maximum intensity difference able to be displayed by the monitor. Subjects were placed at a distance where two circular gratings are seen the left and right of a fixation cross. Pairs of reference and test stimuli were presented for 1s, after which subjects selected the stimulus which appeared to move faster. Pairs of stimuli were repeated if responses were not given within 1 s and the period between stimuli, during which nothing was presented, was 1.5 s. See Stocker and Simoncelli (2006) for the full details of the experiment.

Obtaining RDT-optimal channels using the Blahut algorithm. Our RDT model needed to take a given stimulus (v, c), and establish an RDT-optimal channel representing $p_c(\hat{v}|v)$. This RDT-optimal channel was to be optimal with respect to our prior $p(v)$, cost $d(x, y)$, and capacity constraint $r(c)$. Of these figures, the prior and the rate were determined by our parameters, while the cost was fixed.

It is a well known limitation of Rate Distortion Theory that there are not yet analytic methods for producing RDT-optimal channels for general combinations of prior, cost, and capacity. However, there is an efficient algorithm for computing RDT-optimal channels numerically for cases in which the input and output alphabets are finite i.e., the prior distribution and resulting channel are discrete (Blahut, 1972).

Sims (2016) implemented the Blahut algorithm and used it to produce RDT-optimal channels and applied the channels to a variety of visual perceptual experiments. We took the same approach, using Sims (2016) implementation of the Blahut algorithm in R as a starting point and reimplementing it in MATLAB, with minor modifications.

We used these discrete RDT-optimal channels obtained with the Blahut algorithm to approximate continuous RDT-optimal channels. This meant three things. First, we restricted our alphabet from $\mathbb{R}$ to a finite alphabet of representative values. Next, we approximated the continuous prior with a discrete prior over the new alphabet. We did not change the cost function. Finally, we used the Blahut algorithm to find an optimal channel in this discrete space. To demonstrate that this is a reasonable way to approximate RDT-optimal channels for continuous distributions, we tested the method on a configuration where the optimal channel is known analytically (see Appendix B).

Extracting the prior and capacity function for RDT. We now explain how we fit the parameters of the RDT model to the data of the subject via maximum likelihood estimation. In large part, our strategy is a discrete version of the method used to fit the parameters of the B model by Stocker and Simoncelli (2006).

We have chosen to parameterize the prior and capacity function for RDT by 10 and 7 logarithmically spaced sample points respectively. Given the prior p and capacity function r defined over continuous speeds and contrasts, the RDT model will then approximate $p_c(\hat{v}|v)$ for stimulus (v, c).

Our discretized input and outputs alphabets X and Y will be the same set of 50 logarithmically spaced points from .1 to 30 deg s^{-1}. We convert the parameterized prior p over $\mathbb{R}$ to a discrete prior p' over X. We then define the optimal discrete channel as $\text{Blahut}(p', d, r(c))$. Therefore, our naïve approximation of $p_c(\hat{v}_2 > \hat{v}_1)$ is

$$p(\hat{v_2} > \hat{v_1}) \approx p(y_2 \geq y_1) = \sum_{y_2} p_{c_2}(y_2|x_2) \sum_{y_1 \leq y_2} p_{c_1}(y_1|x_1)$$

where y_1 and y_2 are the closest points in Y to $\hat{v}_1$ and $\hat{v}_2$ respectively. We make this approximation smoother by using bilinear interpolation so that

$$\begin{aligned} p(\hat{v_2} > \hat{v_1}) \approx \bar{\lambda_2}&\Big[\bar{\lambda_1}\Phi\Big(x_a(v_1), x_a(v_2)\Big) + \lambda_1\Phi\Big(x_b(v_1), x_a(v_2)\Big)\Big] \\ &+ \lambda_2\Big[\bar{\lambda_1}\Phi\Big(x_a(v_1), x_b(v_2)\Big) + \lambda_1\Phi\Big(x_b(v_1), x_b(v_2)\Big)\Big] \end{aligned}$$

$$\Phi(x_1, x_2) := \sum_{y_2} p_{c_2}(y_2|x_2) \sum_{y_1 \le y_2} p_{c_1}(y_1|x_1)$$

$$\lambda_1 := \frac{v_1 - x_a(v_1)}{x_b(v_1) - x_a(v_1)} \quad \lambda_2 := \frac{v_2 - x_a(v_2)}{x_b(v_2) - x_a(v_2)}$$

$$x_a(v) := \text{ nearest } x \in X \text{ to the left} \quad x_b(v) := \text{ nearest } x \in X \text{ to the right}$$

We can directly compute the log-likelihood of the data given this model. Optimal parameters for the model are those which maximize this log-likelihood.

Extracting the prior and likelihood width for B. Similarly to the RDT model, for the B model, we write the expression $p(\hat{v_2} > \hat{v_1})$ as a function of the parameters of the model, and then use MLE to estimate the parameters. For B, as in Stocker and Simoncelli (2006), we assume that the prior is smooth relative to the width of the likelihood, that the logarithm of the prior is well approximated by a straight line over the range of velocities corresponding to the width of the likelihood function, and that the likelihood function is well approximate by a Gaussian centered at a peak value m_v which is considered to be the speed measurement. Moreover, we assume that the width of the likelihood function is separable in stimulus speed and contrast so that $\sigma(c, v) = g(v)h(c)$.

We use the result derived in Stocker and Simoncelli (2006) which is that, given these assumptions, the shape of $p(\hat{v}|v)$ matches the shape of the likelihood distribution, which we are assuming is Gaussian so that

$$p_c(\hat{v}|v) \approx \mathcal{N}\Big(\hat{v}; v + a(v)\sigma^2(c, v),\ \sigma^2(c, v)\Big)$$

A difference between our model and the model in Stocker and Simoncelli (2006) is that they found better performance by doing speed preprocessing into the log-speed domain, corresponding to a log-Normal likelihood assumption. for simplicity, we assumed a Gaussian likelihood.

References

Bates, C. J., & Jacobs, R. A. (2020). Efficient data compression in perception and perceptual memory. *Psychological Review*.

Berger, T. (1971). *Rate distortion theory: A mathematical basis for data compression.* Prentice-Hall. Retrieved from `https://books.google.com/books?id=-HV1QgAACAAJ`

Blahut, R. (1972). Computation of channel capacity and rate-distortion functions. *IEEE transactions on Information Theory*, *18*(4), 460–473.

Frei, W., & Baxter, B. (1977). Rate-distortion coding simulation for color images. *IEEE Transactions on Communications*, *25*(11), 1385–1392.

Jozefowiez, J. (2019). Perception and the metaphysics of information. *Learning & Behavior*, 1–2.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information.* W.H. Freeman. Retrieved from `https://books.google.com/books?id=GcKywwEACAAJ`

Sims, C. R. (2016). Rate–distortion theory and human perception. *Cognition*, *152*, 181–198.

Stocker, A. A., & Simoncelli, E. P. (2006). Noise characteristics and prior expectations in human visual speed perception. *Nature neuroscience*, *9*(4), 578–585.

Wei, X.-X., & Stocker, A. A. (2015). A bayesian observer model constrained by efficient coding can explain'anti-bayesian'percepts. *Nature neuroscience*, *18*(10), 1509.

Appendix A

Equivalency Example

Let $X = M = Y = \mathbb{R}$. Suppose our prior is $p(x) = \mathcal{N}(x; \mu_p, \sigma_p^2)$ and our cost is $d(x, y) = (x - y)^2$. Then we have the standard RDT result (see Berger (1971) for example) that the RDT-optimal channels are Gaussian. Specifically,

$$Q_{RDT}(\mathcal{N}(x; \mu_p, \sigma_p^2), (x-y)^2, r) = \Big\{\mathcal{N}\Big(x + e^{-2r}(\mu_p - x), e^{-4r}(e^{2r} - 1)\sigma_p^2\Big)\Big\}$$

For B, we start by assuming the likelihood is of the form $l(m|x) = \mathcal{N}(x; m, \sigma_l^2)$. The posterior (using our previous notation g) is

$$\begin{aligned} g(x|m) &= \frac{1}{\alpha} l(m|x)p(x) \\ &= \mathcal{N}\Big(x; \frac{\sigma_p^2 m + \sigma_l^2 \mu_p}{\sigma_p^2 + \sigma_l^2}, \frac{1}{\frac{1}{\sigma_p^2} + \frac{1}{\sigma_l^2}}\Big) \end{aligned}$$

so that the estimate function is

$$\begin{aligned} f(m) &= \operatorname*{argmin}_y \int_{\mathbb{R}} \mathcal{N}\Big(x; \frac{\sigma_p^2 m + \sigma_l^2 \mu_p}{\sigma_p^2 + \sigma_l^2}, \frac{1}{\frac{1}{\sigma_p^2} + \frac{1}{\sigma_l^2}}\Big)(x - y)^2 dx \\ &= \frac{\sigma_p^2 m + \sigma_l^2 \mu_p}{\sigma_p^2 + \sigma_l^2} \\ &= \frac{\sigma_p^2}{\sigma_p^2 + \sigma_l^2} m + \frac{\sigma_l^2 \mu_p}{\sigma_p^2 + \sigma_l^2} \end{aligned}$$

then because $p(m|x) = \mathcal{N}(m; x, \sigma_l^2)$ and y is a linear function of m, we have that

$$p(y|x) = \mathcal{N}\Big(\frac{\sigma_p^2}{\sigma_p^2 + \sigma_l^2} x + \frac{\sigma_l^2 \mu_p}{\sigma_p^2 + \sigma_l^2}, \Big(\frac{\sigma_p^2}{\sigma_p^2 + \sigma_l^2}\Big)^2 \sigma_l^2\Big)$$

i.e. $Q_B(\mathcal{N}(x; \mu_p, \sigma_p^2), (x-y)^2, \mathcal{N}(x; m, \sigma_l^2)\big) = \Big\{\mathcal{N}\Big(\frac{\sigma_p^2}{\sigma_p^2+\sigma_l^2} x + \frac{\sigma_l^2 \mu_p}{\sigma_p^2+\sigma_l^2}, \Big(\frac{\sigma_p^2}{\sigma_p^2+\sigma_l^2}\Big)^2 \sigma_l^2\Big)\Big\}$. So, then RDT-optimal channels and B-optimal channels are Gaussian, and moreover, setting

$$r = -\frac{1}{2} \ln \frac{\sigma_l^2}{\sigma_p^2 + \sigma_l^2}$$

or equivalently

$$\sigma_l^2 = \frac{\sigma_p^2}{e^{2r} - 1}$$

makes $Q_{RDT} = Q_B$. In other words, the family of estimates parameterized by the RDT capacity r is the same which is parameterized by the B likelihood variance σ_l^2 and we have a formula for translating from one to the other!

Appendix B

Blahut Algorithm

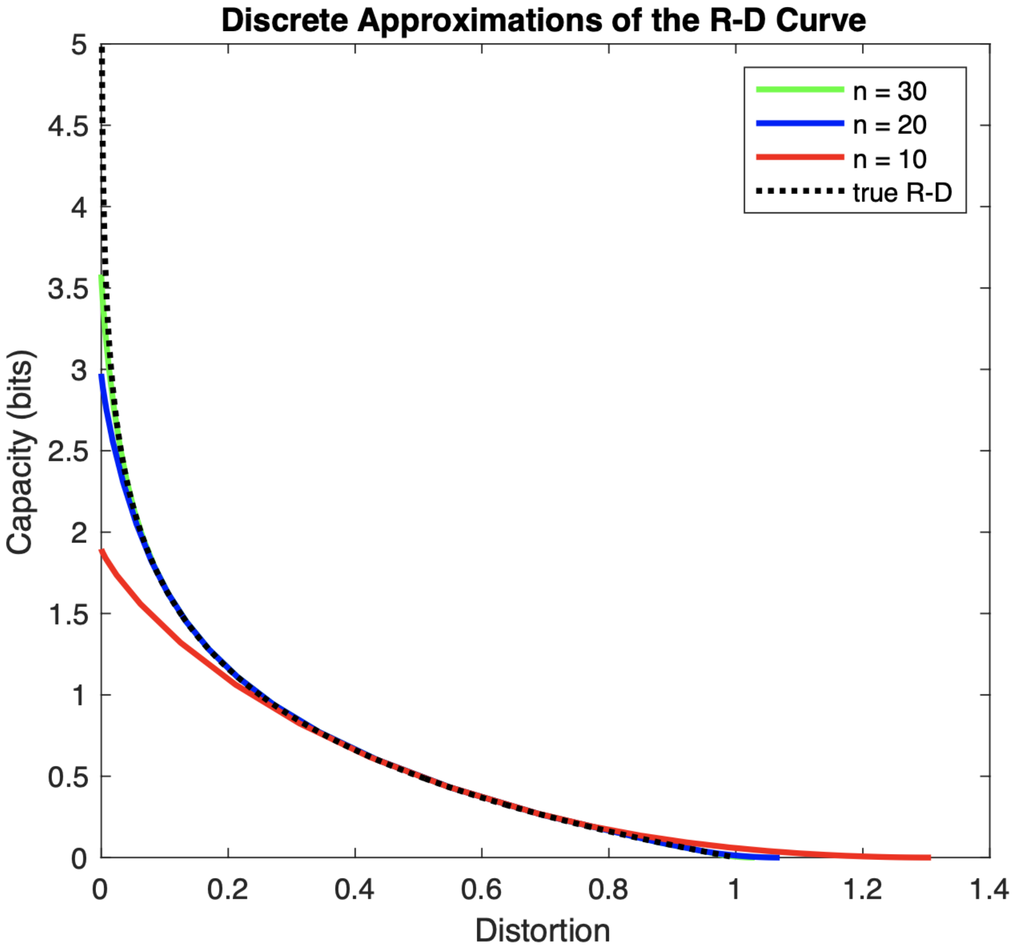


Figure 8. Rate Distortion curves generated by the Blahut algorithm for different alphabet sizes versus the analytic solution. The RDT setup is $p(x) = \mathcal{N}(x; 5, 1)$, $d(x, y) = (x - y)^2$. The black dotted line is the analytic true Rate Distortion curve. The other lines are generated by approximating this RDT setup with a discrete prior. For example, for $n = 30$, the input alphabet is defined to be 30 linearly spaced points from 0 to 10 (0 and 10 are chosen to be the endpoints of our alphabet because the prior is small outside of this range). The prior is a determined by taking the continuous prior at these 10 points, then normalizing so the sum is equal to 1. Then a collection of RDT-optimal channels are computed for this prior and cost using the Blahut algorithm (for different capacity limits). Finally, the capacity and distortion for these channels are plotted.

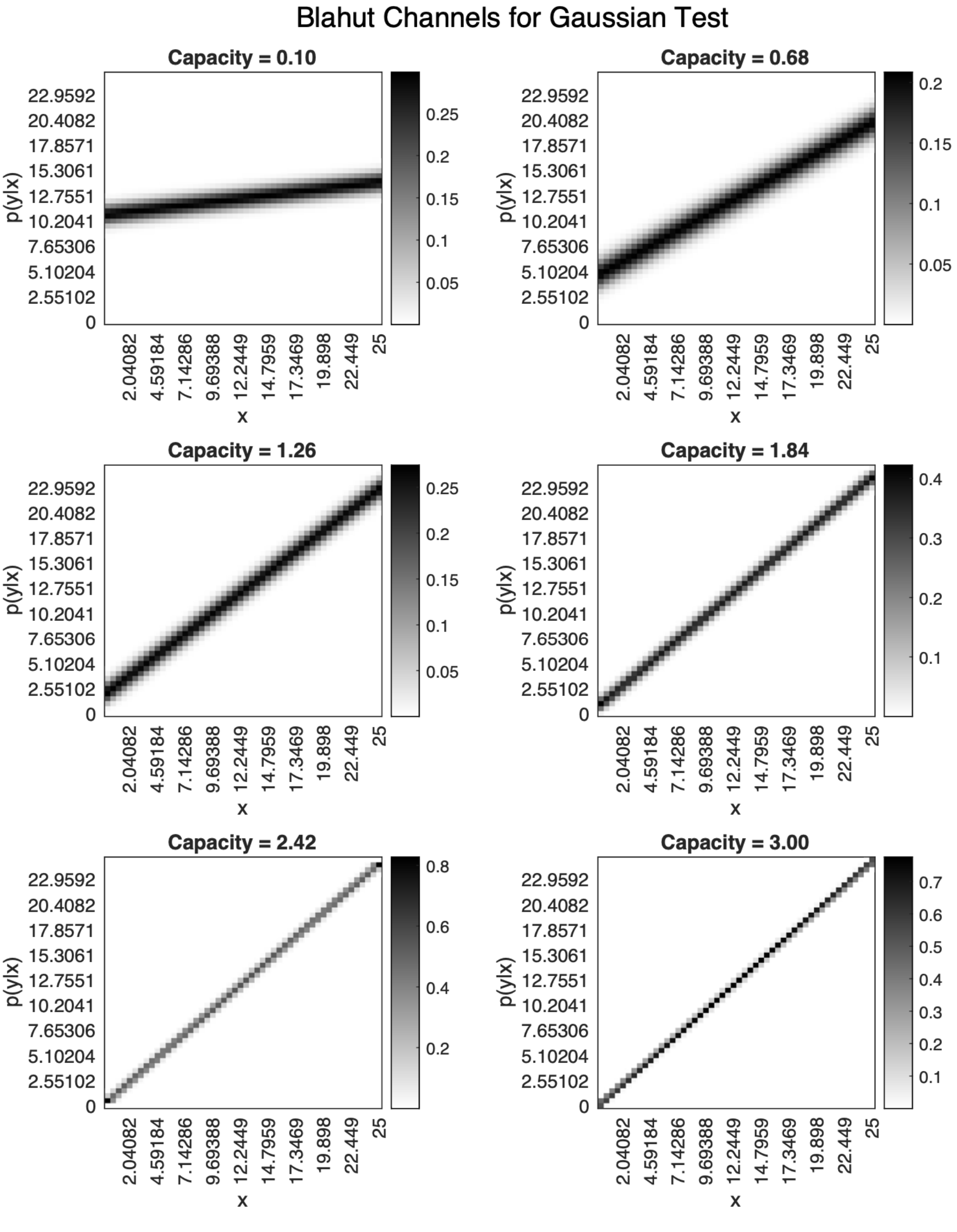
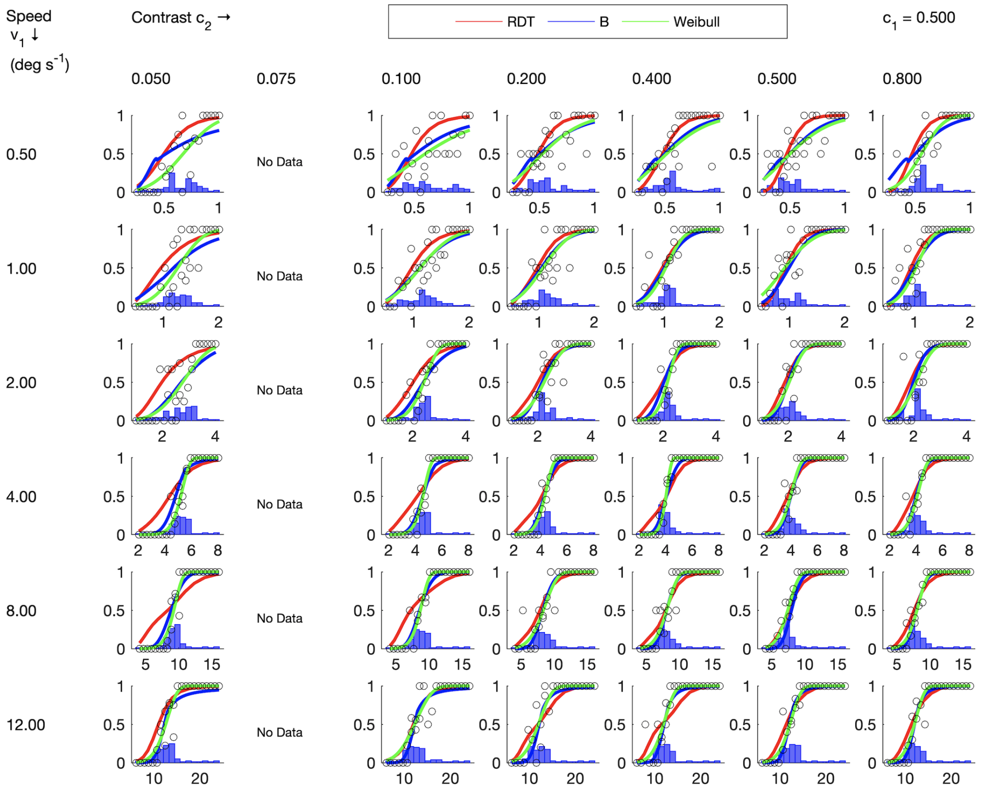


Figure 9. Blahut channels for a discrete approximation of the Gaussian test ($p(x) = \mathcal{N}(x; 5, 1)$, $d(x, y) = (x - y)^2$) with alphabet size of 50 at different capacities. We used this test to validate our approach of using a discrete approximation. For high capacity, the channels are tightly distributed around the input. As the capacity decreases, the channels shift closer to the prior. This is captured mathematically by the analytic solution for this case.

Appendix C

Psychometric Curves



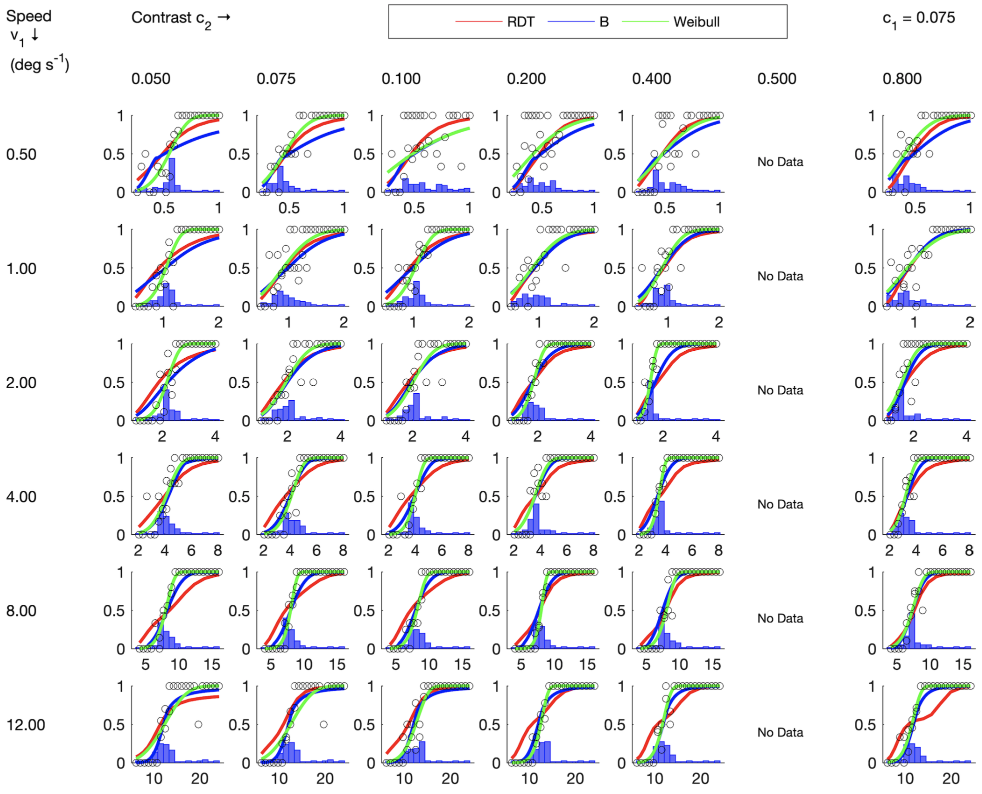


Figure 10. Subject data for the different experimental conditions along with with psychometric curves generated by RDT, B, and Weibull models.