SCIENTIFIC REPORTS

OPEN

Optimal neural inference of stimulus intensities

Travis Monk[1], Cristina Savin[2] & Jörg Lücke[1]

Received: 23 November 2017
Accepted: 14 June 2018
Published online: 03 July 2018

In natural data, the class and intensity of stimuli are correlated. Current machine learning algorithms ignore this ubiquitous statistical property of stimuli, usually by requiring normalized inputs. From a biological perspective, it remains unclear how neural circuits may account for these dependencies in inference and learning. Here, we use a probabilistic framework to model class-specific intensity variations, and we derive approximate inference and online learning rules which reflect common hallmarks of neural computation. Concretely, we show that a neural circuit equipped with specific forms of synaptic and intrinsic plasticity (IP) can learn the class-specific features and intensities of stimuli simultaneously. Our model provides a normative interpretation of IP as a critical part of sensory learning and predicts that neurons can represent nontrivial input statistics in their excitabilities. Computationally, our approach yields improved statistical representations for realistic datasets in the visual and auditory domains. In particular, we demonstrate the utility of the model in estimating the contrastive stress of speech.

The intensity of a sensory stimulus can carry important information. For example, consider the sentence 'Calvin yelled something at Hobbes'. A speaker can change the meaning of the sentence by stressing certain words in it. 'Calvin *yelled* something at Hobbes' emphasizes that Calvin did not speak in a normal voice but that he yelled. 'Calvin yelled something at *Hobbes*' emphasizes that Hobbes, and not somebody else, was the recipient of Calvin's yelling. Stressing other words, or combinations of words, will imply other meanings, a linguistic technique termed contrastive stress[1]. How might neural circuits estimate the contrastive stress of a sentence? More generally, how might neurons learn and represent the intensity of stimuli?

One naive solution would be to infer that a word is stressed if it is louder than other words, since stress and loudness are often correlated[2–4]. This proposal fails because the intensity of a stimulus (utterance) depends on its class (word). As an illustration, consider the logatomes 'pap' and 'zos', taken from the raw Oldenburg Logatome Corpus dataset (OLLO)[5], with example spectrograms shown in Fig. 1A. The total energy in time-frequency space, or their 'brightness' $\hat{y}$, differs across individual logatomes. Furthermore, in this dataset we see that there are systematic differences in intensity for the two logatome classes, with 'zos' generally being louder than 'pap' (Fig. 1B). If stress were determined by a threshold on brightness given by the average intensity across classes, then we would incorrectly label most 'zos' logatomes as stressed and most 'pap' logatomes as de-stressed. This example reveals a key insight about the statistics of natural stimuli: intensity and class information depend on one another. This observation is not restricted to the auditory domain but holds true across sensory modalities (Fig. 1C). Hence, making correct judgements about intensity needs to consider the stimulus class. Conversely, intensity information needs to be considered when estimating the stimulus class.

Traditional models of sensory processing ignore dependencies between intensity and stimulus class[6,7]. Most models treat stimulus intensity as a nuisance variable that is typically removed by some form of *ad hoc* normalization[8–11]. Such preprocessing discards a potentially useful source of information that could have been used to make better inferences about the stimuli. Beyond being computationally inefficient, these models cannot explain a key feature of sensory perception: information about intensity is consciously accessible. For example, we can detect the stress of utterances, or whether a scene is being viewed at midday or at dusk. In summary, existing models ignore an important statistical feature of natural stimuli. Furthermore, it remains unclear how joint inference of intensity and stimulus class can be achieved by neural circuitry.

Here we use a novel and flexible probabilistic generative model to investigate statistical dependencies between the intensity and class of a stimulus. Our model is a rare instance of a generative model that is analytically tractable, yielding closed-form joint and marginal inference for both stimulus class and intensity. Moreover, we find

[1]Neurosensory Science, Cluster of Excellence Hearing4all, University of Oldenburg, Oldenburg, 26129, Germany. [2]Center for Neural Science and Center for Data Science, NYU, New York, 10003, USA. Correspondence and requests for materials should be addressed to T.M. (email: travis.monk@uni-oldenburg.de)

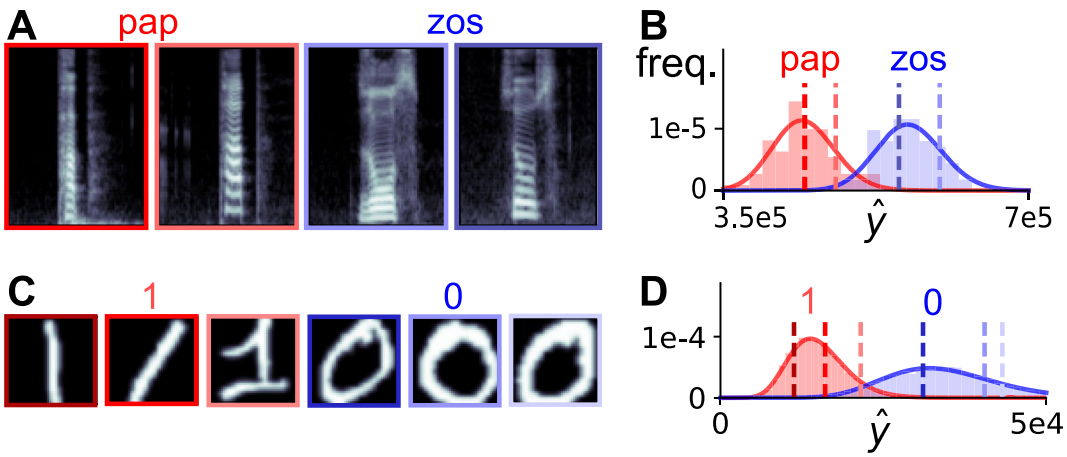

Figure 1. Intensity and class are correlated in natural data. (**A**) Example log-mel spectrograms of logatomes 'pap' and 'zos' from the raw OLLO dataset. (**B**) Histograms of the brightness $\hat{y}$ (the sum of the spectrogram values) across utterances for each class and corresponding Gamma fit. Brightness values corresponding to examples in A are marked by dashed lines. (**C**) Example images of classes '1' and '0' from the raw MNIST dataset. (**D**) Same as B, in the visual domain.

that these computations can be effectively approximated by biologically-plausible neural plasticity and dynamics. We show that a neural circuit equipped with two specific plasticity rules can learn the statistics of sensory inputs. Hebbian plasticity allows the circuit's synapses to capture the features of classes, consistent with other unsupervised learning algorithms[8,12–18]. Intrinsic plasticity (IP)[19,20] adapts the overall excitability of a neuron to reflect the average intensity of each input stimulus class. From a biological perspective, our results provide a principled, normative, and testable computational account for the role of IP in sensory learning. Intensity estimation and IP thus become part of the general approach of perception as probabilistic inference[21–24]. From a machine learning perspective, the derived results provide novel and efficient implementations for a difficult statistical problem, with applications in auditory processing and beyond.

Results

Modeling statistical dependencies between stimulus class and intensity. We start by defining a probabilistic generative model that describes how observations (stimuli) are generated (Fig. 2). According to our model, a stimulus $\mathbf{y}$ (e.g. the spectrogram of a specific utterance) with elements y_d (e.g. points in time-frequency space) belongs to a class c (e.g. the logatome 'pap'). Stimuli are generated by multiplicatively combining class-specific features $\mathbf{W}$ with an intensity variable z. Class features (e.g., prototypical logatomes such as 'pap' and 'zos') are represented by the rows of matrix $\mathbf{W}$. The intensity variable z is drawn from a *class-specific* distribution $P(z|c)$, and Poisson noise is finally added to the observations (Fig. 2A). Here we chose to model shape using a prototypical class representation[8,15,18] as it facilitates a fully probabilistic treatment of class and intensity information. Approaches that provide a decomposition of stimuli into different components[10,13,25] pose additional analytical challenges, but may profit from the results provided here. The multiplicative effect of the intensity is motivated by previous work modeling contrast variations in images[6,7]. Since intensities must be positive, the Gamma distribution is a natural choice for $P(z|c)$. Lastly, the Poisson variability is a canonical noise model for natural data, e.g. photon counts[26] and neural spikes[27]. Its mathematical properties also facilitate a link to neural circuits[8,13,18].

As a visual illustration of the generative model, Fig. 2B shows one simple artificial example in the visual domain. Stimuli belong to 3 classes which vary in their intensity distributions, with class 1 the dimmest and class 3 the brightest on average. Classes also vary in their shapes, modelled here as 2-D images of white boxes of varying sizes on a black background, normalized to sum to a constant. Individual observations are noisy versions of these prototypes, scaled by their intensity. Both the class identity and the intensity of any given stimulus are unknown and need to be inferred. Importantly, it is not only the location of the bright pixels that provides information about class identity; the overall brightness of the image, $\hat{y} = \sum_d y_d$, is also informative about the class. Conversely, the intensity variable z depends not only on $\hat{y}$ but also on the total number of white pixels, which is class-specific. Hence knowing the class c helps to infer the value of z and vice versa.

Similar qualitative features are also present in real-world data. Figure 2C shows example images of written digits '0' and '1' from the MNIST dataset where the position of each pixel is shuffled to destroy shape information (i.e. features). The resulting images still look different across digits: the shuffled '0' is brighter than the shuffled '1'. Hence, when presented with two other shuffled images one can make a decent guess about their classes despite missing spatial structure (Fig. 2D). This example suggests that intensity judgements may be generally useful, even when the assumptions of the model are not satisfied exactly.

Inferring the class and intensity of stimuli under the generative model. Jointly inferring c and z given a stimulus $\mathbf{y}$ can be achieved by applying Bayes's rule: $P(c, z|\mathbf{y}, \theta) \propto P(\mathbf{y}|c, z, \theta)P(z|c, \theta)P(c|\theta)$. θ is shorthand for parameters $\mathbf{W}$ and α, β (the shape and rate parameters of the Gamma distributions, see Fig. 2A). While

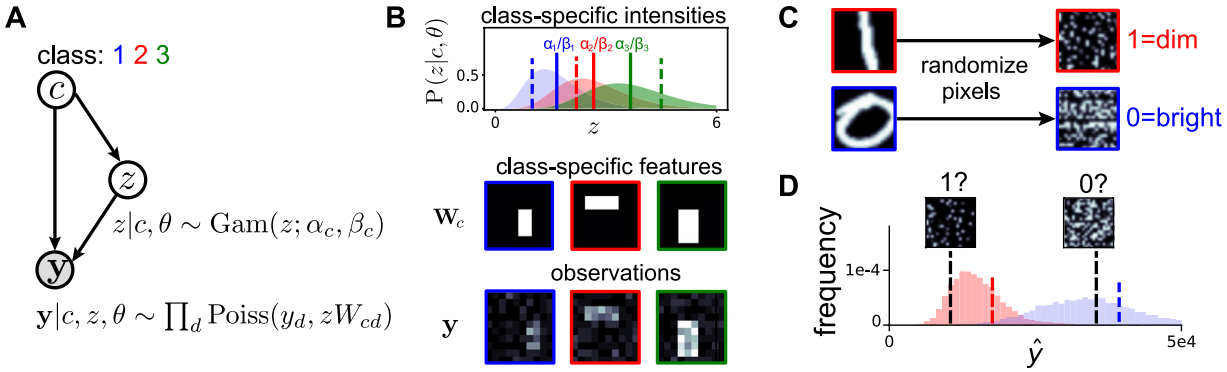

Figure 2. Modeling statistical dependencies between the class and intensity of a stimulus. (**A**) Generative model schematic: given a class c, the intensity z is drawn from a Gamma distribution, with class-specific parameters; the data point $\mathbf{y}$ is then generated by adding Poisson noise to the scaled features $z\mathbf{W}_c$. (**B**) A simple instantiation of the model for 3 classes of rectangles. Top: class-specific intensity distributions with means marked by solid lines. Middle: class-specific feature vectors. Bottom: 3 example data points, one for each class; corresponding intensities shown in top panel using dashed lines. (**C**) Differences between '1's and '0's in MNIST remain even after removing feature information by pixel shuffling. (**D**) We are presented with two new shuffled data points, overlaid on the class specific brightness distribution. Which class do the images belong to? Examples in C marked with colored dashed lines.

such a posterior usually requires approximate or numerical solutions, here it has a closed-form expression (see Supplementary Sec. S1):

$$\mathrm{P}(c,\, z|\mathbf{y},\, \theta) = \frac{(\prod_d W_{cd}^{y_d})\mathrm{NB}\left(\hat{y};\, \alpha_c,\, \frac{1}{\beta_c+1}\right)}{\sum_{c'}(\prod_d W_{c'd}^{y_d})\mathrm{NB}\left(\hat{y};\, \alpha_{c'},\, \frac{1}{\beta_{c'}+1}\right)} \cdot \mathrm{Gam}(z;\, \alpha_c + \hat{y},\, \beta_c + 1), \tag{1}$$

where NB denotes the negative binomial distribution and $\hat{y} = \sum_d y_d$.

Class judgements can be made by marginalizing over z the joint posterior above (see Supplementary Sec. S2):

$$\mathrm{P}(c|\mathbf{y},\, \theta) = \frac{\mathrm{NB}\left(\hat{y};\, \alpha_c,\, \frac{1}{\beta_c+1}\right)\exp(\sum_d y_d \ln W_{cd})}{\sum_{c'}\mathrm{NB}\left(\hat{y};\, \alpha_{c'},\, \frac{1}{\beta_{c'}+1}\right)\exp(\sum_{d'} y_{d'} \ln W_{c'd'})}. \tag{2}$$

Despite its apparent complexity, this posterior is a straightforward generalization of the standard softmax function, which can be implemented neurally using well-understood winner-take-all (WTA) circuit dynamics[8,9,13,18,28]. It optimally combines information about the input shape, implicit in $\mathbf{y}$, and its brightness $\hat{y}$. Moreover, if either cue is not instructive then the corresponding term cancels. If all classes have identical shape (i.e., $\mathbf{W}$ is the same across rows), then the posterior reduces to the negative binomial terms. Conversely, if all classes have the same intensity distribution, then we recover a traditional softmax consistent with previous work[8,11]. Similarly, intensity judgements are obtained by marginalizing the unknown class c (see Supplementary Sec. S2):

$$\langle z \rangle_{\mathrm{P}(z|\mathbf{y},\theta)} = \sum_c s_c \frac{\alpha_c + \hat{y}}{\beta_c + 1}, \tag{3}$$

where s_c denotes the posterior probability of class c (Eq. 2). The expressions for marginal posteriors $\mathrm{P}(c|\mathbf{y}, \theta)$ and $\langle z \rangle_{\mathrm{P}(z|\mathbf{y},\theta)}$ are relatively simple and local, suggesting that optimal inference in our model might be approximated by neural circuitry.

Inferring the class and intensity of stimuli in a neural circuit. The form of the posterior for c (Eq. 2) is reminiscent of well-documented soft-WTA neural dynamics[8,9,13,18,28,29]. The similarity to neural circuits further increases in the limit when NB can be approximated by a Poisson distribution (see Supplementary Sec. S3):

$$\mathrm{P}(c|\mathbf{y},\, \theta) \approx s_c = \frac{\exp(I_c)}{\sum_{c'}\exp(I_{c'})};\; I_c = \sum_d y_d \ln(W_{cd}\lambda_c) - \lambda_c,$$

where $\lambda_c = \alpha_c/\beta_c$.

The inference of class label c can be approximated by a simple feedforward and neural circuit-like architecture (Fig. 3A). A layer of neurons coding for different class values c receive inputs $\mathbf{y}$ via synapses $\mathbf{W}_c$ and interact laterally and competitively to implement the softmax function. The excitability of these neurons is determined by an intrinsic parameter λ_c that reflects average brightness $\hat{y}$ of its preferred inputs. Jointly, the responses of the class

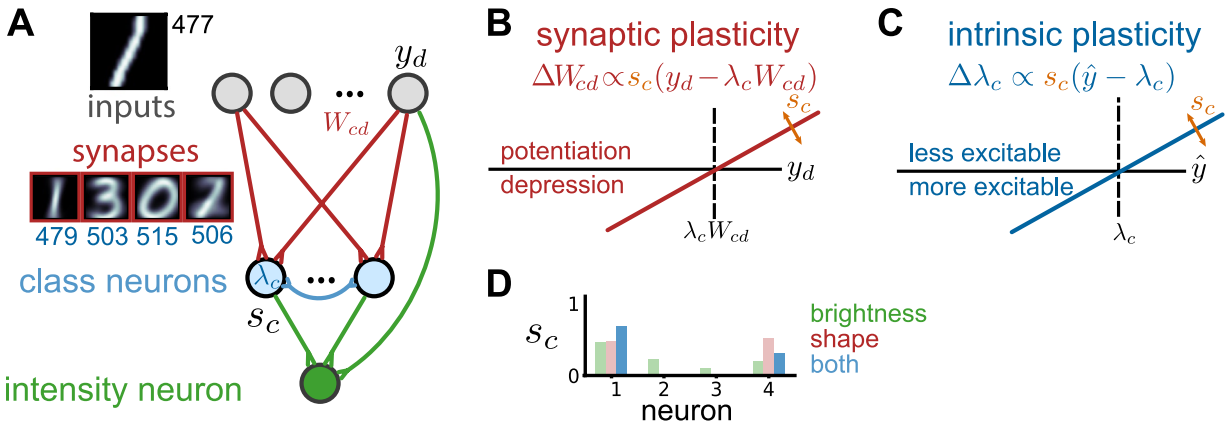

Figure 3. Neural circuit for inference and learning in face of class-specific intensity variations. (**A**) Circuit architecture: input neurons (gray) connect to the first processing layer (blue) via plastic synapses (red), with competition implemented by recurrent interactions. Neurons in this layer vary in their excitability λ, learned by IP (blue numbers). The activities s_c are combined across neurons and with the inputs to compute an estimate of the stimulus intensity. Insets show instantiation of the model after learning with 4 classes, trained on a subset of the MNIST dataset. (**B**) Optimal learning of model parameters via Hebbian and (**C**) IP (see text for details). (**D**) The outputs of the trained circuit for a novel '1' input, either intact (in blue), or disrupted so as to preserve only shape (red) or brightness (green).

neurons s_c encode not only the approximate Bayes-optimal guess for c, but also its associated uncertainty, in the form of the posterior probability[30].

One potentially unrealistic aspect of the neural dynamics is the logarithmic nonlinearity affecting synaptic efficacies (but see[18]). As already shown by Keck *et al.*[8], this issue can be eliminated by linearly approximating the logarithm. After this approximation, the expression for the input current to neuron c becomes $I_c = \sum_d W_{cd} y_d + \hat{y} \log \lambda_c - \lambda_c$ (see Supplementary Sec. S10). As before the primary effect of parameter λ_c is as a neuron-specific threshold. A more subtle effect involves a class-specific modulation of excitability that reflects the total input to the cell. Since the contribution of this term is typically small, we will focus primarily on the λ_c threshold regulation and its changes during learning.

The class posterior values s_c are combined linearly in a second layer to estimate the intensity of a stimulus z (Eq. 3), or quantities related to it such as the contrastive stress. We use expression $\mathcal{E} = \langle z \rangle_{P(z|\mathbf{y},\theta)} - \langle \lambda_c \rangle_{P(c|\mathbf{y},\theta)}$ as a sensible mathematical definition of contrastive stress: it starts from the original intuition of stress reflecting variations in intensity, and also accounts for the class-specific statistics of z. The posterior mean of the intensity s_z optimally combines direct stimulus information from the input layer with class judgements from the class layer to approximate $\mathcal{E}$ (see Fig. 3A and Supplementary Sec. S9):

$$\mathcal{E} \approx s_z = K\left(\hat{y} - \sum_c s_c \lambda_c\right), \tag{4}$$

where $K = 1/(\beta + 1)$ is a constant, and β approximates parameters β_c, assumed to be similar across classes. While we do not necessarily think of contrastive stress estimation as a computation explicitly performed in the cortex, it is interesting to note that the final expression of the intensity estimation can still be performed using simple local operations.

The circuit only makes correct inferences when its synapses and intrinsic parameters reflect the true statistics of inputs. Computationally, these parameters could be learned from data by algorithms such as expectation maximization (EM)[31,32]. Exploiting the mathematical tractability of our model, we derived online parameter update rules that combine to approximate EM in a biologically-plausible way (see Supplementary Secs S4 and S5). The weight updates translate into a form of Hebbian synaptic plasticity, adapting the first layer synapses to reflect the shape-specific information for the corresponding class (Fig. 3B). The λ_c updates implement a form of intrinsic plasticity (IP) which adapts the excitability of the cell to reflect the average intensity of stimuli in that class (Fig. 3C):

$$\Delta W_{cd} = \varepsilon_W s_c (y_d - \lambda_c W_{cd}); \quad \Delta \lambda_c = \varepsilon_\lambda s_c (\hat{y} - \lambda_c), \tag{5}$$

where ε_W and ε_λ are small learning rates. The derived plasticity rules have the same fixed points as optimal EM learning (see Supplementary Sec. S4). They can be intuitively understood as trying to bring the values predicted by the generative model with parameters $\mathbf{W}$ and $\boldsymbol{\lambda}$ closer to the input values. In the case of Hebbian plasticity, the individual inputs y_d are compared with their expected values $\lambda_c W_{cd}$ (Fig. 3B). If the input is larger than expected then the synapse is potentiated, bringing the prediction closer to y_d; if it is lower, synaptic depression occurs. The learning rule converges when the predictions are accurate (on average). Additionally, the magnitude of the synaptic changes is scaled by s_c; the more likely it is that the stimulus belongs to the class, the larger the contribution

of the current stimulus in updating the shape parameters. IP operates in a similar way, but for predictions about stimulus brightness, $\hat{y}$ (Fig. 3C): λ_c increases when the stimulus brightness is larger than expected and vice versa. The primary effect of this change on the neuron's transfer function is as a threshold shift, which acts in a negative feedback loop, as in traditional phenomenological descriptions of homeostatic forms of IP. What is unique to our model is the fact that, similar to Hebbian plasticity, the magnitude of the excitability shift depends on the current activity of the neuron s_c. Furthermore, the change in λ_c also induces a weaker positive feedback loop gain modulation of the total input to the neuron, as a non-homeostatic form of IP. Different experimental setups might preferentially reveal one or the other aspect of IP. Nonetheless, the fact that the change in excitability depends on the overall neural activation suggests possible routes for the experimental validation of the model (see Discussion).

One biologically implausible aspect of the above solution is that neural internal variables λ_c are needed when computing the activity in the second layer. While at first glance such non-locality seems to doom a neural implementation of optimal intensity judgements, the posterior can be approximated using a simple duplication of variables. The key idea is that a copy of λ_c is encoded in the synaptic connections V_c linking the first and second layers. These weights can be learned independently by Hebbian plasticity $\Delta V_c = \varepsilon_V s_c(e - V_c)$, with learning rate ε_V. Under the assumption that during learning the second layer is driven by feedforward inputs, i.e. $s_z = \hat{y}$, weights V_c will converge to λ_c. Hence, it is possible to approximate optimal inference and learning for the Gamma-Poisson generative model with simple nonlinearities and local operations.

The resulting neural circuit self-organizes to optimally combine shape and brightness cues to make inferences. When trained using digits 0–3 in the MNIST dataset, the synapses learn to represent individual digits (Fig. 3A, red shaded images) while learned parameters λ_c reflect their average intensities (Fig. 3A, numbers in blue). When a new input is presented, e.g. a tilted '1' digit with low brightness, the network correctly recognizes that it belongs to the class encoded by neuron 1 (Fig. 3D, blue bars). To highlight the fact that stimulus shape and brightness both matter for this judgement, we can disrupt the input so as to only preserve shape (Fig. 3D, red) or brightness (Fig. 3D, green) cues. Both manipulations negatively affect class identity judgements. Based on shape alone, the stimulus is equally likely to belong to the class represented by neurons 1 or 4. Hence it is the dimness of the stimulus that shifts the balance in favor of neuron 1 in the intact system. Similarly, brightness alone contains less information than shape; it cannot exclude any classes resulting in significantly higher class uncertainty.

Inference and learning for visual data: classifying handwritten digits. The proposed probabilistic model leverages class-specific brightness information to aid the classification of stimuli. Therefore it should outperform solutions that discard intensity information in a preprocessing step[8]. To test this hypothesis, we used digits 0–3 in MNIST and compared digit classification rates obtained using our circuit dynamics to those obtained by a circuit that requires normalized inputs (Fig. 4)[8] (see Methods 1, Supplementary Sec. S7). Both models learn a representation of the (possibly normalized) data in an unsupervised manner, without access to class labels. To assess the quality of the emerging representations in terms of distinguishing different digits, we take the approach used in[8] and train a Bayesian classifier with the class neuron responses s_c as input (see Supplementary Sec. S6). In this way, the classification performance jointly assesses the effects of inference and learning.

The MNIST dataset that we have chosen has limited variations in intensity across classes. Although the digit '1' is well-separated in brightness space, digits '2' and '3' are virtually indistinguishable (Fig. 4A). These statistics are reflected in high confidence '1' judgements when $\hat{y}$ is small, but higher uncertainty for intermediate and high values of $\hat{y}$ (Fig. 4B). Together with the fact that we are considering a relatively easy task (4 digit classes), we cannot expect the boost in performance to be large. Nonetheless, we do see significant performance benefits even in these conditions. Our model not only correctly learns prototypical digit shapes (Fig. 4C) and their corresponding average brightness values (Fig. 4D) but it also makes better digit class judgements (Fig. 4E).

The benefits of accounting for class-specific intensity differences are substantial when the number of classes is small ($C = 4$) but they decrease as the representation becomes more overcomplete ($C = 25$). Since classification performance is very high overall we might expect that limited improvements are due to reaching ceiling performance. To exclude this possibility, we repeated the experiment on the full MNIST dataset and found very similar results (see Supplementary Sec. 8). Our circuit outperforms a solution using input normalization in the complete scenario (solid bars in Fig. 4E, $C = 10$) but there are no differences when the representation is overcomplete ($C = 100$). This result suggests that, as the latent space increases, both circuits gain flexibility in encoding the statistics of the input, and brightness becomes less informative about class. It also suggests that our circuit is particularly efficient at capturing the true input statistics when neural resources are scarce.

We expect that explicitly taking into account class-specific intensity variations would prove particularly computationally advantageous when the correlation between class and brightness is high. To investigate the full potential of intensity-dependent inference and learning, we would need a dataset in which these dependencies are particularly strong. Unfortunately, traditional image datasets aim to minimize such differences by construction. To circumvent this issue, we decided to manipulate the intensity statistics of the standard MNIST dataset to artificially increase the dependence between brightness and stimulus class (by brightening the image foreground in a digit-specific way, see Fig. 4F and Methods 1). We further increase the task difficulty by considering all 10 digits. As before, the model learns accurate representations of the modified stimuli with only $C = 20$ classes (see Supplementary Sec. 8). The classification performance of individual digits is generally high, with the IP circuit generally outperforming classifiers based on single cues, either brightness or shape alone. The exception are the digits '4', '7' and '9' where the network converges to inaccurate average brightness estimates. The poor learning for these classes is possibly due to their strong shape overlap, which could be resolved by increasing the number of class neurons. Nonetheless, after combining performance across all digits (Fig. 4G right) the intact circuit

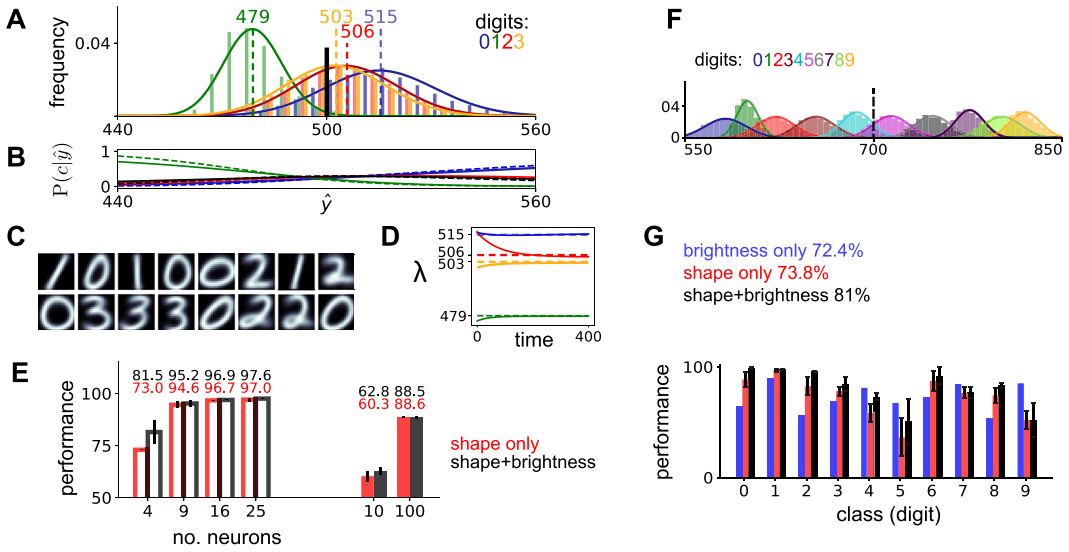

Figure 4. Brightness information improves handwritten digit classification. (**A**) Brightness histrograms for digits 0–3 from MNIST and corresponding gamma fits in solid lines; dashed vertical lines denote the mean brightness for each digit; vertical solid line marks the normalized brightness for the intensity-agnostic alternative neural circuit. (**B**) Posterior distribution of digit class using only brightness as a cue; solid lines are exact value, dashed are neural approximations (see Supplementary Sec. 3). (**C**) Learned weights **W** for $C = 16$ classes. (**D**) Solid lines show the evolution of estimated parameters (averaged across neurons tuned to the same digit); corresponding optimal values in dashed lines. Time measured as number of iterations through the training dataset. (**E**) Comparison of digit classification performance for circuits that ignore (red) or optimally take into account (black) class-specific variations in intensity, estimated using a Bayesian classifier using responses s_c as inputs. Chance performance is 25% for first 4 experiments (digits '0'–'3'), and 10% for filled bars (digits '0'–'9'). Error bars show s.d., estimated across 10 runs. (**F**) Variation of full handwritten digits in which brightness is artificially modulated in a class-specific way and (**G**) corresponding classification performance.

massively outperforms the alternatives (by 7%). These results confirm that, if the statistics of the inputs exhibit strong class-intensity dependencies, then it is computationally advantageous to explicitly account for them.

Inference and learning for auditory data: estimating contrastive stress. While so far we have focused on class judgements, our inference procedure can also make intensity judgements for individual stimuli. While this computation may not be of interest for handwritten digits, it is of critical importance in the auditory domain, e.g. when estimating the contrastive stress of speech[1–4]. Hence, we use auditory data to investigate the utility of our model in making stress judgements. Figure 5A illustrates a particular version of this problem, based on the OLLO logatome database[5] (see Methods 2). A speaker produces a sentence received by a listener. The speaker's vocabulary comprises four logatome classes: 'bup', 'pap', 'zos', and 'ulu'. The sentence includes 10 utterances of these logatomes, with varying levels of emphasis. The listener's goal is to classify the logatomes in the sentence and to estimate their intensity, i.e. to estimate the contrastive stress of the sentence (Fig. 5B).

To accomplish that goal, we trained our network on inputs given as log-mel spectrograms of the individual logatomes (Fig. 5C, see Methods 2 for preprocessing details), reflecting patterns of neural activity as received by the primary auditory cortex[33]. The network had $C = 4$ class neurons. It was trained through a combination of synaptic plasticity and IP, as derived above. At the end of training, individual class neurons were each tuned to individual logatome classes, and weights $\mathbf{W}_c$ resembled general templates for each class of logatome (Fig. 5D; from the upper left box, going clockwise 'ulu', 'pap', 'bup', and 'zos'). We then tested the network's performance on contrastive stress estimation, as well as on classification, using a test sentence constructed from left-out data.

We compared the contrastive stress estimate of our model (Eq. 4, estimates marked 'IP'/red in Fig. 5E), to several alternative estimators (see Supplementary Sec. S9). As lower bounds on performance we considered two *ad hoc* solutions including the naive estimate $\mathcal{E}^N = \hat{y} - \sum_n \hat{y}^{(n)}/N$ discussed in the introduction ('N',/black), and an improved version, which approximates the expected intensity of a logatome as its brightness, i.e. $\langle z \rangle_{P(z|\mathbf{y},\theta)} \approx \hat{y}$, and compares it to the average class intensity $\langle \lambda_c \rangle_{P(c|\mathbf{y},\theta)}$ ('EN'/green). This estimator has access to the ground-truth average brightnesses of individual logatome classes (i.e. all λ_c), but deals sub-optimally with the map between intensity and brightness. Lastly, we used the exact optimal Bayesian estimate, $\mathcal{E}^B$, for an upper bound on performance ('B'/blue). We found that our model exhibits good performance on classification (the x-axis in Fig. 5E), despite the small size of the first layer. It correctly classified all logatomes except the third, where it mistook a weak 'ulu' for a 'bup'. When estimating contrastive stress, the neural circuit implementation $\mathcal{E}^{IP}$ is consistently close to the optimal Bayesian estimate $\mathcal{E}^B$, being the least accurate only for the third logotome, which was misclassified. This suggests a relatively limited detrimental effect due to locality constraints and other neural approximations. As expected, the naive estimator $\mathcal{E}^N$ differs wildly from $\mathcal{E}^B$, particularly for the first, fourth, ninth, and tenth loga-

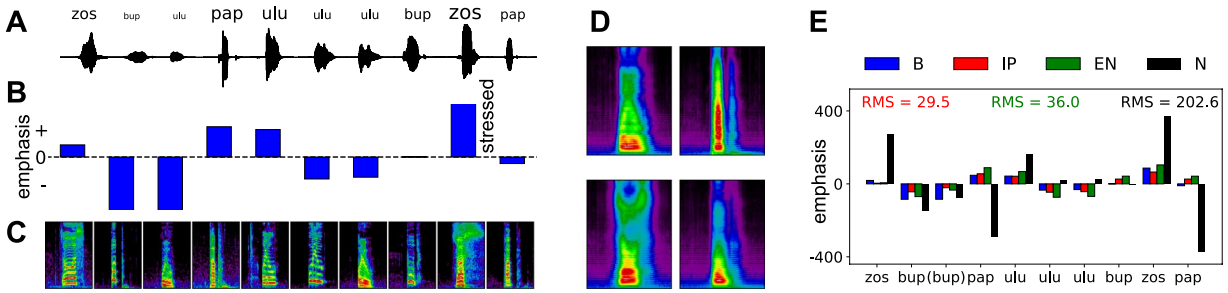

Figure 5. Estimating the contrastive stress of spoken logatomes and classification results. (**A**) Example data, which consists of a waveform sequence combining several logatomes, with different levels of stress, marked by the size of text above. (**B**) The Bayes-optimal stress estimate $\mathcal{E}^B$ for the sentence in panel A. (**C**) Log-mel spectrograms of the waveforms in panel A, provided as inputs to the neural circuit. (**D**) Learned synaptic weights. (**E**) Four estimates of the stress of the test sentence in (**C**) $\mathcal{E}^B$, $\mathcal{E}^{IP}$, $\mathcal{E}^N$, and $\mathcal{E}^{EN}$. The colored text represents the root-mean-square distance of the latter three estimates from the Bayes-optimal $\mathcal{E}^B$. X-axis shows the classification output for each of the logatomes in the sentence, with incorrect classifications in parentheses. These are typical results for this experiment.

tomes. The more sophisticated *ad hoc* estimator $\mathcal{E}^{EN}$ performs closer to the optimum than $\mathcal{E}^N$, but the neural implementation $\mathcal{E}^{IP}$ was found to be the closest overall, as measured by the root-mean-square (RMS) distance between the stress estimate and the Bayes-optimal estimate $\mathcal{E}^B$. This result reflects the fact that the brightness $\hat{y}$ and the expected intensity $\langle z \rangle_{P(z|\mathbf{y},\theta)}$ are different quantities. While the estimates $\mathcal{E}^{IP}$ and $\mathcal{E}^{EN}$ are accurate approximations to the optimal estimate, the latter requires label information in order to calculate the average brightnesses λ_c of the dataset. The circuit also produces an accurate estimate, but learns the λ_c unsupervised (i.e. without labels).

Discussion

Sensory perception needs to reconcile seemingly conflicting goals. On the one hand, salient features in the input should be identified regardless of stimulus intensity[34,35]. On the other hand, brightness information should be preserved for intensity judgements. While neural correlates of the first task have received considerable attention[7,12], how neural circuits estimate stimulus intensity remains unclear. Here, we have argued that there are systematic dependencies between the class and intensity of sensory stimuli. We have derived a set of local inference and learning rules that can exploit class-intensity dependencies to approximate optimal judgements about such stimuli, and illustrated the benefits of the model on examples in the auditory and visual domains. We have further argued that this solution can be well approximated in a plausible neural circuit, where the interaction between specific forms of synaptic and intrinsic plasticity implement approximately optimal learning. It might be possible to derive other neural circuits for optimal learning of class and intensity inference. However, the interplay between synaptic plasticity and IP derived here (A) naturally emerges for such tasks, and (B) represents sensory statistics (shape and brightness) with sensible neural correlates (synaptic weights and neural excitabilities).

Although well-documented experimentally, the diversity of plasticity mechanisms co-active in cortical circuits remains a theoretical puzzle. This lack of functional understanding is particularly obvious when it comes to activity-dependent changes in neural excitability[36], with the role of IP during learning still under debate[17,18,37–40]. In particular, the interaction between IP and synaptic plasticity has so far eluded analytical treatment. The traditional role of IP is viewed as a simple negative feedback loop that regulates average firing rates to keep overall activity close to a target value. Furthermore, recent work has suggested that sparsity-enforcing forms of IP may guide synaptic learning towards efficient neural representations[17], and normative interpretations of (non-) homeostatic forms of IP have been proposed in the context of optimal memory recall[39]. Here we have expanded on these results by deriving the Bayes-optimal interaction between IP and synaptic plasticity as a key component of sensory learning.

Our IP rule differs from past proposals in that the target of regulation is average input currents rather than output firing rates. Furthermore, changes in excitability are both positive- and negative-feedback and gated by output activity. While different experimental manipulations reveal either homeostatic or non-homeostatic forms of IP[41], directly validating these predictions is challenging. Independently manipulating the input and output statistics of the neuron is impossible in traditional IP experiments, which globally interfere with both inputs and outputs, typically by bath application of a drug[19,20]. Nonetheless, techniques that locally manipulate pre- and post-synaptic activity will open new avenues for testing our model experimentally. One very particular aspect of the model is that the excitability of neurons varies across the population, reflecting input statistics. In particular, neural excitability represents the mean intensity of the stimulus class that the neuron represents. This is unlike past IP models which assume that the target firing rate for IP is the same across neurons, set by energetic efficiency constraints[17,37,38]. Our model seems better aligned with biological findings, which suggest that neurons do not share a target firing rate but that cortical circuits have broad average firing rate distributions. Our approach could thus provide a computational account for systematic differences in excitability across neurons seen experimentally[42].

From a computational perspective, our generative model provides a principled treatment of class-specific intensity variations, a ubiquitous feature of realistic datasets. With very limited computational overhead, the use of intensity as an additional cue improves the unsupervised learning of input statistics. The learned intensity statistics facilitate subsequent semi-supervised learning and classification, e.g. of MNIST digits or OLLO logatomes. Moreover, intensity itself can be estimated, and we illustrated one potential application of intensity estimation on acoustic data. The explicit use of syllable stress labels in modern automatic speech recognition (ASR) dictionaries may be taken as evidence for the importance of acoustic stress and intensity information in general[43,44]. For example, if a set of words are related but exhibit different stress patterns (e.g. 'photograph', 'photographer', 'photographic'), then the consideration of syllable stress can improve recognition. However, ASR systems do not autonomously learn to infer stress estimates from data, whereas our approach specifies an unsupervised and statistically-grounded method to do so[45]].

Methods

M1. MNIST experiments. Figure 4 compares the classification performances on two different MNIST datasets which either use normalized inputs (Fig. 4A–E) or which incorporate class-specific brightness information (Fig. 4F,G). The difference between the inference procedures was that one required normalized inputs[8] and classifies based on shape alone, while the other could learn class-specific brightnesses. Here we present the details for data preprocessing and the basic procedures used for training and testing, with corresponding pseudocode in Supplementary Sec. S7.

MNIST details: no class-dependent brightening. The shape-only inference requires inputs $\mathbf{y}^{\mathrm{SA}}$ to be normalized to a constant A, with individual input elements greater than one[8] (Fig. 4A–E):

$$y_d^{\mathrm{SA}} = (A - D)\tilde{y}_d^{\mathrm{SA}} / \sum_{d'=1}^{D} \tilde{y}_{d'}^{\mathrm{SA}} + 1, \tag{6}$$

where D is the dimensionality of $\mathbf{y}^{\mathrm{SA}}$, and $\tilde{\mathbf{y}}^{\mathrm{SA}}$ is the raw data; for the version of the dataset without class-dependent brightening (Fig. 4A–E) we use $D = 400$ (20×20 pixel images) and $A = 500$.

Our model receives input data $\mathbf{y}^{\mathrm{IP}}$ that does not need to be normalized. To facilitate a fair comparison of the networks, we analogously preprocessed the MNIST dataset while preserving the class-specific brightness information in it. Specifically, we computed the brightness of raw MNIST data with respect to the average brightness of all data points: $f = \widehat{\tilde{y}} / \langle \widehat{\tilde{y}} \rangle_{\mathrm{P}(\widehat{\tilde{y}})}$. We then normalized each data point as we did for the shape-only circuit (Eq. 6, with $A = 450$), but amplified or dimmed the foreground of the image, depending on its original brightness, $\mathbf{y}^{\mathrm{IP}} = (\mathbf{y}^{\mathrm{SA}} - 1) f + 1$. The resulting dataset $\mathbf{y}^{\mathrm{IP}}$ has an average brightness of 500 (see Fig. 4A) and all pixels are greater than 1. However, it preserves the brightness information from the raw MNIST dataset. We initialized the weights of both circuits $\mathbf{W}_{\mathrm{IP}}^{\mathrm{init}}$ and $\mathbf{W}_{\mathrm{SA}}^{\mathrm{init}}$ to randomly chosen preprocessed data points from the training set. Learning rates were $\varepsilon_W = 1 \times 10^{-5}$ and $\varepsilon_\lambda = 1 \times 10^{-4}$ for the version using unnormalized inputs and $\varepsilon = 1 \times 10^{-3}$ when learning based on shape alone[8].

MNIST details: brightness-enhanced. We normalized $\mathbf{y}^{\mathrm{SA}}$ as before (Eq. 6), but with $A = 700$. Our input to the IP network $\mathbf{y}^{\mathrm{IP}}$, however, had its brightness enhanced depending on its class. We defined a vector $\nu(l)$ that takes the label of a data point l as input $\nu(l) = [2.3, 3.4, 3.3, 4.0, 4.8, 5.3, 5.9, 6.7, 6.9, 7.5]$. We then normalized MNIST (Eq. 6, with $A = 450$) and amplified the image foregrounds depending on their original brightness and their class (i.e. their label), $\mathbf{y}^{\mathrm{IP}} = (\mathbf{y}^{\mathrm{SA}} - 1)(f + \nu(l) + 1)$. The final dataset $\mathbf{y}^{\mathrm{IP}}$ had an average brightness of 700 (see Fig. 4F) and all pixels greater than 1 (Fig. 4F,G).

We initialized $\mathbf{W}_{\mathrm{IP}}^{\mathrm{init}}$ by taking the average of all data points and adding Poisson noise, $\mathbf{W}_{\mathrm{IP}}^{\mathrm{init}} = \sum_n \mathbf{y}^{(n)}/N + 0.1X; X \sim \mathrm{Poiss}(X; 1)$, where N is the number of data points in the training set. Also, we initialized λ_c by drawing uniform random numbers between 550 and 850. These results indicate that the IP circuit can learn parameters of the dataset for a variety of initialization conditions.

For our proposed model we set the learning rates to $\varepsilon_W = 1 \times 10^{-6}$ and $\varepsilon_\lambda = 1 \times 10^{-5}$. The learning results remained qualitatively similar with changes in these values (see Fig. S5C,D), indicating that the model performance does not depend strongly on these parameters. One constraint is that one should make sure to prevent negative parameter values during training. For the shape-only circuit[8], we set the learning rate to $\varepsilon = 1 \times 10^{-3}$, as before.

Parameter learning in the Gamma-Poisson model. Given a data point $\mathbf{y}^{(n)}$, we calculate the posterior over classes s_c and update the weights W_{cd} and intrinsic parameters λ_c by evaluating the relevant equations in the main text. We iteratively compute these quantities for every data point in the training set, and for some number of iterations over the training set. The behavior of these learning rules is qualitatively illustrated in Fig. 3B,C. Supplementary Sec. S7 presents formal pseudocode for this training process.

Testing the learned representation. The training of the models was done in an unsupervised fashion. However, to evaluate the quality of the learned representations we used the first layer outputs of the circuit $\mathbf{s}_{1:C}$ as inputs to a Bayesian classifier[8] (see Supplementary Sec. S6). We ran these experiments in a semi-supervised setting where the representation was learned using inputs alone, and only a small fraction of the associated class labels ($L = 30$ examples for both MNIST and OLLO; for reference this is approximately 0.5% of the total number of labels in MNIST) were used to train the classifier. Performance was assessed using labeled test data.

M2. OLLO experiment. The Oldenburg Logatome Corpus (OLLO)[5] is a freely-available online database of acoustic data. It comprises logatomes that are nonsensical combinations of consonants (C) and vowels (V). There are 150 different combinations of CVCs and VCVs. Each is spoken by 50 different speakers: 25 men and 25 women; 40 from different regions of Germany and 10 from France. Each speaker produces each logatome in six different versions: 'normal', 'fast', 'slow', 'loud', 'soft', and 'question'. The dataset we used for our experiments is a subset comprising logatomes 'ulu', 'pap', 'zos', 'bup', spoken in the 'normal' version, and only from the 40 German speakers. We used the Python package librosa (visit https://librosa.github.io/) to compute log-mel spectrograms of the audio files, using a sample rate of 16000 Hz, a hop length of 160, and setting the number of audio samples between successive onset measurements to 400. The spectrograms used 128 channels on the frequency axis.

Since our probabilistic model requires inputs $\mathbf{y}^{(n)}$ to have constant dimensionality, we trimmed spectrograms so that they had equivalent temporal durations. In particular, we used the 20 time bins containing the highest energy across all frequencies. We then calculated the time axis center of mass (COM) of those 20 columns and trimmed the spectrograms at 50 columns to either side of the COM. If we could not trim the spectrograms in this manner, i.e. if a logatome was pronounced very close to the beginning or end of a recording, then we discarded the data point. If the trimmed spectrograms contained less than 65% of the energy that was in the original spectrogram, then we also discarded the data point. Finally, we shifted and scaled the data $\mathbf{y}^{(n)}$ such that we can accurately approximate negative binomial distributions as Poisson (see Supplementary Sec. S3). This preprocessing procedure resulted in 310 valid log-mel spectrograms that we used for training (232 datapoints) and testing (78 datapoints). To construct the sentence shown in Fig. 5C, we chose ten random spectrograms from the testing set, with at least two examples of each logatome class. We initialized weights $\mathbf{W}$ by choosing four preprocessed data points and set the intensity parameters to their corresponding brightnesses. Learning rates were $\varepsilon_W = 1 \times 10^{-6}$ and $\varepsilon_\lambda = 1 \times 10^{-2}$.

References

1. Boer, S. E. Meaning and contrastive stress. *Philos. Rev.* **88**(2), 263–298 (1979).
2. Fry, D. B. Duration and intensity as physical correlates of linguistic stress. *J. Acoust. Soc. Am.* **27**, 765–768 (1955).
3. Chrabaszcz, A., Winn, M., Lin, C. Y. & Idsardi, W. J. Acoustic cues to perception of word stress by English, Mandarin, and Russian speakers. *J. Speech Lang. Hear. Res.* **57**, 1468–1479 (2014).
4. Mattys, S. L. The perception of primary and secondary stress in English. *Percept. Psychophys.* **62**(2), 253–265 (2000).
5. Wesker, T. *et al*. Oldenburg logatome speech corpus (OLLO) for speech recognition experiments with humans and machines. In *Proc. of Interspeech* Lisboa, Portugal 1273–1276 (2005).
6. Andrews, D. F. & Mallows, C. L. Scale mixtures of normal distributions. *J. R. Stat. Soc. Series B Methodol.* **36**(1), 99–102 (1974).
7. Wainwright, M. J. & Simoncelli, E. P. Scale mixtures of gaussians and the statistics of natural images. *Adv. Neural Inf. Process. Syst.* **12**, 855–861 (2000).
8. Keck, C., Savin, C. & Lücke, J. Feedforward inhibition and synaptic scaling–two sides of the same coin? *Plos Comp. Biol.* **8**(3), e1002432 (2012).
9. Nessler, B., Pfeiffer, M. & Maass, W. STDP enables spiking neurons to detect hidden causes of their inputs. In *Adv. Neural Inf. Process. Syst.*, 1357–1365 (2009).
10. Schwartz, O. & Simoncelli, E. P. Natural sound statistics and divisive normalization in the auditory system. *Adv. Neural Inf. Process. Syst.*, 166–172 (2000).
11. Holca-Lamarre, R., Lücke, J. & Obermayer, K. Models of acetylcholine and dopamine signals differentially improve neural representations. *Front. Comput. Neurosci.* **11** (2017).
12. Rozell, C. J., Johnson, D. H., Baraniuk, R. G. & Olshausen, B. A. Sparse coding via thresholding and local competition in neural circuits. *Neural Comput.* **20**(10), 2526–2563 (2008).
13. Lücke, J. & Sahani, M. Maximal causes for non-linear component extraction. *J. Mach. Learn. Res.* **9**, 1227–1267 (2008).
14. Lücke, J. Receptive field self-organization in a model of the fine-structure in V1 cortical columns. *Neural Comput.* **21**(10), 2805–2845 (2009).
15. Schmuker, M., Pfeil, T. & Nawrot, M. P. A neuromorphic network for generic multivariate data classification. *Proc. Natl. Acad. Sci.* **111**(6), 2081–2086 (2014).
16. Zylberberg, J., Murphy, J. T. & Deweese, M. R. A sparse coding model with synaptically local plasticity and spiking neurons can account for the diverse shapes of V1 simple cell receptive fields. *Plos Comp. Biol.* **7**(10), e1002250 (2011).
17. Savin, C., Joshi, P. & Triesch, J. Independent component analysis in spiking neurons. *Plos Comp. Biol.* **6**(4), e1000757 (2010).
18. Nessler, B., Pfeiffer, M., Buesing, L. & Maass, W. Bayesian computation emerges in generic cortical microcircuits through spike-timing-dependent plasticity. *Plos Comp. Biol.* **9**(4), e1003037 (2013).
19. Daoudal, G. & Debanne, D. Long-term plasticity of intrinsic excitability: learning rules and mechanisms. *Learn. Memory* **10**(6), 456–465 (2003).
20. Cudmore, R. H. & Turrigiano, G. G. Long-term potentiation of intrinsic excitability in lV visual cortical neurons. *J. Neurophysiol.* **92**(1), 341–348 (2004).
21. Kersten, D., Mamassian, P. & Yuille, A. Object perception as bayesian inference. *Annu. Rev. Psychol.* **55**, 271–304 (2004).
22. Dayan, P. & Abbott, L. F. *Theoretical Neuroscience* (MIT Press, Cambridge 2001).
23. Körding, K. P. *et al*. Causal inference in multisensory perception. *Plos One* **2**(9), https://doi.org/10.1371/journal.pone.0000943 (2007).
24. Turner, R. E. & Sahani, M. Demodulation as probabilistic inference. *IEEE/ACM Trans. Audio, Speech, Language Process.* **19**(8), 2398–2411 (2011).
25. Beck, J., Pouget, A. & Heller, K. A. Complex inference in neural circuits with probabilistic population codes and topic models. *Adv. Neural Inf. Process. Syst* (2012).
26. Raginsky, M., Willett, R. M., Harmany, Z. T. & Marcia, R. F. Compressed sensing performance bounds under Poisson noise. *IEEE Trans. Sig. Process.* **58**(8), 3990–4002 (2010).
27. Wilt, B. A., Fitzgerald, J. E. & Schnitzer, M. J. Photon shot noise limits on optical detection of neuronal spikes and estimation of spike timing. *Biophys. J.* **104**(1), 51–62 (2013).
28. Douglas, R. J. & Martin, K. A. C. Neuronal circuits of the neocortex. *Annu. Rev. Neurosci.* **27**, 419–451 (2004).
29. Rezende, D. J., Wierstra, D. & Gerstner, W. Variational learning for recurrent spiking networks. In *Adv. Neural Inf. Process. Syst.*, 136–144 (2011).
30. Knill, D. C. & Pouget, A. The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci.* **27**(12), 712–719 (2004).
31. Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Stat. Soc. Series B* **39**, 1–38 (1977).

32. Neal, R. & Hinton, G. A view of the EM algorithm that justifies incremental, sparse, and other variants. *Learning in Graphical Models* (ed. Jordan, M. I.), 355–368 (Kluwer, 1998).
33. Mesgarani, N., David, S. V., Fritz, J. B. & Shamma, S. A. Phoneme representation and classification in primary auditory cortex. *J. Acoust. Soc. Am.* **123**(2), 899–909 (2008).
34. Ohzawa, E., Sclar, G. & Freeman, R. D. Contrast gain control in the cat visual cortex. *Nature* **298**, 266–268 (1982).
35. Rabinowitz, N. C., Willmore, B. D., Schnupp, J. W. & King, A. J. Contrast gain control in auditory cortex. *Neuron* **70**(6), 1178–1191 (2011).
36. Turrigiano, G. Homeostatic signaling:the positive side of negative feedback. *Curr. Opin. Neurobiol.* **17**, 318–324 (2007).
37. Triesch, J. Synergies between intrinsic and synaptic plasticity in individual model neurons. In *Adv. Neural Inf. Process. Syst.*, 1417–1424 (2005).
38. Stemmler, M. & Koch, C. How voltage-dependent conductances can adapt to maximize the information encoded by neuronal firing rate. *Nat. Neurosci.* **2**, 521–527 (1999).
39. Savin, C., Dayan, P. & Lengyel, M. Optimal recall from bounded metaplastic synapses: predicting functional adaptations in hippocampal area CA3. *Plos Comp. Biol.* **10**(2), e1003489 (2014).
40. Titley, H. K., Brunel, N. & Hansel, C. Toward a neurocentric view of learning. *Neuron* **95**(1), 19–32 (2017).
41. Turrigiano, G. Too Many Cooks? Intrinsic and Synaptic Homeostatic Mechanisms in Cortical Circuit Refinement. *Annu. Rev. Neurosci.* **34**(1), 89–103 (2011).
42. Buzsáki, G. & Mizuseki, K. The log-dynamic brain: how skewed distributions affect network operations. *Nat. Rev. Neurosci.* **15**(4), 264–278 (2014).
43. Parihar, N., Picone, J., Pearce, D. & Hirsch, H. Performance analysis of the Aurora large vocabulary baseline system. In *Proc. of Eurospeech*, 10–13 (2003).
44. Povey, D. *et al*. The kaldi speech recognition toolkit. *IEEE* 2011 *Work. Autom. Speech Recognit. and Underst* (IEEE Signal Processing Society, 2011).
45. Monk, T., Savin, C. & Lücke, J. Neurons equipped with intrinsic plasticity learn stimulus intensity statistics. In *Adv. Neural Inf. Process. Syst.* 4278–4286 (2016).

Acknowledgements

We acknowledge funding by the DFG within the Cluster of Excellence EXC 1077/1 (Hearing4all) and grant LU 1196/5-1 (J.L. and T.M.) and the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP7/2007–2013) under REA grant agreement No. 291734 (C.S.). We also thank Bernd Meyer for useful discussions.

Author Contributions

Designed research: T.M., J.L. with contributions from C.S. Mathematical derivations: T.M., J.L., contributions from C.S. Numerical experiments: T.M. with contributions from C.S. Results interpretation: J.L., C.S., T.M. Wrote paper: T.M., C.S., J.L.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-28184-5.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.