Recurrent Auto-Encoding Drift Diffusion Model

MOENS Vincent^a, ZENON Alexandre^{a,b}

^a*IoNS, Université Catholique de Louvain
Brussels, Belgium*

^b*INCIA
Université de Bordeaux, France*

Abstract

The Drift Diffusion Model (DDM) is a popular model of behaviour that accounts for patterns of accuracy and reaction time data. In the Full DDM implementation, parameters are allowed to vary from trial-to-trial, making the model more powerful but also more challenging to fit to behavioural data. Current approaches yield typically poor fitting quality, are computationally expensive and usually require assuming constant threshold parameter across trials. Moreover, in most versions of the DDM, the sequence of participants' choices is considered independent and identically distributed (*i.i.d.*), a condition often violated in real data.

Our contribution to the field is threefold: first, we introduce Variational Bayes as a method to fit the full DDM. Second, we relax the *i.i.d.* assumption, and propose a data-driven algorithm based on a Recurrent Auto-Encoder (RAE-DDM), that estimates the local posterior probability of the DDM parameters at each trial based on the sequence of parameters and data preceding the current data point. Finally, we extend this algorithm to illustrate that the RAE-DDM provides an accurate modelling framework for regression analysis. An important result of the approach we propose is that inference at the trial level can be achieved efficiently for each and every parameter of the DDM, threshold included. This data-driven approach is highly generic and self-contained, in the sense that no external input (e.g. regressors or physiological measure) is necessary to fit the data. Using simulations, we show that this method outperforms *i.i.d.*-based approaches (either Markov Chain Monte Carlo or *i.i.d.*-VB) without making any assumption about the nature of the between-trial correlation of the parameters.

Keywords: Drift-Diffusion Model, Importance-Weighted Auto-Encoder, Amortized Inference,

Email address: vincent.moens@uclouvain.be (MOENS Vincent)

29 **1. Introduction**

30 Sequential sampling models (1) are models of behaviour that assume that decisions are the result of
 31 a noisy accumulation of evidence. The most famous of these models, called the Drift Diffusion Model
 32 (DDM), was introduced in the neuroscience field by Roger Ratcliff (2) as a tool for studying animal
 33 decision making and has led to an important amount of publications during the past decades.

34 The DDM is based on a Wiener process, where the evidence in favour of one of several (typically
 35 two) actions is represented by a particle moving towards a boundary with a certain drift rate
 36 and subject to a given noise. The simplest implementation of the DDM assumes the generative
 37 parameters of the model to be fixed across the different trials throughout the experiment. However,
 38 this simple version fails to account for some specific patterns of reaction times (RT), which led to
 39 subsequent improvements (3; 2; 4; 5; 6; 7) in which parameters are free to vary from trial to trial
 40 according to some assumed distribution, giving birth to what is known as the Ratcliff Diffusion
 41 Model (8), Extended DDM (9) or Full DDM (10).

42 While the Full DDM is more realistic (see (11) for a review and discussion), two main drawbacks of
 43 this method can be individuated: first, it is more computationally expensive to fit than its simpler
 44 counterpart by several orders of magnitudes, as a marginal probability density has to be estimated
 45 (or approximated) for each and every trial. So far, several methods used to fit the Full DDM to
 46 behavioural datasets have relied on Numerical Quadrature (NQ) (12; 13; 7; 14; 15) or Markov Chain
 47 Monte Carlo (16). NQ is computationally expensive, which precludes its use for large problems.
 48 While Markov Chain Monte Carlo (MCMC) (17) can be cheaper, it still requires an optimization
 49 time and memory storage that is proportional to the size of the problem: this precludes its use
 50 on very large datasets, which are encountered more and more frequently in cognitive neuroscience
 51 (18; 19; 20).

52 The second important drawback of the Full DDM regards its identifiability. Typically, only trial-to-
 53 trial variability in three (12; 7) or less (5; 21) of the four parameters of the DDM are assessed (drift,
 54 start-point and non-decision time, see below). Variability of the fourth parameter, the threshold,
 55 is usually regarded as non-identifiable, because its trial-to-trial variability would be equivalent to

the variability in the start point position (7; 22). This can be understood intuitively by observing that an increased variability for both these parameters leads to the same behavioural pattern, namely more frequent and faster errors wrt accurate responses. In this perspective, (23) suggest that the choice of the prior distribution shape of the start point (or threshold) determines whether the generative capacity of the model *for a given dataset*. However, the rationale for discarding threshold variability is mainly empirical and lack precise mathematical justification (7): indeed, the lack of identifiability pointed by (23) generalizes to every DDM parameter. We suggest that, even though several models can predict accurately the data at hand, they will differ in their capacity to be generalized to other data and in their model evidence value. Cross-validation and model evidence assessment will therefore be crucial to select model features that should be incorporated in the model. Here, we show that variability in threshold is instead identifiable together with the rest of DDM parameters, especially when the values of the threshold and start point at a specific trial can be predicted from past behavioural observations (time-dependency assumption) and when these values covary with other parameters or physiological measures (covariance assumption). These two restrictions are likely to be observed in practice, as we will see later. In these two cases, one can take advantage of the fact that the values of other variables (latent or observed) constrain the fit of the threshold variability. Although we will focus in the present paper on the first assumption, the approach we propose can model these two phenomena.

Nevertheless, even when the threshold is considered as fixed, identifiability of the Full DDM remains an issue (24). It is usually considered that one can expect the fit of the mean of the DDM parameters to be accurate (7), but not their variance (i.e. trial-to-trial variability) (12). In short, large datasets (e.g. > 100 trials per subject) take unreasonable amount of time to be optimized with these models, for a very limited quality of fit, which worsens dramatically with smaller datasets.

On top of this, another issue with many implementations of the DDM lies in the core assumption of the method used to fit the parameters to the data. Maximum Likelihood Estimate (MLE; i.e. finding the most likely value that the parameters would endorse in order to generate the data at hand) was traditionally used (12; 25; 26; 13; 27; 28; 7; 14), until Bayesian attempts were made (29; 30; 15). Crucially, the validity of the MLE approach is based on some asymptotic properties of the summary statistics of a given dataset when the amount of data tends to infinity (31). However, as we have seen, the computational cost of the numerical methods commonly used to optimize the

86 Full DDM prevents us from fitting large datasets. Hence, we face a dilemma: a large data set
87 will prevent us from computing the ML using classical NQ algorithms, and a small dataset set
88 will violate the assumptions of the MLE. Usage of MCMC only partly solves this problem, as its
89 computational cost is still a function of the size of the dataset, and since it still relies on a strong
90 *i.i.d.* assumption about the generative process.

91 The core idea of this paper is that trial-to-trial dependency of the parameter values and behavioural
92 results can, and should be used to help estimate the posterior probability of the latent variables.
93 This assertion is mainly theoretically justified by the observation that actions belonging to sequences
94 can not be considered as independent (see for instance (32; 33; 34; 35; 36; 37; 38)). Let us for instance
95 consider the toy example where a participant has been asked to press 9 times the same button (e.g.
96 right), and we are performing inference about the DDM parameters at the 10th button press. From
97 a Bayesian point of view, it is very likely that, at this trial, the subject will choose the same action
98 as before. If the participant is asked to choose, and indeed chooses, the right response, we can
99 expect that the RT at this trial will be, on average, lower than the one at the trials before (?):
100 this is a typical sequential effect. However, if the participant is asked to choose the left response at
101 this trial and responds accurately, it is even more likely that this response will be slower than the
102 previous trials, as the current action violates the previous pattern. Another example is post-error
103 slowing, a well-known psychological phenomenon usually interpreted as a change of speed-accuracy
104 trade-off following an error (39; 40). This might be modeled as a resetting of the threshold to a high
105 value following an error. But an error might also bias the decision in the next trial towards one of
106 the responses, or impair the attention of the subject during the subsequent trials. Our aim will be
107 to take advantage of these cross-correlations in a blind manner, i.e. without making any specific
108 assumption about the precise structure of the trial-to-trial correlation. This will be achieved by
109 capturing the most relevant trial-to-trial cross-dependencies with an assumption-free dimensionality
110 reduction algorithm.

111 To do this, we first show that Approximate Bayesian inference techniques can efficiently uncover the
112 parameter sequence: the algorithm we propose is a data-driven, nearly assumption-free model that
113 relaxes many classical assumptions, such as the parametric form of the parameter distributions (14)
114 at the within and between subject levels. Also, it can prevent overfitting by the use of dimensionality
115 reduction and it is scalable to large datasets. Finally, training of this model can be trivially

116 parallelized and accelerated using GPU.

117 More specifically, in this paper, we bring two nested solutions to the aforementioned problems:
 118 first, we show that the Full DDM finds a natural formulation in Bayesian Statistics (Section 2.2
 119 and Section 2.3), and that we can use this formulation to solve efficiently this problem using recent
 120 advances in Approximate inference, mostly with Variational Inference techniques (Section 2.4).
 121 We then present a computationally efficient Variational Bayes model (*i.i.d.*-VB) to compute the
 122 marginal probability density of the Full DDM. This algorithm has the advantage of being scalable,
 123 its computational cost and memory usage being amortized wrt. the size of the problem.

124 Next, we show how, when the *i.i.d.* assumption is relaxed, this basic algorithm can be extended
 125 to account for trial-to-trial dependency in a data-driven manner using a Recurrent Auto-Encoder
 126 (RAE-DDM) (Section 2.6). This approach is assumption-free, in that it does not require the user
 127 to select a specific prior distribution for the parameters nor a precise form of trial-to-trial cross-
 128 correlation.

129 We show in the results (Section 3) that our RAE approach outperforms classical approaches (*i.i.d.*-
 130 VB and Hamiltonian Monte Carlo, HMC) in terms of model evidence and fit of the data at the
 131 trial, subject and population level. An important and compelling result we provide with the RAE-
 132 DDM consists in showing that the trial-to-trial variability of the threshold parameter is possible to
 133 estimate together with the other parameters of the DDM.

134 In Section 4, we discuss our model in detail, highlight situations where it might prove useful, see
 135 its limitations and introduce potential improvements.

136 1.1. Related Work

137 Many routines have been proposed to optimize the DDM. Ratcliff and Tuerlinckx (12), and others
 138 (41; 42; 10) have used a NQ algorithm with a numerical estimate of the gradient to optimize the
 139 likelihood function at the subject level. The computational cost of this method is extremely high,
 140 justifying the development of less flexible, but simpler methods (43; 44; 8; 45). Vandekerckhove et
 141 al. (30) used a Markov Chain Monte Carlo (MCMC) technique to estimate each trial distribution
 142 in a fully Bayesian setting. This is the approach adopted in several Bayesian statistical softwares
 143 such as JAGS (46), WinBUGS (30), or Stan (47), which has broaden the use of this technique

144 for the optimization of the Full DDM in many recent publications (e.g. (48; 49; 50; 51; 52)).
 145 HDDM is a Python toolbox written by Wiecki and colleagues (15), aimed at fitting the DDM
 146 using MCMC. Peculiarly, this toolbox still relies on NQ to compute estimates at the lower level,
 147 rather than estimating each local posterior of the DDM parameters. Our aim will be to alleviate
 148 the computational burden of the integration process without compromising on accuracy and while
 149 keeping the entire original structure of the DDM.

150 The algorithm we propose differs in many ways from the work of Vandekerckhove and colleagues
 151 (30): first, the RAE-DDM is scalable, such that we can make inference about the generative model
 152 of large (or very large) datasets with limited cost, as the gradient of the lower bound to the log-model
 153 evidence is estimated stochastically. Our use of Amortized Variational Inference (53) allows us to
 154 control the convergence of the algorithm using a validation dataset, which is challenging with MCMC
 155 techniques. Finally, the RAE-DDM, besides its ability to account for trial-by-trial variability in
 156 the threshold values, can also provide a more precise estimate of the posterior distribution of local
 157 parameters, as it conditions this posterior on more information than just the RT, the choice and the
 158 prior by adding the previous trials as regressors for the value of the prior and local variable. None
 159 of the approaches presented so far took advantage of this cross-dependency between the values of
 160 latent parameters, choices and reaction times from trial to trial. We are not, however, the first to
 161 propose a form of time-dependant estimate of the behavioural parameters: (33; 34; 35; 54) have for
 162 instance proposed simpler, informative models to uncover precise psychological heuristics that were
 163 always fit using Maximum-Likelihood methods. The assumption-free nature of our model, traded
 164 with its latent parameter sparsity, makes it a compelling tool to give a precise and generalizable
 165 estimate of the expected value of the DDM parameters at each trial as well as the uncertainty of
 166 this estimate.

167 Also, (55; 21) proposed a model (the Neural Diffusion Model) where the neural data and behavioural
 168 parameters are governed by a common distribution that ties together the two kind of measures.
 169 Because the covariance of the two types of parameters is inferred, the fit can indirectly take into
 170 account the time evolution of the DDM parameters through the observed evolution of the neuro-
 171 physiological measure. Notably, this approach requires a large dataset (56) and inference is achieved
 172 with MCMC method. Also, it is still based upon a highly informative model, where the user has
 173 to set (and possibly test) precisely the form of the model that couples the parameters. Finally, this

174 model is complex and has a large number of parameters, making overfitting an issue.

175 Our approach, on the contrary, takes full advantage of the dimensionality reduction property of
 176 RAE to set the number of parameters of the model to a minimum. It also contrasts with the neural
 177 diffusion model by the fact that it is self-contained, meaning that no external input is needed,
 178 although it can easily be included in the model. It also relaxes greatly the need to set a precise
 179 form to the model that governs the data: as we will see, the use of non-linear mapping between the
 180 latent trial-specific prior and its DDM-mapped value flexibly models a large variety of prior shapes
 181 with a limited and controllable risk of overfitting.

182 Our approach to Bayesian inference relies also on a large body of previous work. The Wake-Sleep
 183 algorithm (57) developed almost twenty years ago by Peter Dayan constitutes a cornerstone in the
 184 history of Bayesian inference. This algorithm is a two-side recognition-generative model, which
 185 works in two phases: during the wake phase, the algorithm learns from the data the generative
 186 model by minimizing the error between the predictions and the observed data. During the sleep
 187 phase, the generative model is held fixed and is used to generate fake data that are used to train
 188 the recognition model. Built on the very same idea, Variational Auto-Encoders (VAE) (58) and its
 189 recurrent form (59; 60) have quickly become standard and popular tools to model data in many
 190 areas (61). Advances in this area go fast, and giving an accurate account of all of them is beyond the
 191 scope of this paper. The present paper builds on recent developments of Recurrent Auto-Encoders
 192 that have been concomitantly achieved by several groups (e.g. (62; 63)). In short, the first challenge
 193 with these techniques regards determination of the best approximate posterior shape for the latent
 194 variable $\mathbf{z}_{j,n}$ (64; 65; 66). An efficient solution to improve the performance of a VAE is to replace the
 195 Variational estimator by an Importance Weighted (IW) (67) estimator of the posterior probability.
 196 This simple trick can greatly improve the robustness and convergence of the model. We will show
 197 in what follows how this technique adapts to the problem of the DDM. Another issue consists
 198 in regularizing the parameters of the recognition and generative models: Dropout (68) has been
 199 shown to efficiently prevent overfitting with similar models (69), motivating our choice of relying
 200 on variational Dropout to perform inference about the value of the network weights (70).

201 A final remark concerns identifiability of the model we propose, and of the Full DDM in general. The
 202 impact of the shape of the prior used for inference has led to vivid debates in the DDM community.
 203 Some authors have claimed that a prior probability of an arbitrary complexity can always be found

in order to reflect perfectly the data, implying that the choice of the prior completely determines the capability of the DDM to predict data (23). For instance, following these authors' logic, also disputed by (71), a J^{th} component linear mixture model prior (where J is the number of trials of the dataset) over the sole parameter ξ can reflect perfectly the dataset, even if the noise parameter is set to 0 and other parameters are kept fixed. However, this overfitting issue applies to any hierarchical model where local variables are modeled: as a matter of fact, there should always exist an MLE estimate of any model that would predict perfectly the data at hand with little or no generalization capability. Indeed, two specific – and widely used – solutions exist to deal with overfitting issues. The first approach consists in using Bayesian Model Comparison (72; 73), in which the model evidence (i.e. probability of the data given the model marginalized over the space of parameters) of models of different complexity is compared. The second approach consists in assessing the model predictive capacity on a *test* dataset, separate from the *training* dataset that is used for fitting the model. In the above example, cross-validation of this ad-hoc model will obviously fail to show that the model is able to provide accurate insights about unseen data. Moreover, we propose another solution to this identifiability problem, which is that the values taken by the DDM parameters at each trial is only random to some extent, and can, at least partially be predicted based on the values that these parameters (and related data or regressors) took in the past trials. The approach we propose can achieve these three goals efficiently, as the model is scalable (i.e. it is able to generalize its predictions to unseen data), provides a lower-bound to the log-model evidence, that can be used for Bayesian Model Comparison and can account for autocorrelation of the parameters and data in a blind manner.

Furthermore, the method we propose alleviates the need to specify a precise prior probability distribution, but, by the use of dimensionality reduction, it ensures that the trial-wise latent variable $\mathbf{z}_{j,n}$ captures a maximum of the variability, thus allowing us to keep a highly generalizable model structure with little or no cost caused by the specific prior specifications.

Finally, the capacity of our approach to estimate the threshold variability is a considerable benefit, which has never been considered before. Other authors (e.g. (40; 49)) have used a regression model to estimate the threshold value at the trial level using a measured neurophysiological signal or experimental condition. The RAE-DDM does not preclude the inclusion of these parameters in the model, but would most likely provide a better fit because it also could account for the trial-to-trial

234 dependency in the parameter values that might not be captured by the external regressor.

235 2. Methods

236 2.1. Notation

237 It is convenient to first define the notation we will use. In the following, we will refer to the trial
 238 index as $j \in 1 : J$ and the subject index as $n \in 1 : N$. The trial-specific configuration of the
 239 five DDM parameters, namely the threshold a , the drift-rate ξ , the relative start point w and
 240 the non-decision time τ will be grouped under a single vector $\omega_{j,n}$ (or, if applicable ω_n if it is
 241 assumed that $\omega_{j,n} = \omega_{j',n} \forall j, j' \in 1 : J$). The subject-specific prior distribution parameters of
 242 $\omega_{j,n}$ will be generically grouped under the parameter vector θ_n . As we will see in 2.4, we will
 243 also define an approximate posterior distribution over $\omega_{j,n}$ with parameters $\delta_{j,n}$. In the *i.i.d.*-VB
 244 implementation we propose, these two sets of parameters will have a specific form ($\theta_n \triangleq \{\mu_n, \sigma_n^2\}$
 245 and $\delta_{j,n} \triangleq \{\mu_{j,n}^\omega, \Sigma_{j,n}^\omega\}$). For the purpose of sparsity of the notation and/or to keep the discourse as
 246 general as it may, we will use the generic indicators more often than their implementation-specific
 247 counterparts.

248 The behavioural observations are grouped under $\mathbf{y}_{j,n} \triangleq \{t_{j,n}, c_{j,n}\}$, where $t_{j,n}$ and $c_{j,n}$ are the
 249 corresponding Reaction time (RT) and choice, respectively. Note that, for the purpose of generality,
 250 we distinguish the model boundary b and the choice c , although we will freely use $p(t, c)$ and
 251 $p(t, b)$ interchangeably following the context. In general, capital letters will represent the set of
 252 all the corresponding parameters or random variables (i.e. $\mathbf{Y} \triangleq \{\mathbf{y}_{j,n} \text{ for } j \in 1 : J, n \in 1 : N\}$,
 253 $\Theta \triangleq \{\theta_n \text{ for } n \in 1 : N\}$ etc.). The k^{th} draw of a sample will be indicated by a tilted symbol,
 254 such as $\tilde{\omega}_{j,n,k}$, and the true (and obviously unknown) value of a parameter will be indicated by
 255 $\hat{\omega}_{j,n}$.

256 Finally, if a parameter is deterministically determined by some other variables, such as $\delta_{j,n} =$
 257 $f(\mathbf{y}_{j,n}; \rho)$, we will use interchangeably this notation with the sparser notation $\delta_{j,n}(\mathbf{y}_{j,n}; \rho)$.

258 2.2. DDM

259 The Drift-Diffusion model is a member of the sequential sampling model family. The essence of
 260 these models is that the experimenter models the decision process as an accumulation of evidence

in favour of one of the available responses: the probability of a pair of observations $\{c_{j,n}, t_{j,n}\}$ is formulated as the probability that the evidence reached the threshold b at time t . Importantly, this accumulation of evidence is stochastic: given the same initial state, the same agent will produce a different response time and, possibly, choose a different item. This stochasticity is the main asset of this model (71), since it can explain speed-accuracy trade-offs: indeed, certain parameter settings will produce faster but more entropic response distributions, whereas other will produce slower, more directed responses.

Formally, in the DDM, the evidence accumulation process is assumed to follow a Wiener process (Figure 1):

$$dx = \xi dt + s dW$$

where dx is the evidence step, dt is the time step and $s dW$ a Gaussian noise with a standard deviation s . In this specific model, the agent accumulates evidence in favour of an action c_1 with a rate ξ , called the drift rate, meaning that in the case of a classical Two-alternative forced choice task, she will also accumulate evidence in favour of the alternative action c_2 with a rate $-\xi$. Once the evidence crosses one of the two absorbing boundaries (the boundary b positioned at a distance a corresponding to the action c_1 or the boundary $-b$ at position 0 corresponding to the action c_2), the corresponding action is selected and, after some physiological delay τ , this action is executed. The evidence accumulation can start wherever at $z_0 \in [0; a]$, although it is more convenient to model the relative start point $w \in [0; 1]$ where $w = z_0/a$. Finally, this 5-parameter model is non-metric: multiplying ξ , a , z_0 and s by any positive scalar does not change the behavioural predictions. It is therefore necessary to fix one of them for the model to be used: The accumulation noise $s = 0.1$ is usually used as a metric, but, for the sake of mathematical explicitness and simplicity, we will use the metric $s = 1$ in the following.

Fortunately, the First-Passage Time density of the Wiener process (*Wiener FPT density*) described before has a closed-form formula. For a Two-alternative forced choice task, the joint probability density of crossing a given boundary b (the lower boundary by convention) at a certain point in

RT

Figure 1: DDM structure. The x axis represents the time scale, the y axis the evidence level. Four parameters (threshold a , drift ξ , starting point w and non-decision time τ) are represented. At the two boundaries, the first passage probability density function (sometimes called defective density) is represented. Three accumulation trajectories are displayed.

time t , is given by (74):

$$Wiener(t, b; \omega) = \begin{cases} \frac{\pi}{a^2} \exp(-\xi a w - \frac{\xi^2(t-\tau)}{2}) \sum_{k=1}^{\infty} k \exp(-\frac{k^2 \pi^2(t-\tau)}{2a^2}) \sin(k\pi w) & \text{if } t > \tau \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

If the absorbing boundary is $-b$ (i.e. the upper boundary), the above equation can be adjusted by setting $w' = 1 - w$ and $\xi' = -\xi$. Together, the sum of the First Passage Time Densities at the two boundaries (sometimes called defective probability densities (6; 21)) forms a valid probability distribution, i.e.:

$$\sum_{b \in \mathbf{b}} \int_{t=\tau}^{\infty} Wiener(t, b; \omega) dt = 1 \quad (2)$$

and $Wiener(t, \mathbf{b}; \omega) \geq 0$ for $t \in [\tau, \infty]$ and $b \in \mathbf{b}$

where $\mathbf{b} = \{b, -b\}$. The function in Equation (1) involves an infinite sum, and must therefore be implemented through approximations. Navarro et al. (75) provided a method to truncate this

infinite sum to the appropriate number of terms κ as a function of truncation error. Here, we used their approximation with a numerical precision of 10^{-32} to estimate the parameters value.

The simplest and most intuitive way to fit the DDM to behavioural data is to look for the parameter setting of ω_n that maximizes the likelihood of observing this dataset given the parameters and the model. If tempting, this approach overlooks a crucial detail of the DDM: if the true value of ω varies from trial-to-trial, the fit of a single vector ω_n^* is a biased estimator of $\mathbb{E}[\hat{\omega}_{j,n}]$, and this bias can have an important impact on the interpretation of the parameter values (5). Formally,

$$\mathbb{E}[\omega_{j,n} \text{ for } j \in 1 : J] \neq \omega_n^*$$

where ω_n^* is the value ω_n that maximizes the log-likelihood function.

For instance, a high variability in the value of the starting point w with mean .5 will generally produce a dataset that will be better fitted by a point-estimate of $w^* < .5$ (5). Accounting for the trial-to-trial variability in w can solve this problem. The problem then is to fit a prior θ_n that determines the trial-wise distribution of latent parameters $\omega_{j,n}$. In frequentist paradigms, the parameters $\omega_{j,n}$ are usually integrated out of the model, and θ_n only is fitted. As we will see in 2.3, in a fully Bayesian framework, all the parameters ($\omega_{j,n}$, θ_n and its prior) can be regarded as latent random variables and are estimated accordingly.

Usually, only the subset $\{\xi, w, \tau\}$ is considered to vary from trial to trial, and the threshold is considered to be fixed for each participant. Although it may raise some concerns in readers familiar with the full DDM, we do not make this assumption here, as it is customary in Bayesian statistics and machine learning to consider that each datapoint is generated by its own set of parameters (see for instance (76)). We show in the results section that the performance of the model we propose still outperforms other implementations, and that the fit of the threshold provides meaningful results.

2.3. The Full DDM as an Empirical Bayes problem

The MLE of the Full DDM as formulated by (12) consists in finding the value of the prior θ_n^* that maximizes the log-likelihood function $\log p(\mathbf{y}_{j,n}; \theta_n)$ while having marginalized out the trial-specific parameters $\omega_{j,n}$ (and therefore the posterior probability under an unnormalized, unbounded

uniform prior), and to consider this value as a consistent estimator of the true value of θ_n , namely $\hat{\theta}_n$. This choice can be justified from the asymptotic normality of the posterior distribution theorem (31; 77): as the dataset grows, and considering that the data are generated from the same process, the posterior distribution approaches a multivariate Gaussian distribution with a mean close to $\hat{\theta}_n$ and covariance matrix equal to the inverse of the Observed Fisher Information matrix, i.e. $\theta_n^* \rightarrow \hat{\theta}_n$ as $NJ \rightarrow \infty$ and $p(\theta_n) \approx \mathcal{N}(\theta_n^*, I(\theta_n^*)^{-1})$, where $I(\theta_n^*) = -\nabla_{\theta_n^*}^2 \log p(\mathbf{y}_{j,n}; \theta_n)$. Also, following the Bernstein-von Mises theorem (78; 79; 80), we can see that the choice of the prior becomes irrelevant to the posterior as the amount of data grows, which further justifies the MLE technique for large datasets.

We show in Appendix B that the MLE approach of the Full DDM reduces to what is known as an empirical Bayes estimator of $\hat{\theta}_n$: indeed, the value of θ_n that maximizes the Wiener density can be directly reformulated as a function of the posterior probabilities of each trial parameter $p(\omega_{j,n} | \mathbf{y}_{j,n}; \theta_n)$: this gives the intuition that the problem of fitting the Full DDM could be greatly simplified if one could derive an efficient technique to estimate the set of the trial-wise posterior probability density functions. Figure 2a shows the basic structure of the Full DDM in this perspective.

2.4. Variational Inference

In this section, we will present Variational Inference as a method to find the posterior of the trial-wise DDM parameters $\omega_{j,n}$. We will first describe the most common EM approach to VB (whose derivation for the DDM application is provided in Appendix B), because it is widely used in neuroscience (81; 82; 83; 84; 85). This EM algorithm consists in looping between the evaluation of each trial-specific posterior probability density function given the current estimate of θ_n , before then finding the value of θ_n that maximizes the likelihood (or the lower-bound to the log-model evidence in a fully Bayesian setting) given the current posterior estimates. We will expose its weaknesses, and present another, neater method to maximizes the loss function $\ell(\mathbf{y}_{j,n}; \theta_n)$ that relies on Stochastic Gradient Variational Bayes (SGVB) and inference networks.

More specifically, the EM algorithm relies on computing the posterior over Ω_n (the set of all trial-

by-trial DDM parameter values), which reads

$$p(\Omega_n | \mathbf{Y}_n) = \frac{\prod_{j=1}^J p(\mathbf{y}_{j,n} | \omega_{j,n}) p(\omega_{j,n} | \theta_n)}{p(\mathbf{Y}_n | \theta_n)},$$

assuming that the posterior probability factorizes over each datapoint. This posterior probability is intractable due to its normalizing factor (the model evidence) $p(\mathbf{Y}_n | \theta_n)$. A first approach could be to use simulation techniques to approximate this posterior density function, but the size of the problem precludes this choice, since we would need to compute NJ independent integrals at each step of the EM algorithm. Instead, we can make use of Variational Bayes techniques (VB) (86; 87; 17; 88) to fit an approximate posterior to the true posterior. In the present case, in which we are interested in the MLE value of the prior, this approach is called Variational EM, or VEM.

In short, VB works as follows: we first define a proxy to the posterior density of an arbitrary shape $q(\omega_{j,n}; \delta_{j,n})$, where $\delta_{j,n}$ is a general indicator of the variational parameters that we will optimize in order to match the real posterior as close as possible (see Figure 2b). In order to do so, we can choose to minimize the Kullback-Leibler (KL) divergence between this distribution and the true posterior, since this measure is minimized when the two distributions match exactly:

$$\begin{aligned} \text{if } D_{\text{KL}}[q(\omega_{j,n}; \delta_{j,n}) || p(\omega_{j,n} | \mathbf{y}_{j,n})] &= 0 \\ &\triangleq \int q(\omega_{j,n}; \delta_{j,n}) \log q(\omega_{j,n}; \delta_{j,n}) d\omega_{j,n} - \\ &\quad \int q(\omega_{j,n}; \delta_{j,n}) \log p(\omega_{j,n} | \mathbf{y}_{j,n}) d\omega_{j,n} \\ \text{then } q(\omega_{j,n}; \delta_{j,n}) &= p(\omega_{j,n} | \mathbf{y}_{j,n}). \end{aligned} \tag{3}$$

The first term of the KL divergence is the negative differential entropy of $q(\omega_{j,n}; \delta_{j,n})$:

$$H[q(\omega_{j,n}; \delta_{j,n})] \triangleq - \int q(\omega_{j,n}; \delta_{j,n}) \log q(\omega_{j,n}; \delta_{j,n}) d\omega_{j,n}.$$

For the second term, we can expand $\log p(\omega_{j,n} | \mathbf{y}_{j,n}) = \log p(\mathbf{y}_{j,n}, \omega_{j,n}) - \log p(\mathbf{y}_{j,n})$, and, considering that $\int q(\omega_{j,n}) \log p(\mathbf{y}_{j,n}) d\omega_{j,n} = \log p(\mathbf{y}_{j,n})$, the value of $\delta_{j,n}$ that minimizes the KL divergence simplifies to:

$$\arg \min_{\delta_{j,n}} \underbrace{-H[q(\omega_{j,n}; \delta_{j,n})] - \int q(\omega_{j,n}; \delta_{j,n}) \log p(\mathbf{y}_{j,n}, \omega_{j,n}) d\omega_{j,n}}_{-\mathcal{L}(q(\omega_{j,n}; \delta_{j,n}))} + \log p(\mathbf{y}_{j,n})$$

358 which implies that

$$\begin{aligned} D_{\text{KL}}[q(\boldsymbol{\omega}_{j,n}; \boldsymbol{\delta}_{j,n}) || p(\boldsymbol{\omega}_{j,n} | \mathbf{y}_{j,n})] &= 0 \\ \Leftrightarrow \mathcal{L}(q(\boldsymbol{\omega}_{j,n}; \boldsymbol{\delta}_{j,n})) &= \log p(\mathbf{y}_{j,n}) \end{aligned} \quad (4)$$

359 and therefore

$$\mathcal{L}(q(\boldsymbol{\omega}_{j,n}; \boldsymbol{\delta}_{j,n})) = \mathbb{E}_{q(\boldsymbol{\omega}_{j,n}; \boldsymbol{\delta}_{j,n})} [\log p(\mathbf{y}_{j,n}, \boldsymbol{\omega}_{j,n})] + H[q(\boldsymbol{\omega}_{j,n}; \boldsymbol{\delta}_{j,n})] \quad (5)$$

360 is a lower bound to the log model evidence (or ELBO, for Evidence Lower Bound, also sometimes
361 called negative Free Energy). In essence, Equation (4) tells us that the value of $\boldsymbol{\delta}_{j,n}$ that minimizes
362 $D_{\text{KL}}[q(\boldsymbol{\omega}_{j,n}; \boldsymbol{\delta}_{j,n}) || p(\boldsymbol{\omega}_{j,n} | \mathbf{y}_{j,n})]$ is the one that maximizes the ELBO $\mathcal{L}(q(\boldsymbol{\omega}_{j,n}; \boldsymbol{\delta}_{j,n}))$. Once this
363 value of $\boldsymbol{\delta}_{j,n}$ has been optimized, we can consider that $q(\boldsymbol{\omega}_{j,n}; \boldsymbol{\delta}_{j,n})$ is as close as it could from
364 $p(\boldsymbol{\omega}_{j,n} | \mathbf{y}_{j,n})$ given the shape of the distribution we have chosen, the prior probability $p(\boldsymbol{\omega}_{j,n})$, the
365 data $\mathbf{y}_{j,n}$ and the likelihood function $p(\mathbf{y}_{j,n} | \boldsymbol{\omega}_{j,n})$. Most approaches (89; 90; 91; 92; 93; 94) rely
366 on some form of (possibly natural) gradient ascent method to solve this problem: the gradient of
367 the ELBO wrt. $\boldsymbol{\Delta} = \{\boldsymbol{\delta}_{j,n} \text{ for } j \in 1 : J, n \in 1 : N\}$ is computed, and the optimization proceeds by
368 either optimizing all the variational parameters altogether, or by sequentially optimizing some of
369 them while keeping the others fixed in an EM-like approach. VEM is a special case of the latter,
370 where we are interested in the MLE of the prior.

371 This VEM method suggested above, where $\boldsymbol{\Delta}$ and $\boldsymbol{\theta}_n$ are sequentially optimized, suffers from several
372 issues: the first is that optimizing the approximate posterior for each of the local variables is not
373 trivial, as the ELBO does not have, in general, a closed-form formula (see Appendix B). The second
374 problem is that this approach is highly sensitive to local optima: in other words, the initial draw
375 of $\boldsymbol{\Delta}$ can frame drastically the final result of the optimization process (17). Another important
376 concern is that the cost of such optimization process is unreasonably high, as one needs to iterate
377 across every datapoint in order to make a single update of the prior, which is not guaranteed to
378 be significant. All these obstacles find an elegant solution with the combined use of SGVB and
379 Inference Networks.

380 2.4.1. Stochastic Gradient Descent, Reparametrization trick and SGVB

Let us rewrite the optimization problem we need to solve to find the MLE of θ_n using VB in a concise form:

$$\Theta^* = \arg \max_{\Theta} \max_{\Delta} \frac{1}{NJ} \sum_{n=1}^N \sum_{j=1}^J \mathbb{E}_{q(\omega_{j,n}; \delta_{j,n})} [\log p(\mathbf{y}_{j,n}; \omega_{j,n}) + \log p(\omega_{j,n}; \theta_n)] + H[q(\omega_{j,n}; \delta_{j,n})].$$

The VEM approach detailed above essentially reduces to maximizing the ELBO wrt. Θ and Δ . Intuitively, a more economical way to achieve this would be therefore to maximize both parameters at the same time. The gradient of this expression can be expressed exactly as

$$\nabla_{\Theta, \Delta} \{\mathcal{L}(q(\Omega))\} = \nabla_{\Theta, \Delta} \left\{ \frac{1}{NJ} \sum_{n=1}^N \sum_{j=1}^J \mathbb{E}_{q(\omega_{j,n}; \delta_{j,n})} [\log p(\mathbf{y}_{j,n}; \omega_{j,n}; \theta_n)] + H[q(\omega_{j,n}; \delta_{j,n})] \right\}.$$

With a little loss of generality, we consider that the entropy of the approximate posterior has a closed-form formula, and that samples can be generated from $q(\omega_{j,n}; \delta_{j,n})$ using a differentiable function g st. $\tilde{\omega}_{j,n,k} = g(\delta_{j,n}, \epsilon)$ where $\epsilon \sim p(\epsilon)$ for some arbitrary $p(\epsilon)$. Under these conditions, the above expression has an unbiased, noisy Monte-Carlo estimator given by

$$\begin{cases} \nabla_{\Theta} \{\mathcal{L}(q(\Omega))\} & \approx \frac{1}{NJ} \sum_{n=1}^N \sum_{j=1}^J \frac{1}{K} \sum_{k=1}^K \nabla_{\theta_n} \{\log p(\tilde{\omega}_{j,n,k}; \theta_n)\} \\ \nabla_{\Delta} \{\mathcal{L}(q(\Omega))\} & \approx \frac{1}{NJ} \sum_{n=1}^N \sum_{j=1}^J \frac{1}{K} \sum_{k=1}^K \nabla_{\delta_{j,n}} \{g(\delta_{j,n}, \epsilon_k)\} \nabla_{\tilde{\omega}_{j,n,k}} \{\log p(\mathbf{y}_{j,n}, \tilde{\omega}_{j,n,k}; \theta_n)\} + \\ & \nabla_{\delta_{j,n}} \{H[q(\omega_{j,n}; \delta_{j,n})]\} \end{cases}$$

for a number of K samples. The form we have given to the (expected) log-likelihood makes it clear that

$$\tilde{\nabla}^{\mathcal{L}} \triangleq \frac{1}{SUK} \sum_{n=1}^S \sum_{j=1}^U \sum_{k=1}^K \begin{bmatrix} \nabla_{\theta_n} \{\log p(\tilde{\omega}_{j,n,k}; \theta_n)\} \\ \nabla_{\delta_{j,n}} \{g(\delta_{j,n}, \epsilon_k)\} \nabla_{\tilde{\omega}_{j,n,k}} \{\log p(\mathbf{y}_{j,n}, \tilde{\omega}_{j,n,k}; \theta_n)\} + \nabla_{\delta_{j,n}} \{H[q(\omega_{j,n}; \delta_{j,n})]\} \end{bmatrix}$$

381 is an unbiased, noisy estimator of $\nabla_{\Theta, \Delta} \{\mathcal{L}(q(\Omega))\}$, for any $S \leq N$ and $U \leq J$.

382 The use of the auxiliary variable to compute the gradient estimator is sometimes called Infinitesimal
383 Perturbation Analysis estimator (95; 96; 97) or Reparametrized Gradient (98). It has generally a
384 much lower variance than the gradient obtained with the Score function method (98; 93), and, for
385 a large enough mini-batch size $US \gg 0$, it can be used with a single sample $K = 1$.

386 This kind of noisy gradient can be used to find the minimum of a loss function using Stochastic
387 Gradient Descent (SGD) techniques (99), which have numerous advantages. First, as our notation

suggests, we can subsample some trials in the whole set in order to maximize the ELBO wrt. both Θ and Δ . This already amortizes the cost of the optimization, as we do not need to go through the whole dataset to make a step toward the maximum. Second, we do not need to compute the true posterior probability of each of the sampled trials in order to estimate the gradient of the respective variational parameters $\delta_{j,n}$: by sampling from this approximate posterior and computing the gradient of this sample, we can estimate the gradient of the ELBO wrt. $\delta_{j,n}$.

Finally, as for the VEM approach, there exists a risk of ending in a local optimum. This risk is, however, mitigated by the use of SGD: as only part of the data are used at each time step to climb the likelihood curve, local peaks will usually be avoided because they are caused by a small cluster of data that is partially ignored during this step (100).

Many SGD algorithms have been developed (e.g. (101; 102)). As our aim is not to make an extensive review of their respective advantages and disadvantages, we decided to use the Adam optimizer (58) and refer to the corresponding paper for a detailed algorithmic description and implementation.

2.4.2. Inference Network

Flexibility and sparsity are two of the main desired properties of the approximate posteriors q that we want to fit to our dataset: a loss of flexibility will decrease the performance of the model, or equivalently, worsen the unknown KL divergence between the approximate and true posteriors. Sparsity is however the main asset of Variational Inference, and one would wish to find a cheap form of approximate posterior to perform inference in a reasonable amount of time. A solution to reduce the number of variational parameters consists in using a direct, deterministic mapping from the values of the adjacent nodes of factor graph to a given approximate posterior $\delta_{j,n} \triangleq f(\mathbf{y}_{j,n}, \boldsymbol{\theta}_n; \boldsymbol{\rho})$ where $\boldsymbol{\rho}$ is the vector of parameters of the arbitrary function f that achieves this mapping. This leads the approximate posterior to have the form $q_{\boldsymbol{\rho}}(\mathbf{z}_{j,n} | \mathbf{y}_{j,n}, \boldsymbol{\theta}_n)$ (Figure 2c). Our goal has now become to optimize the mapping function parameters $\boldsymbol{\rho}$ instead of the local variational posteriors $\delta_{j,n}$ which it determines.

This use of an inference network (IN), known as Neural Variational Inference and Learning (53) or amortized variational inference (AVI) (64) has led to numerous applications (93; 103; 64; 70; 104; 65; 105), as it has several advantages. First, from a computational point of view, it alleviates the need

of optimizing each of the local approximate posteriors, which can be problematic for large datasets. Instead, a generic function that maps the points to their approximate posterior, is fitted to the problem at hand. Second, this technique allows the model to find a likely posterior for datapoints that might not have been seen yet, or that have been seen many iterations ago. Finally, and following the same logic, algorithms using IN can spare large amount of memory usage by storing only the parameters of the mapping ρ and not each of the local variational parameters.

Optimizing the ELBO with an IN is straightforward: using the above definition of the approximate posterior, the noisy approximation to $\nabla_{\Theta, \rho} \{\mathcal{L}(q(\Omega))\}$ becomes

$$\tilde{\nabla}^{\mathcal{L}} \triangleq \frac{1}{SUK} \sum_{n,j,k}^{S,U,K} \begin{bmatrix} \nabla_{\theta_n} \{g(f(\mathbf{y}_{j,n}, \theta_n; \rho), \epsilon_k)\} \nabla_{\tilde{\omega}_{j,n,k}} \{\log p(\mathbf{y}_{j,n}, \tilde{\omega}_{j,n,k}; \theta_n)\} + \\ \nabla_{\theta_n} \{H[q(\mathbf{z}_{j,n}; f(\mathbf{y}_{j,n}, \theta_n; \rho))]\} \\ \nabla_{\rho} \{g(f(\mathbf{y}_{j,n}, \theta_n; \rho), \epsilon_k)\} \nabla_{\tilde{\omega}_{j,n,k}} \{\log p(\mathbf{y}_{j,n}, \tilde{\omega}_{j,n,k}; \theta_n)\} + \\ \nabla_{\rho} \{H[q(\mathbf{z}_{j,n}; f(\mathbf{y}_{j,n}, \theta_n; \rho))]\} \end{bmatrix} \quad (6)$$

where $\tilde{\omega}_{j,n,k} = g(f(\mathbf{y}_{j,n}; \rho), \epsilon_k)$. The gradient in Equation (6) can be computed easily using an Automatic Differentiation routine if $g(\cdot)$ is differentiable, which excludes for instance random number generation algorithms based on acceptance-rejection algorithms (106; 107), although this limitation can be circumvented (see (108; 109)).

Approximate Posterior specifications. We now need to specify a convenient shape for the approximate posterior of the DDM parameters. In Appendix C, we opted for the use of Normalizing Flows (69) as a highly flexible albeit simple distribution shape. We found that, for the DDM, this form of posterior distribution outperformed generally simpler approximate posterior (such as Gaussian distributions) in terms of KL divergence to the log-model evidence.

2.5. Preliminary summary

So far, we have presented a method to compute the maximum likelihood of Θ using Variational Inference. Through the use of SGVB, we propose an optimization method that has a limited cost for big datasets: indeed, the number of samples per gradient estimation $S * U * K$ depends more on the complexity of the problem than on the size of the dataset. Also, it alleviates the need to compute the marginal likelihood to get a gradient estimate. Finally, it should be robust to the presence of local optima, thanks to the use of AVI, although this cannot be guaranteed. The

Figure 2: Directed Acyclic Graphical Model (DAG) of the AVI approach to the Maximum-Likelihood (approximate empirical Bayes) Full DDM problem. Black filled circles represent the data (RT and choices), white circles the latent variables, and the greyed circles the subject-specific priors. *a.* Full DDM as a empirical Bayes problem. The aim is to find the value of θ_n that maximizes the marginal likelihood $p(\mathbf{y}_{j,n}|\theta_n)$, where $\omega_{j,n}$ has been marginalized out. This problem reduces to an empirical Bayes problem. *b.* VB can provide a method to find a lower-bound to the log-marginal probability, by the use of approximate posteriors $q(\omega_{j,n}|\delta_{j,n})$. *c.* Amortized Variational Inference (AVI) amortizes the cost of inferring the approximate posterior for each latent variable by using an Inference network with parameters ρ that takes as input the datapoint and the prior in order to deterministically predict the approximate posterior. Note that, contrary to many similar approaches, here θ_n is taken as input to reflect the fact that, with a different prior, the posterior of $\omega_{j,n}$ would have a different shape even if the data $\mathbf{y}_{j,n}$ were identical.

use of IN scales down the computational burden of the problem, and ensures a certain similarity between the approximate posteriors of each datapoint. Moreover, it makes it possible to optimize the algorithm online, as the same mapping function as the one used for observed datapoints can be used for datapoints that have not been observed yet.

Comparison of Algorithm 1 and Algorithm 2 (see below) highlights the difference between the VEM method and AVI. Note that, in the former, the exact posterior probability from Equation (B.6) has been replaced by its approximation q whereas in the latter, the update have been formulated in terms of gradient *ascent* for simplicity of the notation.

Of course, the naive implementation presented in Algorithm 2 does not render the full richness of SGVB: the stopping criterion can, for instance, be refined st. the optimization process is stopped when the performance on a validation dataset starts to decrease (early stopping (17)). This is the method we will present in the Result section 3. Also, the learning rate η can be adapted to match an approximation of a preconditioning matrix, which is indeed what the Adam optimizer, mentioned above, does.

453

Algorithm 1: VEM algorithm.

Data: Choices and reaction times $\mathbf{Y}$

input: threshold u , maximum number of iteration

M

Result: Variational Parameters Δ ,

Maximum-likelihood estimate of Θ

```

1 Initialize randomly  $\Theta, \Delta$ ;
2 Set  $i = 0$ ;
3 Set  $ELBO_0 = -\infty$ ;
4 repeat
5    $i += 1$ ;
6   for  $n \in 1 : N, j \in 1 : J$  do // E step
7     Find  $\arg \min_{\delta_{j,n}} -\mathcal{L}(q(\mathbf{z}_{j,n}; \delta_{j,n}))$  // e.g.
      using SGD
8   for  $n \in 1 : N$  do // M step
9     Set  $\mu_n^* = \frac{1}{T} \sum_{t=1}^T \mathbb{E}_q[\mathbf{z}_{j,n}]$ 
10    Set  $\sigma_n^{2*} =$ 
       $\frac{1}{T} \sum_{t=1}^T \mathbb{E}_q[\mathbf{z}_{j,n}]^2 + \text{Var}_q[\mathbf{z}_{j,n}] - (\mu_n^*)^2$ 
11    Compute  $ELBO_i = \mathcal{L}(q(\mathbf{Z}; \Delta))$ ;
12 until  $i > M$  or  $ELBO_i - ELBO_{i-1} < u$ ;
```

Algorithm 2: SGVB Amortized Inference

Data: Choices and reaction times $\mathbf{Y}$

input: maximum number of iteration M , SGD

learning rate η

Result: Variational Parameters Δ ,

Maximum-likelihood estimate of Θ

```

1 Initialize randomly  $\Theta, \Delta$ ;
2 Set  $i = 0$ ;
3 repeat
4    $i += 1$ ;
5   for  $k \in 1 : S * U * K$  do
6     Select  $n \in 1 : N$  randomly;
7     Select  $j \in 1 : J$  randomly;
8     Draw  $\epsilon \in p(\epsilon)$  randomly;
9     Compute an unbiased, noisy estimate of
       $\tilde{\nabla}_k^{\mathcal{L}} - \nabla_{\rho, \theta_n} \{\mathcal{L}(q(\mathbf{Z}; \rho))\}$  // see
      Equation (6);
10    Update  $\rho = \rho - \eta \frac{1}{SUK} \sum_{k=1}^{SUK} \tilde{\nabla}_{\rho, k}^{\mathcal{L}}$ ;
11    Update  $\theta_n = \theta_n - \eta \frac{1}{SUK} \sum_{k=1}^{SUK} \tilde{\nabla}^{\mathcal{L}} \theta_{n, k}$ ;
12 until  $i > M$ ;
```

454 2.6. Recurrent Auto-encoder

455 We now develop the core idea of this paper, which is that one can take advantage of the sequential
456 nature of data acquisition to predict the trial-wise values of the DDM parameters.

457 The model we will develop is based on Variational Auto-Encoders. Auto-Encoders (AE), and more
458 specifically Variational AE (VAE), are a large family of models (93) that can be seen as two-side
459 algorithms where one side (the *encoder* or *recognition model*) is the equivalent of what we have called
460 an *inference network* so far: it is a (non-)linear mapping with global parameters $\boldsymbol{\rho}$ that determines
461 the shape of the approximate posterior of a local latent variable $\mathbf{z}_{j,n}$ of an arbitrary length d ,
462 based on the data $\mathbf{y}_{j,n}$, with a prior $p(\mathbf{z}_{j,n})$. This variable $\mathbf{z}_{j,n}$ lies between the *encoder* and the
463 *decoder*, which constitutes the *generative model* about which inference is made. Figure 3a sketches
464 the structure of a Variational Auto-encoder.

465 This generative model maps the parameter $\mathbf{z}_{j,n}$ onto the parameters of the distribution of $\mathbf{y}_{j,n}$ ($\boldsymbol{\omega}_{j,n}$
466 in our case) thanks to a (non-)linear deterministic mapping $\boldsymbol{\omega}_{j,n} = f(\mathbf{z}_{j,n}; \boldsymbol{\phi})$.

467 This peculiar model class has the advantage that it allows the user to specify the degree of complexity
468 of the inference process she is trying to achieve by adjusting the hyperparameter d . Indeed, VAE
469 ultimately boil down to dimensionality reduction algorithms, where the bottleneck is situated at
470 the level of the latent variable $\mathbf{z}_{j,n}$.

471 Formally, a VAE for a trial indexed j, n is defined by the following formula of the ELBO:

$$\mathcal{L}(q(\mathbf{z}_{j,n}|\mathbf{y}_{j,n})) = \mathbb{E}_{q_{\boldsymbol{\rho}}(\mathbf{z}_{j,n}|\mathbf{y}_{j,n})} [\log p_{\boldsymbol{\phi}}(\mathbf{y}_{j,n}, \mathbf{z}_{j,n}) - \log q_{\boldsymbol{\rho}}(\mathbf{z}_{j,n}|\mathbf{y}_{j,n})] \quad (7)$$

472 which recalls Equation (5) with the only difference that now an encoder/decoder scheme is used to
473 link the data to the latent variable and the latent variable to the data.

474 We kept on purpose the subject indices in Equation (7), and the hierarchical treatment of this
475 expression will be introduced shortly after. Let us emphasize in Equation (7) the dependency of
476 the probability distributions on $\boldsymbol{\rho}$ and $\boldsymbol{\phi}$: these are the only parameters to fit in order to maximize
477 the ELBO. Ideally, one would like to make inference not only about $\mathbf{z}_{j,n}$ but also about $\boldsymbol{\phi}$, as these
478 parameters also belong to the generative model. Often, however, authors treat these parameters
479 in a Maximum-A-Posteriori (MAP) perspective (93), although inference about these parameters is
480 far from being impossible (70; 65). This problem will be treated in 2.6.1.

481 Chung and colleagues (60) have shown that one can include in this model a specific form of regressor
 482 that deterministically brings to the current trial the information about the state of the model at the
 483 trial before: this approach, known as Recurrent VAE (RVAE), has been the focus of several recent
 484 publications (59; 60) due to its efficiency to trade off accuracy with complexity of the inference
 485 process. Under this assumption, the ELBO now looks like:

$$\begin{aligned} \mathcal{L}(q(\mathbf{z}_{j,n}|\mathbf{y}_{j,n}, \mathbf{h}_{j-1})) &= \mathbb{E}_{q(\mathbf{z}_{j,n}|\mathbf{y}_{j,n}, \mathbf{h}_{j-1})} [\log p_{\phi}(\mathbf{y}_{j,n}, \mathbf{z}_{j,n}|\mathbf{h}_{j-1}) - \log q_{\rho}(\mathbf{z}_{j,n}|\mathbf{y}_{j,n}, \mathbf{h}_{j-1})] \\ &\text{where } \mathbf{h}_{j-1} = f(\mathbf{y}_{j-1,n}, \mathbf{z}_{j-1,n}, \mathbf{h}_{j-2,n}; \phi_{\mathbf{h}}). \end{aligned} \quad (8)$$

488 In Equation (8), we have made explicit the fact that the approximate posterior q and the likelihood
 489 $p(\mathbf{y}_{j,n}, \mathbf{z}_{j,n})$ depend on the previous value of $\mathbf{z}$, $\mathbf{y}$ and $\mathbf{h}$ through a hidden state $\mathbf{h}_{j-1}$. As our
 490 notation suggests, $\mathbf{h}$ is also computed using a (non-)linear transformation of these variables, that
 491 takes as input a subset of the decoder parameters ϕ .

492 This model specification can be understood as follows: we are interested in retrieving a posterior
 493 probability distribution of a variable $\mathbf{z}_{j,n}$ that acts as regression weights of regressors $\mathbf{h}$, which
 494 brings to the current trial a selection of information about the trials before (either parameter
 495 values, external regressors such as experimental conditions, or behavioural measure, RT and choices
 496 included). By specifying the length of $\mathbf{z}_{j,n}$ to its minimum, we ensure that each element of $\mathbf{z}_{j,n}$
 497 explains the maximum of the variance of the output parameters of the DDM, $\omega_{j,n}$. Due to the form
 498 of the regression we use (i.e. multi-layered perceptron), we can get high flexibility in a space of
 499 reduced dimension: very rich patterns can be reflected by such a scheme, even in a low-dimensional
 500 space.

501 Note that we do *not* specify the distribution shape of $\omega_{j,n}$, but the distribution of $\mathbf{z}_{j,n}$, which then
 502 maps non-linearly to $\omega_{j,n}$. This contrasts with all DDM implementations we are aware of, where
 503 the choice of the DDM parameter prior is mostly arbitrary.

504 Regarding the choice of a prior for $\mathbf{z}_{j,n}$, one can assume that it is normally distributed in Equa-
 505 tion (7) with a mean 0 and a covariance I_d , which ensures that the components of $\mathbf{z}_{j,n}$ are somehow
 506 independent and, therefore, that each of them captures the maximum variance in the data. How-
 507 ever, this choice is not optimal for the model in Equation (8), as one would expect the prior of
 508 $\mathbf{z}_{j,n}$ to vary from trial to trial. We used therefore the recommendation of (60) and used a form of
 509 prior $\{\mu_0^{\mathbf{z}_{j,n}}, (\sigma_0^{\mathbf{z}_{j,n}})^2\} = f(\mathbf{h}_{j-1}; \phi_p)$ where ϕ_p is again a subset of the generative model regression

510 weights and biases.

511 The final model specification we need to consider is the way we will capture inter-individual differ-
 512 ences. This can be efficiently achieved using a subject-wise auxiliary latent variable ζ_n of length
 513 d^* with a prior $\mathcal{N}(0, I_{d^*})$ and an approximate posterior $q(\zeta_n | \mathbf{v}_n)$. This regressor was introduced
 514 in the generative model described above to obtain a set of DDM parameters that depended on the
 515 three factors $\mathbf{z}_{j,n}$, $\mathbf{h}_{j-1}$ and ζ_n :

$$\omega_{j,n} = f(\mathbf{z}_{j,n}, \mathbf{h}_{j-1}, \zeta_n; \phi_\omega). \quad (9)$$

516 This simple manipulation ensured that inference at the trial level was a compromise between subject-
 517 wise behavioural specifics and consistent population-wise behavioural patterns. It goes beyond
 518 saying that if some regressor is added to the decoder, it should be added to the encoder too to
 519 improve inference about the local variables, which makes us define the following amortized family
 520 of approximate posterior over $\mathbf{z}_{j,n}$

$$\mathbf{z}_{j,n} \sim q_\rho(\mathbf{z}_{j,n} | \mathbf{y}_{j,n}, \mathbf{h}_{j-1}, \zeta_n). \quad (10)$$

521 It flows directly from Equation (9) and Equation (10) that, if the dataset has been acquired using
 522 various experimental conditions, or if the subject is somehow informed about its performance in
 523 the task by the display of a sensory feedback, these regressors can be easily included in the set of
 524 the predictors of ω , making this model a compellingly generic tool for inference about the vector
 525 ω .

526 *Training the RAE-DDM.* All the tools we have provided so far can be used in order to train a
 527 VAE: the inference network, with possibly the normalizing flow that is presented in Appendix C
 528 can be used to find the posterior probability of $\mathbf{z}_{j,n}$ with an amortized cost. SGVB (by the use of
 529 the reparametrization trick) could be used to optimize the ELBO, with an unbiased gradient of the
 530 form

$$\nabla_{\rho, \phi} \{ \mathcal{L}(q(\mathbf{z}_{j,n} | \mathbf{y}_{j,n}, \mathbf{h}_{j-1})) \} = \frac{1}{JSK} \sum_{j,n,k}^{J,S,K} \left\{ \begin{array}{l} \nabla_\rho \{ \log p_\phi(\mathbf{y}_{j,n}, \tilde{\mathbf{z}}_{j,n,k} | \mathbf{h}_{j-1}) - \log q_\rho(\tilde{\mathbf{z}}_{j,n,k} | \mathbf{y}_{j,n}, \mathbf{h}_{j-1}) \} \\ \nabla_\phi \{ \log p_\phi(\mathbf{y}_{j,n}, \tilde{\mathbf{z}}_{j,n,k} | \mathbf{h}_{j-1}) \} \end{array} \right\} \quad (11)$$

533 where we have made implicit the fact that $\mathbf{z}_{j,n}$ was generated according to a IN approach with
 534 $\tilde{\mathbf{z}}_{j,n,k} = g_\rho(\delta_{j,n}, \epsilon)$. Using this approach, one would generate K independent random sequences

(a) Variational Auto-encoder with Maximal-Likelihood estimate of the top-level hyperprior. Note the difference with 2c, where the latent variable directly conditioned the data $y_{j,n}$.

(b) Recurrent Auto-Encoding Drift Diffusion Model structure. The whole encoder-decoder process is shown for a trial j and a subject n . The model is fully connected, in order to account for dependency between the adjacent levels at the recognition and generative level.

of $\tilde{\mathbf{z}}_{j,n,k}$ for $j \in 1 : J$, but this approach might suffer from several issues: specifically, the error of each sample will propagate through the whole sequence, and the effect of an early "bad" sample will impact all the successive trials, which will severely impair the convergence of the SGD algorithm.

Another issue, closely related to the previous one, regards the sampling of the non-decision time τ : it is hard to put constraints on the output of the generative regression network, and some samples of τ might fall beyond the t , where no gradient can be computed.

To solve these problems, a typical solution can be to move the sampling average *inside the logarithm function*, which transforms the ELBO in an Importance Sampling (IS) estimate of the the log-model evidence:

$$\mathcal{L}(q(\mathbf{z}_{j,n}|\mathbf{y}_{j,n}, \mathbf{h}_{j-1})) \approx \frac{1}{JS} \sum_{j,n} \log \frac{1}{K} \sum_k \frac{p_\phi(\mathbf{y}_{j,n}, \tilde{\mathbf{z}}_{j,n,k}|\mathbf{h}_{j-1})}{q_\rho(\tilde{\mathbf{z}}_{j,n,k}|\mathbf{y}_{j,n}, \mathbf{h}_{j-1})}. \quad (12)$$

When used without recurrent predictions, this is known as an Importance Weighted AE (IWAE) (67), and when used in the context of sequential data, this algorithm can be viewed a special

544 member of the Particle Filtering family of algorithms which has been named Filtering Variational
545 Objective (FIVO) (62).

546 This improved estimate of the ELBO has many advantages, that are detailed in (62): it allows us to
547 weight each sequence of data to compute the approximate log-model evidence, and also to drop and
548 resample sequences that fail to reflect the data in a meaningful way based on the other sequences.
549 Moreover, this approach produces samples with a much lower variance than the SGVB approach,
550 which can be simply understood by observing that the FIVO approach takes more samples than
551 SGVB, which is usually taken with a local sample size of $K = 1$. Finally, this approach solves the
552 problem of the bounded space of τ : we can assign a log-likelihood of $-\infty$ (i.e. a weight of 0) to the
553 particles k whose τ_k fall beyond the t , and provided that at least one the particles does satisfy the
554 conditions for the log-likelihood to be computed, the gradient of the estimator can be computed.
555 However, it might still be the case that the whole set of particles at a given trial produces a
556 sample $\tilde{\tau}_{j,n} > t$. We describe in Appendix C.3 a simple solution to this problem based on Lagrange
557 multipliers.

558 The complete model, which we name the Recurrent Auto-Encoding Drift Diffusion Model (RAE-
559 DDM) is an application of the algorithm presented in (60; 62) to the problem of DDM fitting, and
560 we refer to these articles for the implementation details. The model is sketched in Figure 3b and
561 the training algorithm is displayed in ??.

Algorithm 3: RAE-DDM Training

Data: Choices and reaction times $\mathbf{Y}$

input: maximum number of iteration M , SGD learning rate η

Result: Encoder and decoder weights $\{\phi, \rho\}$, subject approximate posterior distributions $\mathbf{v}_n$

```

1 Initialize randomly  $\{\phi, \rho, \mathbf{v}_n\}$ ;
2 repeat
3   Set  $\tilde{\nabla}_{\rho, \phi, \mathbf{v}_n}^{\mathcal{L}} \leftarrow 0$ ;
4   repeat
5     Select  $n \in 1 : N$  randomly;
6     Initialize hidden state  $\mathbf{h}_0 \leftarrow 0$ , accumulated normalized weight  $\beta^n \leftarrow 1/K$ ;
7     // Forward Pass;
8     Set  $\mathcal{L} \leftarrow 0$ ;
9     Draw  $\tilde{\zeta}_n = g(\epsilon_{\zeta_n}; \mathbf{v}_n)$  with  $\epsilon_{\zeta_n} \sim p(\epsilon_{\zeta_n})$ ;
10    Increment  $\mathcal{L}$  with subject-wise noisy DKL approximation  $\mathcal{L} += \log p(\tilde{\zeta}_n) - \log q(\tilde{\zeta}_n | \mathbf{v}_n)$ ;
11    for  $j \in 1 : J$  do
12      for  $k \in 1 : K$  do
13        Draw  $\tilde{\zeta}_n = g(\epsilon_{\zeta_n}; \mathbf{v}_n)$  with  $\epsilon_{\zeta_n} \sim p(\epsilon_{\zeta_n})$ ;
14        Map  $\delta_{j,n} = f(\rho; \mathbf{y}_{j,n}, \mathbf{h}_{j-1,k}, \tilde{\zeta}_n)$ ;
15        Draw  $\tilde{\mathbf{z}}_{j,n,k} = g(\epsilon_{\tilde{\mathbf{z}}_{j,n,k}}; \delta_{j,n})$  with  $\epsilon_{\tilde{\mathbf{z}}_{j,n,k}} \sim p(\epsilon_{\tilde{\mathbf{z}}_{j,n,k}})$ ;
16        Map  $\{\mu_0^{\mathbf{z}_{j,n}}, (\sigma_0^{\mathbf{z}_{j,n}})^2\} = f(\mathbf{h}_{j-1,k}; \phi_p)$ ;
17        Compute  $H = \log p(\tilde{\mathbf{z}}_{j,n,k} | \mu_0^{\mathbf{z}_{j,n}}, (\sigma_0^{\mathbf{z}_{j,n}})^2) - \log q_{\rho}(\tilde{\mathbf{z}}_{j,n,k})$ ;
18        Map  $\tilde{\omega}_{j,n,k} = f(\tilde{\mathbf{z}}_{j,n,k}, \mathbf{h}_{j-1,k}, \tilde{\zeta}_n; \phi)$ ;
19        Compute unnormalized particle weight  $\log \alpha_k^u = \log \text{Wiener}(\mathbf{y}_{j,n} | \tilde{\omega}_{j,n,k}) + H$ ;
20      Increment ELBO with weighted samples  $\mathcal{L} += \log \sum_k \beta_k^n \alpha_k^u$ ;
21      Update accumulated normalized weight  $\beta_k^n = \beta_k^n \frac{\alpha_k^u}{\sum_{k'} \alpha_{k'}^u}$ ;
22      If resampling criterion is met, resample  $\{\beta_k^n, \tilde{\mathbf{z}}_{j,n,k}\}$  for  $k \in 1 : K$ ;
23      for  $k \in 1 : K$  do
24        Map  $\mathbf{h}_{j,k} = f(\mathbf{y}_{j,n}, \tilde{\mathbf{z}}_{j,n,k}, \mathbf{h}_{j-1}; \phi_h)$ 
25      // Reverse Pass;
26      Backpropagate through lines 7 to 21 to get  $\tilde{\nabla}_{\rho, \phi, \mathbf{v}_n}^{\mathcal{L}} = \tilde{\nabla}_{\rho, \phi, \mathbf{v}_n}^{\mathcal{L}} - \nabla_{\rho, \phi, \mathbf{v}_n} \{\mathcal{L}\}$ 
27    until  $S$  sequences have been collected;
28  Update  $\rho = \rho - \eta \tilde{\nabla}_{\rho}^{\mathcal{L}}$ ;  $\phi = \phi - \eta \tilde{\nabla}_{\phi}^{\mathcal{L}}$ ;  $\mathbf{v}_n = \mathbf{v}_n - \eta \tilde{\nabla}_{\mathbf{v}_n}^{\mathcal{L}}$ ;
29 until some convergence criterion is met;
```

562 2.6.1. Variational Dropout

563 A final consideration regards the overfitting propensity of RNN and, consequently, the potential
 564 overfitting of the RAE-DDM. Dropout is a popular technique to prevent overfitting when training
 565 a neural network (68), and has been suggested as a necessary requirement for Boltzman-machine
 566 like algorithms (69) such as VAE and RAE. It simply consists in hiding part of the network at
 567 each iteration, in order to ensure recursivity in the network and maximum relevance of each of the
 568 inputs. Variational Dropout (VD) (70; 110) is an improved version of the Dropout method where a
 569 specific prior and corresponding variational distribution is assigned to each weight of the network.
 570 Using VD, we can ensure that the model takes maximum advantage of each of the inputs of the
 571 network, and prevents overfitting. Additionally, the use of VD makes the model fully Bayesian, as
 572 we make inference not only about $\mathbf{z}_{j,n}$ but also about the weights $\boldsymbol{\rho}$ and $\boldsymbol{\phi}$. VD implementation of
 573 the RAE-DDM performed slightly better than the vanilla version on the long-run, and prevented
 574 the optimization to fall into spurious local posterior estimates of the DDM parameters. For each set
 575 of weights, including those of the GRU RNN algorithm, and according to (111), we sampled $\tilde{\boldsymbol{\rho}}$ and $\tilde{\boldsymbol{\phi}}$
 576 according to their posterior distribution before achieving the whole sequence of forward/backward
 577 pass described in Algorithm 3, meaning that we used the same sample for each and every trial of
 578 the sequence instead of resampling at each trial.

579 3. Experiment

580 3.1. Data Generation Procedure

581 To provide a proof of concept of the usefulness of accounting directly for the trial-to-trial dependency
 582 of the DDM parameters, we generated a series of datasets with a random Gaussian-Inverse-Gamma
 583 prior over the DDM parameters (see Section 3.1), so that the average mean value of each parameter
 584 was variable across datasets. The values were chosen with the following heuristic: we ensured that
 585 the chosen parameters generated datasets with a good accuracy in the simulated task (see below)
 586 with reasonable RT. We always aim at a high trial-to-trial variability of the ξ and w (lower-bounded
 587 at 0.3), and the variability of τ and a were chosen to be low (bounded at 0). These choices were
 588 motivated by the fact that we did not expect the non-decision time to vary more than by a few
 589 milliseconds. We also wanted to investigate whether the threshold variability was identifiable for

	a	ξ	w	τ
μ_n	$\mathcal{N}(1. + 0.2\epsilon_1, 0.1)$	$\mathcal{N}(1.5 + 0.33\epsilon_2, 0.2)$	$\mathcal{N}(0.05\epsilon_3, 0.1)$	$\mathcal{N}(0.3 + 0.1\epsilon_4, 0.01)$
σ_n^2	$\mathcal{IG}^{-1}(5, u_1 * 0.33)$	$\mathcal{IG}^{-1}(5, 0.3 + u_2)$	$\mathcal{IG}^{-1}(5, 0.3 + u_3)$	$\mathcal{IG}^{-1}(5, u_4 * 0.01)$

Table 1: Generative distributions at the population level (i.e. dataset-specific). $\epsilon \sim \mathcal{N}(0, 1)$ and $u \sim U(0, 1)$ are random noise variables that controlled the distribution of the parameters at the population level. This manipulation ensured heterogeneity in the datasets.

low-to-moderate values of the variance, as this variability is usually assumed to be equal to zero in the DDM literature. The simulated task consisted of a highly generic TAFC task where the subjects needed to choose between a left and a right response to indicate, for instance, the motion direction of dots in a random kinetogram. There was no other manipulation imposed on these datasets. We assumed that the motion direction affected the drift and the drift only. This is a quite common assumption in the DDM literature (7), which is physiologically sensible, as the start point should be set before the instruction is perceived by the agent (112).

In order to avoid biasing the data generation process by using a generative model that would only be identifiable using a non-*i.i.d.* based approach, the only manipulation we allowed ourselves to do was to permute the parameters, initially generated according to *i.i.d.* normal distributions (Figure 4). This permutation induced specific patterns of cross-correlations between successive parameter values, while ensuring that a fit achieved using the *i.i.d.* assumption was not harmed by the assumption of the data generation procedure.

We proceeded in the following way: first, we generated a random matrix of unit-Gaussian distributed random variables $\epsilon \in \mathbb{R}^{4 \times J} \sim \mathcal{N}(0, I)$. Then, we generated a trace of 4 signals following a Gaussian Process (GP) with a Rational quadratic or a Periodic Kernel (113), parametrized with a population-wise random parameter $a \sim [0; 1]$ and with an input vector of equally distant variables u situated between -4 and 4 . This simple manipulation ensured that the frequency of the parameter variability was mostly identical in the whole population but different between datasets, while keeping a rich family of traces form across subjects. Next, we permuted the elements of each column of ϵ in order to minimize the Euclidean distance between each of these elements with the correspondent point generated by the GP. Finally, the data were mapped to their desired distribution location by simply applying the following equation:

Figure 4: Data generation procedure for a prior distribution $\mathcal{N}(2, 0.1)$. All parameters were generated according to a diagonal univariate Gaussian distribution. For each parameter, the order of the appearance of each sample was permuted to minimize the distance wrt. a random GP trace (here with a Rational Quadratic kernel).

$$\omega_{j,n} = \mu_n + \sigma_n \odot \epsilon_{j,n}$$

where σ_n^2 is the Hadamard product operator.

RT and choices were simulated according to (12), where the Wiener process is discretized to an arbitrarily small δt (set to 10^{-6} in our case) and approximated as a binary gambler ruin problem (74). An alternative method could consist in using the Euler-Maruyama algorithm (114; 115), but this algorithm did not prove to be of greater accuracy than the former, for a higher computational cost.

A total of 16 datasets of 64 subjects each and 500 trials per subject were generated with the same procedure, and different arbitrary random seeds (from 1 to 16) for reproducibility.

Each dataset was fitted to the RAE-DDM, to a hierarchical *i.i.d.*-VB version of the DDM (see below) and finally to an HMC version of this last model coded in Stan (47). A single chain per dataset was generated, with 1000 iterations for warm-up and 1000 iterations for sampling. Because the threshold was generated with a random noise (as every other parameter), and because it is customary to fit the DDM while discarding the threshold variability, we also fitted the fixed-threshold of the *i.i.d.*-VB -DDM and similarly for the HMC version in order to assess how good the classical threshold steadiness assumption worked. For convenience, these fits will be named “classical” in what follows, by contrast with the “full” versions that included the threshold variability. This leads to a total of

629 five fits per dataset.

630 3.2. Model Specification

Figure 5: *i.i.d.*-VB and HMC full hierarchical structure. The DDM parameters were distributed according to a subject-specific prior $\theta_n = \{\mu_n, \sigma_n^2\}$. In turn, these parameters were distributed according to a latent Normal $\mathbf{M}$ prior $\mathcal{N}(\mu_0, \sigma_0^2)$ and a latent Inverse-Gamma $\mathbf{G}$ prior $\mathcal{IG}^{-1}(\alpha_0, \beta_0)$ distributions, respectively. The hyperprior of the first distribution was chosen to be a Normal Inverse-Gamma distribution $\mathbf{M}_0$ with parameters $\mathcal{N}(0, 10\sigma_0^2)\mathcal{IG}^{-1}(0.1, 0.1)$. The hyperprior of the second distribution was a set of Exponential distributions with shape parameters $\lambda = 0.01$. These values were chosen in order to be weakly informative. This model was identical for the VB and HMC tests. However, in the VB implementation, the approximate posteriors were modelled according to a 4-layers Normalizing Flow with parameters δ^* , where $\cdot$ indicates the latent variable. The parameters $\delta_{j,n}^\omega$ were generated thanks to an inference network with parameters ρ .

631 *i.i.d.*-VB . For each of the $f(\cdot)$ transform in Section 2.6, we used a two-layer perceptron, except for
632 the mapping of $\mathbf{h}$ for which a Gated Neural Network can be used in order to prevent the vanishing
633 of the early gradient through the network (116). We chose arbitrarily the Gated Recurrent Unit
634 (GRU) (117) for the RNN model. An alternative strategy could consist in using a Long-Short Term
635 Memory (LSTM) (118), and we leave the comparison between these various GNN for future studies.
636 For the sake of sparsity and because it is not relevant to the present study, we do not detail these
637 structures here, and refer to the corresponding papers for further details. We chose a hyperbolic

638 tangent function activation function for all these networks, which proved to be of greater accuracy
639 than the Restricted Linear Unit (119; 120) and softplus activators.

640 The RAE-DDM was configured with 50 nodes per layer in each perceptron, GRU included, for both
641 the *i.i.d.*-VB and RAE-DDM models. For the latter, the length of the latent variable $\mathbf{z}_{j,n}$ was set
642 to 4 for all experiments (so was the length of the subject-specific latent variables ζ_n), as it showed
643 to provide a sufficient predictive accuracy during piloting. A four component (or layer) normalizing
644 flow was used for the *i.i.d.*-VB and RAE-DDM approximate posteriors.

645 Each of the VB models were coded in Julia (121), optimized with a noise-insensitive version of the
646 Adam optimizer (58; 122) for 80000 iterations, in parallel with an implementation of the SGD algo-
647 rithm similar to (123), over 8 CPUs node of various kinds (IvyBridge, Xeon, E5-2650v2, 2.60GHz;
648 Skylake, Xeon, 4116, 2.10GHz; SandyBridge, Xeon, E5-4620, 2.20GHz) with no accelerating device
649 such as GPUs. The code will be made available online.

650 The *i.i.d.* versions of the DDM (VB or HMC) were all tested with a model identical to the one
651 used to generate the data, i.e. where the drift was influenced by the perceived motion.

652 3.3. Testing for model specifics: Convergence and overfitting assessment

653 We tested the convergence of the two *i.i.d.*-VB algorithms (with either fixed or variable thresholds)
654 using a validation dataset during training. Convergence rate was fast, and the performance over
655 the validation dataset worsened slightly during the optimization up to a certain plateau. Note that,
656 in piloting experiments, we found that this effect was mitigated by the use of Variational Dropout.
657 This effect was not observed for the *i.i.d.*-VB models.

658 We calculated *a posteriori* an ad-hoc measure of VB performance named Pareto Smoothed Impor-
659 tance Sampling (PSIS) (124). PSIS provides a parameter $\hat{k}$ that gives an estimate of the goodness
660 of fit of VB once the optimization procedure has been performed, and also provides a way to refine
661 the accuracy of the posterior estimates (we refer to (125) for a full description of the method). In
662 short, if $\hat{k} < 0.5$, one can infer that VB achieved a good performance for the task. Both *i.i.d.*-VB
663 models showed a very good performance, with a maximum value of $\hat{k}$ of 0.51 for the classical ver-
664 sions of *i.i.d.*-VB and 0.41 for the full versions. Figure 7 compares the convergence of the three VB
665 methods.

Figure 6: $\hat{k}$ value for each dataset.

To compare the three VB models, we ran a vanilla Bayesian Model selection algorithm ((72)) to investigate the comparative probability that each model generated the data. This method assigns individuals (datasets in this case) to competing models, and can provide an exceedance probability ϕ which can roughly be interpreted as the probability that a model is better than all others. This analysis clearly favored the RAE-DDM with a total of 13.27 datasets ($\phi=0.999$) assigned to this model, with respect to 1.39 and 1.34 datasets for the classical and full versions of the *i.i.d.*-VB model, respectively. Interestingly, these results did not favour any of the two *i.i.d.*-VB models.

3.4. Fit Comparison

We tested the fit of the five models on the training dataset, as HMC does not provide the opportunity to generalize the inference to unseen datapoints. We report here the results in terms of Pearson's correlation coefficients. Euclidean distance (or log-Euclidean distance for the variances) is reported in the Appendix Appendix D. All results are reported in an unbounded space (a, w are mapped with the inverse mapping described in Appendix A, and all statistics involving the variances were computed with the log-value of the variance). For the RAE-DDM, means and variances were computed according to Equation (B.6). Finally, all the results that regard the drift rate were corrected for motion direction.

Population-wise fit. At the population level, HMC performed well for all parameters average values wrt VB-based methods. From a correlation point of view, the RAE-DDM had an acceptable

Figure 7: Convergence rate on a validation dataset (plain) and training dataset (dashed) for the *i.i.d.*-VB and RAE-DDM approaches.

684 performance, although it tended to underestimate the variance of each parameter.

685 Quite interestingly, the HMC-Full performed well in discriminating datasets with high and low
686 variance of the a and w , which would not have been possible if the variance of these two parameters
687 were equivalent, as claimed in earlier studies ((7)). Moreover, fit of the HMC-Full was always equal
688 or better than the classical version with no threshold variability, which confirms that the Full-DDM
689 with threshold variability is identifiable at the population level.

690 *Subject-wise fit.* All methods provided subject-specific estimates of the mean parameters that cor-
691 related well with their true value, with the notable exception of τ , whose fit was acceptable only
692 for the RAE-DDM. In all mean parameter estimations, the RAE-DDM outperformed or was undif-
693 ferentiable (for w) from other methods.

694 Even more important is the difference between the evaluation of the subjects' specific trial-to-
695 trial variability of the four DDM parameters by the five models. First, it can be clearly seen
696 in Figure 9 that the RAE was able to assess variability with a reasonably high accuracy that
697 was inaccessible to other methods. Second, the models (HMC or *i.i.d.*-VB) where the threshold

Figure 8: Population value correlations. **Blue:** HMC-Classical, **Red:** HMC-Full, **Green:** *i.i.d.*-VB -Classical, **Yellow:** *i.i.d.*-VB -Full, **Purple:** RAE-DDM. Fit of the Classical versions of HMC and *i.i.d.*-VB are lacking in the lower left panel, because these models did not account for threshold variability.

variability were implemented did *not* perform significantly worse than others: on the contrary, the HMC-Full uncovered better the variability of the drift, start point and non-decision time than his classical counterpart. These observations suggest again that relaxing the *i.i.d.* assumption and including the threshold variability in the DDM might be beneficial.

i.i.d.-VB methods had reasonable performance for mean estimates but were poor at estimating trial-wise variability of the DDM parameters.

Trial-wise fit. The RAE-DDM performed better at the trial level than each and every other model (Figure 10). Overall, *i.i.d.*-VB and HMC performed equally well for all parameters. The correlation at the trial level of the parameter a was slightly lower than others. Whether this difference was due to a lower relative variability of this parameter wrt. the others, or to the assumptions about the data-generation process remains to be explored.

Figure 9: Subject value correlations. Heat plots show the value of the Pearson correlation coefficient for each dataset, for each model. The bar plots on top of these show the average of the corresponding heat plots. Error bars show the value of one standard deviation. The stars indicate the p-value range of a GLM model that accounted for the value of the true-to-estimate correlation coefficient value as a function of the fitting method (* for $p < 0.5$, ** for $p < 0.01$, *** for $p < 0.001$ and **** for $p < 0.0001$). It can be observed that when the fit of the RAE-DDM was poor, it was usually also the case for other methods.

709 Nevertheless, the RAE-DDM was able to assess the value of the a at each trial with a relatively
710 good accuracy (Figure 11).

711 4. Discussion

712 In this paper, we present an advanced, scalable, data-driven algorithm to fit the Drift Diffusion
713 Model (DDM) in a fully Bayesian setting that can recover the parameter value at the trial, subject
714 and population levels. This algorithm, named the Recurrent Auto-encoding DDM (RAE-DDM)
715 mainly relies on recent advances in Variational Bayes Inference (VB). The main assumption of this

Figure 10: Trial-value correlations (averaged over subjects for each dataset).

model is that the data are correlated in a forward manner, i.e. that the current value of the data and parameters are conditioned on the preceding value of the data and parameters. It relaxes the *i.i.d.* assumption about the parameter distribution, as well as the assumptions about the distribution of these parameters and, most importantly, about the stationarity of the threshold parameter. We show that the fit of this model outperforms other methods (*i.i.d.*-VB and Hamiltonian Monte Carlo (HMC)), and that inference can be efficiently achieved at every level of the hierarchy with correlation coefficients and distances between true and estimated values that are impossible to obtain with previous approaches.

The RAE-DDM enjoys countless advantages; it can be optimized in parallel or on GPU, it requires a low memory usage thanks to the use of Amortized Variational Inference, and, for the same reason, it is generalizable to unseen data. Its structure allows it to be optimized in an online manner. Furthermore, by the use of dimensionality reduction and variational dropout, this model prevents overfitting.

It can account for the effect of regressors of any kind, without requiring to precisely set the relation that ties these regressors to the parameters. Yet, it provides useful information about the parameter values: since the quality of the fit can be expected to be high for a wide range of datasets, the precise estimation of the trial-wise parameters value allows the user to adequately interpret the parameters fit *a posteriori*, based on the trial-to-trial changes in input regressors. This contrasts with classical, non data-driven modeling approaches, where the model structure has to be set

735 beforehand, fitted and then interpreted based on the quality of fit. Therefore, the RAE-DDM
736 compromises informativeness of the model with assumptions about the model specifications.

737 Importantly, we do not claim that informative models should not be used. We suggest that, if an
738 informative model needs to be tested, which will still be desirable in most instances, the universality
739 of the RAE-DDM can provide a benchmark of the parameters fit and an upper bound to the model
740 evidence that can be compared with the informative model fit, with closer fits between the two
741 approaches providing empirical evidence in favour of the grounds of the informative model.

742 It is noteworthy that, contrary to previous approaches (5; 12; 27; 7; 29; 30; 48; 16), here we did not
743 omit any parameter when modelling the trial-by-trial variability. An important contribution of the
744 present paper is indeed to show that the threshold variability can, and hence should, be included in
745 the model trial-to-trial autocorrelations are included in the model. The present method is the first,
746 to our knowledge, in which this specific feature, previously regarded as unidentifiable, has been
747 implemented. Despite the inclusion of this feature in our algorithm, we observed a better fit of the
748 whole set of parameters, showing that its addition did not compromise, but actually improved the
749 identifiability of our model.

750 Also, while testing the inclusion of the threshold in the models in several datasets fitted with the
751 *i.i.d.*-VB and MCMC approach, we found that not only the RAE-DDM was able to uncover the value
752 of the threshold variability, but that the versions of the HMC or *i.i.d.*-VB models that accounted
753 for threshold variability did not perform worse than others in recovering the parameters of the
754 model. This simple observation should motivate researchers to take with more caution the *a-priori*
755 statement that the threshold variability should be discarded. However, we acknowledge the fact
756 that the exact similarity between the true generative model and the fitted model specifications may
757 have biased these results: as it has been shown by (23), the choice of a different prior distribution
758 for a and w would probably have lead to different, less accurate results. Nevertheless, Bayesian
759 Model Selection and cross-validation should be used to assess the quality of the fit.

760 The first innovation of our models is the use of VB to solve the Full DDM problem. Recent advances
761 in VB have made it nearly as precise as MCMC (103; 64; 126; 80), easy to use (127; 98; 88) and
762 adapted to potentially very large problems (76; 70; 65). We used these advantages of VB to solve
763 two challenging problems: the first one being to recover the DDM parameter variability from trial
764 to trial (7) with a fast algorithm ; the second being to recover an estimate of the trial-to-trial

765 cross-correlation of the DDM parameters and related behavioural data.

766 Another key feature of the RAE-DDM is its ability to make inference at the trial level. The ap-
 767 plications of this feature are countless. For instance, one could assess how parameters change over
 768 time following pharmacological or neurophysiological intervention (e.g. dopamine modulation (128)
 769 or sub-thalamic nucleus stimulation (40)). It also makes it possible to determine the relationship
 770 between neurophysiological data (EEG, fMRI) and trial-by-trial values of DDM parameters, simi-
 771 larly to previous studies in which the post-error slowing phenomenon was related to Frontal Theta
 772 bands and the decision threshold was shown to correlate with this signal (129; 40; 130). The present
 773 methods open the door to a more precise study of such phenomena, by allowing precise estimation
 774 of the generative parameters in a data-driven manner.

775 The highly generic structure of the RAE-DDM opens the door to many further developments. A
 776 potential improvement of the present approaches would be to broaden the models to other sequential
 777 sampling models. Many of them have a closed-form FPT density function formula (131), allowing
 778 our models to extend to these other frameworks. Applicable models include the Collapsing Barrier
 779 DDM (132; 133) or the Time-varying Ornstein-Uhlenbeck model (134; 135; 1; 136).

780 Also, many appealing features could be implemented with the RAE-DDM to model neurophysio-
 781 logical measures together with behavioural data. Similarly to (55), we could set the physiological
 782 measure (e.g. ERP, pupil measure, fMRI measure) to be the output of the generative model.

783 Finally, the model we propose is scalable at the trial level, but the subject-specific latent variables
 784 ζ_n approximate posterior still needs to be fitted independently. Scalable inference at the subject
 785 level would probably greatly improve the performance and efficiency of the model.

786 The assumptions of the RAE-DDM should be discussed here. The first and most obvious assump-
 787 tion is that the parameters and the behavioural results are cross-correlated between trials in a
 788 forward manner. As discussed above, many behavioural observations are known to reflect complex
 789 psychological biases, which justifies the use of previous parameter values as regressors for the current
 790 one. A classical way to test which parameter of the DDM is impacted by the previous behaviour
 791 would be to explicitly model the supposed cross-dependency among parameters (40). Thanks to
 792 its generic and flexible form, the RAE-DDM automatically selects the model configuration that
 793 explains the most variance up to a predefined degree of complexity (represented by the length of

the latent variable $\mathbf{z}_{j,n}$). The results we provide show that the presence of this cross-correlation does not harm the fit, but constrain it adequately, as only very few parameter sequences will explain a specific behavioural pattern across the whole experiment.

The second assumption of the model is that the parameters covary at the trial level in a complex way. In the RAE-DDM, this can be observed by the fact that the latent variable $\mathbf{z}_{j,n}$ is mapped onto $\omega_{j,n}$ with a complex non-linear function. This, again, contrasts with most approaches where the variance of the DDM parameters is assumed to be independent. We think that this latter assumption is likely to be incorrect: for instance, after an error, a subject will likely modify both his bias and threshold value, and therefore these two parameters variability should not reasonably be considered as independent. Once more, the richness of the method we propose does improve rather than impair the fit quality by constraining parameter values to lie within a certain correlated subspace.

In summary, we present a VB-backed hierarchical, sequential and data-driven implementation of the DDM that can retrieve relevant information of the data-generating process at each level. We focused our attention on the possibility of modelling the trial-to-trial dependency of parameters through the development of Recurrent Auto-Encoders. We show that this model is highly flexible, scalable and much more precise than other classical approaches provided that there exists some form of auto-correlation in the values of the parameters and/or data.

Acknowledgement

This work was performed at the Institute of Neuroscience (IoNS) of the Université catholique de Louvain (Brussels, Belgium); it was supported by grants from the ARC (Actions de Recherche Concertées, Communauté Française de Belgique), from the Fondation Médicale Reine Elisabeth (FMRE), and from the Fonds de la Recherche Scientifique (F.R.S.-FNRS) to A.Z. A.Z. was a Senior Research Associate supported by INNOVIRIS and is currently 1st grade researcher for the french National Center for Scientific Research (CNRS).

This research used resources of the Center for High Performance Computing and Mass Storage (CISM, <https://uclouvain.be/en/research/cism>) located at the Université Catholique de Louvain, Belgium, which is supported by the Fonds de la Recherche Scientifique – Fonds National de la

822 Recherche Scientifique (F.R.S.-FNRS). The CISM is member of the "Consortium des Équipements
823 de Calcul Intensif (CÉCI)" (<http://www.cec-hpc.be>).

Figure 11: Randomly selected subject parameter values, and fit for each model. The RAE-DDM estimated the parameter trace with a high accuracy. We emphasize that the fit of the drift (corrected for motion direction in the figure) followed the motion direction pattern that was used to generate the data with a high accuracy, even though this was not directly instructed in the model. Also, other parameters did not follow this pattern, meaning that the algorithm was able to discriminate which parameter value did and did not depend upon the motion direction without explicit user input. This contrasts with the four other methods, where the effect of motion direction was explicitly implemented in the model. It is noteworthy that beside their lack of capacity to account for the pattern of variation of the parameters, the four other methods also failed to provide a good estimate of the parameter variance in this example.

824 Appendix A. Normal prior for the DDM parameters

825 If we are to put a normal prior on the DDM parameters, it is important to transform the threshold,
826 starting point and non-decision time such that they lie in an unbounded space. Consider the set
827 of bounded parameters $\{a, \xi, w, \tau\}$ where $a \in \mathbb{R}^+ \triangleq f^{-1}(a')$, $w \in [0; 1] \triangleq g^{-1}(w')$, $\tau \in [-\infty; RT] \triangleq$
828 $h^{-1}(\tau')$ and $\{a', \xi, w', \tau'\} \in \mathbb{R}$. We can construct a distribution

$$Wiener'(y_{j,n}|a', \xi, w', \tau') \triangleq Wiener(y_{j,n}|a, \xi, w, \tau)$$

829 that satisfies this condition for a , w and τ by choosing appropriate transforms f , g and h . A
830 natural choice for f and g are respectively the exponential or softplus function – the latter being
831 more advantageous since its gradient is bounded to one, which avoids numerical overflow for samples
832 with large values of the threshold – and some sigmoid function such that $\sigma(w') : \mathbb{R} \rightarrow [0; 1]$.

833 Appendix B. Full DDM as an Empirical Bayes problem

834 If we are to use the MLE method for the Full DDM for a set of N subjects with J trials per subject,
835 we therefore need to solve the following optimization problem:

$$\begin{aligned} \Theta^* &= \arg \max_{\Theta} p(\mathbf{Y}; \Theta) \\ &= \arg \max_{\Theta} \frac{1}{NJ} \sum_{n=1}^N \sum_{j=1}^J \ell(\mathbf{y}_{j,n}; \theta_n) \end{aligned} \quad (\text{B.1})$$

where $\ell(\mathbf{y}_{j,n}; \theta_n) = \log \int Wiener(\mathbf{y}_{j,n}; \omega_{j,n}) p(\omega_{j,n}; \theta_n) d\omega_{j,n}$

836 For simplicity, we will consider that the DDM parameters all lie in an unbounded space, so that
837 their prior probability can be formulated as a Gaussian with parameters $\theta_n = \{\mu_n, \sigma_n^2 I\}$, with I
838 being the identity matrix (see Appendix A for justification and implementation).

839 In Equation (B.1), we have used the average log-likelihood with the purpose of showing that, for a
840 sufficiently large number of samples NJ , this average log-likelihood takes the form of a sample of
841 the true expected log-likelihood function:

$$\begin{aligned} \frac{1}{NJ} \sum_{n=1}^N \sum_{j=1}^J \ell(\mathbf{y}_{j,n}; \mu_n, \sigma_n^2) &\approx \mathbb{E} [\ell(\mathbf{y}_{j,n}; \mu_n, \sigma_n^2) | \mathbf{y}_{j,n}] \\ &= \hat{\ell}(\mu_n, \sigma_n^2). \end{aligned} \quad (\text{B.2})$$

Since the integral in Equation (B.1) has no analytical form for a, w and τ , and since we need to evaluate it for every trial (12), finding the MLE of $\{\mu_n, \sigma_n^2\}$ will require us to compute the gradient of each trial using numerical integration over three dimensions (or two in the classical version of the Full DDM), which we can expect to be computationally expensive.

We now show how Bayesian Statistics arise as a natural way to deal with this problem. The maxima of the likelihood function are located where the value of the gradient of the expression in Equation (B.2) is equal to 0:

$$\begin{aligned} \{\mu_n^*, \sigma_n^{2*}\} &= \arg \max_{\mu_n, \sigma_n^2} \frac{1}{NJ} \sum_{n=1}^N \sum_{j=1}^J \ell(\mathbf{y}_{j,n}; \mu_n, \sigma_n^2) \\ \Leftrightarrow \frac{1}{NJ} \sum_{n=1}^N \sum_{j=1}^J \nabla_{\mu_n, \sigma_n^2} \ell(\mathbf{y}_{j,n}; \mu_n, \sigma_n^2) \Big|_{\mu_n=\mu_n^*, \sigma_n^2=\sigma_n^{2*}} &= 0 \end{aligned} \quad (\text{B.3})$$

We can unpack this expression for each trial and make use of the Fisher's identity (137) to show that:

$$\begin{aligned} \nabla_{\mu_n, \sigma_n^2} \log \int Wiener(\mathbf{y}_{j,n} | \omega_{j,n}) p(\omega_{j,n} | \mu_n, \sigma_n^2) d\omega_{j,n} \\ = \int \frac{\overbrace{p(\omega_{j,n} | \mu_n, \sigma_n^2) \nabla_{\mu_n, \sigma_n^2} \log p(\omega_{j,n} | \mu_n, \sigma_n^2)}^{\nabla_{\mu_n, \sigma_n^2} p(\omega_{j,n} | \mu_n, \sigma_n^2)} Wiener(\mathbf{y}_{j,n} | \omega_{j,n}) d\omega_{j,n}}{\int Wiener(\mathbf{y}_{j,n} | \omega_{j,n}) p(\omega_{j,n} | \mu_n, \sigma_n^2) d\omega_{j,n}} \\ = \int \nabla_{\mu_n, \sigma_n^2} \log p(\omega_{j,n} | \mu_n, \sigma_n^2) \frac{Wiener(\mathbf{y}_{j,n} | \omega_{j,n}) p(\omega_{j,n} | \mu_n, \sigma_n^2)}{\int Wiener(\mathbf{y}_{j,n} | \omega_{j,n}) p(\omega_{j,n} | \mu_n, \sigma_n^2) d\omega_{j,n}} d\omega_{j,n} \\ = \int \nabla_{\mu_n, \sigma_n^2} \log p(\omega_{j,n} | \mu_n, \sigma_n^2) \frac{Wiener(\mathbf{y}_{j,n} | \omega_{j,n}) p(\omega_{j,n} | \mu_n, \sigma_n^2)}{p(\mathbf{y}_{j,n} | \mu_n, \sigma_n^2)} d\omega_{j,n} \\ = \int \nabla_{\mu_n, \sigma_n^2} \log p(\omega_{j,n} | \mu_n, \sigma_n^2) p(\omega_{j,n} | \mathbf{y}_{j,n}, \mu_n, \sigma_n^2) d\omega_{j,n} \\ = \mathbb{E}_{p(\omega_{j,n} | \mathbf{y}_{j,n}, \mu_n, \sigma_n^2)} [\nabla_{\mu_n, \sigma_n^2} \log p(\omega_{j,n} | \mu_n, \sigma_n^2)] \end{aligned} \quad (\text{B.4})$$

where we have used the Leibniz rule, the log-derivative trick and finally the Bayes theorem $p(\omega_{j,n} | \mathbf{y}_{j,n}) = \frac{p(\mathbf{y}_{j,n} | \omega_{j,n}) p(\omega_{j,n})}{\int p(\mathbf{y}_{j,n} | \omega_{j,n}) p(\omega_{j,n}) d\omega_{j,n}}$. Equation (B.4) shows that the DDM problem can be solved if we can compute (or approximate) the posterior probability $p(\omega_{j,n} | \mathbf{y}_{j,n}, \mu_n, \sigma_n^2)$. Indeed, since the gradient formulated in Equation (B.4) must be zero at the mode, we can expand this formula and apply it to all trials

855 to show the following:

$$\begin{aligned} & \frac{1}{NJ} \sum_{n=1}^N \sum_{j=1}^J \mathbb{E}_{p(\omega_{j,n}|\mathbf{y}_{j,n}, \boldsymbol{\mu}_n, \sigma_n^2)} [\nabla_{\boldsymbol{\mu}_n, \sigma_n^2} \log p(\omega_{j,n}; \boldsymbol{\mu}_n, \sigma_n^2)] = 0 \\ \Leftrightarrow & \begin{cases} \frac{1}{NJ} \sum_{n=1}^N \sum_{j=1}^J \mathbb{E}_{p(\omega_{j,n}|\mathbf{y}_{j,n}, \boldsymbol{\mu}_n, \sigma_n^2)} \left[-\frac{(\boldsymbol{\mu}_n - \omega_{j,n})}{\sigma_n^2} \right] \\ \frac{1}{NJ} \sum_{n=1}^N \sum_{j=1}^J \mathbb{E}_{p(\omega_{j,n}|\mathbf{y}_{j,n}, \boldsymbol{\mu}_n, \sigma_n^2)} \left[\frac{(\boldsymbol{\mu}_n - \omega_{j,n})^2}{\sigma_n^3} - 1/\sigma_n^2 \right] \end{cases} = 0. \end{aligned} \quad (\text{B.5})$$

856 Solving for $\boldsymbol{\mu}_n$ and σ_n^2 , we have:

$$\begin{aligned} \boldsymbol{\mu}_n^* &= \frac{1}{T} \sum_{t=1}^T \mathbb{E}_{p(\omega_{j,n}|\mathbf{y}_{j,n}, \boldsymbol{\mu}_n, \sigma_n^2)} [\omega_{j,n}] \\ \sigma_n^{2*} &= \frac{1}{T} \sum_{t=1}^T \mathbb{E}_{p(\omega_{j,n}|\mathbf{y}_{j,n}, \boldsymbol{\mu}_n, \sigma_n^2)} [\omega_{j,n}]^2 + \text{Var}_{p(\omega_{j,n}|\mathbf{y}_{j,n}, \boldsymbol{\mu}_n, \sigma_n^2)} [\omega_{j,n}] - (\boldsymbol{\mu}_n^*)^2. \end{aligned} \quad (\text{B.6})$$

857 This can be viewed as an Empirical Bayes problem (138; 139) where a MLE of the prior distribution
858 $\mathcal{N}(\boldsymbol{\mu}_n, \sigma_n^2)$ has to be retrieved from the data. Using a Hierarchical Bayesian terminology, $\omega_{j,n}$ is
859 the latent (or local) variable and $\{\boldsymbol{\mu}_n, \sigma_n^2\}$ are the prior or global parameters (Figure 2).

860 Equation (B.6) shows that $\boldsymbol{\mu}_n^*$ and σ_n^{2*} will have a closed form formula as long as we can estimate
861 the first and second moments of $p(\omega_{j,n}|\mathbf{y}_{j,n}, \boldsymbol{\mu}_n, \sigma_n^2)$ for each trial. If the posterior probability
862 of $\omega_{j,n}$ were easy to estimate, this cross-dependency between the posterior probability over $\omega_{j,n}$
863 and the value of $\boldsymbol{\mu}_n^*$ and σ_n^{2*} could suggests an expectation maximization (EM) algorithm: we
864 could iteratively estimate the posterior $p(\omega_{j,n}|\mathbf{y}_{j,n}, \boldsymbol{\mu}_n, \sigma_n^2)$ under the current belief of $\boldsymbol{\mu}_n$ and σ_n^2 ,
865 and then, we could estimate $\boldsymbol{\mu}_n^*$ and σ_n^{2*} using Equation (B.6), until some convergence criterion
866 is reached. This posterior probability does not, however, have a closed-form: additional work is
867 therefore necessary to infer the value of the posterior probability.

868 Appendix C. Approximate posterior Specification

869 Appendix C.1. Approximate posterior shape and reparametrization trick

Sampling the first three parameters of the DDM is trivial if these are distributed according to a (possibly diagonal) Gaussian distribution. Let $\{\boldsymbol{\mu}_{j,n}^\omega, L_{j,n}^\omega\}$ be the mean and lower-Cholesky decomposition of the amortized approximate posterior Gaussian distribution $q(\omega_{j,n}|\mathbf{y}_{j,n}, \boldsymbol{\theta}_n; \boldsymbol{\chi})$, where $\{\boldsymbol{\mu}_{j,n}^\omega, L_{j,n}^\omega\} = f(\mathbf{y}_{j,n}, \boldsymbol{\theta}_n; \boldsymbol{\chi})$, then:

$$\tilde{\omega}_{j,n,k} = \boldsymbol{\mu}_{j,n}^\omega + L_{j,n}^\omega \boldsymbol{\epsilon} \quad \text{with } \boldsymbol{\epsilon} \sim \mathcal{N}(0, 1).$$

870 One can easily see that this expression is differentiable and therefore suited for SGVB or IWAE.

871 Sampling the non-decision time is more tricky: whereas the domain of the prior is unbounded ($\tau_{j,n} \in$
872 $[-\infty; \infty]$), the domain of the approximate posterior we would wish to derive is not ($\tau_{j,n} \in [-\infty; t_t]$),
873 as an unbounded probability distribution would generate samples that might give a gradient of 0
874 to the ELBO, which would in these case take a value of $-\infty$.

875 One solution to this problem would be to use a Truncated Gaussian ($\mathcal{N}_{-\infty}(\mu_{j,n}^\tau, \sigma_{j,n}^\tau)$) distribution,
876 where the upper bound would be set to the RT of interest, and the lower bound would be set to
877 $-\infty$. Two main reasons might discourage us from adopting this strategy: the first would be that
878 efficient random number generation algorithms for the truncated Gaussian are not differentiable,
879 because they rely on acceptance-rejection algorithms (106). However, as the cumulative density
880 function (CDF) of the univariate truncated Gaussian can be computed efficiently, we can compute
881 an estimator of the gradient $\nabla_{\mu_{j,n}^\tau, \sigma_{j,n}^\tau} g(\epsilon; \mu_{j,n}^\tau, \sigma_{j,n}^\tau)$ as (108):

$$\nabla_{\mu_{j,n}^\tau, \sigma_{j,n}^\tau} g(\eta; \mu_{j,n}^\tau, \sigma_{j,n}^\tau) = \frac{\nabla_{\mu_{j,n}^\tau, \sigma_{j,n}^\tau} P(\tau_{j,n} | \mu_{j,n}^\tau, \sigma_{j,n}^\tau)}{p(\tau_{j,n} | \mu_{j,n}^\tau, \sigma_{j,n}^\tau)} \quad (\text{C.1})$$

882 where P is the CDF of $\mathcal{N}_{(-\infty, t]}(\mu_{j,n}^\tau, \sigma_{j,n}^\tau)$ and p the PDF. We refer to the code available online
883 for the derivation of this gradient. Note that, whereas for the multivariate Gaussian, the gradient
884 and random numbers were computed at the same time, here, one must first generate the random
885 number and then, on this basis, compute the gradient estimator of g given the sample that has
886 been drawn.

887 The second issue why not to use a univariate truncated Gaussian distribution for τ might be found
888 in the strong mean-field assumption that it requires. Using a multivariate truncated Gaussian is,
889 however, much more challenging than factorizing $q(\tau)$ from the other parameters of the DDM, as
890 the CDF of this distribution is extremely hard to derive and expensive to compute. Moreover, current
891 implementations of this function are not differentiable wrt the distribution parameters (see for
892 instance (140)). We will introduce how this issue can be solved by the use of Normalizing Flows
893 short after.

894 Appendix C.2. Normalizing Flow

895 If the deterministic mapping from the datapoint to the approximate posterior shape described in
 896 Section 2.4.2 facilitates the optimization of the ELBO by reducing the size of the problem, it does
 897 not make the approximate posterior more flexible. Here, we used a Normalizing Flow (NF) to
 898 transform an initial approximate posterior (typically a Gaussian or uniform distribution) through
 899 a sequence of deterministic invertible mappings so that its shape gets closer to the true shape of
 900 the posterior (64). The core idea of Normalizing Flows is quite simple: it consists in defining an
 901 arbitrarily complex approximate posterior $q^0(\omega_{j,n}|\delta_{j,n}, \chi)$, where both $\delta_{j,n}$ and χ are the trial and
 902 subject-wise outputs of the inference network.

903 Consider an initial approximate posterior distribution $q(\omega_0)$. We would like to transform this prob-
 904 ability distribution with a deterministic mapping $\omega_K = f(\omega_0, \chi)$ with K successive transformations,
 905 where $\chi = \{\chi_k \text{ for } k \in 1 : K\}$ are the parameters controlling how f changes the shape of q . In
 906 order to keep the volume of $q(\omega_0)$ so that it still integrates to 1 over $\omega_{j,n}$ after the sequence of
 907 transformations, the resulting probability density function must be multiplied by the absolute value
 908 of the determinant of the Jacobian of the inverse transform wrt ω_K , or alternatively divided by the
 909 absolute value of the determinant of the transform wrt ω_0 :

$$q(\omega_1) = q(\omega_0) \left| \det \frac{\delta f^{-1}}{\delta \omega_1} \right| = q(\omega_0) \left| \det \frac{\delta f}{\delta \omega_0} \right|^{-1} \quad (\text{C.2})$$

910 and similarly for the following transformations. After K transformations, the ELBO to the model
 911 evidence looks like

$$\mathcal{L}(q(\omega_K)) = \mathbb{E}_{q(\omega_K)} [\log p(\mathbf{y}_{j,n}, \omega_K)] - \mathbb{E}_{q(\omega_K)} \left[\log q(\omega_0) + \sum_{k=1}^K \log \left| \det \frac{\delta f_k}{\delta \omega_{k-1}} \right| \right] \quad (\text{C.3})$$

where $\omega_K = f_K \circ f_{K-1} \dots f_1(\omega_0)$

For the sake of simplicity, we chose to use a series of planar transforms with parameters $\chi_k = \{\mathbf{u}_k, \mathbf{w}_k, b_k\}$:

$$f_{\chi_k}(\omega_{k-1}) = \omega_{k-1} + \mathbf{u}_k \tanh(\mathbf{w}_k^T \omega_{k-1} + b_k).$$

912 This mapping is only valid if $f_k(\omega_{k-1})$ is invertible. This can be easily achieved if $\mathbf{u}$ and $\mathbf{w}$ respect
 913 specific conditions, which can be found in (64).

DDM parameters correlation. The issue of accounting for posterior correlation between τ and the other parameters of the DDM finds a convenient solution with the use of planar NF. To account for the covariance between the parameters, we sampled the parameters $\{a, \xi, w\}$ according to a diagonal Gaussian distribution, and τ according to a truncated Gaussian distribution as stated above. We then transformed the shape of this factorized distribution by a series of planar transformations constituted of expansions and contractions around the initial draw of the parameters $\tilde{\omega}_{j,n,k}$. The important point here is that every parameter *except* τ can be transformed in such a way, but always *with τ as an input for the transformation*:

$$\omega_{k+1} = \omega_k + \mathbf{u}_k \psi(\mathbf{w}_k^T \omega_k + b_k)$$

914 where $\mathbf{u}_k$ and $\mathbf{w}_k$ are two four dimensional vectors, respectively, which are the output of the inference
915 network, with the notable exception that the component of $\mathbf{u}_k$ corresponding to τ is equal to 0,
916 which ensures that τ lies in the space bounded by the RT. Invertibility of this transformation can
917 still be ensured by the algorithm provided by (64). This form of approximate posterior proved to
918 be efficient in retrieving covariance between τ and the other parameters.

919 *Appendix C.3. FIVO and non-decision-time sampling*

920 The use of the FIVO lead to a convenient sampling algorithm in which we were able to deal with
921 $\tilde{\tau}_{j,n}$ falling above the RT $t_{j,n}$.

It might however still be the case that each particle of the FIVO was a degenerated sample. To deal with this problem, we artificially modified the gradients of the encoder and decoder parameters by adding a component to the samples of the ELBO such that, if all particles were degenerated at a trial j , then the SGD optimizer was instructed to correct the network weights to counter this effect. In practice, the ELBO was transformed to look like:

$$\begin{aligned} \mathcal{L}(q(\mathbf{z}_{j,n} | \mathbf{y}_{j,n}, \mathbf{h}_{j-1})) \approx & \frac{1}{JS} \sum_{j,n} \log \frac{1}{K} \sum_k \exp \left\{ \log p_{\phi}(\mathbf{y}_{j,n}, \tilde{\mathbf{z}}_{j,n,k} | \mathbf{h}_{j-1}) \right. \\ & \left. - \log q_{\rho}(\tilde{\mathbf{z}}_{j,n,k} | \mathbf{y}_{j,n}, \mathbf{h}_{j-1}) - \frac{1}{2} \max(0, \tilde{\tau}_k - t_{j,n})^2 \right\} \end{aligned} \quad (\text{C.4})$$

where $\tilde{\tau}_k$ is the sampled non-decision time of the k particle. This is a simple case of the use of Lagrange multipliers for constraint optimization: in the case where $\tilde{\tau}_k > t$ for some but not all

particles k , then the weight of these particles was set to 0 because $\log p_{\phi}(\mathbf{y}_{j,n}, \tilde{\mathbf{z}}_{j,n,k} | \mathbf{h}_{j-1}) \rightarrow -\infty$. Therefore, the Lagrange correction had no impact on the gradient estimate. The only case where this correction had an impact was therefore when all particles satisfied the above condition, in which case the weights were all set to $1/K$ and the gradient used for backpropagation was set to:

$$\nabla_{\rho, \phi} \left\{ \mathcal{L} \left(q(\mathbf{z}_{j,n} | \mathbf{y}_{j,n}, \mathbf{h}_{j-1}) \right) \right\} \approx \frac{1}{JS} \sum_{j,n} \log \frac{1}{K} \sum_k \nabla_{\rho, \phi} \left\{ \exp \left\{ -\log q_{\rho}(\tilde{\mathbf{z}}_{j,n,k} | \mathbf{y}_{j,n}, \mathbf{h}_{j-1}) - \frac{1}{2} \max(0, \tilde{\tau}_k - t_{j,n})^2 \right\} \right\}. \quad (\text{C.5})$$

Appendix D. Euclidean Distances

We report here the Euclidean distances between the true values of the DDM parameters and their model estimates.

Overall, the RAE-DDM performed better or equally than the HMC approach at the trial and subject level.

At the population level (Figure D.12), the mean and variance estimates of the RAE-DDM were slightly more biased than the HMC estimates. This confirms the results shown in Figure 8 that the RAE-DDM had a low discrimination power between datasets.

At the subject level, however, the RAE-DDM performed equally well or better than the HMC Figure D.13. The same was true at the trial level Figure D.14.

We also report here the corresponding correlation plot Figure D.15 of the example subject trace in 11.

Figure D.12: Population true to estimate (log-)Euclidean distances.

Figure D.13: Subject-wise true to estimate (log-)Euclidean sum of distances.

Figure D.14: Trial-wise true to estimate (log-)Euclidean sum of distances. The RAE-DDM performed equally well or better than other approaches.

Figure D.15: Correlation and distances of the same random subject as in Figure 11. RAE-DDM showed a high accuracy for both these measures, displayed in the upper left corner of each panel. p -values state for the significance of the Pearson's correlation coefficient displayed above.

- 934 [1] Roger Ratcliff and Philip L Smith. A comparison of sequential sampling models for two-choice
935 reaction time. *Psychological review*, 111(2):333–67, apr 2004.
- 936 [2] Roger Ratcliff. A theory of memory retrieval. *Psychological Review*, 85(2):59–108, 1978.
- 937 [3] Donald Laming. *Information theory of choice-reaction times.*, volume 14. 1968.
- 938 [4] Roger Ratcliff. Continuous versus discrete information processing modeling accumulation of
939 partial information. *Psychological review*, 95(2):238–255, 1988.
- 940 [5] R. Ratcliff and J. N. Rouder. Modeling Response Times for Two-Choice Decisions. *Psycho-*
941 *logical Science*, 9(5):347–356, 1998.
- 942 [6] P L Smith and T Van Zandt. Time-dependent Poisson counter models of response latency
943 in simple judgment. *The British journal of mathematical and statistical psychology*, 53 (Pt
944 2):293–315, nov 2000.
- 945 [7] Roger Ratcliff and Gail McKoon. The diffusion decision model: theory and data for two-choice
946 decision tasks. *Neural computation*, 20:873–922, 2008.
- 947 [8] Eric-Jan Wagenmakers, Han L J van der Maas, and Raoul P P P Grasman. An EZ-diffusion
948 model for response time and accuracy. *Psychonomic Bulletin & Review*, 14(1):3–22, 2007.
- 949 [9] Rafal Bogacz, Eric Brown, Jeff Moehlis, Philip Holmes, and Jonathan D. Cohen. The physics
950 of optimal decision making: A formal analysis of models of performance in two-alternative
951 forced-choice tasks. *Psychological Review*, 113(4):700–765, oct 2006.
- 952 [10] Thomas V Wiecki and Michael J Frank. A computational model of inhibitory control in
953 frontal cortex and basal ganglia. *Psychological review*, 120(2):329–55, apr 2013.
- 954 [11] Roger Ratcliff, Philip L. Smith, Scott D. Brown, and Gail McKoon. Diffusion Decision Model:
955 Current Issues and History. *Trends in Cognitive Sciences*, 20(4):260–281, 2016.
- 956 [12] Roger Ratcliff and Francis Tuerlinckx. Estimating parameters of the diffusion model: Ap-
957 proaching to dealing with contaminant reaction and parameter variability. *Psychonomic*
958 *Bulletin & Review*, 9(3):438–481, 2002.

- 959 [13] Francis Tuerlinckx. The efficient computation of the cumulative distribution and probability
960 density functions in the diffusion model. *Behavior research methods, instruments, & comput-*
961 *ers : a journal of the Psychonomic Society, Inc*, 36(4):702–16, 2004.
- 962 [14] Roger Ratcliff. Parameter variability and distributional assumptions in the diffusion model.
963 *Psychological Review*, 120(1):281–292, 2013.
- 964 [15] Thomas V Wiecki, Imri Sofer, and Michael J Frank. HDDM: Hierarchical Bayesian estimation
965 of the Drift-Diffusion Model in Python. *Frontiers in neuroinformatics*, 7(August):14, 2013.
- 966 [16] Michael D. Nunez, Joachim Vandekerckhove, and Ramesh Srinivasan. How attention influ-
967 ences perceptual decision making: Single-trial EEG correlates of drift-diffusion model param-
968 eters. *Journal of Mathematical Psychology*, 76:117–130, 2017.
- 969 [17] Christopher Bishop. *Pattern Recognition And Machine Learning*. Springer-Verlag New York,
970 2006.
- 971 [18] Harriet R Brown, Peter Zeidman, Peter Smittenaar, Rick a Adams, Fiona McNab, Robb B
972 Rutledge, and Raymond J Dolan. Crowdsourcing for cognitive science - the utility of smart-
973 phones. *PloS one*, 9(7):e100662, jan 2014.
- 974 [19] Fiona McNab, Peter Zeidman, Robb B. Rutledge, Peter Smittenaar, Harriet R. Brown,
975 Rick A. Adams, and Raymond J. Dolan. Age-related changes in working memory and the
976 ability to ignore distraction. *Proceedings of the National Academy of Sciences*, 112(20):6515–
977 6518, 2015.
- 978 [20] Robb B. Rutledge, Peter Smittenaar, Peter Zeidman, Harriet R. Brown, Rick A. Adams,
979 Ulman Lindenberger, Peter Dayan, and Raymond J. Dolan. Risk Taking for Potential Reward
980 Decreases across the Lifespan. *Current Biology*, 26(12):1634–1639, 2016.
- 981 [21] Brandon M. Turner, Leendert van Maanen, and Birte U. Forstmann. Informing Cognitive
982 Abstractions through Neurophysiology: The Neural Drift Diffusion Model. *Psychological*
983 *Review*, 122(2):312–336, 2015.
- 984 [22] Roger Ratcliff. Measuring psychometric functions with the diffusion model. *Journal of Ex-*
985 *perimental Psychology: Human Perception and Performance*, 40(2):870–888, 2014.

- 986 [23] Matt Jones and Ehtibar N Dzhafarov. Unfalsifiability and mutual translatability of major
987 modeling schemes for choice reaction time. *Psychological review*, 121(1):1–32, 2014.
- 988 [24] Veronika Lerche and Andreas Voss. Model complexity in diffusion modeling: Benefits of
989 making the model more parsimonious. *Frontiers in Psychology*, 7(SEP):1–14, 2016.
- 990 [25] A Heathcote, S Brown, and D J K Mewhort. Quantile maximum likelihood estimation of
991 response time distributions. *Psychonomic Bulletin & Review*, 9(2):394–401, 2002.
- 992 [26] Scott Brown and Andrew Heathcote. QMLE: Fast, robust, and efficient estimation of distri-
993 bution functions based on quantiles. *Behavior Research Methods, Instruments, & Computers*,
994 35(4):485–492, nov 2003.
- 995 [27] Joachim Vandekerckhove and Francis Tuerlinckx. Fitting the Ratcliff diffusion model to
996 experimental data. *Psychonomic bulletin & review*, 14(6):1011–1026, 2007.
- 997 [28] Joachim Vandekerckhove, Francis Tuerlinckx, and Michael D. Lee. A Bayesian approach to
998 diffusion process models of decision-making. *Social Sciences*, (949):1429–1434, 2008.
- 999 [29] Joachim Vandekerckhove, Francis Tuerlinckx, and M. Lee. A Bayesian approach to diffusion
1000 process models of decision-making. *Proceedings of the 30th annual conference of the cognitive*
1001 *science society*, pages 1429–1434, 2008.
- 1002 [30] Joachim Vandekerckhove, Francis Tuerlinckx, and Michael D Lee. Hierarchical diffusion mod-
1003 els for two-choice response times. *Psychological methods*, 16(1):44–62, 2011.
- 1004 [31] Andrew Gelman, John B Carlin, Hal S Stern, David B. Dunson, Aki Vehtari, and Donald B
1005 Rubin. *Bayesian Data Analysis, Third Edition (Chapman & Hall/CRC Texts in Statistical*
1006 *Science)*. CRC, 2013.
- 1007 [32] Nathaniel D Daw. Trial-by-trial data analysis using computational models. *Attention and*
1008 *Performance*, pages 1–26, 2009.
- 1009 [33] Angela J Yu and Jonathan D Cohen. Sequential effects: Superstition or rational behavior?
1010 *Advances in Neural Information Processing Systems*, 21:1873–80, 2009.

- 1011 [34] Juan Gao, KongFatt Wong-Lin, Philip Holmes, Patrick Simen, and Jonathan D. Cohen.
1012 Sequential Effects in Two-Choice Reaction Time Tasks: Decomposition and Synthesis of
1013 Mechanisms. *Neural Computation*, 21(9):2407–2436, sep 2009.
- 1014 [35] M. H. Wilder, M. Jones, and M. C. Mozer. Sequential effects reflect parallel learning of
1015 multiple environmental regularities. *Advances in Neural Information Processing Systems*,
1016 22:2053–61, 2009.
- 1017 [36] Joshua Benjamin Miller and Adam Sanjurjo. Surprised by the Gambler’s and Hot Hand
1018 Fallacies? A Truth in the Law of Small Numbers. *SSRN Electronic Journal*, pages 1–44,
1019 2015.
- 1020 [37] Gui Xue, Chi-Hung Juan, Chi-Fu Chang, Zhong-Lin Lu, and Qi Dong. Lateral prefrontal
1021 cortex contributes to maladaptive decisions. *Proceedings of the National Academy of Sciences
1022 of the United States of America*, 109(12):4401–6, mar 2012.
- 1023 [38] Robin Shao, Delin Sun, and Tatia M C Lee. The interaction of perceived control and Gambler’s
1024 fallacy in risky decision making: An fMRI study. *Human Brain Mapping*, 37(3):1218–1234,
1025 2016.
- 1026 [39] Wim Notebaert, Femke Houtman, Filip Van Opstal, Wim Gevers, Wim Fias, and Tom
1027 Verguts. Post-error slowing: An orienting account. *Cognition*, 111(2):275–279, may 2009.
- 1028 [40] James F Cavanagh, Thomas V Wiecki, Michael X Cohen, Christina M Figueroa, Johan
1029 Samanta, Scott J Sherman, and Michael J Frank. Subthalamic nucleus stimulation reverses
1030 mediofrontal influence over decision threshold. *Nature Neuroscience*, 14(11):1462–1467, 2011.
- 1031 [41] Trisha Van Zandt. How to fit a response time distribution. *Psychonomic Bulletin & Review*,
1032 7(3):424–465, sep 2000.
- 1033 [42] Andreas Voss, Jochen Voss, and Veronika Lerche. Assessing cognitive processes with diffusion
1034 model analyses: A tutorial based on fast-dm-30. *Frontiers in Psychology*, 6(MAR):1–14, mar
1035 2015.
- 1036 [43] Andreas Voss and Jochen Voss. A fast numerical algorithm for the estimation of diffusion
1037 model parameters. *Journal of Mathematical Psychology*, 52(1):1–9, feb 2008.

- 1038 [44] Eric-Jan Wagenmakers, Raoul P.P.P. Grasman, and Peter C.M. Molenaar. On the relation
1039 between the mean and the variance of a diffusion model response time distribution. *Journal*
1040 *of Mathematical Psychology*, 49:195–204, 2005.
- 1041 [45] Raoul P P P Grasman, Eric-Jan Wagenmakers, and Han L J Van Der Maas. On the mean
1042 and variance of response times under the diffusion model with an application to parameter
1043 estimation. *Journal of Mathematical Psychology*, 53(2):55–68, apr 2009.
- 1044 [46] Dominik Wabersich and Joachim Vandekerckhove. Extending JAGS: A tutorial on adding
1045 custom distributions to JAGS (with a diffusion model example). *Behavior Research Methods*,
1046 46(1):15–28, 2014.
- 1047 [47] Andrew Gelman, Daniel Lee, and Jiqiang Guo. Stan: A Probabilistic Programming Language
1048 for Bayesian Inference and Optimization. *Journal of Educational and Behavioral Statistics*,
1049 40(5):530–543, 2015.
- 1050 [48] Peter R. Murphy, Joachim Vandekerckhove, and Sander Nieuwenhuis. Pupil-Linked Arousal
1051 Determines Variability in Perceptual Decision Making. *PLoS Computational Biology*, 10(9),
1052 2014.
- 1053 [49] M. J. Frank, C. Gagne, E. Nyhus, S. Masters, T. V. Wiecki, J. F. Cavanagh, and D. Badre.
1054 fMRI and EEG Predictors of Dynamic Decision Parameters during Human Reinforcement
1055 Learning. *Journal of Neuroscience*, 35(2):485–494, jan 2015.
- 1056 [50] Rafael Polanía, Marius Moisa, Alexander Opitz, Marcus Grueschow, and Christian C. Ruff.
1057 The precision of value-based choices depends causally on fronto-parietal phase coupling. *Nature*
1058 *Communications*, 6, 2015.
- 1059 [51] Samuel R. Mathias. Unified analysis of accuracy and reaction times via models of decision
1060 making. volume 050001, page 050001, 2016.
- 1061 [52] Mads Lund Pedersen, Michael J Frank, and Guido Biele. The drift diffusion model as the
1062 choice rule in reinforcement learning. *Psychonomic Bulletin & Review*, 24(4):1234–1251, aug
1063 2017.
- 1064 [53] Andriy Mnih and Karol Gregor. Neural Variational Inference and Learning in Belief Networks.
1065 *ArXiv stat.ML*, 32(October):1–20, 2014.

- 1066 [54] Matt Jones, Tim Curran, Michael C Mozer, and Matthew H Wilder. Supplemental Material
1067 for Sequential Effects in Response Time Reveal Learning Mechanisms and Event Represen-
1068 tations. *Psychological review*, 120(3):628–666, 2013.
- 1069 [55] Brandon M. Turner, Birte U. Forstmann, Eric Jan Wagenmakers, Scott D. Brown, Per B.
1070 Sederberg, and Mark Steyvers. A Bayesian framework for simultaneously modeling neural
1071 and behavioral data. *NeuroImage*, 72:193–206, 2013.
- 1072 [56] Brandon M. Turner, Birte U. Forstmann, Bradley C. Love, Thomas J. Palmeri, and Leendert
1073 Van Maanen. Approaches to analysis in model-based cognitive neuroscience. *Journal of*
1074 *Mathematical Psychology*, 76:65–79, 2017.
- 1075 [57] Peter Dayan. Helmholtz machines and wake-sleep learning. *Handbook of Brain Theory and*
1076 *Neural Network. MIT ...*, 44(0), 2000.
- 1077 [58] Diederik P. Kingma and Jimmy Lei Ba. Adam: a Method for Stochastic Optimization.
1078 *International Conference on Learning Representations 2015*, pages 1–15, 2015.
- 1079 [59] Otto Fabius and Joost R. van Amersfoort. Variational Recurrent Auto-Encoders. (2013):1–5,
1080 dec 2014.
- 1081 [60] Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron Courville, and Yoshua
1082 Bengio. A Recurrent Latent Variable Model for Sequential Data. pages 1–9, jun 2015.
- 1083 [61] Aaron Goodfellow, Ian, Bengio, Yoshua, Courville. Deep Learning. *MIT Press*, 2016.
- 1084 [62] Chris J. Maddison, Dieterich Lawson, George Tucker, Nicolas Heess, Mohammad Norouzi,
1085 Andriy Mnih, Arnaud Doucet, and Yee Whye Teh. Filtering Variational Objectives. (Nips):1–
1086 11, 2017.
- 1087 [63] Tuan Anh Le, Maximilian Igl, Tom Jin, Tom Rainforth, and Frank Wood. Auto-Encoding
1088 Sequential Monte Carlo. 2017.
- 1089 [64] Danilo Jimenez Rezende and Shakir Mohamed. Variational Inference with Normalizing Flows.
1090 *Proceedings of the 32nd International Conference on Machine Learning*, 37:1530–1538, may
1091 2015.

- 1092 [65] Diederik P. Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max
1093 Welling. Improving Variational Inference with Inverse Autoregressive Flow. (Nips), 2016.
- 1094 [66] Jakub M. Tomczak and Max Welling. Improving Variational Auto-Encoders using House-
1095 holder Flow. 2, 2016.
- 1096 [67] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance Weighted Autoencoders.
1097 pages 1–14, 2015.
- 1098 [68] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhut-
1099 dinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of*
1100 *Machine Learning Research*, 15:1929–1958, 2014.
- 1101 [69] D J Rezende, S Mohamed, and D Wierstra. Stochastic backpropagation and approximate
1102 inference in deep generative models. *Proceedings of The 31st . . .*, 32:1278–1286, 2014.
- 1103 [70] Diederik P. Kingma, Tim Salimans, and Max Welling. Variational Dropout and the Local
1104 Reparameterization Trick. (Mcmc):1–14, 2015.
- 1105 [71] Philip L. Smith, Roger Ratcliff, and Gail McKoon. The diffusion model is not a deterministic
1106 growth model: Comment on Jones and Dzhaminov (2014). *Psychological Review*, 121(4):679–
1107 688, 2014.
- 1108 [72] Klaas Enno Stephan, Will D. Penny, Jean Daunizeau, Rosalyn J. Moran, and Karl J. Friston.
1109 Bayesian model selection for group studies. *NeuroImage*, 46(4):1004–1017, 2009.
- 1110 [73] L. Rigoux, K. E. Stephan, K. J. Friston, and J. Daunizeau. Bayesian model selection for
1111 group studies - Revisited. *NeuroImage*, 84:971–985, 2014.
- 1112 [74] William Feller. An Introduction to Probability Theory and Its Applications. *Wiley*, 2:509,
1113 1968.
- 1114 [75] Daniel J. Navarro and Ian G. Fuss. Fast and accurate calculations for first-passage times in
1115 Wiener diffusion models. *Journal of Mathematical Psychology*, 53(4):222–230, 2009.
- 1116 [76] Matt Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic Variational Infer-
1117 ence. (2), 2012.

- 1118 [77] a W van der Vaart. *Asymptotic Statistics*. 1998.
- 1119 [78] Lucien Le Cam. *Asymptotic Methods in Statistical Decision Theory*. Springer Series in Statis-
1120 tics. Springer New York, New York, NY, 1986.
- 1121 [79] Sungjin Ahn, Anoop Korattikara, and Max Welling. Bayesian Posterior Sampling via Stochas-
1122 tic Gradient Fisher Scoring. *International Conference on Machine Learning*, cs.LG, 2012.
- 1123 [80] Stephan Mandt, Matthew D. Hoffman, and David M. Blei. Stochastic Gradient Descent as
1124 Approximate Bayesian Inference. *Mathematical Biosciences and Engineering*, 13(3):613–629,
1125 apr 2017.
- 1126 [81] Karl Friston, Jérémie Mattout, Nelson Trujillo-Barreto, John Ashburner, and Will Penny.
1127 Variational free energy and the Laplace approximation. *NeuroImage*, 34(1):220–234, 2007.
- 1128 [82] Christoph Mathys. A Bayesian foundation for individual learning under uncertainty. *Frontiers*
1129 *in Human Neuroscience*, 5(May):1–20, 2011.
- 1130 [83] Kay H. Brodersen, Christoph Mathys, Justin R. Chumbley, Jean Daunizeau, Cheng Soon
1131 Ong, Joachim M. Buhmann, and Klaas E. Stephan. Bayesian mixed-effects inference on
1132 classification performance in hierarchical data sets. *Journal of Machine Learning Research*,
1133 13:3133–3176, 2012.
- 1134 [84] Christoph D Mathys, Ekaterina I Lomakina, Jean Daunizeau, Sandra Iglesias, Kay H Broder-
1135 sen, Karl J Friston, and Klaas E Stephan. Uncertainty in perception and the Hierarchical
1136 Gaussian Filter. *Frontiers in Human Neuroscience*, 8(November):825, nov 2014.
- 1137 [85] Karl Friston, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, John O’Doherty,
1138 and Giovanni Pezzulo. Active inference and learning. *Neuroscience and Biobehavioral Reviews*,
1139 68:862–879, 2016.
- 1140 [86] Tommi S Jaakkola and Michael I Jordan. A variational approach to Bayesian logistic regres-
1141 sion models and their extensions. *Aistats*, (AUGUST 2001), 1996.
- 1142 [87] David M Blei. Variational Inference. *Cs.Princeton.Edu*, pages 1–12, 2002.
- 1143 [88] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational Inference: A Review for
1144 Statisticians. *arXiv*, pages 1–33, 2016.

- 1145 [89] Masa-Aki Sato. Online Model Selection Based on the Variational Bayes. *Neural Comput.*,
1146 13(7):1649–1681, 2001.
- 1147 [90] John Winn, Cm Bishop, and T Jaakkola. Variational Message Passing. *Journal of Machine*
1148 *Learning Research*, 6:661–694, 2005.
- 1149 [91] David Knowles and Thomas P. Minka. Non-conjugate variational message passing for multi-
1150 nomial and binary regression. *Nips*, pages 1–9, 2011.
- 1151 [92] James Hensman, Magnus Rattray, and Neil D. Lawrence. Fast Variational Inference in the
1152 Conjugate Exponential Family. pages 1–20, 2012.
- 1153 [93] Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. dec 2013.
- 1154 [94] Evan Archer, Il Memming Park, Lars Buesing, John Cunningham, and Liam Paninski. Black
1155 box variational inference for state space models. *International Conference on Learning Rep-*
1156 *resentations*, pages 1–11, 2016.
- 1157 [95] Michael C. Fu. Chapter 19 Gradient Estimation. In *Handbooks in Operations Research and*
1158 *Management Science*, volume 13, pages 575–616. 2006.
- 1159 [96] Pierre L’Ecuyer. A Unified View of Infinitesimal Perturbation Analysis and Likelihood Ratios.
1160 Technical report, US Dept of the Army, feb 1989.
- 1161 [97] P. Glasserman. Performance continuity and differentiability in Monte Carlo optimization. In
1162 *1988 Winter Simulation Conference Proceedings*, pages 518–524. IEEE.
- 1163 [98] Francisco J. R. Ruiz, Michalis K. Titsias, and David M. Blei. Overdispersed Black-Box
1164 Variational Inference. 2016.
- 1165 [99] Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. *The Annals of*
1166 *Mathematical Statistics*, 22(3):400–407, 1951.
- 1167 [100] L Bottou. Stochastic Gradient Learning in Neural Networks. *Proceedings of Neuro-Nimes*,
1168 91(8), 1991.
- 1169 [101] John Duchi, Elad Hazan, and Yoram Singer. Adaptive Subgradient Methods for Online
1170 Learning and Stochastic Optimization. *Journal of Machine Learning Research*, 12:2121–2159,
1171 2011.

- 1172 [102] Matthew D. Zeiler. ADADELTA: An Adaptive Learning Rate Method. dec 2012.
- 1173 [103] Rajesh Ranganath, Dustin Tran, and David M Blei. Hierarchical Variational Models. *arXiv*,
1174 pages 1–9, 2014.
- 1175 [104] Dustin Tran, Rajesh Ranganath, and David M. Blei. Variational Gaussian Process. *Iclr*,
1176 pages 1–14, 2015.
- 1177 [105] Dustin Tran, Matthew D. Hoffman, Rif A. Saurous, Eugene Brevdo, Kevin Murphy, and
1178 David M. Blei. Deep Probabilistic Programming. pages 1–18, 2017.
- 1179 [106] George Marsaglia. Generating a Variable from the Tail of the Normal Distribution. *Techno-*
1180 *metrics*, 6(1):101, feb 1964.
- 1181 [107] George Marsaglia and Wai Wan Tsang. A simple method for generating gamma variables.
1182 *ACM Transactions on Mathematical Software*, 26(3):363–372, 2000.
- 1183 [108] David A. Knowles. Stochastic gradient variational Bayes for gamma approximating distribu-
1184 tions. pages 1–14, 2015.
- 1185 [109] Christian A. Naesseth, Francisco J. R. Ruiz, Scott W. Linderman, and David M. Blei. Repa-
1186 rameterization Gradients through Acceptance-Rejection Sampling Algorithms. 54, 2016.
- 1187 [110] Dmitry Molchanov, Arsenii Ashukha, and Dmitry Vetrov. Variational Dropout Sparsifies
1188 Deep Neural Networks. 2017.
- 1189 [111] Yarin Gal and Zoubin Ghahramani. A Theoretically Grounded Application of Dropout in
1190 Recurrent Neural Networks. 19994575, dec 2015.
- 1191 [112] Martijn J Mulder, Eric-Jan Wagenmakers, Roger Ratcliff, Wouter Boekel, and Birte U
1192 Forstmann. Bias in the brain: a diffusion model analysis of prior probability and poten-
1193 tial payoff. *The Journal of neuroscience : the official journal of the Society for Neuroscience*,
1194 32(7):2335–43, feb 2012.
- 1195 [113] Carl E. Rasmussen and Christopher K. I. Williams. *Gaussian processes for machine learning.*,
1196 volume 14. 2004.

- 1197 [114] P. E. Kloeden and E. Platen. *Numerical solution of stochastic differential equations*, vol-
1198 ume 23. 1992.
- 1199 [115] Stijn Verdonck, Kristof Meers, and Francis Tuerlinckx. Efficient simulation of diffusion-based
1200 choice RT models on CPU and GPU. *Behavior Research Methods*, mar 2015.
- 1201 [116] Yuhuang Hu, Adrian Huber, Jithendar Anumula, and Shih-Chii Liu. Overcoming the vanish-
1202 ing gradient problem in plain recurrent networks. (Section 2):1–20, 2018.
- 1203 [117] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi
1204 Bougares, Holger Schwenk, and Yoshua Bengio. Learning Phrase Representations using RNN
1205 Encoder-Decoder for Statistical Machine Translation. 2014.
- 1206 [118] Sepp Hochreiter and J Uergen Schmidhuber. Long Short-Term Memory. *Neural Computation*,
1207 9(8):1735–1780, 1997.
- 1208 [119] Richard H. R. Hahnloser, Rahul Sarpeshkar, Misha A. Mahowald, Rodney J. Douglas, and
1209 H. Sebastian Seung. Erratum: Digital selection and analogue amplification coexist in a
1210 cortex-inspired silicon circuit. *Nature*, 405(6789):947–951, jun 2000.
- 1211 [120] Vinod Nair and Geoffrey E Hinton. Rectified Linear Units Improve Restricted Boltzmann
1212 Machines. *Proceedings of the 27th International Conference on Machine Learning*, (3):807–
1213 814, 2010.
- 1214 [121] Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B. Shah. Julia: A Fresh Approach
1215 to Numerical Computing. *SIAM Review*, 59(1):65–98, jan 2017.
- 1216 [122] Vincent Moens. Hierarchical Adaptive Forgetting Variational Filtering. *In prep.*, 2018.
- 1217 [123] John Langford, Alexander J. Smola, and Martin Zinkevich. Slow Learners are Fast. *Advances*
1218 *in Neural Information Processing Systems - NIPS’09*, 1(2099):2331—2339, 2009.
- 1219 [124] Aki Vehtari, Andrew Gelman, and Jonah Gabry. Pareto Smoothed Importance Sampling.
1220 (October), 2015.
- 1221 [125] Yuling Yao, Aki Vehtari, Daniel Simpson, and Andrew Gelman. Yes, but Did It Work?:
1222 Evaluating Variational Inference. pages 1–22, 2018.

- 1223 [126] Tim Salimans, Diederik P Kingma, and Max Welling. Markov Chain Monte Carlo and Vari-
1224 ational Inference : Bridging the Gap. *International Conference on Machine Learning*, 2015.
- 1225 [127] Rajesh Ranganath, Sean Gerrish, and David M Blei. Black Box Variational Inference. *Aistats*,
1226 33, 2013.
- 1227 [128] Yu-Ting Huang, Dejan Georgiev, Tom Foltynie, Patricia Limousin, Maarten Speekenbrink,
1228 and Marjan Jahanshahi. Different effects of dopaminergic medication on perceptual decision-
1229 making in Parkinson’s disease as a function of task difficulty and speed–accuracy instructions.
1230 *Neuropsychologia*, 75:577–587, aug 2015.
- 1231 [129] James F. Cavanagh, Michael J. Frank, Theresa J. Klein, and John J.B. B Allen. Frontal
1232 theta links prediction errors to behavioral adaptation in reinforcement learning. *NeuroImage*,
1233 49(4):3198–3209, 2010.
- 1234 [130] James F. Cavanagh and Michael J. Frank. Frontal theta as a mechanism for cognitive control.
1235 *Trends in Cognitive Sciences*, 18(8):414–421, aug 2014.
- 1236 [131] Vaibhav Srivastava, Samuel F. Feng, Jonathan D. Cohen, Naomi Ehrlich Leonard, and Amitai
1237 Shenhav. A martingale analysis of first passage times of time-dependent Wiener diffusion
1238 models. *Journal of Mathematical Psychology*, 77:94–110, 2017.
- 1239 [132] Peter I. Frazier and Angela J Yu. Sequential hypothesis testing under stochastic deadlines.
1240 *Advances in Neural Information Processing Systems*, pages 1–8, 2008.
- 1241 [133] Milica Milosavljevic, Jonathan Malmaud, and Alexander Huth. The Drift Diffusion Model
1242 can account for the accuracy and reaction time of value-based choices under high and low
1243 time pressure. *Judgment and Decision Making*, 5(6):437–449, 2010.
- 1244 [134] David Cox and Hilton David Miller. *The Theory of Stochastic Processes*. 1965.
- 1245 [135] Mario Lefebvre. First hitting place distributions for the Ornstein-Uhlenbeck process. *Statistics*
1246 *& Probability Letters*, 34(3):309–312, jun 1997.
- 1247 [136] Adele Diederich. Dynamic Stochastic Models for Decision Making under Time Constraints.
1248 *Journal of Mathematical Psychology*, 41(3):260–274, sep 1997.

- 1249 [137] Olivier Cappé, Eric Moulines, and Tobias Ryden. *Inference in Hidden Markov Models*. 2006.
- 1250 [138] Herbert Robbins. The Empirical Bayes Approach to Statistical Decision Problems. *The*
1251 *Annals of Mathematical Statistics*, 35(1):1–20, 1964.
- 1252 [139] Manfred Opper and Cédric Archambeau. The Variational Gaussian Approximation Revisited.
1253 *Neural Computation*, 21(3):786–792, 2009.
- 1254 [140] Z. I. Botev. The normal law under linear restrictions: simulation and estimation via minimax
1255 tilting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(1):125–
1256 148, jan 2016.