The virtual loss function in the summary perception of motion and its limited adjustability

Tianyuan Teng
Academy for Advanced Interdisciplinary Studies,
Peking University, Beijing, China
Peking-Tsinghua Center for Life Sciences,
Peking University, Beijing, China

Sheng Li
School of Psychological and Cognitive Sciences and
Beijing Key Laboratory of Behavior and Mental Health,
Peking University, Beijing, China
PKU-IDG/McGovern Institute for Brain Research,
Peking University, Beijing, China

Hang Zhang
Peking-Tsinghua Center for Life Sciences,
Peking University, Beijing, China
School of Psychological and Cognitive Sciences and
Beijing Key Laboratory of Behavior and Mental Health,
Peking University, Beijing, China
PKU-IDG/McGovern Institute for Brain Research,
Peking University, Beijing, China
Chinese Institute for Brain Research, Beijing, China

Humans can grasp the "average" feature of a visual ensemble quickly and effortlessly. However, it is largely unknown what is the exact form of the summary statistic humans perceive and it is even less known whether this form can be changed by feedback. Here we borrow the concept of loss function to characterize how the summary perception is related to the distribution of feature values in the ensemble, assuming that the summary statistic minimizes a virtual expected loss associated with its deviation from individual feature values. In two experiments, we investigated a random-dot motion estimation task to infer the virtual loss function implicit in ensemble perception and see whether it can be changed by feedback. On each trial, participants reported the average moving direction of an ensemble of moving dots whose distribution of moving directions was skewed. In Experiment 1, where no feedback was available, participants' estimates fell between the mean and the mode of the distribution and were closer to the mean. In particular, the deviation from the mean and toward the mode increased almost linearly with the mode-to-mean distance. The pattern was best modeled by an inverse Gaussian loss function, which punishes large errors less heavily than the quadratic loss function does. In Experiment 2, we tested whether this virtual loss function can be altered by feedback. Two groups of participants either received the mode or the mean as the correct answer. After extensive training up to five days, both groups' estimates moved slightly towards the mode. That is, feedback had no specific influence on participants' virtual loss function. To conclude, the virtual loss function in the summary perception of motion is close to inverse Gaussian, and it can hardly be changed by feedback.

Introduction

Looking out of the window in an early spring, you may see green leaves on the trees, and it may take you a while before you realize that every leaf has a slightly different color. Humans can quickly extract summary statistics from a visual scene. The documentation of such ability dated back to Peterson and Beach's (1967) classic review "Man as an Intuitive Statistician" and, more recently, has grown into a field known as ensemble perception (Alvarez, 2011; Ariely, 2001; Chong & Treisman, 2003; Chong & Treisman, 2005), which includes a variety of visual dimensions, such as orientation (Girshick, Landy, & Simoncelli, 2011; Tomassini, Morgan, & Solomon, 2010), motion (Hol & Treue, 2001; Webb, Ledgeway, & Rocchi, 2011),

Citation: Teng, T., Li, S., & Zhang, H. (2021). The virtual loss function in the summary perception of motion and its limited adjustability. *Journal of Vision*, *21*(5):2, 1–16, https://doi.org/10.1167/jov.21.5.2.

https://doi.org/10.1167/jov.21.5.2 Received August 26, 2020; published May 4, 2021 ISSN 1534-7362 Copyright 2021 The Authors

color (Chetverikov, Campana, & Kristjansson, 2017), shape (de Gardelle & Summerfield, 2011), size (Chong & Treisman, 2005), and facial expression (Haberman & Whitney, 2010). One fundamental question is, what summary statistic of a visual distribution do humans perceive as the average (i.e., estimate of central tendency)?

One natural candidate for the average is the (arithmetic) mean of the distribution, which was widely presumed in previous studies of ensemble perception (Ariely, 2001; Solomon, Morgan, & Chubb, 2011). An alternative and more sophisticated hypothesis is *robust averaging,* that humans may underweight outliers in their summary perception of the distribution, which receives support from several lines of studies (de Gardelle & Summerfield, 2011; Juni, Singh, & Maloney, 2010; Vandormael, Herce Castanon, Balaguer, Li, & Summerfield, 2017). Compared with the computation of the mean that assigns equal weight to each sample, robust averaging can result in a more reliable estimate of central tendency for samples that are contaminated by non-Gaussian noises (Cohen, Singh, & Maloney, 2008; Huber, 2004; Juni et al., 2010).

What remains largely unknown for ensemble perception is the exact functional form that determines the weight for averaging assigned to each individual sample, depending on the location of the sample in the distribution. Here we introduce loss function, one of the key components of Bayesian Decision Theory (see Maloney & Zhang, 2010 for a review), to characterize the summary statistic in participants' ensemble perception. In this framework, the average feature perceived by a participant in a distribution of features can be considered as a point estimate for an unobservable random variable that follows the distribution. We assume that the participant's perceived average feature effectively minimizes her expected loss associated with the deviation between the point estimate and the random variable. In other words, different forms of loss function would result in different summary statistics. For example, minimizing quadratic loss ($Loss(error) = |error|^2$) would correspond to perceiving the mean of the distribution as the average, while minimizing hit loss ($Loss(error) = 0\ if\ error = 0$; $Loss(error) = 1\ if\ error \neq 0$) would correspond to perceiving the mode as the average. We understand that the participant's goal in ensemble perception is to reach a summary statistic of the distribution (such as mean, median, or mode), and thus the deviations between the participant's estimate and individual samples from the distribution do not incur any real loss. However, the concept of loss function is convenient for us to specify an otherwise recursive functional of how the influence of an individual sample on a summary statistic may change with the distance of the sample to the summary statistic. To avoid confusion with real loss function, we will use the term "virtual loss function" instead of "loss function" to characterize ensemble perception. In the present article, we will consider different families of virtual loss functions and see which form best predicts participants' ensemble perception in the absence of feedback. This form of virtual loss function will be referred to as the participant's "default" virtual loss function.

After identifying human participants' default summary statistic (virtual loss function), we ask a further question: Can people learn an arbitrary summary statistic that is chosen by the experimenter? This question is arguably important, given that the most rewarding summary statistic in different environments can be different and it would be profitable for people to adjust their ensemble perception accordingly. However, probably because it resides on the border of two different areas—ensemble perception and perceptual learning—this question has received surprisingly little treatment. As far as we know, only a few studies (Bauer, 2009; Fan, Turk-Browne, & Taylor, 2016) had investigated learning in ensemble perception, but with a different focus. For example, Fan et al. (2016) focused on the possible increase of precision of ensemble perception over training. The symmetric distribution they used, where the mean, median, and mode were all the same, cannot be used to distinguish between the different hypotheses about summary statistics. Bauer (2009) used skewed distributions, but all four sets of stimuli in his experiment had similar positive skewness so that participants could simply apply a positive or negative shift to calibrate their estimates to the designated correct answer—arithmetic or geometric mean of the distribution. In other words, it is unknown whether people can really learn an arbitrary summary statistic. Here we provide such a test, using an experimental design that implies distinct responses for different summary statistics and that precludes use of any simple calibration strategies.

In our two experiments, participants were required to report the average motion direction of an ensemble of dots that moved in different directions. We were not interested in motion perception itself, but used random dot motion as a convenient way to present thousands of samples within a few hundred milliseconds. In the literature of ensemble perception, both random dot motion (Dakin, Mareschal, & Bex, 2005; Watamaniuk & McKee, 1998; Watamaniuk, Sekuler, & Williams, 1989) and simultaneous presentation of orientation (Dakin, 2001; Girshick et al., 2011) are widely used stimuli (see Whitney & Yamanashi Leib, 2018 for a review). Previous studies of ensemble perception involve integrating features over time (Albrecht & Scholl, 2010; Joo, Shin, Chong, & Blake, 2009; Yamanashi Leib, Fischer, Liu, Qiu, Robertson, & Whitney, 2014), as well as that over space (Ariely, 2001; Chong & Treisman, 2003). According to two recent studies

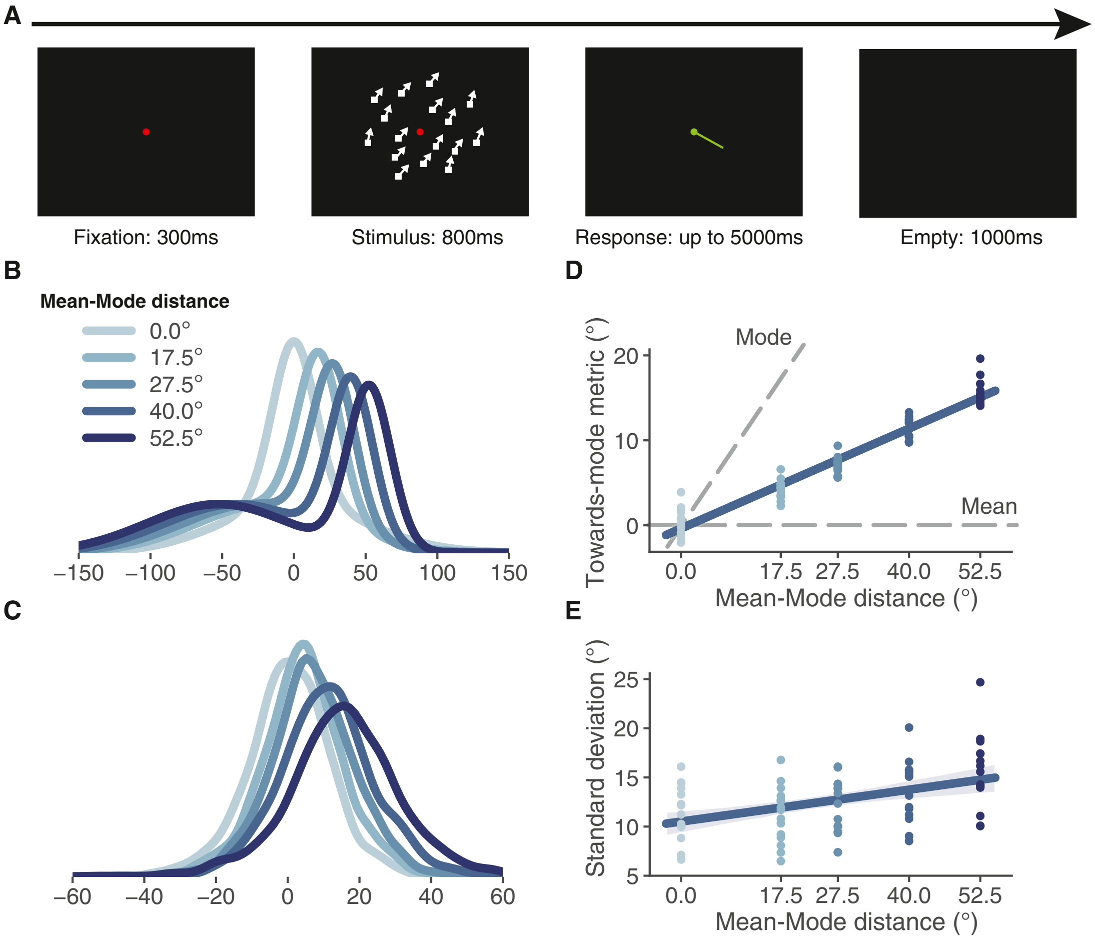

Figure 1. Task, design, and results of Experiment 1. (A) Time course of one trial. (B) Five possible forms of generative distributions. In each trial, each dot's moving direction was sampled from one Gaussian mixture distribution, whose Mean-Mode distance had five possible levels. Dark color represents larger Mean-Mode distance. (C) The distributions of participants' responses under the five Mean-Mode distance levels. (D) The towards-mode metric in participants' responses increased almost linearly with the Mean-Mode distance. The two dashed lines represent the predicted towards-mode metrics if participants report the mean or mode of the stimulus distribution. (E) The standard deviation of participants' responses also increased with the Mean-Mode distance. Dots denote data from individual participants. The line and its shading denote linear regression fit and its 95% confidence interval.

(Florey, Dakin, & Mareschal, 2017; Gorea, Belkoura, & Solomon, 2014), these two types of integration are similar in efficiency and may share common sampling mechanisms.

On each trial, the motion direction of each dot was randomly drawn from a skewed distribution, where the mean, median, and mode were dissociable. In Experiment 1, no feedback was available, and participants' estimates fell between the mean and mode, best modeled by an inverse Gaussian loss function that is consistent with robust averaging. In Experiment 2, two separate groups of participants were trained to report either the mean or the mode of specific distributions and we compared their report before and after training for both the trained and untrained distributions. After up to five days of training, participants' report became slightly closer to the mode of the motion distribution, no matter which feedback (mean or mode) they received. This suggests that people cannot flexibly adjust their virtual loss function according to feedback, even after thousands of trials of training.

Experiment 1

In Experiment 1, we aimed to identify the virtual loss function implicit in participants' ensemble perception. On each trial, participants saw an ensemble of moving dots and were required to reproduce the overall moving direction (Figure 1A). The moving direction of each dot was randomly generated from a skewed distribution that was a mixture of two Gaussian distributions of different means and variances. Across trials we varied the distance between the two Gaussian distributions to manipulate the disparity between the mean and the mode of the generative distribution (Figure 1B). We were interested in whether the mean or mode is perceived as the average moving direction, and if neither is the case, what virtual loss function can characterize participants' responses.

In our modeling, we also considered the possibility that participants may not process all the dots but instead base their estimates on random samples from the population (Dakin et al., 2005; Marchant, Simons, & de Fockert, 2013). Random sampling errors would not lead to systematic bias but might contribute to additional variations in participants' estimates.

Methods

Participants

Fifteen participants (aged 18–25, eight female) participated in Experiment 1. One participant was the first author. The other participants were naïve to the purpose of the experiment. The study was approved by the Institutional Review Board of School of Psychological and Cognitive Sciences at Peking University. All participants provided written informed consent in accordance with the Declaration of Helsinki and were compensated for their time.

Stimuli and procedure

Stimuli were presented on a Display++ monitor (Cambridge Research Systems; 31.5-inch [67.7 × 38.1 cm]; resolution 1920 × 1080 px; refresh rate 120 Hz) in a dark room, controlled by Matlab and Psychtoolbox-3 (Brainard, 1997; Pelli, 1997). Participants were seated ~60 cm in front of the screen, with their head stabilized by a chinrest.

Each trial started with a red fixation dot (diameter 0.34 cm, ≈0.3 deg) on a black background for 1000 ms, followed by 800 ms of random-dot kinematogram (RDK). Subsequently, a green bar appeared at the center of the screen. Participants were asked to use the mouse to adjust the pointing direction of the responding bar to reproduce the overall moving direction of the dots and then press the space key to confirm their response. If they did not confirm their response within 5000 ms, this trial would be forced to end and a warning message "Time-out!" would appear on the screen. We recorded the final pointing direction of the responding bar.

The RDK was composed of white dots (diameter 0.11 cm, ≈0.1 deg) whose initial positions were randomized within a square window (width 17.25 cm, ≈15 deg). The density of moving dots was set to be 13.8 dots/cm^2/second (≈16.7 dots/deg^2/second), which resulted in about 25 dots at a time on the screen. Each dot followed a two-dimensional random walk in a square area (width and height 17.25 cm, ≈15 deg of visual angle). On each subsequent frame (refreshed every 8.33 ms), each dot was displaced by 0.046 cm (i.e., moving speed 5.75 cm/s, ≈5 deg/s), whose moving direction was randomly and independently resampled from a Gaussian mixture distribution (Figure 1B). When the dot moved out of the square, it would re-enter the square from the opposite side. A circular window (diameter 17.25 cm, ≈15 deg) was applied over the square so that only dots within the circular window were visible. Please see supplemental video files for demos of RDK stimuli. Part of the stimulus code was adapted from the open resource from Shadlen lab (https://shadlenlab.columbia.edu/resources/VCRDM.html).

The generative distribution varied from trial to trial, each of which was a mixture of two equally weighted Gaussian distributions (standard deviations [*SD*s] 15° and 50°). By varying the distance between the centers of the two Gaussian distributions (0°, 35°, 55°, 80°, or 105°), we obtained five levels of distance between the mean and the mode of the mixture distribution, which was, respectively, approximately 0°, 17.5°, 27.5°, 40°, or 52.5°. The mean of the mixture distribution was sampled from 5° to 355° in steps of 10°, resulting in 36 different values. The mode of the distribution was clockwise to the mean in half of the trials and counterclockwise in the other half. All different conditions were randomly mixed. Thus, before a trial, participants had no clues to which directions the dots would be moving.

There were 36 (Mean directions) × 5 (Mean-Mode distance levels) × 2 (Mode relative to Mean: clockwise or counter-clockwise) × 2 (repetitions) = 720 experimental trials in total, divided into five blocks. Participants completed eight practice trials before the main experiment and completed the whole experiment in ~75 minutes.

Statistical analysis

For each trial, we defined "towards-mode metric" as the deviation of participants' response from the mean of the stimulus distribution towards the mode of the distribution. A towards-mode metric of 0 implies that the participant reported the mean moving direction of the stimulus distribution. A larger towards-mode metric implies a larger deviation from the mean and towards the mode. We used the mean and the *SD* of towards-mode metrics to, respectively, quantify the bias and variability in participants' responses.

We applied linear mixed model (LMM) analyses separately to the mean and *SD* of towards-mode metrics using the lme4 package in R, which included a

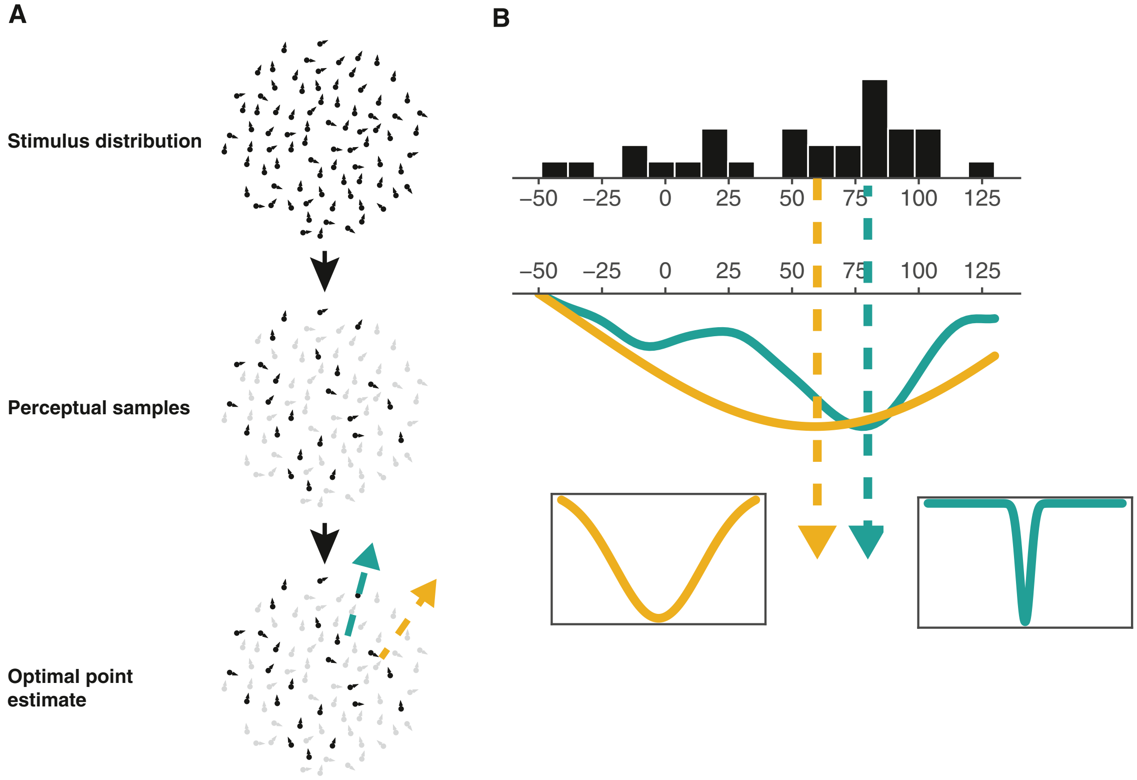

Figure 2. Illustration of sampling-based optimal decision models. (A) Assumptions. The observer draws a fixed number of perceptual samples from the motion stimulus distribution, based on which she derives a point estimate of the overall motion direction. The estimate is an optimal decision that minimizes expected loss. The yellow and green arrows denote the optimal point estimates based on two different loss functions (as shown in B). (B) How different loss functions correspond to different optimal estimates. Black bars denote effective sub-samples from the stimulus distribution. Underneath are the overall losses of different decisions under two different loss functions (a wider and a narrower inverse Gaussian loss functions). The wider loss function corresponds to an optimal estimate closer to the distribution mean, while the narrower loss function corresponds to an optimal estimate closer to the mode.

fixed effect of the mean-mode distance and a maximal random effect design (Barr, Levy, Scheepers, & Tily, 2013). The degrees of free and p values were estimated by Satterthwaite method.

Modeling

We modeled participants' estimate of the overall motion direction as the optimal choice that minimizes expected loss (Figure 2). We constructed four alternative models, all of which have the following three components. First, we assumed that on each trial the participant randomly drew a fixed number of samples (termed "perceptual samples") from the empirical distribution of motion directions and estimated the overall motion direction based on these samples. Note that drawing an infinitely large number of samples is a limiting case of a fixed number of samples. Given that the empirical distribution of motion directions on each trial consisted of approximately 2300 independent motion instances and was thus practically indistinguishable from the generative distribution, we simulated participants' perceptual samples by directly sampling from the generative distribution.

Second, given a specific loss function, we assumed that participants would choose a point estimate of the distribution that minimizes expected loss (Ma & Jazayeri, 2014; Maloney & Mamassian, 2009; Maloney & Zhang, 2010):

$$Optimal\ estimate = \underset{a}{argmin} \frac{\sum_{i=1}^{N} Loss\,(a, s_i)}{N}, \quad (1)$$

where a denotes a specific choice of point estimate, s_i denotes the motion direction of the ith sample, N is the number of perceptual samples the participant draws (effective sample size), and Loss(.) denotes the virtual loss function, which specifies the magnitude of virtual loss incurred had a deviated from s_i. Here the expected virtual loss is approximated by the mean virtual loss across the available perceptual samples. The optimal estimate that minimizes expected virtual loss would depend on the loss function, as well as the perceptual samples (Figure 2B, lower panel). We considered two alternative families of virtual loss functions. The first was the Lp loss family,

$$Loss\,(a, s; p) = |a - s|^p \quad (2)$$

The quadratic and hit loss functions we described in the Introduction are two special cases of the L_p loss family ($p = 2$ and $p = 0$), termed L_2 and L_0 , respectively, whose optimal estimates correspond to the mean and the mode of the distribution respectively. The second virtual loss family we considered was inverse Gaussian,

$$Loss\,(a, s; \sigma) = 1 - e^{-\frac{(a-s)^2}{2\sigma^2}}, \quad (3)$$

where σ controls the width of the inverted-bell-shaped loss function.

Last, a Gaussian error term, $Normal(0, \quad \sigma_{late}{}^2)$, was added to the optimal point estimate to model the late noise (e.g., motor noise and memory noise) in participants' responses.

In total, we considered four different models whose assumptions differ in two dimensions: sample size (limited vs. infinite) and virtual loss function family (inverse Gaussian vs. Lp). The 2 by 2 combinations of models were abbreviated as Ltd-InvGau, Ltd-Lp, Inf-InvGau, and Inf-Lp. Among them, the Ltd-InvGau model has three free parameters: N (effective sample size), σ (width parameter of the virtual loss function), and σ_{late} (width parameter of the late error distribution). The Ltd-Lp model also has three free parameters: N, p (shape parameter of Lp loss function), and σ_{late}. The Inf-InvGau model has two free parameters: σ and σ_{late}. The Inf-Lp model has two free parameters: p and σ_{late}.

For each participant and each model, we combined Monte-Carlo simulation and grid search to find the maximum likelihood estimation of the model parameters. Grid search settings: effective sample size N varies from 2 to 200 in step size 6; late noise σ_{late} varies from 0° to 20° in step size 0.5°; shape parameter of Lp loss function p varies from 0.1 to 3 in step size 0.1; width parameter of Inverse Gaussian loss function σ varies from 10° to 150° in step size 5°. For each combination of parameters and Mean-Mode distance level (arbitrarily setting Mean = 0), we generated 6000 simulated responses, on the basis of which we calculated the likelihood function of participants' responses. In particular, we fit a Gaussian distribution (with mean and variance as free parameters) to the 6000 simulated responses as a numerical approximation of the likelihood function. The combination of parameters that maximize the summed log likelihood across trials were chosen as the participant's estimated parameters for the model.

The Akaike information criterion with a correction for sample sizes, AICc (Akaike, 1974; Hurvich & Tsai, 1989), was used for model selection. For a specific model, the ΔAICc was computed for each participant and each task as the difference of AICc between the model and the minimum AICc among the four models. The best model on the group level was the model with the lowest ΔAICc summed across participants. The group-level Bayesian model selection (Daunizeau, Adam, & Rigoux, 2014; Rigoux, Stephan, Friston, & Daunizeau, 2014; Stephan, Penny, Daunizeau, Moran, & Friston, 2009) was used to provide an additional omnibus measure of model advantage.

In our experimental design, statistical and modeling analyses described above, we used Gaussian distributions as an approximation for von Mises distributions in the circular space and omitted possible wrap-around issues. We also performed additional modeling analyses in the circular space that compensate for wrap-around and obtained similar results (see Supplementary Figure S6 for details).

Results and discussion

The distribution of all participants' towards-mode metrics is plotted in Figure 1C. We found that participants' estimate of the overall motion direction fell between the mean and the mode of the stimulus distribution. On average the towards-mode metric (the deviation of participants' response from the mean of the stimulus distribution toward the mode of the distribution) was 29% of the Mean-Mode distance. The towards-mode metric increased almost linearly with the Mean-Mode distance (Figure 1D), whose slope according to an LMM analysis was significantly greater than zero ($t(14) = 26.18$, $p < 0.01$) and less than one ($t(14) = -62.58$, $p < 0.01$).

The SD of participants' responses (Figure 1E) also increased with the Mean-Mode distance ($t(14) = 4.76$, $p < 0.01$), which apparently could not be explained by late noises. If the variability in participants' responses was merely due to an additive late noise, it would have been constant rather than changing with the Mean-Mode distance level of the stimulus. Therefore our finding suggests that the precision of participants' summary perception decreased when the variability of the stimulus distribution increased.

In the analyses above, we collapsed participants' towards-mode metric across motion distributions with different mean directions. We also examined whether the towards-mode metric was influenced by the mean direction (Supplementary Figures S3 and S4). According to a linear mixed model analysis (Supplementary LMM S1) on participants' response, when the mean of the motion distribution was close to the horizontal axis, participants' response was

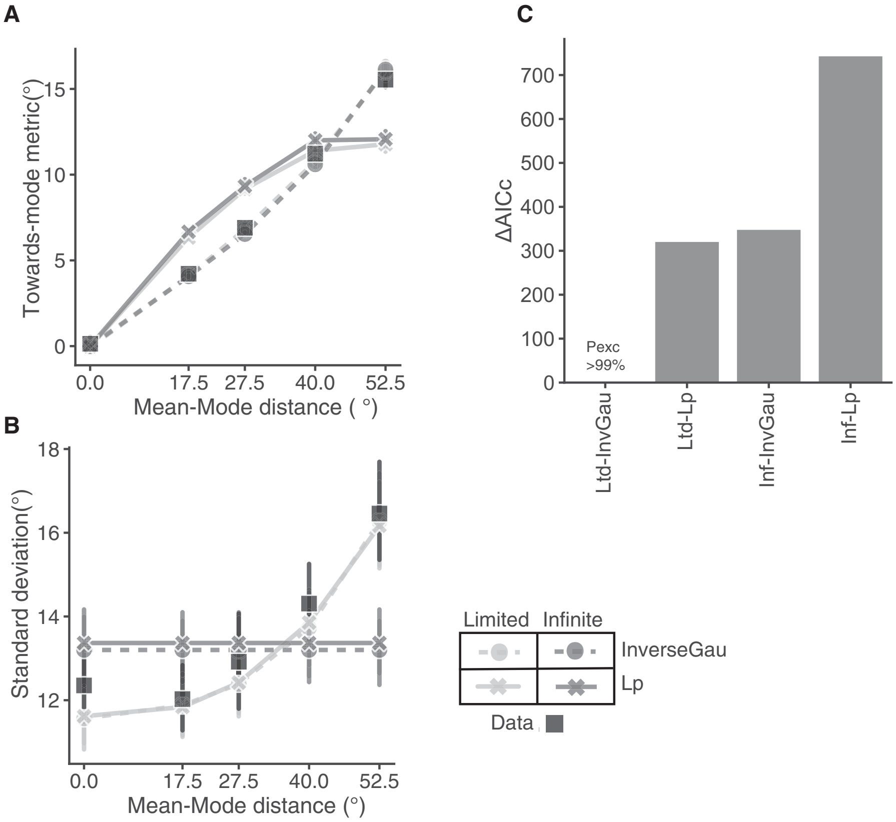

Figure 3. Modeling results of Experiment 1. Data versus model fits for the towards-mode metric (A) and standard deviation (B) of participants' responses. Error bars denote SE. Only the Ltd-InvGau model fits well to both the bias and the standard deviation patterns. (C) The ΔAICc summed over all participants. Smaller ΔAICc indicates better fit. The probability for the Ltd-InvGau model to outperform all the other models, P_{exc}, was greater than 99%.

repulsed away from the horizontal axis, which was consistent with the reference repulsion effect reported in the literature of motion or orientation perception (Rauber & Treue, 1998; Wei & Stocker, 2015). However, no repulsion effect was found for the vertical axis. The response variability did not vary with the mean direction of the motion distribution (Supplementary LMM S2). In Supplementary LMM S1, we also checked that the initial direction of the responding bar had little influence on the participant's response (Supplementary Figure S5).

We fit four alternative models—Ltd-InvGau, Ltd-Lp, Inf-InvGau, and Inf-Lp—to each participant's responses using maximum likelihood estimation and plot the model predictions versus data in Figures 3A and 3B. Only the Ltd-InvGau model well predicted both the towards-mode metric and variability in participants' responses, whereas the other models exhibited patterned deviations. In particular, models assuming an inverse Gaussian loss function but not those assuming an Lp loss function could predict the linear increase of the towards-mode metric with the Mean-Mode distance. Models assuming limited but not infinite sample size could predict the increase of response variability with the Mean-Mode distance, because the former but not the latter would introduce random sampling error that increases with the variance of the stimulus distribution.

A model comparison analysis of the four models using AICc also showed that the Ltd-InvGau model fit best to the data (Figure 3C, exceedance probability >99%). The fitted parameters of the Ltd-InvGau model is shown in Supplementary Table S1. According to the median parameters of Ltd-InvGau, participants' sample size was 53, the SD of the inverse Gaussian loss function was 75°, and the standard deviation of late noise was 10.5°. That people use only a limited number of perceptual samples is consistent with previous findings in ensemble perception (Dakin et al., 2005; Marchant et al., 2013).

The inverted-bell-shaped inverse Gaussian loss function that we identified in our data does not punish large errors as heavily as the quadratic (i.e. L2) loss function does. It agrees with the loss function Kording and Wolpert (2004) found in sensorimotor learning and is also consistent with previous findings of the underweighting of outliers in ensemble perception (de Gardelle & Summerfield, 2011;

Haberman & Whitney, 2010; Vandormael et al., 2017). We will discuss its implications further in the General discussion.

Experiment 2

In Experiment 1, we gave participants no feedback and estimated their default ensemble perception (virtual loss function). A natural question follows: Can participants' bias in ensemble perception be changed by feedback?

Although feedback has been commonly used in perceptual learning studies (Dosher & Lu, 2017), the question we ask here is different. Traditional perceptual learning studies focused on improving perceptual discriminability, whereas we focused on the bias of perceptual decisions, especially when the desired bias may vary from distribution to distribution following an abstract rule (i.e., minimizing virtual loss).

In Experiment 2, we provided two groups of participants with different feedbacks and tested whether participants could adjust their responses accordingly. One group of participants received the mean of the stimulus distribution as the feedback direction. The other group received the mode as the feedback. Participants were instructed to find a proper way to interpret the motion stimulus and reduce their error relative to the feedback. In training sessions feedback was available only at the Mean-Mode distance of 27.5°, but were tested at Mean-Mode distances of both 17.5° and 27.5° in pretests and posttests. The inclusion of an untrained Mean-Mode distance level in the tests was intended to test the generalizability of the training.

One note: Based on a limited number of samples, the mode of a continuous distribution can only be estimated with uncertainty and kernel density estimation is required. Such estimation may seem to be difficult for human participants. However, reasonably good performance was found in previous research where participants were required to estimate the mode for a multimodal, continuous distribution based on 70 samples (Sun, Li, & Zhang, 2019).

Methods

Participants

Twenty-eight naïve participants (aged 18–24, 20 female) participated in Experiment 2. They were assigned to either the Mean-feedback group (16 participants) or Mode-feedback group (12 participants). The study was approved by the Institutional Review Board of School of Psychological and Cognitive Sciences at Peking University. All participants provided written informed consent in accordance with the Declaration of Helsinki and were compensated for their time.

Stimuli and procedure

The apparatus was the same as that of Experiment 1. Similar to Experiment 1, the RDK in Experiment 2 consisted of white dots (diameter 0.11 cm, ≈0.1 deg of visual angle), whose initial positions were randomized within a square window (width 17.25 cm, ≈15 deg). The density of moving dots was set to be 27.6 dots/cm^2/s (≈33.4 dots/deg^2 /s), which resulted in about 50 dots at a time on a deep grey back ground. Each dot followed a two-dimensional random walk in a square area (width and height 17.25 cm, ≈15 deg). On each subsequent frame (refreshed every 8.33 ms), 4% of the dots disappeared and were relocated to random positions. The remaining dots were displaced by 0.12 cm (i.e., moving speed 11.5 cm/s, ≈10 deg/s), whose moving direction was randomly and independently re-sampled from a Gaussian mixture distribution. When the dot moved out of the square, it would be transformed to the opposite side of the square. A circular window (diameter 17.25 cm, ≈15 deg) was applied over the square so that only dots within the circular window were visible. Please see supplemental video files for demos of RDK stimuli. After viewing the RDK for 1500 ms, participants were asked to report the overall moving direction of the RDK and then received a 500-ms feedback of the correct answer (Figure 4). The higher density of dots and longer presentation time in Experiment 2 was motivated by the consideration that the stimuli in Experiment 1 might have not provided enough motion samples for participants.

Each participant completed one pretest, five training, and one posttest sessions in five different days. On the first day, participants first completed eight practice trials to be familiarized with the task. They then completed a no-feedback pretest session and a short training session. In the following three days, they completed three long training sessions. On the last day, participants first completed a short training session and then a no-feedback posttest session that had the same design as the pretest session.

Participants were trained only at the Mean-Mode distance of 27.5° but pretested and posttested at both 27.5° and 17.5°. As in Experiment 1, all different conditions in each session were randomly mixed. In each test session, there were 36 (Mean directions) × 2 (Mean-Mode distance levels) × 2 (Mean-Mode relative directions: clockwise or counter-clockwise) × 2 (repetitions) = 288 trials. Some of our early participants complained that the sessions on the first and last days (test + short training) were too long and tiring. To improve participants' experience, we slightly reduced the length of the short training sessions on

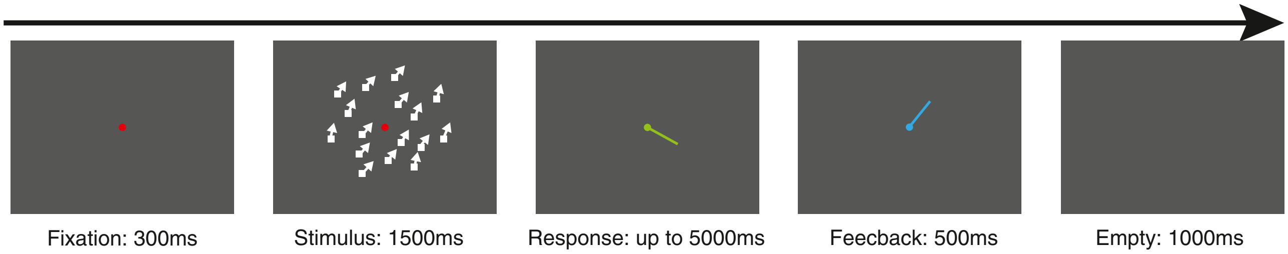

Figure 4. Task of Experiment 2 during training. Time course of one trial during training. No feedback was available in pretest and posttest, for which the procedure was the same as Experiment 1.

the first day and last day for later participants. As the result, the first nine participants of the mode feedback group completed 2448 training trials in total, and the remaining 19 participants completed 2304 training trials in total.

Statistical analysis

Similar to Experiment 1, we applied LMM analyses separately to the mean and SD of towards-mode metrics. The fixed effects of the LMMs included the main and interaction effects of experimental session (pretest session, the five training sessions, and posttest session were coded as 1 to 7), Mean-Mode distance level, and feedback group. The random effect structure was kept maximal. The significance of the fixed effects was interpreted using the lmerTest package in R (Kuznetsova, Brockhoff, & Christensen, 2017), where the degrees of freedom and p values were estimated by the Satterthwaite method. For significant interactions, we applied “emmeans” package to do post hoc tests.

Modeling

The goal of our modeling was to identify the latent dimensions that had changed across pretests and posttests. Based on the winning model of Experiment 1 (Ltd-InvGau model), we constructed eight models that differed in their flexibility across pretests and posttests in three parameters (dimensions): sample size N, loss function width σ, and noise SD σ_{late}. If one specific dimension (e.g., loss function width) was set to be “variable,” two different parameters would be used for the dimension to model the pretest and posttest data (e.g. σ^{pre} and σ^{post}). In contrast, if the dimension was set to be “fixed,” a single parameter would be used for the pretests and posttests. Each dimension can be either “variable” or “fixed,” thus resulting in $2 \times 2 \times 2 = 8$ models.

The models are named according to their assumption of flexibility on each dimension, where “V” represents “variable” and “F” represents “fixed.” For example, [F-sample, V-loss, F-noise] represents a model with fixed effective sample size, variable loss function width, and fixed late noise. The number of parameters in each model equals 3 plus the number of “F” in the model name.

For each participant and model, we fit the model to the participant’s towards-mode metrics in the pretest and posttest sessions. The model fitting and comparison procedures were the same as those of Experiment 1.

Similar to Experiment 1, we performed additional modeling analyses in the circular space that compensate for wrap-around and obtained similar results (see Supplementary Figure S7 for details).

Results and discussion

Participants’ towards-mode metric is plotted against different experimental conditions in Figure 5A. We performed a linear mixed model analysis on towards-mode metric to identify the possible differences between the two feedback groups in learning effects. Consistent with our results in Experiment 1, participants’ towards-mode metric was larger for larger Mean-Mode distance (i.e., 27.5° > 17.5°, $F(1, 37.8) = 277.96$, $p < 0.001$). Meanwhile, towards-mode metric increased with increasing experimental sessions ($F(1, 27.9) = 6.28$, $p = 0.02$). The increase of towards-mode metric across experimental sessions was larger at the trained 27.5° than at the untrained 17.5° Mean-Mode distance level (interaction $F(1, 32.6) = 5.59$, $p = 0.02$), which echoed our finding in Experiment 1 that participants’ towards-mode metric scaled with the Mean-Mode distance (see Figure 1D).

Meanwhile, the interaction between the feedback type and the experiment session did not reach significance ($F(1, 27.9) = 1.85$, $p = 0.18$). In other words, whether the feedback during training was the mean or the mode of the motion distribution had little influence on participants’ response.

According to a similar linear mixed model analysis on the SD of towards-bias metric (Figure 5B), participant’s response variability decreased across the experimental sessions ($F(1, 28.0) = 19.6$, $p < 0.01$).

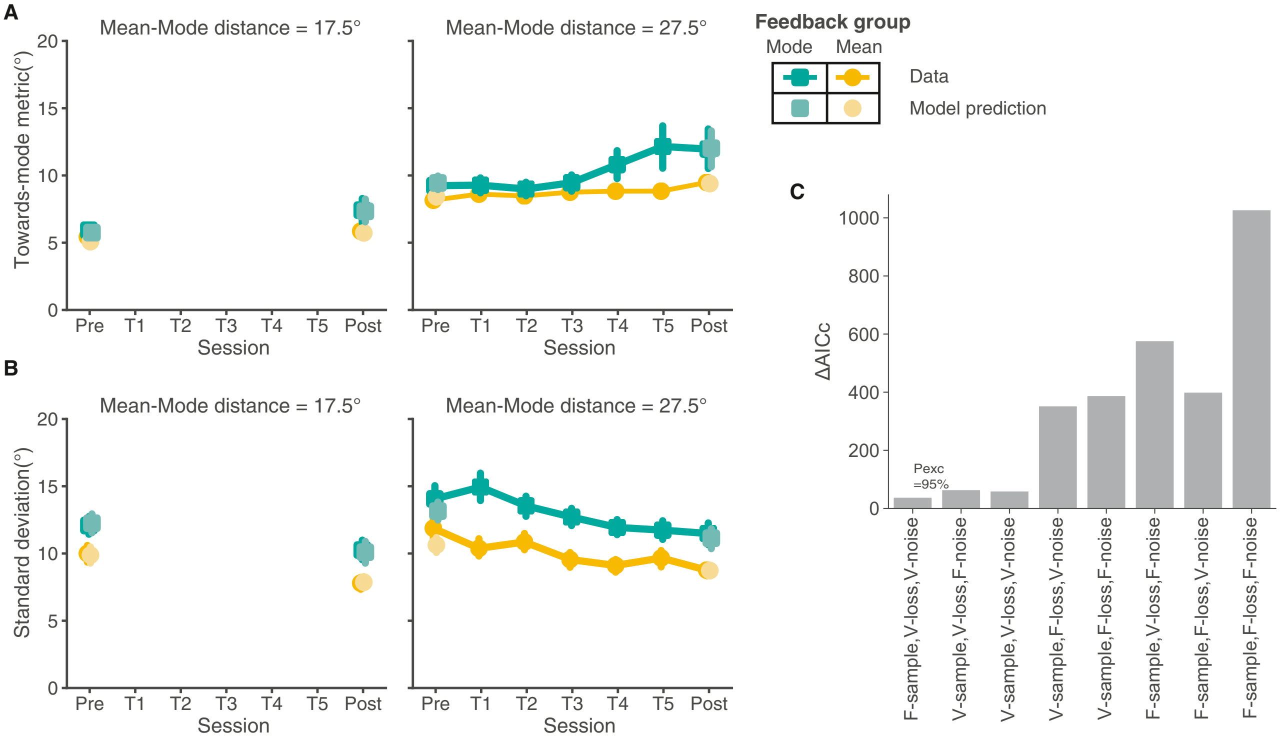

Figure 5. Results of Experiment 2. (A) The change of towards-mode metric in participants' responses separately for the Mode-feedback (green) and Mean-feedback (yellow) groups. There was no significant interaction between the feedback groups and the experimental sessions. Compared to the pretest, both groups' towards-mode metric moved slightly closer to the mode of the stimulus distribution in the posttest. (B) The change of participants' response SD. (C) Results of model comparisons based on the pretest and posttest responses. We considered eight models whose assumptions are combinations of the following three factors: whether (1) sample size, (2) loss function, and (3) late noise have been fixed (denoted "F") or variant (denoted "V") across pretest and posttest. Lower AICc indicates better fit. The probability for the winning model ([F-sample, V-loss, V-noise]) to outperform all the other models, P_{exc}, was 95.3%.

Consistent with our finding in Experiment 1 (Figure 1E), the response variability was larger at the trained 27.5° than at the untrained 17.5° Mean-Mode distance level ($F(1, 28.0) = 15.23$, $p < 0.01$). The Mode-feedback group's response variability was overall larger than the Mean-feedback group ($F(1, 28.0) = 15.23$, $p < 0.01$).

It is not ideal that nine participants in the Mode-feedback group have slightly smaller number of training trials. To exclude the possible influence of different training length, we performed additional trial-level linear mixed model analyses (Supplementary LMM S3) to reveal how participants' performance might change trial by trial with the number of training trials. All statistical conclusions were the same as those of the session-level analysis reported above.

Participants' towards-mode metric had increased over training, a change that was not specific to Mean- or Mode-feedback. Did this non-specific change arise from perception or a mapping between perception and motor response? Given that perceptual-motor mappings can be quickly shaped by feedback (Shadmehr, Smith, & Krakauer, 2010; Wolpert, Ghahramani, & Flanagan, 2001), the lack of specific learning effects after several days of training suggests that the change was probably perceptual.

To further test whether the change of participants' toward-mode bias and response variability over training was due to the change in loss function, we constructed eight models that were all based on Ltd-InvGau, the winning model of Experiment 1, but differed in whether the three parameters of Ltd-InvGau (effective sample size, loss function's width, and noise SD) were allowed to vary ("variable", denoted by prefix "V") or kept constant ("fixed", denoted by prefix "F") in fitting pretest and posttest data. According to model comparisons based on AICc, the best-fitting model ([F-sample,V-loss,V-noise]) assumed inflexible effective sample size, flexible loss function width, and flexible late noise (Figure 5C, exceedance probability = 95.3%). In fact, the second and third best models also assumed flexible loss function width. The predictions of the best model agreed well with the observed towards-mode metric and response SD (Figures 5A and 5B). The fitted parameters of the winning model is shown in Supplementary Table S2.

To summarize, we found that different feedbacks had similar effects on participants' toward-mode metric—the responses of both the feedback groups moved slightly toward the mode after extensive training. We will discuss this limited adjustability of virtual loss function in General discussion.

General discussion

We investigated ensemble perception using skewed motion distributions and found that the overall motion direction participants reported falls between the mean and mode of the stimulus distribution, and that participants' bias toward the mode and their response variability increase with the Mean-Mode distance. These patterns can be well predicted by a sampling-based optimal decision model that assumes an inverse-Gaussian loss function, which effectively underweights extreme values in the stimulus distribution. In a second experiment, we further examined whether participants' ensemble perception, in terms of virtual loss function, can be changed by feedback. We trained two groups of participants for five days with either the mean or mode of the stimulus distribution as feedback and found no feedback-specific learning effects but that participants in both groups moved their estimates slightly toward the mode.

Excluding an alternating-response hypothesis

That the *SD*, as well as the mean of participants' towards-mode metric, increased with the Mean-Mode distance of the stimulus might be explained by the following alternating-response hypothesis: Participants had perceived two (or more) discrete directions from the mixture distribution, such as the two centers of the two Gaussian components, and alternatively reported different directions in different trials. If so, we would expect to see bimodality (or multimodality) in the distribution of their responses (Laquitaine & Gardner, 2018). However, our data patterns did not support this hypothesis. Supplementary Figure S1 shows the distribution of towards-mode metrics in Experiment 1, separately for each participant, each Mean-Mode distance condition, and Mean-Mode relative direction (clockwise or counter-clockwise). Almost all the distributions appeared to be unimodal. We used the bimodality coefficient (BC) to measure how likely a distribution is bimodal or multimodal instead of being unimodal. A BC higher than 0.555 suggests that the distribution is bimodal or multimodal, whereas a BC lower than 0.555 suggests unimodality (Freeman & Dale, 2013). Of the 135 distributions in Supplementary Figure S1, 131 distributions had BC lower than 0.555 (group-averaged BC: 0.324). We computed the mean BCs for each participant and performed a group-level *t*-test against the null hypothesis "group-averaged BC was higher than 0.555," which indicates that the distribution of towards-mode metric was unimodal ($t(14) = -39.89$, $p < 0.01$). Moreover, as a more direct evidence against the alternating-response hypothesis, there were no increased responses at either of the two modes of the two Gaussian components (marked by dash lines in Supplementary Figure S1).

Similarly, in Experiment 2, there was little bimodality or multimodality in the distributions of individual participants' towards-mode metrics (Supplementary Figure S2). All 224 distributions in Supplementary Figure S2 had a BC lower than 0.555 (group-averaged BC: 0.341, $t(27) = -36.13$, $p < 0.01$). Again, this is against the alternating-response hypothesis and suggests that a single moving direction was perceived in each trial.

Of course, if participants had perceived the two centers of the two Gaussian components in the mixture distribution but used a weighted average of them as their response, no multimodality would be observed likely. However, had participants been able to perceive and integrate the two discrete directions, in Experiment 2 it would not have been so hard for them to adjust their responses to match the predefined correct answer (i.e., the Mode or Mean of the mixture distribution). Therefore we considered it unlikely that participants had perceived two (or more) discrete directions of motion from the mixture distribution.

Loss function

In studies where Bayesian observer models are used to model human perception (Stocker & Simoncelli, 2006), action (Kording & Wolpert, 2004), and working memory (Ding, Cueva, Tsodyks, & Qian, 2017) but where loss functions are not explicitly specified, both L0 and L2 are common choices of loss functions in modeling practice, such as in the maximum a posteriori and Bayes least-squares models of Jazayeri and Shadlen (2010). However, we found that the loss function implicit in participants' ensemble perception of motion agrees neither with L0 nor with L2 but lies in between. A caveat to interpreting our results is that the loss function we measured is "virtual," which is applied to a stimulus distribution instead of a posterior distribution of beliefs as in Bayesian observer models. Despite this difference, the virtual loss function we studied here may still capture a common essential aspect of human behavior: how people summarize an arbitrary distribution.

Our modeling approach was partly inspired by Kording and Wolpert's (2004) work in sensorimotor loss function, where they asked participants to play a virtual shooting game under skewed distributions of sensorimotor errors. Their data allowed them to reject both L0 and L2 loss functions but were inconclusive about whether the Lp or the inverse Gaussian loss function fits better. In contrast, our data clearly favored the inverse Gaussian over the Lp loss function: The former outperformed the latter in predicting participants' response patterns (Figures 3A and 3B), as well as in goodness-of-fit (Figure 3C, exceedance probability >99%).

The inverse Gaussian loss function found in our study of motion perception corresponds to a summary statistic between the mean and mode, which agrees with findings of Webb and colleagues in motion perception (Webb, Ledgeway, & McGraw, 2007; Webb et al., 2011) as well as with the loss function found in sensorimotor planning (Kording & Wolpert, 2004) or visual working memory (Sims, 2015). An exception is Sun, Li, and Zhang (2019), where participants were explicitly required to report the mean and mode of skewed visuospatial distributions and the mean reported by participants biased toward the tail instead of the mode of the distribution.

One important feature of the inverse Gaussian loss function is that it does not penalize large errors as much as the L2 (quadratic) loss function. In other words, the choice that minimize inverse Gaussian loss would underweight outliers. Indeed, humans are widely documented to underweight outliers in ensemble perception, such as the facial expression (Haberman & Whitney, 2010), color (de Gardelle & Summerfield, 2011), number (Vandormael et al., 2017), and orientation (Li, Herce Castanon, Solomon, Vandormael, & Summerfield, 2017), known as robust averaging (Huber, 2004; Juni et al., 2010). Our findings add to evidence for the hypothesis of robust averaging and further advance our understanding of robust averaging in the following two aspects. First, we have observed that the inverse Gaussian loss function can better characterize the human summary perception of motion than the Lp loss function, though both loss functions can implement robust averaging. Second, such robust averaging is "robust" itself and changes very little under different external goals, even after extensive training. The theoretical implications of these observations deserve future research.

Sampling and effective sample size

We assumed that instead of applying virtual loss function to all samples, participants may only take a limited number of samples into account. The introduction of limited sample size can explain why the variability of participants' ensemble perception increases with the Mean-Mode distance of the stimulus distribution.

Similar to previous studies of ensemble perception (Dakin et al., 2005; Marchant et al., 2013), the sampling process we modeled is at the computational-theory level, whose capacity is quantified by effective sample size. We cannot exclude other forms of sampling process that provides equivalent amount of information (e.g. taking a larger number of noisier samples). Whether and how participants really sample from the stimulus distribution is a question for future research. For example, participants might sample from the trajectory of a single dot, or instead from multiple dots, simultaneously or sequentially. Though in Experiment 1 the estimated effective sample size (median 53) was smaller than the number of samples in one dot's trajectory (median 96), in Experiment 2 the former (median 32) was much larger than the latter (median 17), thus largely excluding the possibility that participants sampled only from one single dot's trajectory. Eye tracking and manipulation of spatial attention would be two promising methods to further investigate the algorithm people use to gather information in the summary perception of motion.

We also cannot exclude the possibility that participants may have perceived multiple consecutive motion samples as one motion sample, due to limited temporal precision of their visual system. If such pooling had occurred, the motion samples participants actually perceived would follow a distribution that has the same mean as the presented distribution but whose mode is closer to the mean. This might explain why participants failed to report the exact mode of the presented distribution in the Mode-feedback group, but could not explain why participants also failed to learn the mean in the Mean-feedback group. That is, pooling is unlikely to cause the lack of specific learning effects. But pooling may provide an explanation for the nonspecific change we observed: If participants used the same way to integrate the samples they perceived but over training each percept pooled a smaller number of samples because of increased temporal precision, their responses would slightly move toward the mode regardless of feedback.

We have omitted modeling the visual noise in perceiving individual motion samples, partly for simplicity and partly because the effect of visual noise may not be empirically separable from that of effective sample size, virtual loss function, or late noise. Similar to late noise, visual noise alone could not explain why the towards-mode metric and its variability increase with the Mean-Mode distance of the distribution. Similar to pooling, visual noise may bias the mode but not the mean estimated from samples.

Limited adjustability of loss function

Given the large body of evidence that human decisions are adaptive (Cheadle et al., 2014; Dayan & Niv, 2008; Keramati, Smittenaar, Dolan, & Dayan, 2016), one might expect participants to be able to adjust the summary statistic of their ensemble perception according to the rewarding structure of the environment. However, we found little evidence for such adjustments. Why do people have difficulty adjusting their virtual loss functions? We consider a few possibilities below.

One possibility is that people may be insensitive to any higher-order probabilistic information beyond the mean and variance (i.e., the first two moments) of the motion distribution, as suggested by Waskom, Asfour, and Kiani (2018). But this is unlikely to be true in our case, otherwise participants' estimates of the overall motion direction would not have systematically deviated from the mean of the stimulus distribution, neither would their responses be shaped by feedback at all.

A second possibility is that participants' ensemble perception may be determined by hard-wired neural circuits that are hardly subject to the rewarding structure of the environment. From the perspective of population coding, the winner-take-all and vector-averaging decoding algorithms roughly correspond to ensemble perception at the mode and mean of the stimulus distribution, respectively (Zohary, Scase, & Braddick, 1996). Parallel to our rejection of the L0 and L2 loss functions, Webb and colleagues found that neither of the two decoding algorithms can explain participants' psychophysical data in motion perception, which implies an ensemble perception between the mean and the mode of the stimulus distribution (Webb et al., 2007; Webb et al., 2011). Meanwhile, they found that the overall direction participants perceive would vary with the duration and the temporal or spatial dynamics of motion stimuli. However, they did not consider the decoding algorithm itself to be adjustable but explained the changed motion perception under different motion conditions as a result of neural temporal dynamics. Our results were consistent with their conjecture that the neural read-out of the global motion direction might not be adjustable.

It is also possible that people may have the ability to adjust the virtual loss function implicit in their ensemble perception but simply do not have enough motivation to do so. It should be noted that the initial bias is much closer to the mean than to the mode of the stimulus distribution. As a result, participants in the Mean-feedback group may not be well motivated to improve.

Finally, other than the training received in our laboratory, participants are frequently exposed to motion stimuli in their daily life. Several hours of laboratory training is probably not intense enough to reverse many years of perceptual experience.

Keywords: summary perception, ensemble perception, loss function, robust averaging, plasticity

Acknowledgments

Supported by grants 31871101 and 31571117 from National Natural Science Foundation of China.

Funding from Peking-Tsinghua Center for Life Sciences, and grant 2017YFB1002503 from National Key R&D Program of China.

Data from the two experiments are available at https://osf.io/x4j3k/.

Commercial relationships: none.
Corresponding authors: Sheng Li and Hang Zhang.
Email: sli@pku.edu.cn; hang.zhang@pku.edu.cn.
Address: Peking University, 52 Haidian Road, Haidian District, Beijing, 100080, China.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723, doi:10.1109/TAC.1974.1100705.

Albrecht, A. R., & Scholl, B. J. (2010). Perceptually averaging in a continuous visual world: Extracting statistical summary representations over time. *Psychological Science, 21*(4), 560–567, doi:10.1177/0956797610363543.

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences, 15*(3), 122–131, doi:10.1016/j.tics.2011.01.003.

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science, 12*(2), 157–162, doi:10.1111/1467-9280.00327.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278, doi:10.1016/j.jml.2012.11.001.

Bauer, B. (2009). The danger of trial-by-trial knowledge of results in perceptual averaging studies. *Attention, Perception, & Psychophysics, 71*(3), 655–665, doi:10.3758/APP.71.3.655.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*(4), 433–436, doi:10.1163/156856897X00357.

Cheadle, S., Wyart, V., Tsetsos, K., Myers, N., de Gardelle, V., Herce Castanon, S., . . . Summerfield, C. (2014). Adaptive gain control during human perceptual choice. *Neuron, 81*(6), 1429–1441, doi:10.1016/j.neuron.2014.01.020.

Chetverikov, A., Campana, G., & Kristjansson, A. (2017). Representing color ensembles. *Psychological Science, 28*(10), 1510–1517, doi:10.1177/0956797617713787.

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research, 43*(4), 393–404, doi:10.1016/s0042-6989(02)00596-5.

Chong, S. C., & Treisman, A. (2005). Statistical processing: Computing the average size in perceptual groups. *Vision Research, 45*(7), 891–900, doi:10.1016/j.visres.2004.10.004.

Cohen, E. H., Singh, M., & Maloney, L. T. (2008). Perceptual segmentation and the perceived orientation of dot clusters: The role of robust statistics. *Journal of Vision, 8*(7), 6.1–13, doi:10.1167/8.7.6.

Dakin, S. C. (2001). Information limit on the spatial integration of local orientation signals. *Journal of the Optical Society of America A, 18*(5), 1016–1026, doi:10.1364/JOSAA.18.001016.

Dakin, S. C., Mareschal, I., & Bex, P. J. (2005). Local and global limitations on direction integration assessed using equivalent noise analysis. *Vision Research, 45*(24), 3027–3049, doi:10.1016/j.visres.2005.07.037.

Daunizeau, J., Adam, V., & Rigoux, L. (2014). Vba: A probabilistic treatment of nonlinear models for neurobiological and behavioural data. *PLoS Computational Biology, 10*(1), e1003441, doi:10.1371/journal.pcbi.1003441.

Dayan, P., & Niv, Y. (2008). Reinforcement learning: The good, the bad and the ugly. *Current Opinion in Neurobiology, 18*(2), 185–196, doi:10.1016/j.conb.2008.08.003.

de Gardelle, V., & Summerfield, C. (2011). Robust averaging during perceptual judgment. *Proceedings of the National Academy of Sciences of the United States of America, 108*(32), 13341–13346, doi:10.1073/pnas.1104517108.

Ding, S., Cueva, C. J., Tsodyks, M., & Qian, N. (2017). Visual perception as retrospective bayesian decoding from high- to low-level features. *Proceedings of the National Academy of Sciences of the United States of America, 114*(43), E9115–E9124, doi:10.1073/pnas.1706906114.

Dosher, B., & Lu, Z. L. (2017). Visual perceptual learning and models. *Annual Review of Vision Science, 3*, 343–363, doi:10.1146/annurev-vision-102016-061249.

Fan, J. E., Turk-Browne, N. B., & Taylor, J. A. (2016). Error-driven learning in statistical summary perception. *Journal of Experimental Psychology: Human Perception and Performance, 42*(2), 266–280, doi:10.1037/xhp0000132.

Florey, J., Dakin, S. C., & Mareschal, I. (2017). Comparing averaging limits for social cues over space and time. *Journal of Vision, 17*(9), 17–17, doi:10.1167/17.9.17.

Freeman, J. B., & Dale, R. (2013). Assessing bimodality to detect the presence of a dual cognitive process. *Behavior Research Methods, 45*(1), 83–97.

Girshick, A. R., Landy, M. S., & Simoncelli, E. P. (2011). Cardinal rules: Visual orientation perception reflects knowledge of environmental statistics. *Nature Neuroscience, 14*(7), 926–932, doi:10.1038/nn.2831.

Gorea, A., Belkoura, S., & Solomon, J. A. (2014). Summary statistics for size over space and time. *Journal of Vision, 14*(9), 22–22, doi:10.1167/14.9.22.

Haberman, J., & Whitney, D. (2010). The visual system discounts emotional deviants when extracting average expression. *Attention, Perception, & Psychophysics, 72*(7), 1825–1838, doi:10.3758/APP.72.7.1825.

Hol, K., & Treue, S. (2001). Different populations of neurons contribute to the detection and discrimination of visual motion. *Vision Research, 41*(6), 685–689, doi:10.1016/s0042-6989(00)00314-x.

Huber, P. J. (2004). *Robust statistics* (*Vol. 523*). New York: John Wiley & Sons.

Hurvich, C. M., & Tsai, C. L. (1989). Regression and time-series model selection in small samples. *Biometrika, 76*(2), 297–307, doi:10.1093/biomet/76.2.297.

Jazayeri, M., & Shadlen, M. N. (2010). Temporal context calibrates interval timing. *Nature Neuroscience, 13*(8), 1020–1026, doi:10.1038/nn.2590.

Joo, S. J., Shin, K., Chong, S. C., & Blake, R. (2009). On the nature of the stimulus information necessary for estimating mean size of visual arrays. *Journal of Vision, 9*(9), 7–7, doi:doi.org/10.1167/9.9.7.

Juni, M. Z., Singh, M., & Maloney, L. T. (2010). Robust visual estimation as source separation. *Journal of Vision, 10*(14), 2, doi:10.1167/10.14.2.

Keramati, M., Smittenaar, P., Dolan, R. J., & Dayan, P. (2016). Adaptive integration of habits into depth-limited planning defines a habitual-goal-directed spectrum. *Proceedings of the National Academy of Sciences of the United States of*

America, 113(45), 12868–12873, doi:10.1073/pnas.1609094113.

Kording, K. P., & Wolpert, D. M. (2004). The loss function of sensorimotor learning. *Proceedings of the National Academy of Sciences of the United States of America, 101*(26), 9839–9842, doi:10.1073/pnas.0308394101.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). Lmertest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1–26, doi:10.18637/jss.v082.i13.

Laquitaine, S., & Gardner, J. L. (2018). A switching observer for human perceptual estimation. *Neuron, 97*(2), 462–474 e466, doi:10.1016/j.neuron.2017.12.011.

Li, V., Herce Castanon, S., Solomon, J. A., Vandormael, H., & Summerfield, C. (2017). Robust averaging protects decisions from noise in neural computations. *PLoS Computational Biology, 13*(8), e1005723, doi:10.1371/journal.pcbi.1005723.

Ma, W. J., & Jazayeri, M. (2014). Neural coding of uncertainty and probability. *Annual Review of Neuroscience, 37*(1), 205–220, doi:10.1146/annurev-neuro-071013-014017.

Maloney, L. T., & Mamassian, P. (2009). Bayesian decision theory as a model of human visual perception: Testing bayesian transfer. *Visual Neuroscience, 26*(1), 147–155, doi:10.1017/S0952523808080905.

Maloney, L. T., & Zhang, H. (2010). Decision-theoretic models of visual perception and action. *Vision Research, 50*(23), 2362–2374, doi:10.1016/j.visres.2010.09.031.

Marchant, A. P., Simons, D. J., & de Fockert, J. W. (2013). Ensemble representations: Effects of set size and item heterogeneity on average size perception. *Acta Psychologica, 142*(2), 245–250, doi:10.1016/j.actpsy.2012.11.002.

Pelli, D. G. (1997). The videotoolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*(4), 437–442, doi:10.1163/156856897x00366.

Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological bulletin, 68*(1), 29–&, doi:10.1037/h0024722.

Rauber, H. J., & Treue, S. (1998). Reference repulsion when judging the direction of visual motion. *Perception, 27*(4), 393–402, doi:10.1068/p270393.

Rigoux, L., Stephan, K. E., Friston, K. J., & Daunizeau, J. (2014). Bayesian model selection for group studies - revisited. *NeuroImage, 84*, 971–985, doi:10.1016/j.neuroimage.2013.08.065.

Shadmehr, R., Smith, M. A., & Krakauer, J. W. (2010). Error correction, sensory prediction, and adaptation in motor control. *Annual Review of Neuroscience, 33*, 89–108, doi:10.1146/annurev-neuro-060909-153135.

Sims, C. R. (2015). The cost of misremembering: Inferring the loss function in visual working memory. *Journal of Vision, 15*(3), 2–2, doi:10.1167/15.3.2.

Solomon, J. A., Morgan, M., & Chubb, C. (2011). Efficiencies for the statistics of size discrimination. *Journal of Vision, 11*(12), 13, doi:10.1167/11.12.13.

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian model selection for group studies. *NeuroImage, 46*(4), 1004–1017, doi:10.1016/j.neuroimage.2009.03.025.

Stocker, A. A., & Simoncelli, E. P. (2006). Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience, 9*(4), 578–585, doi:10.1038/nn1669.

Sun, J., Li, J., & Zhang, H. (2019). Human representation of multimodal distributions as clusters of samples. *PLoS Computational Biology, 15*(5), e1007047, doi:10.1371/journal.pcbi.1007047.

Tomassini, A., Morgan, M. J., & Solomon, J. A. (2010). Orientation uncertainty reduces perceived obliquity. *Vision Research, 50*(5), 541–547, doi:10.1016/j.visres.2009.12.005.

Vandormael, H., Herce Castanon, S., Balaguer, J., Li, V., & Summerfield, C. (2017). Robust sampling of decision information during perceptual choice. *Proceedings of the National Academy of Sciences of the United States of America, 114*(10), 2771–2776, doi:10.1073/pnas.1613950114.

Waskom, M. L., Asfour, J., & Kiani, R. (2018). Perceptual insensitivity to higher-order statistical moments of coherent random dot motion. *Journal of Vision, 18*(6), 9, doi:10.1167/18.6.9.

Watamaniuk, S. N., & McKee, S. P. (1998). Simultaneous encoding of direction at a local and global scale. *Perception & Psychophysics, 60*(2), 191–200, doi:10.3758/bf03206028.

Watamaniuk, S. N., Sekuler, R., & Williams, D. W. (1989). Direction perception in complex dynamic displays: The integration of direction information. *Vision Research, 29*(1), 47–59, doi:10.1016/0042-6989(89)90173-9.

Webb, B. S., Ledgeway, T., & McGraw, P. V. (2007). Cortical pooling algorithms for judging global motion direction. *Proceedings of the National Academy of Sciences of the United States of America, 104*(9), 3532–3537, doi:10.1073/pnas.0611288104.

Webb, B. S., Ledgeway, T., & Rocchi, F. (2011). Neural computations governing spatiotemporal pooling of visual motion signals in humans. *Journal of Neuroscience, 31*(13), 4917–4925, doi:10.1523/JNEUROSCI.6185-10.2011.

Wei, X. X., & Stocker, A. A. (2015). A bayesian observer model constrained by efficient coding can explain ’anti-bayesian’ percepts. *Nature Neuroscience, 18*(10), 1509–1517, doi:10.1038/nn.4105.

Whitney, D., & Yamanashi Leib, A. (2018). Ensemble perception. *Annual Review of Psychology, 69*, 105–129, doi:10.1146/annurev-psych-010416-044232.

Wolpert, D. M., Ghahramani, Z., & Flanagan, J. R. (2001). Perspectives and problems in motor learning. *Trends in Cognitive Sciences, 5*(11), 487–494, doi:10.1016/s1364-6613(00)01773-3.

Yamanashi Leib, A., Fischer, J., Liu, Y., Qiu, S., Robertson, L., & Whitney, D. (2014). Ensemble crowd perception: A viewpoint-invariant mechanism to represent average crowd identity. *Journal of Vision, 14*(8), 26–26, doi:10.1167/14.8.26.

Zohary, E., Scase, M. O., & Braddick, O. J. (1996). Integration across directions in dynamic random dot displays: Vector summation or winner take all? *Vision Research, 36*(15), 2321–2331, doi:10.1016/0042-6989(95)00287-1.