A distributional Bayesian learning theory for visual perceptual learning

Training subjects to discriminate fine details of stimuli can improve perceptual capabilities, a phenomenon known as perceptual learning (PL). Experiments have discovered intriguing psychophysical findings despite the simple stimuli (e.g., Gabor patches) and training procedures (e.g., 2AFC task) involved. Providing explanations for these learning effects could shed light on sensory plasticity in adulthood and constraints in sensory processing. A notable feature of (visual) PL is its slow learning speed, which requires days of training and hundreds or thousands of trials for the performance to saturate. This is sometimes attributed to the low signal-to-noise ratio in the sensory activities, which poses a challenging classification problem for decision neurons that readout these activities. How could the decision neurons learn when the signal is so weak and brief, especially when supervision signal is delayed? We hypothesize that, rather than relying on the short-lived activities at each trial, the brain may learn according to the *distribution* of sensory activities summarized by sensory neurons over multiple stimulus presentations. We also assume that the decision neurons combine sensory activities using uncertain readout weights modeled as probabilistic (Bayesian) synapses (e.g. with a mean and sd). During PL, the weights are updated by averaging over the stimulus distribution of the presented category; during perception, the decision neuron acts according to the probability of the perceived category computed by a sample of the posterior weights. This model can explain several behavioral findings obtained by Dosher and Lu (1998, 2005): the uniform downward shift of threshold-versus-noise contrast (TVC) curves, the power-law decrease of the signal threshold with training, and the asymmetric transfer between noisy and clean displays. This theory thus offers an alternative to the Hebbian reweighting model (Dosher and Lu, 2010) and connects the theoretical literature of probabilistic synapses to visual perceptual learning.

Stimulus and task We present the theory in the context of a simple 2AFC visual orientation discrimination task, i.e. whether the stimulus is clockwise or counter-clockwise w.r.t. a reference. We model the class of the stimulus by a binary variable $y \in \{-1, +1\}$. The noisy responses $\mathbf{x}$ of orientation-selective neurons can be modeled as a random variable with bell-shaped mean and Poisson-like variability as shown in Fig. A. Specifically, we model the class-conditional activity distribution by $p_{\mathbf{X}|Y}(\mathbf{x}|y) = \mathcal{N}\left(\bar{\mathbf{x}}_y, \mathbf{\Sigma}(\bar{\mathbf{x}}_y)\right)$,where $\bar{\mathbf{x}}_{-1}$ and $\bar{\mathbf{x}}_{+1}$ are the class-conditional population mean activities for $y = -1$ and $y = +1$, respectively, and $\mathbf{\Sigma}()$ is a Poisson-like external noise covariance that depends on the mean (or can also be isotropic). The task is to predict the hidden class associated to the stimulus that induced activity $\mathbf{x}$.

Perception model The decision neuron in the brain predicts the class label according to the belief

$$q(\hat{y}|\mathbf{w}, \mathbf{x}) = \text{Bernoulli}(\phi(\mathbf{w} \cdot \mathbf{x})) \tag{1}$$

where $\hat{y} \in \{-1, +1\}$ is the *perceived* class, $\mathbf{w}$ are the readout weights to be trained, and $\phi()$ is a nonlinearity mapping to a probability. If there is no additional (internal) readout noise for a given $\mathbf{x}$ and $\mathbf{w}$, then ϕ is the Heaviside step function with a hard threshold. If there is an additional additive Gaussian internal noise with standard deviation σ_i, then ϕ is the cumulative density function of $\mathcal{N}(0,\sigma_i^2)$ (probit). This model can be regarded as a readout model (e.g., Dosher and Lu, 1998; Zhang et al., 2010; Sotiropoulos et al., 2011).

Distributional learning In most existing readout models, the weights $\mathbf{w}$ are deterministic and may be learned using known plausible rules (e.g., Hebbian). We found that, without further modifications, the perception model above with deterministic weights failed to capture key experimental results (also see Fig. H). Instead, we model the weights as *random* variables, adopting the core hypothesis of the probabilistic synapses framework (Aitchison et al., 2021). Treating naive subjects as having a zero-mean Gaussian prior $p_{\mathbf{W}}(\mathbf{w})$, one could try to model trial-by-trial, stimulus-driven PL as approximating

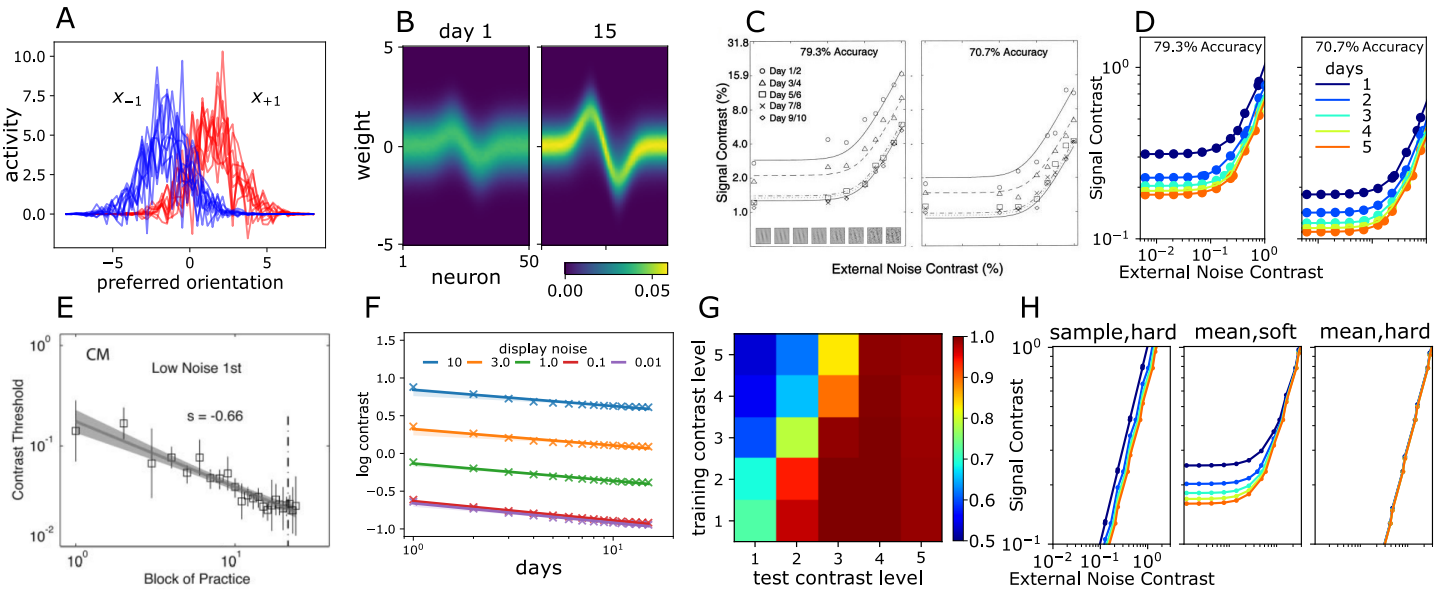
$$p_{\text{s}}\left(\mathbf{w}|\left\{\mathbf{x}_i, y_i\right\}_{i=1}^{n}\right) \propto \prod_{i=1}^{n} p_{\mathbf{W}}(\mathbf{w}) q(y_i|\mathbf{x}_i, \mathbf{w}), \tag{2}$$

where p_{s} indicates that learning depends on each *single* stimulus, and n is the trial number. This rule yields almost deterministic weights that could not reproduce experimental results. Further, computing the complicated posterior above given very brief sensory input may be implausible. This approach is also problematic when the label is presented *after* the stimulus as in most phychophysical experiments, since it requires that the label and the exact stimulus be available simultaneously.

Instead, we consider the possibility that the sensory neurons memorize distributional statistics of their activities over time (Harris et al., 2001), which can be retrieved by briefly presented stimuli. In other words, learning may not be driven by the brief activities induced by a single stimulus but by the distributional statistics. These statistics may last for a longer period (e.g., if encoded in recurrent connectivity) after each stimulus and be ready for interaction with supervision signals. We thus postulate that the weights are adjusted according to distribution-dependent rule

$$p_{\mathrm{d}}\left(\mathbf{w}\middle|\left\{y_i\right\}_{i=1}^{n}\right) \propto p_{\mathbf{W}}(\mathbf{w}) \prod_{i}^{n} \int d\mathbf{x}_i q\left(y_i\middle|\mathbf{w},\mathbf{x}_i\right) p_{\mathbf{X}|y}(\mathbf{x}_i|y_i), \tag{3}$$

where p_{d} indicates its *distributional* dependence. The key difference compared with (2) is the marginalization over $p_{\mathbf{X}|y}$, which requires responses statistics of the presented class. When $p_{\mathbf{X}|y}$ is Gaussian, this marginalization is closed-form for the aforementioned choices of ϕ (1). We trained the perception model (3) on the task conducted by Dosher and Lu (1998, 2005). Given a posterior p_{d}, the decision strategy that replicated three experimental findings required: a) *sampl*ing from p_{d} and b) a soft probit ϕ in (1) with probabilistic action; see Figure below.



A, Samples from the class-conditional $p_{\mathbf{X}|y}$ in model simulation. **B**, Marginal $p_{\mathrm{d}}\left(w_m\middle|\left\{y_i\right\}_{i=1}^{n}\right)$ in (3) for $m \in \{1, \ldots 50\}$ sensory neurons after day 1 and day 15 with 45 trials per day. Heatmap shows posterior density. The posterior mean approaches the optimal decision boundary but remain uncertain. In contrast, the single-stimulus driven $p_{\mathrm{s}}\left(w_m\middle|\left\{y_i\right\}_{i=1}^{n}\right)$ quickly collapse to almost deterministic weights (not shown). **C**, TVC curves after human subjects are trained on the 2AFC task, reproduced from (Dosher and Lu, 1998). There is a uniform decrease in signal threshold at all noise levels over the course of training. **D**, TVC curves from model simulation when decisions are made by sampling from p_{d} and acting probabilistically on q with a probit (soft) ϕ. **E**, The log-signal threshold of human subjects decreases linearly with log-training days, reproduced from (Dosher and Lu, 2005). **F**, Signal threshold versus training days from model simulation, repeated for multiple noise levels. **G**, Accuracy for each training and testing (transfer) noise contrast levels; transfer from low noise to high noise is more substantial than from high noise to low noise, similar to reported by Dosher and Lu (2005). **H**, Other decision strategies that did not replicate key findings in C: sampling from p_{d} and using a Heaviside (hard) ϕ, or using the deterministic *mean* of p_{d}.