Supplementary information

Fluctuating internal states mediate neural-behavioral covariations in V1

Baowang Li, Jason Samonds, Yuzhi Chen,
Thibaud Taillefumier, Nicholas Priebe, and Eyal Seidemann

1 Decision models

1.1 Additive model

We denote by V_{i} the membrane voltage of the recorded neuron whose receptive field location determines the "in side" of the field of view. Similarly, we denote by V_{o} the membrane voltage of a neuron with symmetric receptive-field location in the opposite hemifield, in the "out side" of the field of view. In the additive model, we consider that these voltages are stochastically related via

$$\begin{aligned} V_{\mathrm{i}} &= X_{\mathrm{i}} + A_{\mathrm{i}} + N_{\mathrm{i,e}}\,, \\ V_{\mathrm{o}} &= X_{\mathrm{o}} + A_{\mathrm{o}} + N_{\mathrm{o,e}}\,, \end{aligned}$$

where X_{i} and X_{o} denote stimulus strengths, $N_{\mathrm{i,e}}$ and $N_{\mathrm{o,e}}$ denote (early) sensory noise. We further consider that variable internal states contribute to voltage fluctuations as nonnegative additive random variables A_{i} and A_{o}, some distribution p_A such that $A \geq 0$ with $\mathbb{E}\,[A] = 1$. This internal state can be thought as representing attention or arousal levels, and for conciseness, we will refer to A_{i} and A_{o} as attention variables.
Given some criterion c, the decision variable is $D = \mathbb{1}_{\{V_{\mathrm{i}} - V_{\mathrm{o}} \geq c + N_{\mathrm{d}}\}}$, where N_{d} denotes some downstream noise. All noise distributions will be assumed Gaussian centered, which means that we need to consider two standard deviation parameters σ_{e} and σ_{d}.

1.2 Multiplicative model

As for the additive model, V_{i} and V_{o} refers to the voltage on the "in side" and the "out side", respectively. In the multiplicative model, we consider that these voltages follows

$$\begin{aligned} V_{\mathrm{i}} &= A_{\mathrm{i}}(X_{\mathrm{i}} + N_{\mathrm{i,e}}) + N_{\mathrm{i,l}}\,, \\ V_{\mathrm{o}} &= A_{\mathrm{o}}(X_{\mathrm{o}} + N_{\mathrm{o,e}}) + N_{\mathrm{o,l}}\,, \end{aligned}$$

where X_{i} and X_{o} denote stimulus strengths, $N_{\mathrm{i,e}}$ and $N_{\mathrm{o,e}}$ denote early sensory noise, and $N_{\mathrm{i,l}}$ and $N_{\mathrm{o,l}}$ denote late sensory noise. The hallmark of the multiplicative model is to consider that variable internal states are modeled by gain random variables A_{i} and A_{o}, with some distribution p_A such that $A \geq 0$ with $\mathbb{E}\,[A] = 1$. As for the additive model, this internal state can be thought as representing attention or arousal levels, and for conciseness, we will refer to A_{i} and A_{o} as attention variables. One could allow for gain variables to also contribute additively to voltage fluctuations by including constant offset terms within parentheses in the above equations. Such inclusions, however, will not impact our results, and for this reason, we only consider A_{i} and A_{o} as pure gain variables in the multiplicative model. Given some criterion c, the decision variable is $D = \mathbb{1}_{\{V_{\mathrm{i}} - V_{\mathrm{o}} \geq c + N_{\mathrm{i,d}} + N_{\mathrm{o,d}}\}}$, where $N_{\mathrm{i,d}}$ and $N_{\mathrm{o,d}}$ denote some hemisphere-specific downstream noise. All noise distributions will be assumed independent Gaussian centered, which means that we need to consider three standard deviation parameters σ_{e}, σ_{l}, and σ_{d}.

1.3 Probabilities of decision outcomes

For simplicity, we will always assume no decision bias, i.e. $c = 0$. There is no lack of generality in doing so as considering a decision bias is equivalent to shifting the means μ_{V_i} and μ_{V_o}, which features as free parameters in our derivations. The probability of the decision outcomes D_i and D_o admits an explicit integral expression involving the joint probabilities of the voltages $p(v_\mathrm{i}, v_\mathrm{o} \,|\, (X_\mathrm{i}, X_\mathrm{o}))$. To see this, first observe that assuming knowledge of $(X_\mathrm{i}, X_\mathrm{o})$ and denoting the decision variable as $W = V_\mathrm{i} - V_\mathrm{o} + N_\mathrm{i,d} - N_\mathrm{o,d}$, we have

$$p(W, V_\mathrm{i}, V_\mathrm{o}) = p(W \,|\, V_\mathrm{i}, V_\mathrm{o}) p(V_\mathrm{i}, V_\mathrm{o}) = g_{\sqrt{2}\sigma_\mathrm{d}}(W - (V_\mathrm{i} - V_\mathrm{o})) p(V_\mathrm{i}, V_\mathrm{o})\,,$$

by independence of the hemisphere-specific decision noises. Then, we have

$$\begin{aligned} p\big[D_\mathrm{i} \,|\, (X_\mathrm{i}, X_\mathrm{o})\big] &= \int \mathrm{d}v_\mathrm{i} \int \mathrm{d}v_\mathrm{o} \int \mathrm{d}w\, \mathbb{1}_{\{w \geq 0\}} p(w, v_\mathrm{i}, v_\mathrm{o} \,|\, (X_\mathrm{i}, X_\mathrm{o}))\,, \\ &= \int \mathrm{d}v_\mathrm{i} \int \mathrm{d}v_\mathrm{o}\, p(v_\mathrm{i}, v_\mathrm{o} \,|\, (X_\mathrm{i}, X_\mathrm{o})) \int \mathrm{d}w\, \mathbb{1}_{\{w \geq 0\}} g_{\sqrt{2}\sigma_\mathrm{d}}(w - (v_\mathrm{i} - v_\mathrm{o}))\,, \\ &= \int \mathrm{d}v_\mathrm{i} \int \mathrm{d}v_\mathrm{o}\, p(v_\mathrm{i}, v_\mathrm{o} \,|\, (X_\mathrm{i}, X_\mathrm{o}))\, (1 + \mathrm{erf}\,[(v_\mathrm{i} - v_\mathrm{o})/\sigma_\mathrm{d}])\, /2\,, \end{aligned}$$

so that we also have

$$p\big[D_\mathrm{o} \,|\, (X_i, X_o)\big] = 1 - p\big[D_\mathrm{i} \,|\, (X_i, X_o)\big] = \int \mathrm{d}v_\mathrm{i} \int \mathrm{d}v_\mathrm{o}\, p(v_\mathrm{i}, v_\mathrm{o} \,|\, (X_\mathrm{i}, X_\mathrm{o}))\, (1 - \mathrm{erf}\,[v_\mathrm{i} - v_\mathrm{o})/\sigma_\mathrm{d}])\, /2\,.$$

where $\mathrm{erf}\,[\cdot]$ denotes the error function.

1.4 Symmetric probability model for variable internal state

For simplicity, we consider a model with variable internal states for which each hemisphere alternates between a low-attention state $A_\mathrm{ll} > 0$ and a high-attention state $A_\mathrm{hh} > 0$ with $A_\mathrm{ll} < A_\mathrm{hh}$. To probe the role of correlations across hemispheres while maintaining symmetry, we consider that the attention states of both hemispheres can be recapitulated by the following two-dimensional probabilistic model:

$$\begin{aligned} \mathbb{P}\,[A_\mathrm{ll}] = \mathbb{P}\,[A_\mathrm{i} = A_\mathrm{ll}, A_\mathrm{o} = A_\mathrm{ll}] = p\,, \qquad & \mathbb{P}\,[A_\mathrm{hl}] = \mathbb{P}\,[A_\mathrm{i} = A_\mathrm{hh}, A_\mathrm{o} = A_\mathrm{ll}] = q\,, \\ \mathbb{P}\,[A_\mathrm{lh}] = \mathbb{P}\,[A_\mathrm{i} = A_\mathrm{ll}, A_\mathrm{o} = A_\mathrm{hh}] = q\,, \qquad & \mathbb{P}\,[A_\mathrm{hh}] = \mathbb{P}\,[A_\mathrm{i} = A_\mathrm{hh}, A_\mathrm{o} = A_\mathrm{hh}] = 1 - p - 2q\,. \end{aligned}$$

For $q = 0$, there is perfect correlation across hemispheres, i.e., $(A_\mathrm{i}, A_\mathrm{o})$ alternates between $(A_\mathrm{ll}, A_\mathrm{ll})$ and $(A_\mathrm{hh}, A_\mathrm{hh})$ with probability p, whereas for $q = 1/2$, which implies $p = 0$, there is perfect anticorrelation across hemispheres, i.e., $A_\mathrm{i} = A_\mathrm{hh}$ whenever $A_\mathrm{o} = A_\mathrm{ll}$ and vice versa. More generally, the correlation across hemispheres satisfies

$$\rho = \frac{p - (p+q)^2}{(p+q)(1-p-q)}\,, \tag{1}$$

so that zero correlation is achieved for $q = \sqrt{p}(1 - \sqrt{p}) \leq 1/4$, with independence for $q = p = 1/4$.

1.5 Choice discriminability via δ-metric

Here, we assume that the gain variables are fixed to deterministic values for each hemisphere: $A_\mathrm{i} = a_i$ and $A_\mathrm{o} = a_\mathrm{o}$. Under these assumptions, the models under consideration are linear, allowing one to carry out explicit calculations for Gaussian noise. Specifically, one can find a closed-form expression for a simple

choice-discriminability metric associated to the alternative forced choice D. Denoting the "choice in" by $D_{\mathrm{i}} = D$ and the "choice out" by $D_{\mathrm{o}} = 1 - D$ for conciseness, this metric is defined as a difference between conditional expectations

$$\delta(T) = \mathbb{E}\left[V_{\mathrm{i}} \,|\, D_{\mathrm{i}}, T\right] - \mathbb{E}\left[V_{\mathrm{i}} \,|\, D_{\mathrm{o}}, T\right] = \frac{\mathbb{E}\left[V_{\mathrm{i}} D_{\mathrm{i}} \,|\, T\right]}{\mathbb{P}\left[D_{\mathrm{i}} \,|\, T\right]} - \frac{\mathbb{E}\left[V_{\mathrm{i}} D_{\mathrm{o}} \,|\, T\right]}{\mathbb{P}\left[D_{\mathrm{o}} \,|\, T\right]},$$

where the conditioning is with respect to the decision outcome D. This metric is closely related to the classical δ-metric, which is derived from δ by normalizing with respect to the standard deviations of the involved distributions. We consider the δ-metric for three stimulus conditions. In all these conditions, we assume that the presence of target-in the receptive field of a neuron leads to an elevation of the mean voltage level for this neuron, i.e., the presence or absence of a stimulus corresponds to values x_1 and x_0, respectively, such that $x_1 > x_0$. Then, the three considered stimulus conditions are:

"target-in": $T_{\mathrm{i}} = \{X_{\mathrm{i}} = x_1, X_{\mathrm{o}} = x_0\}$.

"target-out": $T_{\mathrm{o}} = \{X_{\mathrm{i}} = x_0, X_{\mathrm{o}} = x_1\}$.

"no-stimulus": $T_{\mathrm{n}} = \{X_{\mathrm{i}} = x_0, X_{\mathrm{o}} = x_0\}$.

2 Choice discriminability with fixed attention state

Independent of the stimulus condition, given fixed attention states $A_{\mathrm{i}} = a_i$ and $A_{\mathrm{o}} = a_{\mathrm{o}}$, V_{i} and V_{o} follow independent Gaussian distributions so that

$$p(V_{\mathrm{i}}, V_{\mathrm{o}} \,|\, (X_{\mathrm{i}}, X_{\mathrm{o}})) = g_{\sigma_{V_{\mathrm{i}}}}(v_{\mathrm{i}} - \mu_{V_{\mathrm{i}}}) g_{\sigma_{V_{\mathrm{o}}}}(v_{\mathrm{o}} - \mu_{V_{\mathrm{o}}}),$$

where $g_{\sigma_{V_{\mathrm{i}}}}$ and $g_{\sigma_{V_{\mathrm{o}}}}$ denotes the density of centered Gaussian variables with variance $\sigma^2_{V_{\mathrm{i}}}$ and $\sigma^2_{V_{\mathrm{o}}}$, respectively. The means and variances are given by:

additive model : $\mu_{V_{\mathrm{i}}} = X_{\mathrm{i}} + a_{\mathrm{i}}$, $\mu_{V_{\mathrm{o}}} = X_{\mathrm{o}} + a_{\mathrm{o}}$, $\sigma^2_{V_{\mathrm{i}}} = \sigma^2_{\mathrm{e}}$, $\sigma^2_{V_{\mathrm{o}}} = \sigma^2_{\mathrm{e}}$.

multiplicative model: $\mu_{V_{\mathrm{i}}} = a_{\mathrm{i}} X_{\mathrm{i}}$, $\mu_{V_{\mathrm{o}}} = a_{\mathrm{o}} X_{\mathrm{o}}$, $\sigma^2_{V_{\mathrm{i}}} = a^2_{\mathrm{i}} \sigma^2_{\mathrm{e}} + \sigma^2_{\mathrm{l}}$, $\sigma^2_{V_{\mathrm{o}}} = a^2_{\mathrm{o}} \sigma^2_{\mathrm{e}} + \sigma^2_{\mathrm{l}}$.

For a fixed attention state, independent of the nature of the model, the decision process based on a competition between the voltage signals of both hemispheres can be represented schematically. We utilize such a representation in Fig 1 in relation to the definition of the δ-metric as the difference between two choice-conditioned voltage means for V_{i}. In the Gaussian context, one can turn this schematic view into analytical expressions allowing one to explore the parametric dependence of the δ-metric. For instance, one can express the probability of the decision outcomes D_{i} and D_{o} explicitly. The derivation relies on performing the change of variable $(x, y) = (v_{\mathrm{i}} - v_{\mathrm{o}}, v_{\mathrm{i}} + v_{\mathrm{o}})$ to evaluate two iterated integrals. For instance,

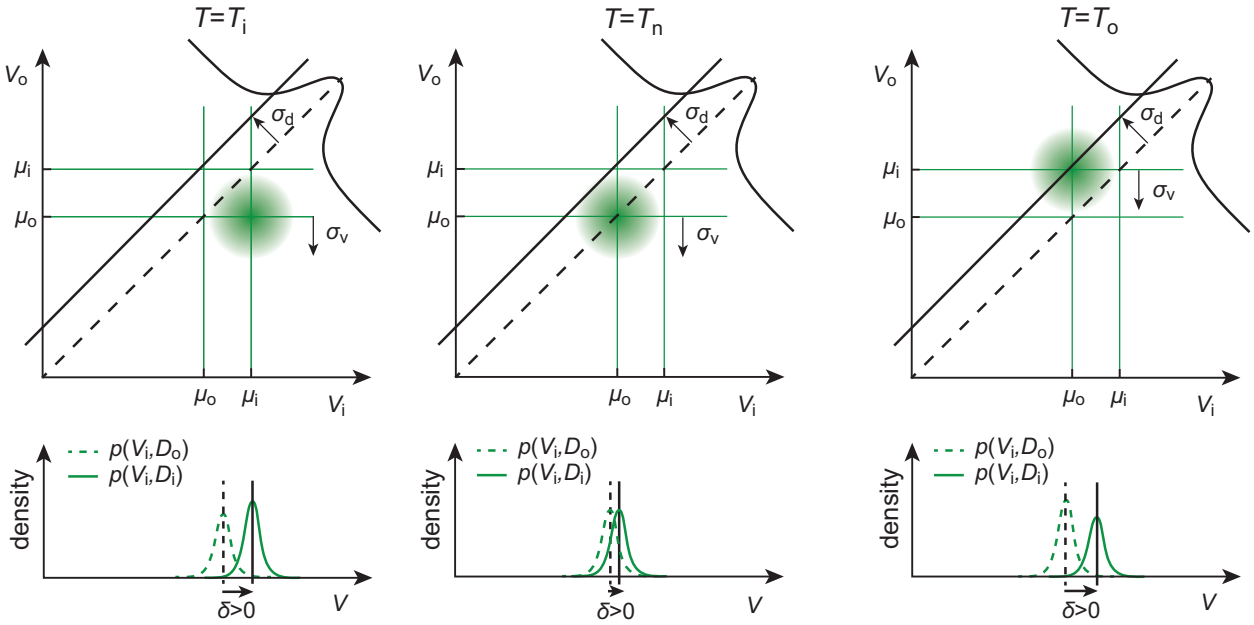

Figure 1: **Schematic for decision with fixed gain.** We consider three stimulus conditions: "target-in", i.e., $T = T_\mathrm{i}$, "no-stimulus", i.e., $T = T_\mathrm{n}$, and "target-out", i.e., $T = T_\mathrm{o}$. For fixed gain, the model is purely Gaussian for all stimulus conditions, which impact the distribution of voltages $(V_\mathrm{i}, V_\mathrm{o})$ via variables means. The assumption is that the presence of a stimulus is assumed to cause an increase in mean voltages as depicted in the two-dimensional phase space (top panels). Within this picture, the decision outcome is determined by the position of the voltage point $(V_\mathrm{i}, V_\mathrm{o})$ with respect to the decision boundary, a slope-one line with Gaussian-distributed intercept with variance σ_d^2 (solid black line). Recorded voltage data correspond to the projection of the phase space on the V_i-space conditioning to the decision outcome (bottom panels): for $D = D_\mathrm{i}$ and $D = D_\mathrm{o}$, this corresponds to marginalizing the distribution of voltages $(V_\mathrm{i}, V_\mathrm{o})$ restricted to the portion of the phase plane below or above the fluctuating decision boundary (solid black line in top panels), respectively. For large downstream decision variance, the joint probability densities $p(V_\mathrm{i}, D_\mathrm{i} \,|\, T)$ and $p(V_\mathrm{i}, D_\mathrm{i} \,|\, T)$ should integrate to about the same the choice probabilities, leading to nonnegative but low δ-metric values.

the probability of "choice in" is given as

$$\begin{aligned}
\mathbb{P}\left[D_{\mathrm{i}} \,|\, (X_{\mathrm{i}}, X_{\mathrm{o}})\right] &= \frac{1}{4\pi\sigma_{V_{\mathrm{i}}}\sigma_{V_{\mathrm{o}}}} \iint \exp\left(-\frac{(v_{\mathrm{i}}-\mu_{V_{\mathrm{i}}})^2}{2\sigma_{V_{\mathrm{i}}}^2} - \frac{(v_{\mathrm{o}}-\mu_{V_{\mathrm{o}}})^2}{2\sigma_{V_{\mathrm{o}}}^2}\right)\left(1+\operatorname{erf}\left[\frac{v_{\mathrm{i}}-v_{\mathrm{o}}}{\sigma_{\mathrm{d}}}\right]\right) \mathrm{d}v_{\mathrm{i}}\mathrm{d}v_{\mathrm{o}}\,, \\
&= \frac{1}{2\pi\sigma_{V_{\mathrm{i}}}\sigma_{V_{\mathrm{o}}}} \iint \exp\left(-\frac{(x+y-\mu_{V_{\mathrm{i}}})^2}{2\sigma_{V_{\mathrm{i}}}^2} - \frac{(y-x-\mu_{V_{\mathrm{o}}})^2}{2\sigma_{V_{\mathrm{o}}}^2}\right)\left(1+\operatorname{erf}\left[\frac{x}{\sigma_{\mathrm{d}}}\right]\right) \mathrm{d}x\mathrm{d}y\,, \\
&= \frac{1}{\sqrt{2\pi}\sqrt{\sigma_{V_{\mathrm{i}}}^2+\sigma_{V_{\mathrm{o}}}^2}} \int \exp\left(-\frac{(2x+\mu_{V_{\mathrm{o}}}-\mu_{V_{\mathrm{i}}})^2}{2(\sigma_{V_{\mathrm{i}}}^2+\sigma_{V_{\mathrm{o}}}^2)}\right)\left(1+\operatorname{erf}\left[\frac{x}{\sigma_{\mathrm{d}}}\right]\right) \mathrm{d}x\,, \\
&= \frac{1}{2}\left(1+\operatorname{erf}\left[\frac{\mu_{V_{\mathrm{i}}}-\mu_{V_{\mathrm{o}}}}{\sqrt{2\left(\sigma_{V_{\mathrm{i}}}^2+\sigma_{V_{\mathrm{o}}}^2+2\sigma_{\mathrm{d}}^2\right)}}\right]\right)\,,
\end{aligned}$$

where the last equality follows from the identity

$$\int \sqrt{\frac{a}{\pi}} e^{-(ax+b)^2} \operatorname{erf}[x]\, \mathrm{d}x = -\operatorname{erf}\left[\frac{b}{\sqrt{a^2+1}}\right]\,.$$

The complementary probability of "choice out" directly follows as

$$\mathbb{P}\left[D_{\mathrm{o}} \,|\, (X_{\mathrm{i}}, X_{\mathrm{o}})\right] = 1 - \mathbb{P}\left[D_{\mathrm{i}} \,|\, (X_{\mathrm{i}}, X_{\mathrm{o}})\right] = \frac{1}{2}\left(1-\operatorname{erf}\left[\frac{\mu_{V_{\mathrm{i}}}-\mu_{V_{\mathrm{o}}}}{\sqrt{2}\sqrt{\sigma_{V_{\mathrm{i}}}^2+\sigma_{V_{\mathrm{o}}}^2+2\sigma_{\mathrm{d}}^2}}\right]\right)\,.$$

Then, one can use similar calculations to derive the closed-form expression of the expectations $\mathbb{E}\left[V_{\mathrm{i}} D_{\mathrm{i}} \,|\, T\right]$ and $\mathbb{E}\left[V_{\mathrm{i}} D_{\mathrm{o}} \,|\, T\right]$. For instance, we find that

$$\begin{aligned}
&\mathbb{E}\left[V_{\mathrm{i}} D_{\mathrm{i}} \,|\, (X_{\mathrm{i}}, X_{\mathrm{o}})\right] \\
&= \frac{1}{4\pi\sigma_{V_{\mathrm{i}}}\sigma_{V_{\mathrm{o}}}} \iint v_{\mathrm{i}} \exp\left(-\frac{(v_{\mathrm{i}}-\mu_{V_{\mathrm{i}}})^2}{2\sigma_{V_{\mathrm{i}}}^2} - \frac{(v_{\mathrm{o}}-\mu_{V_{\mathrm{o}}})^2}{2\sigma_{V_{\mathrm{o}}}^2}\right)\left(1+\operatorname{erf}\left[\frac{v_{\mathrm{i}}-v_{\mathrm{o}}}{\sigma_{\mathrm{d}}}\right]\right) \mathrm{d}v_{\mathrm{i}}\mathrm{d}v_{\mathrm{o}}\,, \\
&= \frac{1}{4\pi\sigma_{V_{\mathrm{i}}}\sigma_{V_{\mathrm{o}}}} \iint (x+y) \exp\left(-\frac{(x+y-\mu_{V_{\mathrm{i}}})^2}{2\sigma_{V_{\mathrm{i}}}^2} - \frac{(y-x-\mu_{V_{\mathrm{o}}})^2}{2\sigma_{V_{\mathrm{o}}}^2}\right)\left(1+\operatorname{erf}\left[\frac{x}{\sigma_{\mathrm{d}}}\right]\right) \mathrm{d}v_{\mathrm{i}}\mathrm{d}v_{\mathrm{o}}\,, \\
&= \iint \frac{\mu_{V_{\mathrm{o}}}\sigma_{V_{\mathrm{i}}}^2 + \mu_{V_{\mathrm{i}}}\sigma_{V_{\mathrm{o}}}^2 + 2\sigma_{V_{\mathrm{i}}}^2 x}{\sqrt{2\pi}\left(\sigma_{V_{\mathrm{i}}}^2+\sigma_{V_{\mathrm{o}}}^2\right)^{3/2}} \exp\left(-\frac{(2x+\mu_{V_{\mathrm{o}}}-\mu_{V_{\mathrm{i}}})^2}{2(\sigma_{V_{\mathrm{i}}}^2+\sigma_{V_{\mathrm{o}}}^2)}\right)\left(1+\operatorname{erf}\left[\frac{x}{\sigma_{\mathrm{d}}}\right]\right) \mathrm{d}v_{\mathrm{i}}\mathrm{d}v_{\mathrm{o}}\,, \\
&= \frac{\mu_{V_{\mathrm{i}}}}{2}\left(1+\operatorname{erf}\left[\frac{\mu_{V_{\mathrm{i}}}-\mu_{V_{\mathrm{o}}}}{\sqrt{2}\sqrt{\sigma_{V_{\mathrm{i}}}^2+\sigma_{V_{\mathrm{o}}}^2+2\sigma_{\mathrm{d}}^2}}\right]\right) + \\
&\qquad \frac{\sigma_{V_{\mathrm{i}}}^2}{\sqrt{2\pi}\sqrt{\sigma_{V_{\mathrm{i}}}^2+\sigma_{V_{\mathrm{o}}}^2+2\sigma_{\mathrm{d}}^2}} \exp\left(-\frac{(\mu_{V_{\mathrm{i}}}-\mu_{V_{\mathrm{o}}})^2}{2\left(\sigma_{V_{\mathrm{i}}}^2+\sigma_{V_{\mathrm{o}}}^2+2\sigma_{\mathrm{d}}^2\right)}\right)\,.
\end{aligned}$$

Normalizing by the choice probability $\mathbb{P}\left[D_{\mathrm{i}} \,|\, (X_{\mathrm{i}}, X_{\mathrm{o}})\right]$ yields the conditional expectation

$$\mathbb{E}\left[V_{\mathrm{i}} \,|\, D_{\mathrm{i}}, (X_{\mathrm{i}}, X_{\mathrm{o}})\right] = \mu_{V_{\mathrm{i}}} + \frac{\sqrt{2}\sigma_{V_{\mathrm{i}}}^2}{\sqrt{\pi}\sqrt{\sigma_{V_{\mathrm{i}}}^2+\sigma_{V_{\mathrm{o}}}^2+2\sigma_{\mathrm{d}}^2}}\left(\frac{\exp\left(-\frac{(\mu_{V_{\mathrm{i}}}-\mu_{V_{\mathrm{o}}})^2}{2\left(\sigma_{V_{\mathrm{i}}}^2+\sigma_{V_{\mathrm{o}}}^2+2\sigma_{\mathrm{d}}^2\right)}\right)}{1+\operatorname{erf}\left[\frac{\mu_{V_{\mathrm{i}}}-\mu_{V_{\mathrm{o}}}}{\sqrt{2}\sqrt{\sigma_{V_{\mathrm{i}}}^2+\sigma_{V_{\mathrm{o}}}^2+2\sigma_{\mathrm{d}}^2}}\right]}\right)\,,$$

whereas similar calculations for the "choice out" yield

$$\mathbb{E}\left[V_{\mathrm{i}} \,|\, D_{\mathrm{o}}, (X_{\mathrm{i}}, X_{\mathrm{o}})\right] = \mu_{V_{\mathrm{i}}} - \frac{\sqrt{2}\sigma^2_{V_{\mathrm{i}}}}{\sqrt{\pi}\sqrt{\sigma^2_{V_{\mathrm{i}}} + \sigma^2_{V_{\mathrm{o}}} + 2\sigma^2_{\mathrm{d}}}} \left(\frac{\exp\left(-\frac{(\mu_{V_{\mathrm{i}}} - \mu_{V_{\mathrm{o}}})^2}{2\left(\sigma^2_{V_{\mathrm{i}}} + \sigma^2_{V_{\mathrm{o}}} + 2\sigma^2_{\mathrm{d}}\right)} \right)}{1 - \mathrm{erf}\left[\frac{\mu_{V_{\mathrm{i}}} - \mu_{V_{\mathrm{o}}}}{\sqrt{2}\sqrt{\sigma^2_{V_{\mathrm{i}}} + \sigma^2_{V_{\mathrm{o}}} + 2\sigma^2_{\mathrm{d}}}} \right]} \right) .$$

As expected, conditioning on the decision outcome D introduces biases in the evaluation of the expected values of V_{i} and V_{o}, which otherwise evaluate to $\mu_{V_{\mathrm{i}}}$ and $\mu_{V_{\mathrm{o}}}$, respectively, in the absence of conditioning. One can check that these biases vanish in the limit of large downstream decision noise $\sigma_{\mathrm{d}} \to \infty$, as the decision no longer depends on the voltages V_{i} and V_{o}. Moreover, these biases simplify when $b = \mu_{V_{\mathrm{i}}} - \mu_{V_{\mathrm{o}}}$ is large compared with the total noise level. Indeed, when $(\mu_{V_{\mathrm{i}}} - \mu_{V_{\mathrm{o}}})^2 \gg \sigma^2_{V_{\mathrm{i}}} + \sigma^2_{V_{\mathrm{o}}} + 2\sigma^2_{\mathrm{d}}$, we have

$$\mathbb{E}\left[V_{\mathrm{i}} \,|\, D_{\mathrm{i}}, (X_{\mathrm{i}}, X_{\mathrm{o}})\right] \underset{b\to\infty}{\sim} \mu_{V_{\mathrm{i}}} \qquad \text{and} \qquad \mathbb{E}\left[V_{\mathrm{i}} \,|\, D_{\mathrm{o}}, (X_{\mathrm{i}}, X_{\mathrm{o}})\right] \underset{b\to\infty}{\sim} \mu_{V_{\mathrm{i}}} \left(1 - \frac{\sigma^2_{V_{\mathrm{i}}}}{\sigma^2_{V_{\mathrm{i}}} + \sigma^2_{V_{\mathrm{o}}} + 2\sigma^2_{\mathrm{d}}} \right) ,$$

$$\mathbb{E}\left[V_{\mathrm{i}} \,|\, D_{\mathrm{o}}, (X_{\mathrm{i}}, X_{\mathrm{o}})\right] \underset{b\to-\infty}{\sim} \mu_{V_{\mathrm{i}}} \qquad \text{and} \qquad \mathbb{E}\left[V_{\mathrm{i}} \,|\, D_{\mathrm{i}}, (X_{\mathrm{i}}, X_{\mathrm{o}})\right] \underset{b\to-\infty}{\sim} \mu_{V_{\mathrm{i}}} \left(1 - \frac{\sigma^2_{V_{\mathrm{i}}}}{\sigma^2_{V_{\mathrm{i}}} + \sigma^2_{V_{\mathrm{o}}} + 2\sigma^2_{\mathrm{d}}} \right) .$$

Thus conditioning gets more impactful when most of the variability originates from $\sigma_{V_{\mathrm{i}}}$ as opposed to $\sigma_{V_{\mathrm{o}}}$ and $\sigma_{V_{\mathrm{d}}}$.

Finally, the corresponding δ-metric admits the following closed-form expression

$$\delta(T) = \frac{2\sqrt{2}\sigma^2_{V_{\mathrm{i}}}}{\sqrt{\pi}\sqrt{\sigma^2_{V_{\mathrm{i}}} + \sigma^2_{V_{\mathrm{o}}} + 2\sigma^2_{\mathrm{d}}}} \left(\frac{\exp\left(-\frac{(\mu_{V_{\mathrm{i}}} - \mu_{V_{\mathrm{o}}})^2}{2\left(\sigma^2_{V_{\mathrm{i}}} + \sigma^2_{V_{\mathrm{o}}} + 2\sigma^2_{\mathrm{d}}\right)} \right)}{1 - \left(\mathrm{erf}\left[\frac{\mu_{V_{\mathrm{i}}} - \mu_{V_{\mathrm{o}}}}{\sqrt{2}\sqrt{\sigma^2_{V_{\mathrm{i}}} + \sigma^2_{V_{\mathrm{o}}} + 2\sigma^2_{\mathrm{d}}}} \right] \right)^2} \right) ,$$

where the involved means $\mu_{V_{\mathrm{i}}}$ and $\mu_{V_{\mathrm{o}}}$ and variances $\sigma^2_{V_{\mathrm{i}}}$ and $\sigma^2_{V_{\mathrm{o}}}$ depend on the stimulus condition and on the decision model. Interestingly, the above expression is symmetric with respect to $\mu_{V_{\mathrm{i}}}$ and $\mu_{V_{\mathrm{o}}}$. Thus, if the noise levels are identical across hemisphere, i.e., if the variances are independent of the stimulus condition such that $\sigma^2_{V_{\mathrm{i}}} = \sigma^2_{V_{\mathrm{o}}} = \sigma^2_V$, we expect the δ-metric to be identical for "choice in" and "choice out". By contrast, the δ-metric attains its minimum in the no-stimulus condition for which we have $\mu_{V_{\mathrm{i}}} = \mu_{V_{\mathrm{o}}}$, which leads to

$$\delta(T_{\mathrm{n}}) = \frac{2\sigma_V}{\sqrt{\pi}\sqrt{1 + (\sigma_{\mathrm{d}}/\sigma_V)^2}} < \delta(T_{\mathrm{i}}) = \delta(T_{\mathrm{o}}) .$$

3 Choice discriminability with fluctuating attention state

3.1 Heuristic picture

We explain the heuristics of our approach by first considering the additive model with perfect correlation in Fig. 3. For perfect correlation, $q = 0$ so that either both hemispheres are in the low attention state, i.e., $(A_{\mathrm{i}}, A_{\mathrm{o}}) = (A_{\mathrm{ll}}, A_{\mathrm{ll}})$, a state we denote by A_{ll} as a shorthand, or both hemispheres are in the high attention state, i.e., $(A_{\mathrm{i}}, A_{\mathrm{o}}) = (A_{\mathrm{hh}}, A_{\mathrm{hh}})$, a state we denote by A_{hh} as a shorthand. Let us further assume that $p = 1/2$ for simplicity so that the low-attention state and the high-attention state are equiprobable. In the

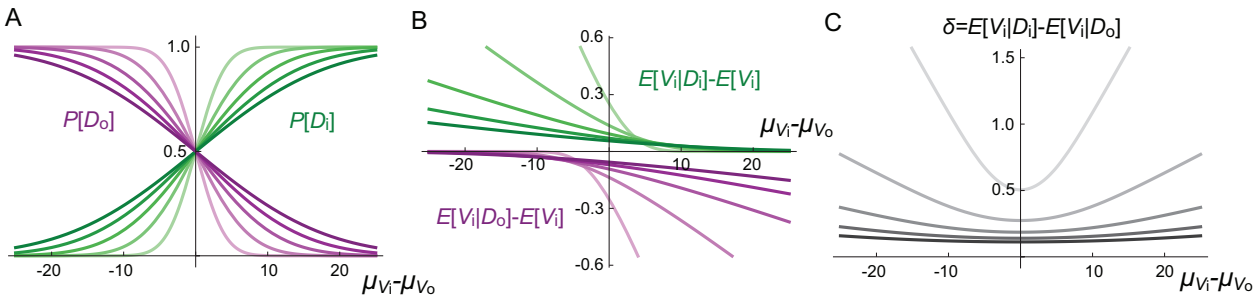

Figure 2: **Decision statistics for the deterministic Gaussian gain model. A.** Probabilities of decision as functions of the unconditional mean difference $\mu_{V_\mathrm{i}} - \mu_{V_\mathrm{o}}$. **B.** Conditioning bias as a function of the unconditional mean difference $\mu_{V_\mathrm{i}} - \mu_{V_\mathrm{o}}$. **C.** δ-metric as a function of the unconditional mean difference $\mu_{V_\mathrm{i}} - \mu_{V_\mathrm{o}}$. For large downstream decision variance σ^2_d, the δ-metric becomes essentially independent of the stimulus-driven mean activity as specified by $\mu_{V_\mathrm{i}} - \mu_{V_\mathrm{o}}$. Parameters: $\mu_{V_\mathrm{i}} = 0$, $\sigma^2_\mathrm{V} = 1$, $\sigma^2_\mathrm{d} = 1, 2, \ldots, 5$, increasing variance corresponds to increasing downstream decision variance.

additive model, the effect of changing attention levels is to shift the voltage distribution of $(V_\mathrm{i}, V_\mathrm{o})$ along the slope-one decision boundary. Then, by symmetry, marginalizing over V_o given the low-attention state or the high-attention state leads to two conditional distributions $p(V_\mathrm{i}, D_\mathrm{i} \,|\, A_\mathrm{ll}, T_\mathrm{i})$ and $p(V_\mathrm{i}, D_\mathrm{o} \,|\, A_\mathrm{ll}, T_\mathrm{o})$ that are translated version of one another. This is consistent with the fact that choices are made on the basis of $V_\mathrm{i} - V_\mathrm{o}$, i.e., along a direction that is orthogonal to the decision boundary. Finally, the δ-metric results from averaging the conditional distributions $p(V_\mathrm{i}, D_\mathrm{i} \,|\, A_\mathrm{ll}, T_\mathrm{i})$ and $p(V_\mathrm{i}, D_\mathrm{o} \,|\, A_\mathrm{ll}, T_\mathrm{o})$ over the equiprobable attention states. Such averaging leaves the δ-metric metric unchanged compared to the case of a single, fixed attention state.

One could have spared the arguments given above by simply remarking that the decision variable cancels out additive variables that are perfectly correlated across hemispheres: if $A_\mathrm{i} = A_\mathrm{o} = A$, we have $D = \mathbb{1}_{W \geq 0}$ with $W = (V_\mathrm{i} + A + N_\mathrm{i,d}) - (V_\mathrm{i} + A + N_\mathrm{o,d}) = (V_\mathrm{i} + N_\mathrm{i,d}) - (V_\mathrm{i} + N_\mathrm{o,d})$. Therefore, modeling gain as a purely additive shared component across hemispheres cannot alter decision making. The main interest for making the arguments developed above is to consider them in relation to the multiplicative model with perfect correlation depicted in Fig. 4. In the low-attention state A_ll, the biased sensory input that follows the presentation of a stimulus is not large enough to significantly overcome the downstream noise, leading to approximatively equiprobable decision outcomes, similarly as for the additive model in Fig. 3. This corresponds to the fact that the densities $p(V_\mathrm{i}, D_\mathrm{o} \,|\, A_\mathrm{ll}, T)$ and $p(V_\mathrm{i}, D_\mathrm{i} \,|\, A_\mathrm{ll}, T)$ in Fig. 4 both approximately integrate to probability one half, independent of the stimulus T. By contrast, in the high-attention state A_hh, the amplification of sensory input is sufficient to significantly bias decision toward the correct outcome. This is schematically illustrated by the fact that in the presence of a stimulus, the density $p(V_\mathrm{i}, D_\mathrm{i} \,|\, A_\mathrm{hh}, T_\mathrm{i})$ and $p(V_\mathrm{i}, D_\mathrm{o} \,|\, A_\mathrm{hh}, T_\mathrm{o})$, which correspond to hits, integrate to a much larger probability than the density $p(V_\mathrm{i}, D_\mathrm{o} \,|\, A_\mathrm{ll}, T_\mathrm{i})$ and $p(V_\mathrm{i}, D_\mathrm{i} \,|\, A_\mathrm{ll}, T_\mathrm{o})$, which corresponds to misses.

Following classical Bayesian analysis, the imbalance in choice probabilities in the the high-attention state A_hh compared to the approximatively equiprobable probabilities in the low attention state A_ll can significantly impact the evaluation of decision metrics such as the δ-metric. For instance, if one assume that low- and high-attention states are equiprobable as in Fig. 4, one can qualitatively explain how the δ-metric can become negative in the "target-out" condition. The key is to recognize that the overall choice probability densities $p(V_\mathrm{i}, D_\mathrm{i} \,|\, T_\mathrm{o})$ and $p(V_\mathrm{i}, D_\mathrm{o} \,|\, T_\mathrm{o})$ are obtained by mixing the gain specific densities with equal contributions. As a result of this mixture, and because the miss probability is low in the high-attention state compared to the low attention state, the vast majority of "choices in" occurs in the low attention state, so

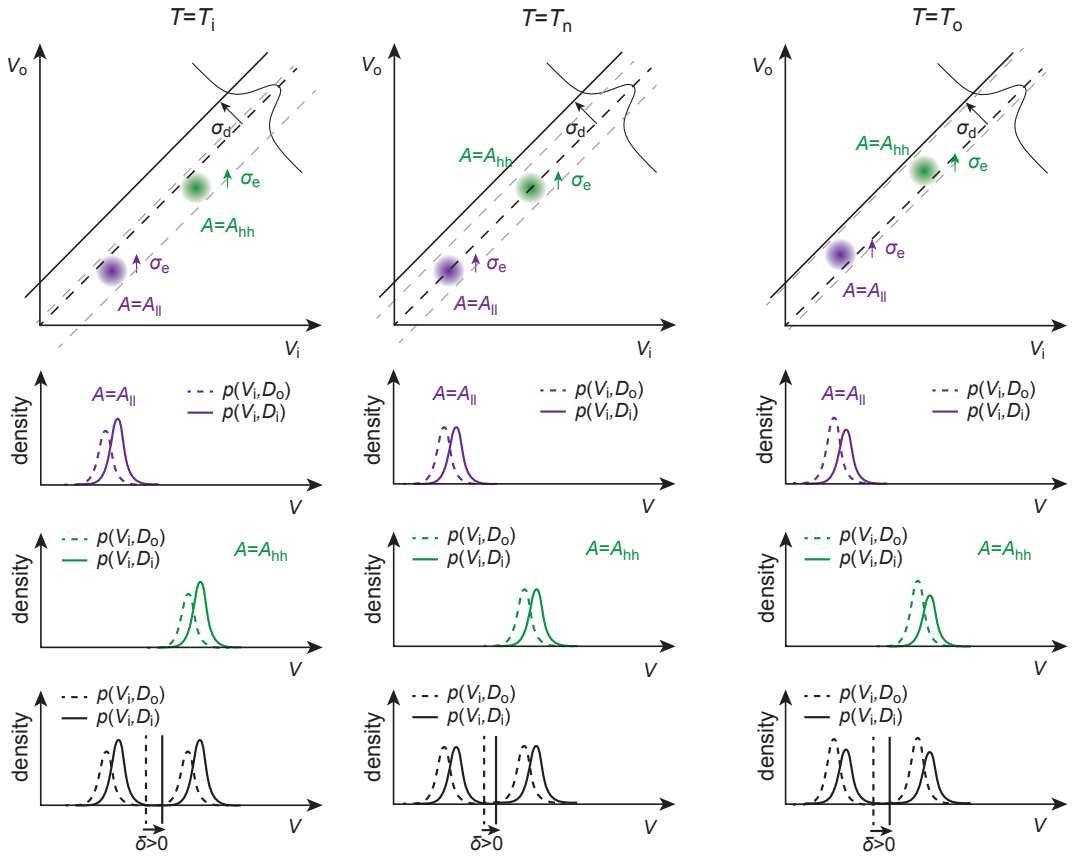

Figure 3: **Schematic for the additive decision model with perfectly correlated fluctuating gain.** We consider three stimulus conditions: "target-in", i.e., $T = T_{\rm i}$, "no-stimulus", i.e., $T = T_{\rm n}$, and "target-out", i.e., $T = T_{\rm o}$; and two equiprobable perfectly correlated attention states: "low gain", i.e., $A = A_{\rm ll}$, and "high gain", i.e., $A = A_{\rm hh}$. The effect of varying attention states is simply to jointly shift the mean values of voltage values $(V_{\rm i}, V_{\rm o})$ along the decision boundary in the two-dimensional phase space (top panels). When conditioned on an attention state, either $A_{\rm ll}$ or $A_{\rm hh}$, the model is identical to the case of fixed attention state in Fig. 1. Given an attention state, the recorded voltage data can be obtained by the same marginalization operation of the distribution of $(V_{\rm i}, V_{\rm o})$ restricted to each decision region of the two-dimensional phase space as in Fig. 1. Because the δ-metric is invariant with respect to translation along the decision boundary (solid black lines in the top panels), the conditional δ-metric is the same for both attention state in each stimulus condition. The full unconditioned δ-metric is obtained as an average of the two (identical) conditional metric. Thus, considering an additive fluctuating attention states have no impact on the δ-metric compared with the fixed attention state.

that the conditional expectation $\mathbb{E}\left[V_{\rm i} \,|\, D_{\rm i}, T_{\rm o}\right]$ is only marginally larger than $\mathbb{E}\left[V_{\rm i} \,|\, D_{\rm i}, A_{\rm ll}, T_{\rm o}\right]$. By contrast, the majority of "choices out" happens in the high-attention state, so that the conditional expectation $\mathbb{E}\left[V_{\rm i} \,|\, D_{\rm o}, T_{\rm o}\right]$ can considerably exceed $\mathbb{E}\left[V_{\rm i} \,|\, D_{\rm o}, A_{\rm ll}, T_{\rm o}\right]$, and even approaches $(2\mathbb{E}\left[V_{\rm i} \,|\, D_{\rm o}, A_{\rm hh}, T_{\rm o}\right] + \mathbb{E}\left[V_{\rm i} \,|\, D_{\rm o}, A_{\rm ll}, T_{\rm o}\right])/3$ when the high-attention miss probability is close to zero. In the latter extreme scenario, we expect to have

$$\begin{aligned}
\delta(T_{\rm o}) &= \mathbb{E}\left[V_{\rm i} \,|\, D_{\rm i}, T_{\rm o}\right] - \mathbb{E}\left[V_{\rm i} \,|\, D_{\rm o}, T_{\rm o}\right] \,, \\
&\simeq \mathbb{E}\left[V_{\rm i} \,|\, D_{\rm i}, A_{\rm ll}, T_{\rm o}\right] - (2\mathbb{E}\left[V_{\rm i} \,|\, D_{\rm o}, A_{\rm hh}, T_{\rm o}\right] + \mathbb{E}\left[V_{\rm i} \,|\, D_{\rm o}, A_{\rm ll}, T_{\rm o}\right])/3 \,, \\
&\simeq \frac{1}{3}\delta(T_{\rm o}, A_{\rm ll}) - \frac{2}{3}(\mathbb{E}\left[V_{\rm i} \,|\, D_{\rm o}, A_{\rm hh}, T_{\rm o}\right] - \mathbb{E}\left[V_{\rm i} \,|\, D_{\rm i}, A_{\rm ll}, T_{\rm o}\right]) \,,
\end{aligned}$$

where we note that $\delta(T_{\rm o}, A_{\rm ll})$ is typically small in the low-attention state. Thus, $\delta(T_{\rm o})$ is negative whenever the high-attention state amplification is strong enough so that the voltage expectation given $D_{\rm o}$ and $A_{\rm hh}$ is larger than the voltage expectation given $D_{\rm i}$ and $A_{\rm ll}$.

3.2 Quantitative argument

Here, we turn our heuristic arguments into quantitative ones for the proposed symmetric model with fluctuating gain level and varying degree of correlation across hemispheres. By symmetry, we denote the miss probability conditioning on the attention state by

$$\begin{aligned}
m_{\rm ll} &= \mathbb{P}\left[D_{\rm o} \,|\, A_{\rm ll}, T_{\rm i}\right] = \mathbb{P}\left[D_{\rm i} \,|\, A_{\rm ll}, T_{\rm o}\right] \,, \\
m_{\rm lh} &= \mathbb{P}\left[D_{\rm o} \,|\, A_{\rm lh}, T_{\rm i}\right] = \mathbb{P}\left[D_{\rm i} \,|\, A_{\rm hl}, T_{\rm o}\right] \,, \\
m_{\rm hl} &= \mathbb{P}\left[D_{\rm o} \,|\, A_{\rm hl}, T_{\rm i}\right] = \mathbb{P}\left[D_{\rm i} \,|\, A_{\rm lh}, T_{\rm o}\right] \,, \\
m_{\rm hh} &= \mathbb{P}\left[D_{\rm o} \,|\, A_{\rm hh}, T_{\rm i}\right] = \mathbb{P}\left[D_{\rm i} \,|\, A_{\rm hh}, T_{\rm o}\right] \,.
\end{aligned}$$

All the above miss probabilities can be evaluated explicitly by the same treatment as for the case of a fixed, single attention state. Our goal is to evaluate explicitly the δ-metric in the above fluctuating setting. To do so, we observe that

$$\mathbb{E}\left[V_{\rm i} \,|\, D, T\right] = \frac{\mathbb{E}\left[V_{\rm i} D \,|\, T\right]}{\mathbb{P}\left[D \,|\, T\right]} = \sum_{\alpha,\beta\in\{\rm l,h\}} \frac{\mathbb{E}\left[V_{\rm i} D, A_{\alpha\beta} \,|\, T\right]}{\mathbb{P}\left[D \,|\, T\right]} = \frac{1}{\mathbb{P}\left[D \,|\, T\right]} \sum_{\alpha,\beta\in\{\rm l,h\}} \mathbb{E}\left[V_{\rm i} D \,|\, A_{\alpha\beta}, T\right] \mathbb{P}\left[A_{\alpha\beta}\right] \,,$$

where we assume that the gain fluctuates independently of the stimulus condition: $\mathbb{P}\left[A \,|\, T\right] = \mathbb{P}\left[A\right]$. All the expectations featured above can be evaluated explicitly by the same treatment as for the case of a fixed, single attention state. Thus, it only remains to evaluate the conditional probabilities $\mathbb{P}\left[D \,|\, T\right]$. In the presence of a stimulus, this involves estimating the marginal miss probability m, satisfying $m = \mathbb{P}\left[D_{\rm o} \,|\, T_{\rm i}\right] = \mathbb{P}\left[D_{\rm i} \,|\, T_{\rm o}\right] = (1 - \mathbb{P}\left[D_{\rm i} \,|\, T_{\rm i}\right]) = (1 - \mathbb{P}\left[D_{\rm o} \,|\, T_{\rm o}\right])$ with

$$m = \sum_{\alpha,\beta\in\{\rm l,h\}} \mathbb{P}\left[D_{\rm o} \,|\, A_{\alpha\beta}, T_{\rm i}\right] \mathbb{P}\left[A_{\alpha\beta} \,|\, T_{\rm i}\right] = m_{\rm ll}\, p + (m_{\rm hl} + m_{\rm lh}) q + m_{\rm hh}(1 - p - 2q) \,,$$

whereas in the absence of a stimulus, we have $\mathbb{P}\left[D_{\rm i} \,|\, T_{\rm n}\right] = \mathbb{P}\left[D_{\rm o} \,|\, T_{\rm n}\right] = 1/2$ by symmetry. From there, we can expressed explicitly the δ-metric for all stimulus conditions via the formulas

$$\delta(T) = \sum_{\alpha,\beta\in\{\rm l,h\}} \left(\mathbb{E}\left[V_{\rm i} D_{\rm i} \,|\, A_{\alpha\beta}, T\right] \frac{\mathbb{P}\left[A_{\alpha\beta}\right]}{\mathbb{P}\left[D_{\rm i} \,|\, T\right]} - \mathbb{E}\left[V_{\rm i} D_{\rm o} \,|\, A_{\alpha\beta}, T\right] \frac{\mathbb{P}\left[A_{\alpha\beta}\right]}{\mathbb{P}\left[D_{\rm o} \,|\, T\right]} \right) \,.$$

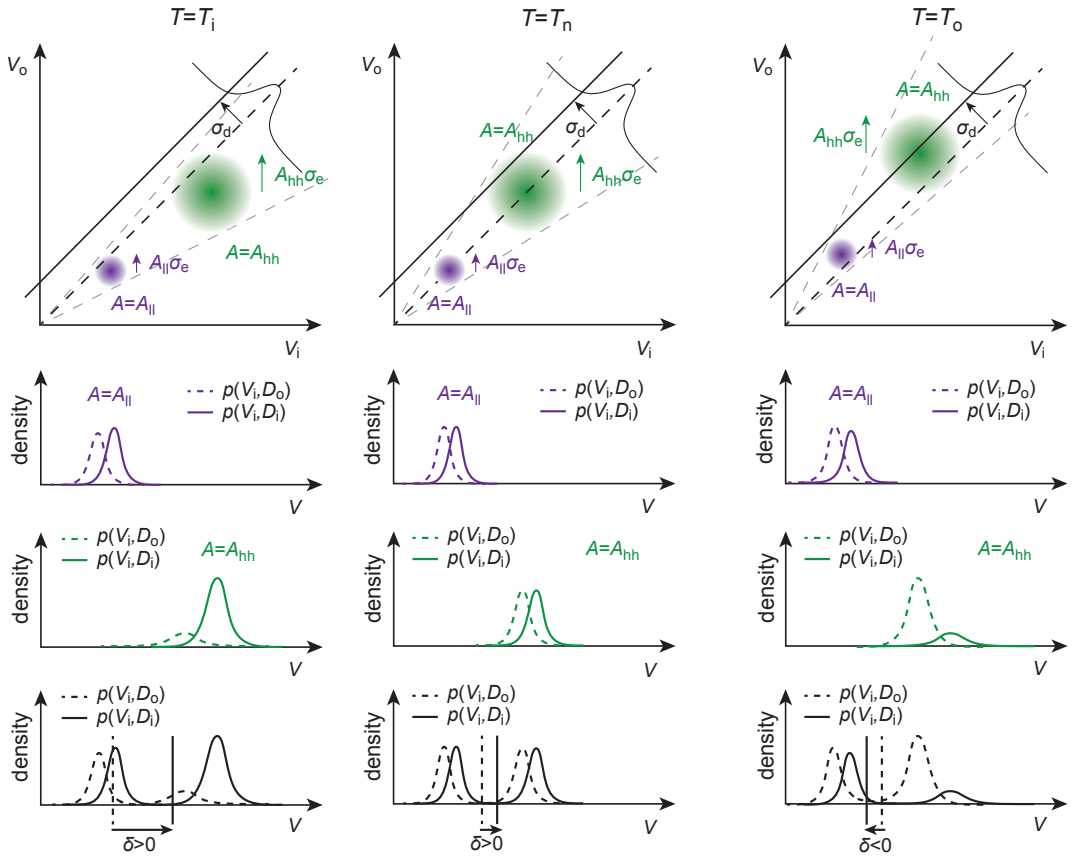

Figure 4: **Schematic for the multiplicative decision model with perfectly correlated fluctuating gain.** We consider three stimulus conditions: "target-in", i.e., $T = T_{\rm i}$, "no-stimulus", i.e., $T = T_{\rm n}$, and "target-out", i.e., $T = T_{\rm o}$; and two equiprobable perfectly correlated attention states: "low gain", i.e., $A = A_{\rm ll}$, and "high gain", i.e., $A = A_{\rm hh}$. The effect of varying attention states is to scale the distributions of voltage values $(V_{\rm i}, V_{\rm o})$ along the decision boundary in the two-dimensional phase space (top panels). Because of this scaling, the bias entailed by the presence of a stimulus is magnified in the high attention states $A_{\rm hh}$. Specifically, in the high attention state, the voltage values $(V_{\rm i}, V_{\rm o})$ are more dispersed but stay further away from the decision boundary (solid straight line) on average. This is by contrast with the additive model of Fig. 3. As in Fig. 3, given an attention state, the recorded voltage data can be obtained by the same marginalization operation of the distribution of $(V_{\rm i}, V_{\rm o})$ restricted to each decision region of the two-dimensional phase space as in Fig. 1. Given knowledge of the attention state, the expected value of $V_{\rm i}$ is always larger in the "choice in" condition compared with the "choice out" condition. This is similar as for the additive model of Fig. 3. However, the key difference with Fig. 3 is that in the multiplicative model, there are much less misses in the high-attention state compared with the low attention state. As a result, when conditioned on the decision outcome alone and without knowledge of the attention state, the expected value of $V_{\rm i}$ can be less for "choice in" condition than for "choice out" when $T = T_{\rm o}$, leading to a negative δ-metric. This is because when $T = T_{\rm o}$, "choices in" are misses, which are more likely to happen in the low attention state, for which the gain variable A—thus the voltage $V_{\rm i}$—is low.

For instance, one can check that for the "no-stimulus" condition, we have

$$\begin{aligned}
\delta(T_{\mathrm{n}}) &= \sum_{\alpha,\beta\in\{\mathrm{l},\mathrm{h}\}} \left(\mathbb{E}\left[V_{\mathrm{i}} D_{\mathrm{i}} \,|\, A_{\alpha\beta}, T_{\mathrm{n}}\right] \frac{\mathbb{P}\left[A_{\alpha\beta}\right]}{\mathbb{P}\left[D_{\mathrm{i}} \,|\, T_{\mathrm{n}}\right]} - \mathbb{E}\left[V_{\mathrm{i}} D_{\mathrm{o}} \,|\, A_{\alpha\beta}, T_{\mathrm{n}}\right] \frac{\mathbb{P}\left[A_{\alpha\beta}\right]}{\mathbb{P}\left[D_{\mathrm{o}} \,|\, T_{\mathrm{n}}\right]} \right) , \\
&= 2 \sum_{\alpha,\beta\in\{\mathrm{l},\mathrm{h}\}} \big(\mathbb{E}\left[V_{\mathrm{i}} D_{\mathrm{i}} \,|\, A_{\alpha\beta}, T_{\mathrm{n}}\right] - \mathbb{E}\left[V_{\mathrm{i}} D_{\mathrm{o}} \,|\, A_{\alpha\beta}, T_{\mathrm{n}}\right] \big) \mathbb{P}\left[A_{\alpha\beta}\right] , \\
&= \sum_{\alpha,\beta\in\{\mathrm{l},\mathrm{h}\}} \left(\frac{\mathbb{E}\left[V_{\mathrm{i}} D_{\mathrm{i}} \,|\, A_{\alpha\beta}, T_{\mathrm{n}}\right]}{\mathbb{P}\left[D_{\mathrm{i}} \,|\, A_{\alpha\beta}, T_{\mathrm{n}}\right]} - \frac{\mathbb{E}\left[V_{\mathrm{i}} D_{\mathrm{o}} \,|\, A_{\alpha\beta}, T_{\mathrm{n}}\right]}{\mathbb{P}\left[D_{\mathrm{o}} \,|\, A_{\alpha\beta}, T_{\mathrm{n}}\right]} \right) \mathbb{P}\left[A_{\alpha\beta}\right] , \\
&= \sum_{\alpha,\beta\in\{\mathrm{l},\mathrm{h}\}} \delta(T_{\mathrm{n}}, A_{\alpha\beta}) \mathbb{P}\left[A_{\alpha\beta}\right] ,
\end{aligned}$$

so that the δ-metric is guaranteed to be positive as the average of the δ-metrics $\delta(T_{\mathrm{n}}, A_{\alpha\beta}) > 0$ obtained for the case of a fixed, single attention state $A_{\alpha,\beta}$. The expressions for "stimulus in" and "stimulus out" are slightly more complicated as they necessitate evaluating the marginal miss probability m. Specifically, we have

$$\begin{aligned}
\delta(T_{\mathrm{i}}) &= \sum_{\alpha,\beta\in\{\mathrm{l},\mathrm{h}\}} \left(\mathbb{E}\left[V_{\mathrm{i}} D_{\mathrm{i}} \,|\, A_{\alpha\beta}, T_{\mathrm{i}}\right] \frac{\mathbb{P}\left[A_{\alpha\beta}\right]}{\mathbb{P}\left[D_{\mathrm{i}} \,|\, T_{\mathrm{i}}\right]} - \mathbb{E}\left[V_{\mathrm{i}} D_{\mathrm{o}} \,|\, A_{\alpha\beta}, T_{\mathrm{i}}\right] \frac{\mathbb{P}\left[A_{\alpha\beta}\right]}{\mathbb{P}\left[D_{\mathrm{o}} \,|\, T_{\mathrm{i}}\right]} \right) , \\
&= \frac{1}{1-m} \sum_{\alpha,\beta\in\{\mathrm{l},\mathrm{h}\}} \mathbb{E}\left[V_{\mathrm{i}} D_{\mathrm{i}} \,|\, A_{\alpha\beta}, T_{\mathrm{i}}\right] \mathbb{P}\left[A_{\alpha\beta}\right] - \frac{1}{m} \sum_{\alpha,\beta\in\{\mathrm{l},\mathrm{h}\}} \mathbb{E}\left[V_{\mathrm{i}} D_{\mathrm{o}} \,|\, A_{\alpha\beta}, T_{\mathrm{i}}\right] \mathbb{P}\left[A_{\alpha\beta}\right] ,
\end{aligned}$$

$$\begin{aligned}
\delta(T_{\mathrm{o}}) &= \sum_{\alpha,\beta\in\{\mathrm{l},\mathrm{h}\}} \left(\mathbb{E}\left[V_{\mathrm{i}} D_{\mathrm{i}} \,|\, A_{\alpha\beta}, T_{\mathrm{o}}\right] \frac{\mathbb{P}\left[A_{\alpha\beta}\right]}{\mathbb{P}\left[D_{\mathrm{i}} \,|\, T_{\mathrm{o}}\right]} - \mathbb{E}\left[V_{\mathrm{i}} D_{\mathrm{o}} \,|\, A_{\alpha\beta}, T_{\mathrm{o}}\right] \frac{\mathbb{P}\left[A_{\alpha\beta}\right]}{\mathbb{P}\left[D_{\mathrm{o}} \,|\, T_{\mathrm{o}}\right]} \right) , \\
&= \frac{1}{m} \sum_{\alpha,\beta\in\{\mathrm{l},\mathrm{h}\}} \mathbb{E}\left[V_{\mathrm{i}} D_{\mathrm{i}} \,|\, A_{\alpha\beta}, T_{\mathrm{o}}\right] \mathbb{P}\left[A_{\alpha\beta}\right] - \frac{1}{1-m} \sum_{\alpha,\beta\in\{\mathrm{l},\mathrm{h}\}} \mathbb{E}\left[V_{\mathrm{i}} D_{\mathrm{o}} \,|\, A_{\alpha\beta}, T_{\mathrm{o}}\right] \mathbb{P}\left[A_{\alpha\beta}\right] .
\end{aligned}$$

In the following, we use the above expressions to make quantitative predictions about the δ-metric. It is also not too hard to use these expressions to derive the heuristic arguments that explains why $\delta(T_{\mathrm{o}})$ can be negative for fluctuating gain shared across both hemispheres. For perfect correlation across hemisphere, we have $q = 0$ and for equiprobable low- and high-attention states we have $p = 1/2$, so the marginal hit probability is $m = (m_{\mathrm{ll}} + m_{\mathrm{hh}})/2$. This implies that

$$\delta(T_{\mathrm{o}}) = \frac{1}{2m}(\mathbb{E}\left[V_{\mathrm{i}} D_{\mathrm{i}} \,|\, A_{\mathrm{ll}}, T_{\mathrm{o}}\right] + \mathbb{E}\left[V_{\mathrm{i}} D_{\mathrm{i}} \,|\, A_{\mathrm{hh}}, T_{\mathrm{o}}\right]) - \frac{1}{2(1-m)}(\mathbb{E}\left[V_{\mathrm{i}} D_{\mathrm{o}} \,|\, A_{\mathrm{ll}}, T_{\mathrm{o}}\right] + \mathbb{E}\left[V_{\mathrm{o}} D_{\mathrm{i}} \,|\, A_{\mathrm{hh}}, T_{\mathrm{o}}\right]) .$$

In general, we have $m_{\mathrm{ll}} = \mathbb{P}\left[D_{\mathrm{o}} \,|\, A_{\mathrm{ll}}, T_{\mathrm{i}}\right] = \mathbb{P}\left[D_{\mathrm{i}} \,|\, A_{\mathrm{ll}}, T_{\mathrm{o}}\right] < 1/2$ and $m_{\mathrm{hh}} = \mathbb{P}\left[D_{\mathrm{o}} \,|\, A_{\mathrm{hh}}, T_{\mathrm{i}}\right] = \mathbb{P}\left[D_{\mathrm{i}} \,|\, A_{\mathrm{hh}}, T_{\mathrm{o}}\right] < 1/2$. We typically consider that decision bias entailed by the presence of a stimulus is small in the low-attention state so that $\mathbb{P}\left[D_{\mathrm{o}} \,|\, A_{\mathrm{ll}}, T\right] = \mathbb{P}\left[D_{\mathrm{i}} \,|\, A_{\mathrm{ll}}, T\right] \simeq 1/2 \simeq m_{\mathrm{ll}}$, independent of the stimulus T. This assumption becomes exact in the limit of large downstream noise $\sigma_{\mathrm{d}} \to \infty$ or vanishing gain amplification $A_{\mathrm{ll}} \to 0$. By contrast, the probability $m_{\mathrm{h,h}}$ becomes negligible for large enough amplification $A_{\mathrm{hh}} \to \infty$. This involves that $\mathbb{E}\left[V_{\mathrm{i}} D_{\mathrm{i}} \,|\, A_{\mathrm{hh}}, T_{\mathrm{o}}\right]$ is also negligible in this limit, so that

$$\delta(T_{\mathrm{o}}) = (\mathbb{E}\left[V_{\mathrm{i}} \,|\, D_{\mathrm{i}}, A_{\mathrm{ll}}, T_{\mathrm{o}}\right]) - \mathbb{E}\left[V_{\mathrm{i}} \,|\, D_{\mathrm{o}}, A_{\mathrm{ll}}, T_{\mathrm{o}}\right]) - \mathbb{E}\left[V_{\mathrm{o}} D_{\mathrm{i}} \,|\, D_{\mathrm{i}}, A_{\mathrm{hh}}, T_{\mathrm{o}}\right] ,$$

where the parenthesized term is the conditional δ-metric in the low attention state A_{ll}. This term becomes negligible for small enough gain $A_{\mathrm{ll}} \to 0$ and large enough downstream noise. Therefore, the only meaningful contribution comes from the rightmost negative term, leading to an overall negative unconditioned δ-metric.

3.3 Parametric dependences of the δ-metric

In Fig 5, we assume perfect correlation across hemispheres ($q = 0$) and we represent the dependence of the δ-metrics $\delta(T_{\rm i})$, $\delta(T_{\rm n})$, and $\delta(T_{\rm o})$ on three parameters: the high-attention multiplicative gain $a_{\rm hh}$ in Fig 5A, the probability to be in the high-attention state $p_{\rm hh} = 1 - p$ in Fig 5B, and the variance of the downstream noise $\sigma_{\rm d}$ in Fig 5C. This graphical illustration shows that the occurrence of negative $\delta(T_{\rm o})$ is due to having fluctuating attention states of large enough magnitudes and frequency.

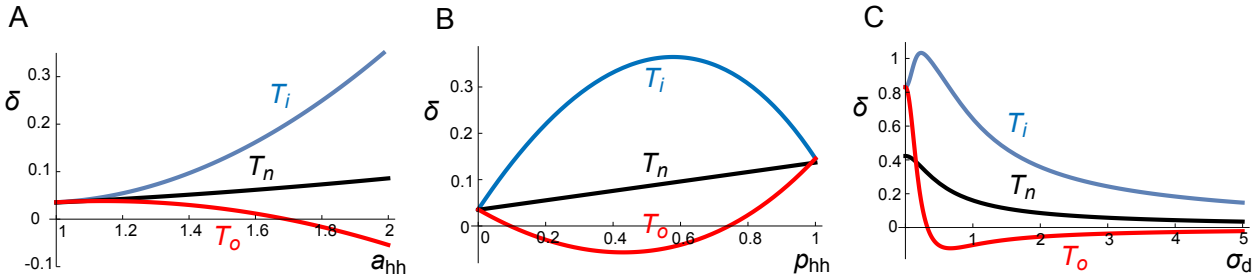

Figure 5: **Parametric dependence of the δ-metric for the multiplicative model. A.** Dependence on gain $a_{\rm hh}$. For consistently low attention state $a_{\rm hh} = a_{\rm ll} = 1$, the δ-metric is almost constant at a low values, independent of the stimulus condition (actually, we have $\delta(T_{\rm n}) < \delta(T_{\rm i}) = \delta(T_{\rm o})$ but this distinction is negligible). When increasing the gain of the high attention state $a_{\rm hh} > a_{\rm ll} = 1$, the δ-metric separates for the three stimulus conditions, with negative values for "target out" when the gain is large enough. **B.** Dependence on high-attention state probability $p_{\rm hh}$. The separation of the δ-metrics for the three stimulus conditions is due to gain fluctuations and is maximal for equiprobable low- and high-attention states. Observe that when the high-attention state is certain ($p_{\rm hh} = 1$), δ-metrics revert to similar values. However, by contrast with the low attention state ($p_{\rm hh} = 0$), the gain is large enough in the high attention state so that the δ-metrics for "target in" and "target out", which are equal by symmetry, are noticeably larger than for the "no-stimulus" condition: $\delta(T_{\rm n}) < \delta(T_{\rm i}) = \delta(T_{\rm o})$. **C.** Dependence on the downstream noise $\sigma_{\rm d}$. The limit of vanishing downstream noise $\sigma_{\rm d} \to 0$ is akin to the limit of infinite gain $a_{\rm hh}, a_{\rm ll}$ for which the δ-metrics for "target in" and "target out", which are equal by symmetry, are significantly larger than for the "no-stimulus condition". Increasing the downstream-noise variance $\sigma_{\rm d}$ generally reduces the δ-metrics leading to a regime for which the "target out" delta metric is negative for large enough values of $\sigma_{\rm d}$. Parameters: $\mu_{V_{\rm i}} = 2$, $\mu_{V_{\rm o}} = 1$, $\sigma_{V_{\rm i}} = 0.25$, $\sigma_{V_{\rm o}} = 0.25$, $\sigma_{\rm d} = 2$, $a_{\rm ll} = 1$, $a_{\rm hh} = 2$ and $p_{\rm ll} = 1/2$. For all panels, both hemispheres are assumed perfectly correlated, i.e., $q = 0$.

In Fig 6, we relax the assumption of perfect correlations across hemispheres and explore under which conditions the δ-metric can become negative for the "target-out" condition. Our symmetric model for correlations involves two probability parameters p and q which are such that $1 - p - 2q \geq 0$. Thus the cross-hemisphere correlation $\rho(p, q)$ given via formula (1) is defined over a triangular region in the (p, q)-plane (see Fig. 6A). Fig 6B-D represent the δ-metrics $\delta(T_{\rm i})$, $\delta(T_{\rm n})$, and $\delta(T_{\rm o})$ as functions of the parameters p and q. The functional form of the δ-metric given above specifies the level sets of δ-metrics as quadrics. The "no-stimulus" case, for which the δ-metric is always positive, corresponds to the degenerate case when these level sets are affine functions. This graphical illustrations reveals that the occurrence of negative $\delta(T_{\rm o})$ is only consistent with $q \simeq 0$ and $p \simeq 1/2$, i.e., with a fluctuating attention state and with close to perfect correlation across hemispheres.

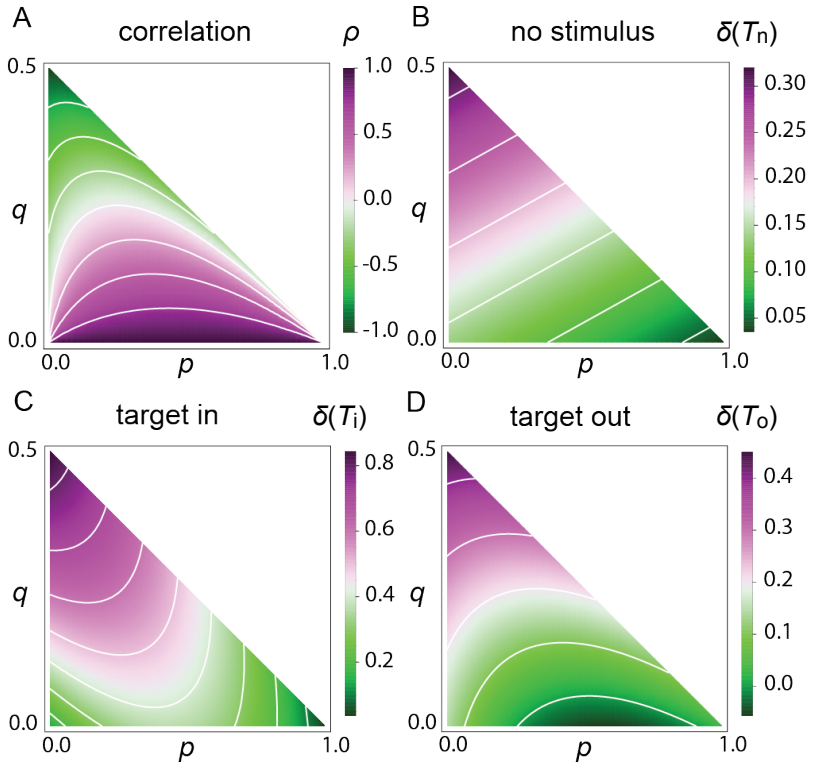

Figure 6: **Impact of correlations across hemispheres on δ-metric. A.** Correlation of the gain variables $\rho = \mathbb{C}[A_{\rm i}, A_{\rm o}] / \sqrt{\mathbb{V}[A_{\rm i}]\,\mathbb{V}[A_{\rm o}]}$ as a function of the probability parameters $p = \mathbb{P}[A_{\rm ll}]$ and $q = \mathbb{P}[A_{\rm lh}] = \mathbb{P}[A_{\rm hl}]$. **B, C, D.** Delta metric $\delta(T_{\rm i})$ as a function of the same parameters p and q for the "no-stimulus" condition (**B**), and the "target-in" condition (**C**), and the "target-out" condition (**D**). For all stimulus conditions, the δ-metric increases with negative correlation across hemisphere, whereas the δ-metric generically satisfy $\delta(T_{\rm i}) \geq \delta(T_{\rm n}) \geq \delta(T_{\rm o})$. Negative δ-metric only occurs for $T = T_{\rm o}$ and for strongly, fluctuating attention state, i.e., $q \simeq 0$ and $p \simeq 1/2$. Parameters: $\mu_{V_{\rm i}} = 2$, $\mu_{V_{\rm o}} = 1$, $\sigma_{V_{\rm i}} = 0.25$, $\sigma_{V_{\rm o}} = 0.25$, $\sigma_{\rm d} = 2$, $a_{\rm ll} = 1$, $a_{\rm hh} = 2$.

3.4 Numerical results for d-prime metrics and choice probabilities

Our analysis was derived for a two-state gain model but generalizes straightforwardly to the case of a continuously varying attention state. To see this, we consider that the multiplicative gain A modeling attention is shared across hemispheres and follows a distribution with density

$$p(a) = \frac{1}{\sigma_{\mathrm{a}}\sqrt{2\pi a^3}} \exp\left(-\frac{(1-a)^2}{2\sigma_{\mathrm{a}}^2 a}\right) \quad \text{for} \quad a \geq 0\,. \tag{2}$$

The above distribution is such that gain values A are positive with probability one, have unit expectation value, i.e. $\mathbb{E}[A] = 1$, and have variance σ_{a}^2. Highly fluctuating attention states correspond to large values $\sigma_{\mathrm{a}}^2 \gg 0$. One can extend the analytical results obtained for the δ-metric to continuously varying attention states by mere integration over the attention states. In Fig. 7A, we explore numerically the resulting expressions by varying σ_{a}, which quantifies the degree of gain fluctuation. As expected, we found that for the chosen set of parameters, $\delta(T_o)$ becomes negative for large enough σ_a.

In principle, one could further extend the analysis to the case of the d-prime metrics, a classical metric for choice discriminability. This would require, however, to compute conditional variances in addition to conditional expectations, which yields unwieldy expressions. Therefore, we only evaluate the d-prime metrics numerically in Fig. 7B, for the same set of parameters as in Fig. 7A. One can check that as expected, the d-prime metrics become negative in the "choice-out" condition for the same value σ_{a} as for the δ-metric, which is a scaled version of d-prime.

Our final goal is to check that our analysis remains qualitatively valid for another key choice-related metrics called choice probabilities. In our context, choice probabilities quantifies how well an ideal observer can predict the behavioral choice from reading out a neuron's voltage distribution. Formally, given the stimulus condition T, this involves considering the choice-related conditional variables $U_{\mathrm{i}} \,|\, T = V_{\mathrm{i}} \,|\, (D_{\mathrm{i}}, T)$ and $U_{\mathrm{o}} \,|\, T = V_{\mathrm{i}} \,|\, (D_{\mathrm{o}}, T)$, where U_{i} refers to the voltage V_{i} conditioned on "choice-in" and U_{o} refers to the refers to the voltage V_{i} conditioned on "choice-out". Then, the choice probabilities $\mathrm{CP}\,(T)$ is defined as the probability that U_{i} exceeds U_{o}:

$$\mathrm{CP}\,(T) = \mathbb{P}\left[U_{\mathrm{i}} \geq U_{\mathrm{o}} \,|\, T\right] = \mathbb{P}\left[U_{\mathrm{i}} \,|\, T \geq U_{\mathrm{o}} \,|\, T\right]\,.$$

In Fig. 7C, we compute numerically the choice probabilities associated to our multiplicative model for the same parameters as in Fig. 7A,B. Similar to Fig. 7A,B, one can check that for large enough σ_{a}, the choice probability $\mathrm{CP}\,(T_{\mathrm{o}})$ performs below chance level, a paradoxical effect akin to observing negative d-prime or δ-metric. Observe, however, that the paradoxical choice probabilities are observed for smaller values σ_{a}.

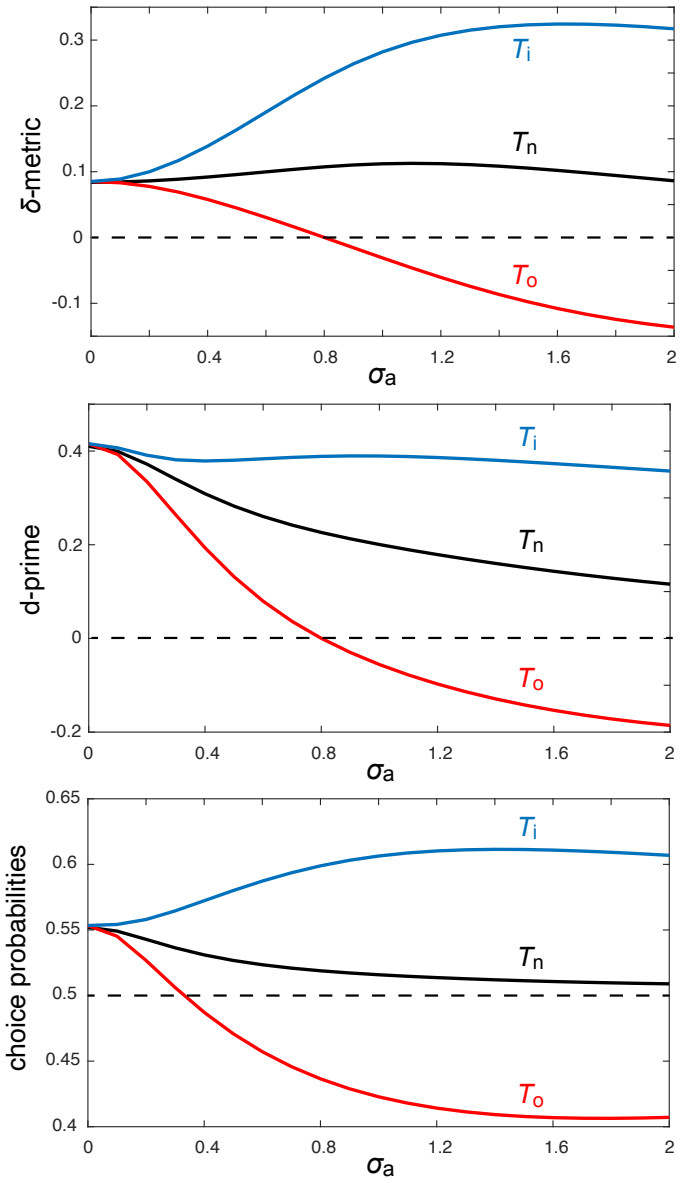

Figure 7: δ**-metric,** d**-prime metric, and choice probabilities.** The multiplicative gain is drawn from a distribution with density given by (2), which depends on a single variance parameter σ_{a}^2. For all metrics, a paradoxical effect is observed in the "choice-out" condition for large enough fluctuations in the attention state: $\delta(T_{\mathrm{o}})$ and $d'(T_{\mathrm{o}})$ become negative, whereas CP (T_{o}) becomes smaller than chance level (0.5). Parameters: $\mu_{V_{\mathrm{i}}} = 0.75$, $\mu_{V_{\mathrm{o}}} = 0.5$, $\sigma_{V_{\mathrm{i}}} = 0.1$, $\sigma_{V_{\mathrm{o}}} = 0.1$, $\sigma_{\mathrm{d}} = 1$.