Published in: *Advances in Neural Information Processing Systems 27* (2014), 1880–1888.

Optimal prior-dependent neural population codes under shared input noise

Agnieszka Grabska-Barwińska
Gatsby Computational Neuroscience Unit
University College London
agnieszka@gatsby.ucl.ac.uk

Jonathan W. Pillow
Princeton Neuroscience Institute
Department of Psychology
Princeton University
pillow@princeton.edu

Abstract

The brain uses population codes to form distributed, noise-tolerant representations of sensory and motor variables. Recent work has examined the theoretical optimality of such codes in order to gain insight into the principles governing population codes found in the brain. However, the majority of the population coding literature considers either conditionally independent neurons or neurons with noise governed by a stimulus-independent covariance matrix. Here we analyze population coding under a simple alternative model in which latent "input noise" corrupts the stimulus before it is encoded by the population. This provides a convenient and tractable description for irreducible uncertainty that cannot be overcome by adding neurons, and induces stimulus-dependent correlations that mimic certain aspects of the correlations observed in real populations. We examine prior-dependent, Bayesian optimal coding in such populations using exact analyses of cases in which the posterior is approximately Gaussian. These analyses extend previous results on independent Poisson population codes and yield an analytic expression for squared loss and a tight upper bound for mutual information. We show that, for homogeneous populations that tile the input domain, optimal tuning curve width depends on the prior, the loss function, the resource constraint, and the amount of input noise. This framework provides a practical testbed for examining issues of optimality, tuning width, noise, and correlations in realistic neural populations.

1 Introduction

A substantial body of work has examined the optimality of neural population codes [1–18]. However, the classical literature has focused predominantly on codes with independent Poisson neurons and on analyses of unbiased decoders using Fisher information. Real neurons, by contrast, exhibit noise correlations (dependencies not introduced by the stimulus), and Fisher information does not accurately quantify information when performance is biased or close to threshold [6, 14, 17]. Moreover, the classical population codes with independent Poisson noise predict unreasonably good performance with even a small number of neurons. A variety of studies have shown cases where a small number of independent neurons can outperform an entire animal [19, 20]. For example, a population of only 220 Poisson neurons with tuning width of 60 deg (full width at half height) and tuning amplitude of 10 spikes can match the human orientation discrimination threshold of ≈ 1 deg. (See Supplement S1 for derivation.) Even fewer neurons would be required if tuning curve amplitude were higher.

The mismatch between this predicted efficiency and animals' actual behaviour has been attributed to the presence of information-limiting correlations between neurons [21, 22]. However, deviation

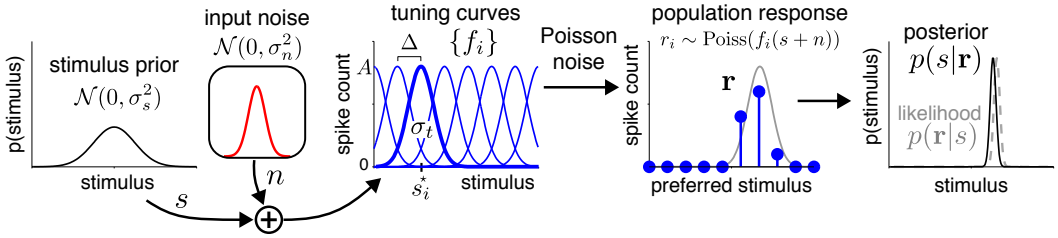

Figure 1: Bayesian formulation of neural population coding with input noise.

from independence renders most analytical treatments infeasible, necessitating the use of numerical methods (Monte Carlo simulations) for quantifying the performance of such codes [6, 14].

Here we examine a family of population codes for which the posterior is approximately Gaussian, which makes it feasible to perform a variety of analytical treatments. In particular, we consider a population with Gaussian tuning curves that "tile" the input domain, and Gaussian stimulus priors. This yields a Gaussian-shaped likelihood and a Gaussian posterior with variance that depends only on the total spike count [2, 15]. We use this formulation to derive tractable expressions for neurometric functions such as mean squared error (MSE) and mutual information (MI), and to analyze optimality without resorting to Fisher information, which can be inaccurate for short time windows or small spike counts [6, 14, 17]. Secondly, we extend this framework to incorporate shared "input noise" in the stimulus variable of interest (See Fig. 1). This form of noise differs from many existing models, which assume noise to arise from shared connectivity, but with no direct relationship to the stimulus coding [5, 14, 17, 23]. (See [15, 24] for related approaches).

This paper is organised as follows. In Sec. 2, we describe an idealized Poisson population code with tractable posteriors, and review classical results based on Fisher Information. In Sec. 3, we provide a Bayesian treatment of these codes, deriving expressions for mean squared error (MSE) and mutual information (MI). In Sec. 4, we extend these analyses to a population with input noise. Finally, in Sec. 5, we examine the patterns of pairwise dependencies introduced by input noise.

2 Independent Poisson population codes

Consider an idealized population of Poisson neurons that encode a scalar stimulus s with Gaussian-shaped tuning curves. Under this (classical) model, the response vector $\boldsymbol{r} = (r_1, \dots r_N)^\top$ is conditionally Poisson distributed:

$$r_i|s \sim \text{Poiss}(f_i(s)), \qquad p(\boldsymbol{r}|s) = \prod_{i=1}^{N} \tfrac{1}{r_i!} f_i(s)^{r_i} e^{-f_i(s)}, \qquad \textit{(Poisson encoding)} \quad (1)$$

where tuning curves $f_i(s)$ take the form

$$f_i(s) = \tau A \exp\left(-\tfrac{1}{2\sigma_t^2}(s - \overset{\star}{s}_i)^2\right), \qquad \textit{(tuning curves)} \quad (2)$$

with equally-spaced centers or "preferred stimuli" $\overset{\star}{\boldsymbol{s}} = (\overset{\star}{s}_1, \dots \overset{\star}{s}_N)$, tuning width σ_t, amplitude A, and time window for counting spikes τ. We assume that the tuning curves "tile", i.e., sum to a constant over the relevant stimulus range:

$$\sum_{i=1}^{N} f_i(s) \approx \lambda. \qquad \textit{(tiling property)} \quad (3)$$

We can determine λ by integrating the summed tuning curves (eq. 3) over the stimulus space, giving $\int ds \sum_{i=1}^{N} f_i(s) = NA\sqrt{2\pi}\sigma_t = S\lambda$, which gives:

$$\lambda = a\sigma_t/\Delta, \qquad \textit{(expected total spike count)} \quad (4)$$

where $\Delta = S/N$ is the spacing between tuning curve centers, and $a = \sqrt{2\pi}A\tau$ is an "amplitude constant" that depends on true tuning curve amplitude and the time window for integrating spikes.

Note, that tiling holds almost perfectly if tuning curves are broad compared to their spacing (e.g. $\sigma_t > \Delta$). However, our results hold for a much broader range of σ_t (see Supplementary Figs S3 and S4 for a numerical analysis.)

Let $R = \sum r_i$ denote the total spike count from the entire population. Due to tiling, R is a Poisson random variable with rate λ, regardless of the stimulus: $p(R|s) = \frac{1}{R!}\lambda^R e^{-\lambda}$.

For simplicity, we will consider stimuli drawn from a zero-mean Gaussian prior with variance σ_s^2:

$$s \sim \mathcal{N}(0, \sigma_s^2), \qquad p(s) = \tfrac{1}{\sqrt{2\pi}\sigma_s} e^{-\frac{s^2}{2\sigma_s^2}}. \qquad \textit{(stimulus prior)} \quad (5)$$

Since $\prod_i e^{-f_i(s)} = e^{-\lambda}$ due to the tiling assumption, the likelihood (eq. 1 as a function of s) and posterior both take Gaussian forms:

$$p(\boldsymbol{r}|s) \propto \prod_i f_i(s)^{r_i} \propto \mathcal{N}\left(\tfrac{1}{R}\boldsymbol{r}^\top \boldsymbol{\star}, \tfrac{1}{R}\sigma_t^2\right) \qquad \textit{(likelihood)} \quad (6)$$

$$p(s|\boldsymbol{r}) = \mathcal{N}\Big(\frac{\boldsymbol{r}^\top \boldsymbol{\star}}{R+\rho}, \frac{\sigma_t^2}{R+\rho}\Big), \qquad \textit{(posterior)} \quad (7)$$

where $\rho = \sigma_t^2/\sigma_s^2$ denotes the ratio of the tuning curve variance to prior variance. The maximum of the likelihood (eq. 6) is the so-called center-of-mass estimator estimator, $\frac{1}{R}\boldsymbol{r}^\top \boldsymbol{\star}$, while the mean of the posteror (eq. 7) is biased toward zero by an amount that depends on ρ. Note that the posterior variance does not depend on which neurons emitted spikes, only the total spike count R, a fact that will be important for our analyses below.

2.1 Capacity constraints for defining optimality

Defining optimality for a population code requires some form of constraint on the capacity of the neural population, since clearly we can achieve arbitrarily narrow posteriors if we allow arbitrarily large total spike count R. In the following, we will consider two different biologically plausible constraints:

- An *amplitude constraint*, in which we constrain the tuning curve amplitude. Under this constraint, expected spike count λ will grow as the tuning width σ_t increases (see eq. 4), since more neurons will respond to any stimulus when tuning is broader.
- An *energy constraint*, in which we fix λ while allowing σ_t and amplitude A to vary. Here, we can make tuning curves wider so that more neurons respond, but must reduce the amplitude so that total expected spike count remains fixed.

We will show that the optimal tuning depends strongly on which kind of constraint we apply.

2.2 Analyses based on Fisher Information

The Fisher information provides a popular, tractable metric for quantifying the efficiency of a neural code, given by $\mathbb{E}[-\frac{\partial^2}{\partial s^2}\log p(\boldsymbol{r}|s)]$, where expectation is taken with respect to encoding distribution $p(\boldsymbol{r}|s)$. For our idealized Poisson population, the total Fisher information is:

$$I_F(s) = \sum_{i=1}^N \frac{f_i'(s)^2}{f_i(s)} = \sum_{i=1}^N A\frac{(s-\overset{\star}{s}_i)^2}{\sigma_t^4}\exp\Big(-\frac{(s-\overset{\star}{s}_i)^2}{2\sigma_t^2}\Big) = \frac{a}{\sigma_t\Delta} = \frac{\lambda}{\sigma_t^2}, \qquad \textit{(Fisher info)} \quad (8)$$

which we can derive, as before, using the tiling property (eq. 3). (See Supplemental Sec. S2 for derivation.) The first of the two expressions at right reflects I_F for the amplitude constraint, where λ varies implicitly as we vary σ_t. The second expresses I_F under the energy constraint, where λ is constant so that the amplitude constant a varies implicitly with σ_t. For both constraints, I_F increases with decreasing σ_t [5].

Fisher information provides a well-known bound on the variance of an unbiased estimator $\hat{s}(\boldsymbol{r})$ known as the Cramér-Rao (CR) bound, namely $\mathrm{var}(\hat{s}|s) \geq 1/I_F(s)$. Since FI is constant over s in our idealized setting, this leads to a bound on the mean squared error ([6, 11]):

$$\text{MSE} \triangleq \mathbb{E}\left[(\hat{s}(\boldsymbol{r}) - s)^2\right]_{p(\boldsymbol{r},s)} \geq \mathbb{E}\left[\frac{1}{I_F(s)}\right]_{p(s)} = \frac{\sigma_t\Delta}{a} = \frac{\sigma_t^2}{\lambda}, \quad (9)$$

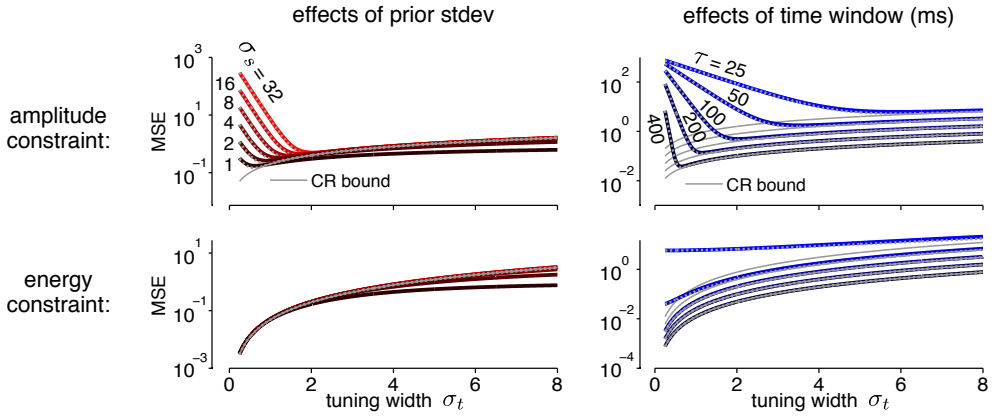

Figure 2: Mean squared error as a function of the tuning width σ_t, under amplitude constraint (top row) and energy constraint (bottom row), for spacing $\Delta = 1$ and amplitude $A = 20$ sp/s. and **Top left:** MSE for different prior widths σ_s (with $\tau = 100$ms), showing that optimal σ_t increases with larger prior variance. Cramér-Rao bound (gray solid) is minimized at $\sigma_t = 0$, whereas bound (eq. 12, gray dashed) accurately captures shape and location of the minimum. **Top right:** Similar curves for different time windows τ for counting spikes (with σ_s=32), showing that optimal σ_t increases for lower spike counts. **Bottom row:** Similar traces under energy constraint (where A scales inversely with σ_t so that $\lambda = \sqrt{2\pi}\tau A\sigma_t$ is constant). Although the CR bound grossly understates the true MSE for small counting windows (right), the optimal tuning is maximally narrow in this configuration, consistent with the CR curve.

which is simply the inverse of Fisher Information (eq. 8).

Fisher information also provides a (quasi) lower bound on the mutual information, since an efficient estimator (i.e., one that achieves the CR bound) has entropy upper-bounded by that of a Gaussian with variance $1/I_F$ (see [3]). In our setting this leads to the lower bound:

$$\mathrm{MI}(s, \boldsymbol{r}) \;\triangleq\; \mathrm{H}(s) - \mathrm{H}(s|\boldsymbol{r}) \;\geq\; \tfrac{1}{2}\log\left(\sigma_s^2 \frac{a}{\sigma_t \Delta}\right) = \tfrac{1}{2}\log\left(\sigma_s^2 \frac{\lambda}{\sigma_t^2}\right). \tag{10}$$

Note that neither of these FI-based bounds apply exactly to the Bayesian setting we consider here, since Bayesian estimators are generally biased, and are inefficient in the regime of low spike counts [6]. We examine them here nonetheless (gray traces in Figs. 2 and 3) due to their prominence in the prior literature ([5, 11, 13]), and to emphasize their limitations for characterizing optimal codes.

2.3 Exact Bayesian analyses

In our idealized population, the total spike count R is a Poisson random variable with mean λ, which allows us to compute the MSE and MI by taking expectations w.r.t. this distribution.

Mean Squared Error (MSE)

The mean squared error, which equals the average posterior variance (eq. 7), can be computed analytically for this model:

$$\mathrm{MSE} = \mathbb{E}\left[\frac{\sigma_t^2}{R+\rho}\right]_{p(R)} = \sum_{R=0}^{\infty}\left(\frac{\sigma_t^2}{R+\rho}\right)\frac{\lambda^R}{R!}e^{-\lambda} = \sigma_t^2\, e^{-\lambda}\,\Gamma(\rho)\,\gamma^*\left(\rho, -\lambda\right), \tag{11}$$

where $\rho = \sigma_t^2/\sigma_s^2$ and $\gamma^*(a,z) = z^{-a}\frac{1}{\Gamma(a)}\int_0^z t^{a-1}e^{-t}dt$ is the holomorphic extension of the lower incomplete gamma function [25] (see SI for derivation). When the tuning curve is narrower than the prior (i.e., $\sigma_t^2 \leq \sigma_s^2$), we can obtain a relatively tight lower bound:

$$\mathrm{MSE} \;\;\geq\;\; \frac{\sigma_t^2}{\lambda}\left(1 - e^{-\lambda}\right) + (\sigma_s^2 - \sigma_t^2)e^{-\lambda}. \tag{12}$$

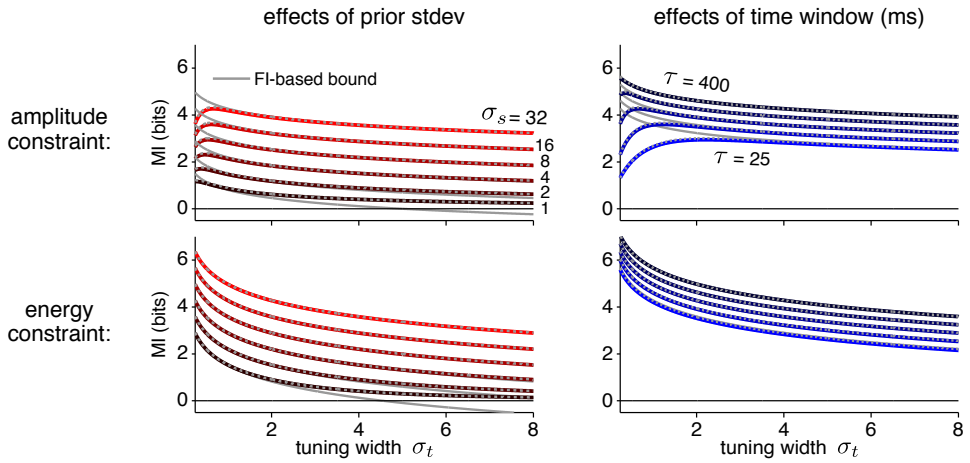

Figure 3: Mutual information as a function of tuning width σ_t, directly analogous to plots in Fig. 2. Note the problems with the lower bound on MI derived from Fisher information (top, gray traces) and the close match of the derived bound (eq. 14, dashed gray traces). The effects are similar to Fig. 2, except that MI-optimal tuning widths are slightly smaller (upper left and right) than for MSE-optimal codes. For both loss functions, optimal width is minimal under an energy constraint.

Figure 2 shows the MSE (and derived bound) as a function of the tuning width σ_t over the range where tiling approximately holds. Note the high accuracy of the approximate formula (12, dashed gray traces) and that the FI-based bound does not actually lower-bound the MSE in the case of narrow priors (darker traces).

For the amplitude-constrained setting (top row, obtained by substituting $\lambda = a\sigma_t/\Delta$ in eqs. 11 and 12), we observe substantial discrepancies between the true MSE and FI-based analysis. While FI suggests that optimal tuning width is near zero (down to the limits of tiling), analyses reveal that the optimal σ_t grows with prior variance (left) and decreasing time window (right). These observations agree well with the existing literature (e.g. [14, 15]). However, if we restrict the average population firing rate (energy constraint, bottom plots), the optimal tuning curves once again approach zero. In this case, FI provides correct intuitions and better approximation of the true MSE.

Mutual Information (MI)

For a tiling population and Gaussian prior, mutual information between stimulus and response is:

$$\mathrm{MI}(s,\boldsymbol{r}) \quad = \quad \tfrac{1}{2}\mathbb{E}\left[\log\left(1+R\tfrac{\sigma_s^2}{\sigma_t^2}\right)\right]_{p(R)}, \tag{13}$$

which has no closed-form solution, but can be calculated efficiently with a discrete sum over R from 0 to some large integer (e.g., $R = \lambda + n\sqrt{\lambda}$ to capture n standard deviations above the mean). We can derive an upper bound using the Taylor expansion to log while preserving the exact zeroth order term:

$$\mathrm{MI}(s,\boldsymbol{r}) \leq \tfrac{1-e^{-\lambda}}{2}\log\left(1+(\tfrac{\lambda}{1-e^{-\lambda}})\tfrac{\sigma_s^2}{\sigma_t^2}\right) \quad = \quad \tfrac{1-e^{-a\sigma_t/\Delta}}{2}\log\left(1+\tfrac{a}{1-e^{-a\sigma_t/\Delta}}\tfrac{\sigma_s^2}{\sigma_t\Delta}\right) \tag{14}$$

Once again, we investigate the efficiency of population coding, but in terms of the maximal MI. Figure 3 shows MI as a function of the neural tuning width σ_t. We observe a similar effect as for the MSE: the optimal tuning widths are now different from zero, but only for the amplitude constraint. Under the energy constraint (as with FI) the optimum is maximally narrow tuning.

3 Poisson population coding with input noise

We can obtain a more general family of correlated population codes by considering "input noise", where the stimulus s is corrupted by an additive noise n (see Fig. 1):

$$s \sim \mathcal{N}(0, \sigma_s^2) \quad \textit{(prior)} \tag{15}$$

$$n \sim \mathcal{N}(0, \sigma_n^2) \quad \textit{(input noise)} \tag{16}$$

$$r_i|s, n \sim \text{Poiss}(f_i(s+n)) \quad \textit{(population response)} \tag{17}$$

The use of Gaussians allows us to marginalise over n analytically, resulting in a Gaussian form for the likelihood and Gaussian posterior:

$$p(\boldsymbol{r}|s) \propto \mathcal{N}\left(\tfrac{1}{R}\boldsymbol{r}^\top \boldsymbol{\check{s}}, \tfrac{1}{R}\sigma_t^2 + \sigma_n^2\right) \quad \textit{(likelihood)} \tag{18}$$

$$p(s|\boldsymbol{r}) = \mathcal{N}\left(\frac{\boldsymbol{r}^\top \boldsymbol{\check{s}}}{\sigma_t^2/\sigma_s^2 + R(\sigma_n^2/\sigma_s^2+1)}, \frac{(\sigma_t^2 + R\sigma_n^2)\sigma_s^2}{\sigma_t^2 + R(\sigma_n^2+\sigma_s^2)}\right) \quad \textit{(posterior)} \tag{19}$$

Note that even in the limit of many neurons and large spike counts, the posterior variance is non-zero, converging to $\sigma_n^2\sigma_s^2/(\sigma_n^2+\sigma_s^2)$, a limit defined by the prior and input noise variance. [22].

3.1 Population coding characteristics: FI, MSE, & MI

Fisher information for a population with input noise can be computed using the fact that the likelihood (eq. 18) is Gaussian:

$$I_F(s) = \mathbb{E}\left[\frac{R}{\sigma_t^2 + R\sigma_n^2}\right]_{p(R)} = \frac{\lambda e^{-\lambda}}{\sigma_n^2}\Gamma(1+\rho)\gamma^*(1+\rho, -\lambda) \tag{20}$$

where $\rho = \sigma_t^2/\sigma_n^2$ and $\gamma^*(\cdot,\cdot)$ once again denotes holomorphic extension of lower incomplete gamma function. Note that for $\sigma_n = 0$, this reduces to (eq. 8).

It is straightforward to employ the results from Sec. 2.3 for the exact Bayes analyses of a Gaussian posterior (19):

$$\begin{aligned}\text{MSE} &= \sigma_s^2\mathbb{E}\left[\frac{\sigma_t^2 + R\sigma_n^2}{\sigma_t^2 + R(\sigma_n^2+\sigma_s^2)}\right]_{p(R)} = \sigma_s^2\phi\,\mathbb{E}\left[\frac{1}{\phi+R}\right]_{p(R)} + \tfrac{\sigma_s^2\sigma_n^2}{\sigma_s^2+\sigma_n^2}\mathbb{E}\left[\frac{R}{\phi+R}\right]_{p(R)} \\ &= \left[\phi\Gamma(\phi)\gamma^*(\phi,-\lambda) + \tfrac{\sigma_n^2}{\sigma_s^2+\sigma_n^2}\lambda\Gamma(1+\phi)\gamma^*(1+\phi,-\lambda)\right]\sigma_s^2 e^{-\lambda}, \text{ and}\end{aligned} \tag{21}$$

$$\text{MI} = \tfrac{1}{2}\mathbb{E}\left[\log\left(1 + \frac{R\sigma_s^2}{\sigma_t^2 + R\sigma_n^2}\right)\right]_{p(R)}, \tag{22}$$

where $\phi = \sigma_t^2/(\sigma_s^2+\sigma_n^2)$. Although we could not determine closed-form analytical expressions for MI, it can be computed efficiently by summing over a range of integers $[0, \ldots R_{max}]$ for which $P(R)$ has sufficient support. This is still a much faster procedure than estimating these values from Monte Carlo simulations.

3.2 Optimal tuning width under input noise

Fig. 4 shows the optimal tuning width under the amplitude constraint, that is, the value of σ_t that achieves minimal MSE (left) or maximal MI (right) as a function of the prior width σ_s, for several different time windows τ. Blue traces show results for a Poisson population, while green traces correspond to a population with input noise ($\sigma_n = 1$).

For both MSE and MI loss functions, optimal tuning width decreases for narrower priors. However, under input noise (green traces), the optimal tuning width saturates at the value that depends on the available number of spikes. As the prior grows wider, the growth of the optimal tuning width depends strongly on the choice of loss function: optimal σ_t grows approximately logarithmically with σ_s for minimizing MSE (left), but it grows much slower for maximizing MI (right). Note that for realistic prior widths (i.e. for $\sigma_s > \sigma_n$), the effects of input noise on optimal tuning width are far more substantial under MI than under MSE.

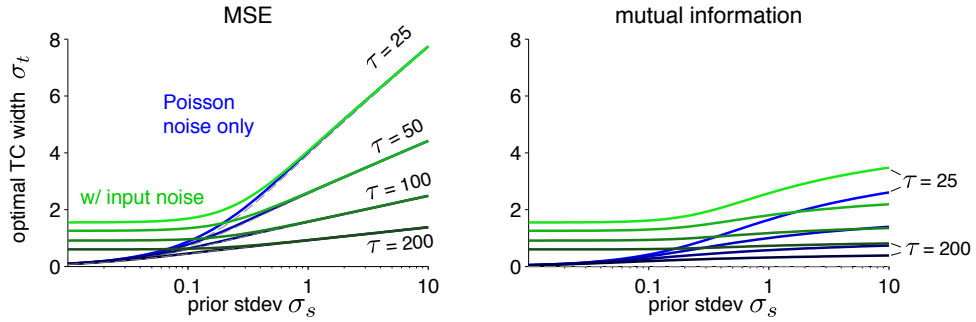

Figure 4: Optimal tuning width σ_t (under amplitude constraint only) as a function of prior width σ_s, for classic Poisson populations (blue) and populations with input-noise (green, $\sigma_n^2 = 1$). Different traces correspond to different time windows of integration, for $\Delta = 1$ and $A = 20$ sp/s. As σ_n increases, the optimal tuning width increases under MI, and under MSE when $\sigma_s < \sigma_n$ (traces not shown). For MSE, predictions of the Poisson and input-noise model converge for priors $\sigma_s > \sigma_n$.

We have not shown plots for energy-constrained population codes because the optimal tuning width sits at the minimum of the range over which tiling can be said to hold, regardless of prior width, input noise level, time window, or choice of loss function. This can be seen easily in the expressions for MI (eqs. 13 and 22), in which each term in the expectation is a decreasing function of σ_t for all $R > 0$. This suggests that, contrary to some recent arguments (e.g., [14, 15]), narrow tuning (at least down to the limit of tiling) really is best if the brain has a fixed energetic budget for spiking, as opposed to a mere constraint on the number of neurons.

4 Correlations induced by input noise

Input noise alters the mean, variance, and pairwise correlations of population responses in a systematic manner that we can compute directly (see Supplement for derivations). In Fig. 5 we show the effects of input noise with standard deviation $\sigma_n = 0.5\Delta$, for neurons with the tuning amplitude of $A = 10$. The tuning curve (mean response) becomes slightly flatter (Fig. 5A), while the variance increases, especially at the flanks (Fig. 5B). Fig. 5C shows correlations between the two neurons with tuning curves and variance are shown in panels A-B: one pair with the same preferred orientation at zero (red) and a second with a 4 degree difference in preferred orientation (blue). From these plots, it is clear that the correlation structure depends on both the tuning as well as the stimulus. Thus, in order to describe such correlations one needs to consider the entire stimulus range, not simply the average correlation marginalized over stimuli.

Figure 5D shows the pairwise correlations across an entire population of 21 neurons given a stimulus at $s = 0$. Although we assumed Gaussian tuning curves here, one can obtain similar plots for arbitrary unimodal tuning curves (see Supplement), which should make it feasible to test our predictions in real data. However, the time scale of the input noise and basic neural computations is about 10 ms. At such short spike count windows, available number of spikes is low, and so are correlations induced by input noise. With other sources of second order statistics, such as common input gains (e.g. by contrast or adaptation), these correlations might be too subtle to recover [22].

5 Discussion

We derived exact expressions for mean squared error and mutual information in a Bayesian analysis of: (1) an idealized Poisson population coding model; and (2) a correlated, conditionally Poisson population coding model with shared input noise. These expressions allowed us to examine the optimal tuning curve width under both loss functions, under two kinds of resource constraints. We have confirmed that optimal σ_t diverges from predictions based on Fisher information, if the overall spike count is allowed to grow with tuning width (i.e., because more neurons respond to the stimulus when tuning curves become broader). We refer to this as an "amplitude constraint", because the amplitude is fixed independently of tuning width. This differs from an "energy constraint", in

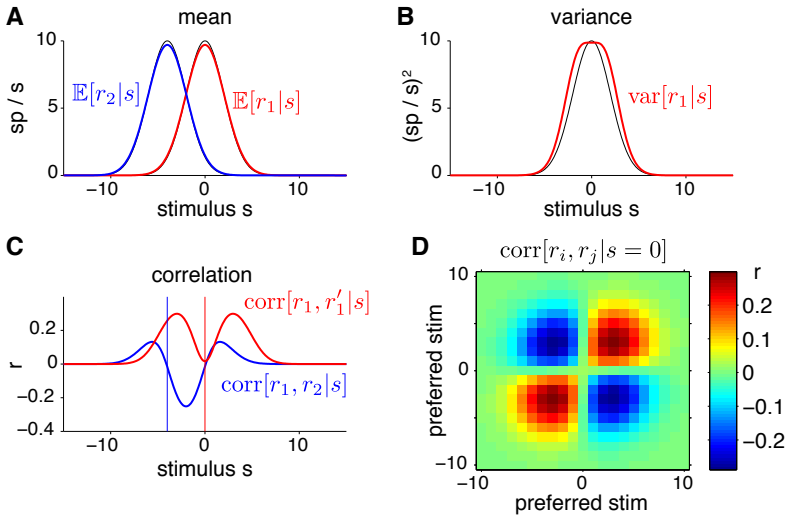

Figure 5: Response statistics of neural population with input noise, for standard deviation $\sigma_n = 0.5$. **(A)** Expected spike responses of two neurons: $\mathring{s}_1 = 0$ (red) and $\mathring{s}_2 = -4$ (blue). The common noise effectively smooths blurs the tuning curves with a Gaussian kernel of width σ_n. **(B)** Variance of neuron 1, its tuning curve replotted in black for reference. Input noise has largest influence on variance at the steepest parts of the tuning curve. **(C)** Cross-correlation of the neuron 1 with two others: one sharing the same preference (red), and one with $\mathring{s} = -4$ (blue). Note that correlation of two identically tuned neurons is largest at the steepest part of the tuning curve. **(D)** Spike count correlations for entire population of 21 neurons given a fixed stimulus $s = 0$, illustrating that the pattern of correlations is signal dependent.

which tuning curve amplitude scales with tuning width so that average total spike count is constant. Under an energy constraint, predictions from Fisher information match those of an exact Bayesian analysis, and we find that optimal tuning width should be narrow (down to the limit at which the tiling assumption applies), regardless of the duration, prior width, or input noise level.

We also derived explicit expressions for the response correlations induced by the input noise. These correlations depend on the shape and amplitude of tuning curves, and on the amount of input noise (σ_n). However, for a reasonable assumption that noise distribution is much narrower than the width of the prior (and tuning curves), under which the mean firing rate changes little, we can derive predictions for the covariances directly from the measured tuning curves. An important direction for future work will be to examine the detailed structure of correlations measured in large populations. We feel that the input noise model — which describes exactly those correlations that are most harmful for decoding — has the potential to shed light on the factors affecting the coding capacity in optimal neural populations [22].

Finally, we can return to the introductory example involving orientation discrimination, to ask how the number of neurons necessary to reach the human discrimination threshold of δs=1 degree changes in the presence of input noise. As σ_n approaches δs, the number of neurons required goes rapidly to infinity (See Supp. Fig. S1).

Acknowledgments

This work was supported by the McKnight Foundation (JP), NSF CAREER Award IIS-1150186 (JP), NIMH grant MH099611 (JP) and the Gatsby Charitable Foundation (AGB).

References

[1] HS Seung and H. Sompolinsky. Simple models for reading neuronal population codes. *Proceedings of the National Academy of Sciences*, 90(22):10749–10753, 1993.

[2] R. S. Zemel, P. Dayan, and A. Pouget. Probabilistic interpretation of population codes. *Neural Comput*, 10(2):403–430, Feb 1998.

[3] Nicolas Brunel and Jean-Pierre Nadal. Mutual information, fisher information, and population coding. *Neural Computation*, 10(7):1731–1757, 1998.

[4] Kechen Zhang and Terrence J. Sejnowski. Neuronal tuning: To sharpen or broaden? *Neural Computation*, 11(1):75–84, 1999.

[5] A. Pouget, S. Deneve, J. Ducom, and P. E. Latham. Narrow versus wide tuning curves: What's best for a population code? *Neural Computation*, 11(1):85–90, 1999.

[6] M. Bethge, D. Rotermund, and K. Pawelzik. Optimal short-term population coding: When fisher information fails. *Neural computation*, 14(10):2317–2351, 2002.

[7] P. Series, P. E. Latham, and A. Pouget. Tuning curve sharpening for orientation selectivity: coding efficiency and the impact of correlations. *Nature Neuroscience*, 7(10):1129–1135, 2004.

[8] W. J. Ma, J. M. Beck, P. E. Latham, and A. Pouget. Bayesian inference with probabilistic population codes. *Nature Neuroscience*, 9:1432–1438, 2006.

[9] Marcelo A Montemurro and Stefano Panzeri. Optimal tuning widths in population coding of periodic variables. *Neural computation*, 18(7):1555–1576, 2006.

[10] R. Haefner and M. Bethge. Evaluating neuronal codes for inference using fisher information. *Neural Information Processing Systems*, 2010.

[11] D. Ganguli and E. P. Simoncelli. Implicit encoding of prior probabilities in optimal neural populations. In *Adv. Neural Information Processing Systems*, volume 23, May 2010.

[12] Xue-Xin Wei and Alan Stocker. Efficient coding provides a direct link between prior and likelihood in perceptual bayesian inference. In *Adv. Neur. Inf. Proc. Sys. 25*, pages 1313–1321, 2012.

[13] Z Wang, A Stocker, and A Lee. Optimal neural tuning curves for arbitrary stimulus distributions: Discrimax, infomax and minimum l_p loss. In *Adv. Neur. Inf. Proc. Sys. 25*, pages 2177–2185, 2012.

[14] P. Berens, A.S. Ecker, S. Gerwinn, A.S. Tolias, and M. Bethge. Reassessing optimal neural population codes with neurometric functions. *Proceedings of the National Academy of Sciences*, 108(11):4423, 2011.

[15] Steve Yaeli and Ron Meir. Error-based analysis of optimal tuning functions explains phenomena observed in sensory neurons. *Frontiers in computational neuroscience*, 4, 2010.

[16] J. M. Beck, P. E. Latham, and A. Pouget. Marginalization in neural circuits with divisive normalization. *J Neurosci*, 31(43):15310–15319, Oct 2011.

[17] Stuart Yarrow, Edward Challis, and Peggy Seriès. Fisher and shannon information in finite neural populations. *Neural Computation*, 24(7):1740–1780, 2012.

[18] D Ganguli and E P Simoncelli. Efficient sensory encoding and Bayesian inference with heterogeneous neural populations. *Neural Computation*, 26(10):2103–2134, Oct 2014. Published online: 24 July 2014.

[19] E. Zohary, M. N. Shadlen, and W. T. Newsome. Correlated neuronal discharge rate and its implications for psychophysical performance. *Nature*, 370(6485):140–143, Jul 1994.

[20] Keiji Miura, Zachary Mainen, and Naoshige Uchida. Odor representations in olfactory cortex: distributed rate coding and decorrelated population activity. *Neuron*, 74(6):1087–1098, 2012.

[21] Jakob H. Macke, Manfred Opper, and Matthias Bethge. Common input explains higher-order correlations and entropy in a simple model of neural population activity. *Phys. Rev. Lett.*, 106(20):208102, May 2011.

[22] R. Moreno-Bote, J. Beck, I. Kanitscheider, X. Pitkow, P.E. Latham, and A. Pouget. Information-limiting correlations. *Nat Neurosci*, 17(10):1410–1417, Oct 2014.

[23] K. Josic, E. Shea-Brown, B. Doiron, and J. de la Rocha. Stimulus-dependent correlations and population codes. *Neural Comput*, 21(10):2774–2804, Oct 2009.

[24] G. Dehaene, J. Beck, and A. Pouget. Optimal population codes with limited input information have finite tuning-curve widths. In *CoSyNe*, Salt Lake City, Utah, February 2013.

[25] NIST Digital Library of Mathematical Functions. http://dlmf.nist.gov/, Release 1.0.9 of 2014-08-29.

[26] David C Burr and Sally-Ann Wijesundra. Orientation discrimination depends on spatial frequency. *Vision Research*, 31(7):1449–1452, 1991.

[27] P. Seriès, A. A. Stocker, and E. P. Simoncelli. Is the homunculus 'aware' of sensory adaptation? *Neural Computation*, 21(12):3271–3304, Dec 2009.

[28] R. L. De Valois, E. W. Yund, and N. Hepler. The orientation and direction selectivity of cells in macaque visual cortex. *Vision research*, 22(5):531–544, 1982.

Supplementary Material for

Optimal prior-dependent neural population codes under shared input noise

Agnieszka Grabska-Barwińska & Jonathan W. Pillow
NIPS, 2014

S1 Humans vs. Poisson population codes

Here we unpack the rough comparison of decoding performance in humans and ideal (conditionally independent) Poisson population codes provided in the Introduction.

Burr & Wijesundra 1991 [26] reports orientation discrimination thresholds (δs) as low as 0.5 deg in human observers, where threshold is defined as the angular difference at which observes achieve 81.6% correct performance in a 2AFC task.

We can relate this threshold to sensitivity (d') and Fisher information using the formula (eq. 4.4 in [27]):

$$(\delta s) \geq d'_\rho \frac{1}{\sqrt{I_F}} \tag{23}$$

where I_F is the Fisher information and d'_ρ is the sensitivity (d') for two stimuli that can be correctly discriminated with an error probability of ρ, given by

$$d'_\rho = \sqrt{2}\,\Phi^{-1}(1-\rho), \tag{24}$$

where $\Phi^{-1}(\cdot)$ is the inverse normal cumulative density function (cdf). For probability of correct $1-\rho = 0.816$, this gives $d'_\rho \approx 1.27$. Plugging this value into (Eq. 23), it's clear that to obtain human-level discrimination performance, we need FI of at least:

$$I_f \geq (d'_\rho)^2/(ds)^2 \approx (1.27/0.5)^2 = 6.45. \tag{25}$$

Now, consider a population of 500 V1 neurons with tuning curves spaced evenly around the circle ($\Delta = 0.72$ deg), with a maximum spike rate of 50 spikes/sec, and a full bandwidth at half-height of 60 degrees (near the upper end of the range reported in monkeys [28]). This corresponds to a Gaussian tuning width of $\sigma_t = 30\sqrt{-1/(2 * \log .5)} \approx 25$ deg. This population (which clearly tiles) achieves Fisher Information (eq. 8) of approximately $I_F = \sqrt{2\pi} \times 50/(0.72 \times 25) = 7$, for all stimuli, so it is sufficient to reproduce human performance. For a population of 2000 neurons with identical characteristics, an efficient decoder could achieve a discrimination threshold twice as low as a human observer, or $(\delta s) = d'_\rho/\sqrt{I_F} \approx 0.25$ deg.

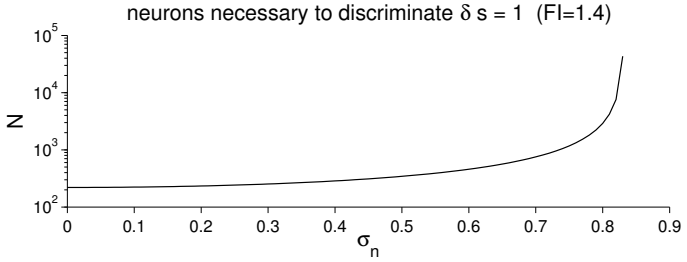

Figure S1: Number of neurons necessary to discriminate $\delta s = 1$ degree with 80% probability correct as a function of noise σ_n. We optimised Eq. 20 to match I_F of the noiseless case.

S2 Derivation of Fisher information for tiling Poisson neurons

Fisher information for Poisson neurons and Gaussian tuning curves:

$$I_F(s) = \sum_{i=1}^{N} \frac{f_i'(s)^2}{f_i(s)} = \sum_{i=1}^{N} A\frac{(s-\overset{*}{s}_i)^2}{\sigma_t^4} \exp\left(-\frac{(s-\overset{*}{s}_i)^2}{2\sigma_t^2}\right)$$

The average I_F per neuron equals,

$$I_F^i = \frac{1}{S}\int_{-S/2}^{S/2} I_F^i(s)ds = \frac{A}{S\sigma_t^4}\int_{-S/2}^{S/2} ds(s - \overset{\star}{s}_i)^2 \exp\left(-\frac{(s-\overset{\star}{s}_i)^2}{2\sigma_t^2}\right)$$

Assuming $S \gg \sigma_t$ (i.e. for neurons well away from the ends of the s-domain), we get

$$I_F^i = \frac{A}{S\sigma_t^4}\sqrt{2\pi}\sigma_t^3 = \frac{A\sqrt{2\pi}}{S\sigma_t}$$

Thus, each neuron contributes a similar average I_F, summing to:

$$I_F = N\frac{A\sqrt{2\pi}}{S\sigma_t} = \frac{A\sqrt{2\pi}}{\Delta\sigma_t}$$

S3 Derivation of MSE for Poisson population code

The formula for mean squared error (MSE) in a standard Poisson population code (eq. 11) can be derived using the following series representation of the holomorphic extension of the lower incomplete Gamma function ([25], equation 8.7.1):

$$\gamma^*(a,z) = \frac{1}{\Gamma(a)}\sum_{k=0}^{\infty}\frac{(-z)^k}{k!(k+a)}. \tag{26}$$

If we substitute $\rho = (\sigma_t^2/\sigma_s^2)$ and $-\lambda$ for a and z, respectively, then (beginning from the r.h.s. of eq. 11), we have:

$$\sigma_t^2\, e^{-\lambda}\,\Gamma(\rho)\,\gamma^*(\rho,-\lambda) = \sigma_t^2 e^{-\lambda}\sum_{R=0}^{\infty}\frac{\lambda^R}{R!(R+\rho)} = \sigma_t^2\mathbb{E}\left[\frac{1}{R+\rho}\right]_{p(R)} = \text{MSE}, \tag{27}$$

as stated in the main text, where $p(R)$ is the Poisson distribution with mean λ.

S4 Derivation of I_F for the input noise

$$I_F = \mathbb{E}\left[\frac{R}{\sigma_t^2 + R{\sigma_n^2}^2}\right]_{p(R)} \tag{28}$$

$$= \sum_{R=0}^{\infty}\frac{R}{\sigma_t^2+R\sigma_n^2}\frac{\lambda^R}{R!}e^{-\lambda} \tag{29}$$

$$= e^{-\lambda}\sum_{R=1}^{\infty}\frac{1}{\sigma_t^2+R\sigma_n^2}\frac{\lambda^R}{(R-1)!} \tag{30}$$

$$= e^{-\lambda}\sum_{R=1}^{\infty}\frac{1}{\sigma_t^2+\sigma_n^2+(R-1)\sigma_n^2}\frac{\lambda^R}{(R-1)!} \tag{31}$$

$$= \lambda e^{-\lambda}\sum_{R=0}^{\infty}\frac{1}{\sigma_t^2+\sigma_n^2+R\sigma_n^2}\frac{\lambda^R}{R!} \tag{32}$$

$$= \frac{1}{\sigma_n^2}\lambda e^{-\lambda}\sum_{R=0}^{\infty}\frac{1}{\sigma_t^2/\sigma_n^2+1+R}\frac{\lambda^R}{R!} \tag{33}$$

$$= \frac{1}{\sigma_n^2}\lambda e^{-\lambda}\Gamma(1+\rho)\gamma^*(1+\rho,-\lambda) \tag{34}$$

where $\rho = \sigma_t^2/\sigma_n^2$. Combining this derivation with the above derivation for MSE (Supplement section S3) yields the terms necessary for the exact MSE under input noise (eq. 21).

S5 Optimal tuning widths for very broad priors.

The effect of the prior width on optimal tuning width is much stronger for MSE than for MI. Empirically, we have noticed that for broad priors, the shared input noise model yields similar optimal tuning widths as the noiseless input model. From the approximation introduced in (eq. 12), we can see that the main contribution to the MSE for broad priors is $\sigma_s^2 e^{-\lambda}$. Thus, by setting $\lambda \propto 2\log\sigma_s$, we reduce the contribution of that factor to a constant. See results in Fig. S2.

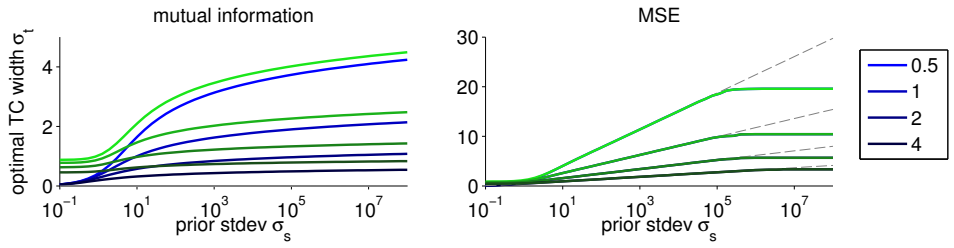

Figure S2: Optimal tuning widths. Different shades of color correspond to increasing tuning amplitude, A. Blue depicts the optimal σ_t for the noise-less case (as in Figs 2–3), whereas green corresponds to the input noise $\sigma_n = 1$. As we increase σ_n (not shown here), the optimal tuning curves increase systematically for MI, and for MSE when $\sigma_s < \sigma_n$. However, for MSE, predictions of the noise-free and full model still converge for priors $\sigma_s > \sigma_n$. The dashed gray lines are optimal tuning curves obtained from the approximate lower bounds for the noiseless case (eq. 12 and 14). Δ=1 degree.

S6 Effects of imperfect tuning curve tiling

One of our main assumptions was that the tuning curves should "tile". However, the cases we considered often ventured into a range $\sigma_t \sim \Delta$, for which the tiling is not supposed to hold. We thus estimated the effect of the uneven tiling for our neuromorphic functions. We conclude that our estimates are true on average for a broad range of σ_t, reaching well below Δ. That is, despite the local dependence on s, our metrics hold true on average.

We performed Monte Carlo simulations, whereby per each true stimulus s (x-axis in Fig. 3), we simulated 1000 network responses. We estimated posteriors per each response, their MSE and entropy. In Fig. 3, 3rd and 4th column, we report the mean of these estimates as a function of s. One can see that even for $\sigma_t = \Delta/2$, both MSE and MI are relatively constant in s. However, even for the most drastic cases of $\sigma_t = \Delta/4$ and $\sigma_t = \Delta/10$, where fluctuations around the mean are clearly distinguishable, the average still matches our analytical predictions (depicted with red line in Fig. 3).

We compare the numerical and analytical predictions directly in Fig. 4. We see a robust match between the two, with small discrepancies showing only for the lowest values of σ_t/Δ in MI (top). In fact, discrepancies in MI most likely result from a finite sample of responses. At A=1, only neurons very close to the stimulus are likely to fire, with a big range of stimuli in between the peaks of tuning curves which have a much lower firing rate, and a much higher MI gain, for any of neurons that spikes for such "non-preferred" stimuli.

S7 Correlations induced by input noise

We start by expressing changes induced by the noise in the most generic way (i.e. independent of the likelihood function). By writing response as composed of the deterministic and the noisy part: $r_i = f_i + \eta$, where by definition $\mathbb{E}[\eta] = 0$, we can write:

$$\mathbb{E}[r_i|s] = \int \tfrac{1}{\sqrt{2\pi}\sigma_n} e^{-\frac{n^2}{2\sigma_n^2}} f_i(s+n)\, dn \tag{35a}$$

$$\mathbb{E}[r_i^2|s] = \int \tfrac{1}{\sqrt{2\pi}\sigma_n} e^{-\frac{n^2}{2\sigma_n^2}} \left(f_i(s+n)^2 + \mathbb{E}[\eta_i^2]\right) dn \tag{35b}$$

$$\mathbb{E}[r_i r_j|s] = \int \tfrac{1}{\sqrt{2\pi}\sigma_n} e^{-\frac{n^2}{2\sigma_n^2}} \left(f_i(s+n) f_j(s+n) + \mathbb{E}[\eta_i \eta_j]\right) dn \tag{35c}$$

As we see, the mean will change in the same way regardless of the neurons' noise distribution, let us call it $\tilde{f}_i \equiv \mathbb{E}[r_i|s]$. The higher order statistics however, will depend on the noise model. For independent Poisson neurons:

$$\mathrm{var}[r_i|s] = \int f_i(s+n)^2 \tfrac{1}{\sqrt{2\pi}\sigma_n} e^{-\frac{n^2}{2\sigma_n^2}} dn + \tilde{f}_i(s) - \tilde{f}_i(s)^2 \tag{36}$$

$$\mathrm{cov}[r_i, r_j|s] = \int f_i(s+n) f_j(s+n) \tfrac{1}{\sqrt{2\pi}\sigma_n} e^{-\frac{n^2}{2\sigma_n^2}} dn - \tilde{f}_i(s)\tilde{f}_j(s) \tag{37}$$

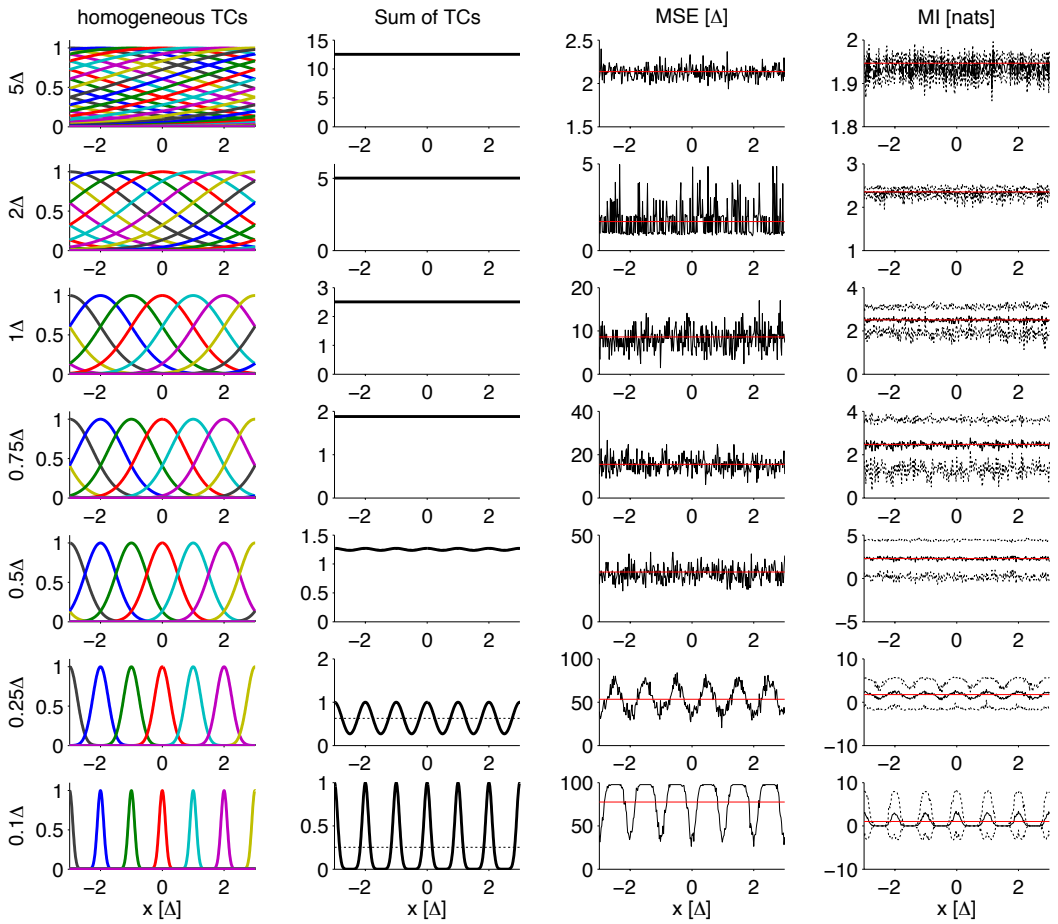

Figure S3: Analyses of imperfect tuning curve tiling. Each row shows a different value of tuning width σ_t; A=1. Only neurons from the middle of the coding range are shown. Average MSE (3rd column) and mutual information MI (4th column) were estimated from 1000 sample responses. The red lines in the MSE and MI columns depict our analytical predictions. Dashed lines in the fourth column demarcate mean$\pm$std (for the MSE, the standard deviation is too large to fit on the plots).

For Gaussian tuning curves, these statistics are easy to compute, using

$$\tilde{f}(s) = \frac{A\sigma_t}{\sqrt{\sigma_n^2+\sigma_t^2}} e^{-\frac{\left(s-\overset{\star}{s}_i\right)^2}{2(\sigma_n^2+\sigma_t^2)}}$$

and

$$\int f_i(s+n) f_j(s+n) \frac{1}{\sqrt{2\pi}\sigma_n} e^{-\frac{n^2}{2\sigma_n^2}}\, dn = \frac{(A)^2\sigma_t}{\sqrt{2\sigma_n^2+\sigma_t^2}} e^{-\frac{(\overset{\star}{s}_i-\overset{\star}{s}_j)^2}{4\sigma_t^2} - \frac{(s-(\overset{\star}{s}_i+\overset{\star}{s}_j)/2)^2}{2(\sigma_n^2+\sigma_t^2/2)}} .$$

In Fig. 5, we illustrate how neural statistics change due to input noise. For $\sigma_t = 2$ and a strong input noise of $\sigma_n = \frac{1}{4}\sigma_s$, the effects on the mean and variance are barely visible if the expected spike count is 1 (top plot). The covariance between identically tuned neurons (top right) also doesn't exceed 2% of the variance, leading to correlations not exceeding 0.05. Only for higher firing rates, the effects of noise become more visible, with covariance of identically tuned neurons reaching 20% of variance ($\rho \sim 0.4$) for $A = 10$. However, the time scale of the input noise and basic neural computations (i.e. excluding temporal integration) is about 10 ms. At such short spike count windows, available number of spikes is low.

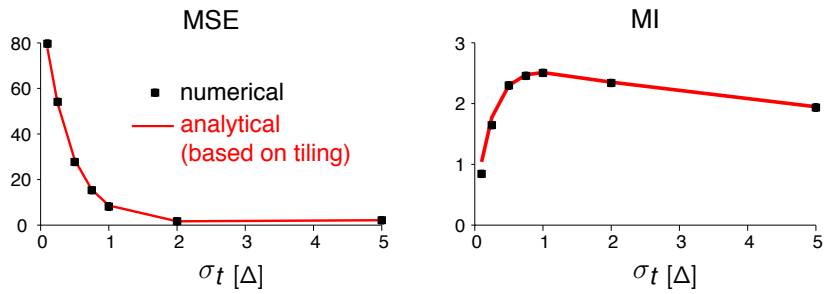

Figure S4: Comparison of exact performance measures (computed numerically by Monte Carlo simulations, 1000 samples, black pluses) with formulas derived under the assumption of a perfectly tiling tuning curves (red line), for mutual information (above, (eq. 13), and mean square error (below, (eq. 11).