Sparseness, Anti-Sparseness and Anything in Between: The Operating Point of a Neuron Determines its Computational Repertoire

Terry Elliott[1]

Department of Electronics and Computer Science,

University of Southampton,

Highfield,

Southampton, SO17 1BJ,

United Kingdom.

Running Title: ICA in a neuron with threshold and gain.

March 27, 2014.

[1]Tel.: +44 (0)23 8059 6000, Fax.: +44 (0)23 8059 2783, E.-mail: te@ecs.soton.ac.uk.

Abstract

A recent model of intrinsic plasticity coupled to Hebbian synaptic plasticity proposes that adaptation of a neuron's threshold and gain in a sigmoidal response function to achieve a sparse, exponential output firing rate distribution facilitates the discovery of heavy-tailed or super-Gaussian sources in the neuron's inputs. We show that the exponential output distribution is irrelevant to these dynamics and that, furthermore, while sparseness is sufficient, it is not necessary. The intrinsic plasticity mechanism drives the neuron's threshold large and positive, and we prove that in such a regime, the neuron will find super-Gaussian sources; equally, however, if the threshold is large and negative (an "anti-sparse" regime), it will also find super-Gaussian sources. Away from such extremes, the neuron can also discover sub-Gaussian sources. By examining a neuron with a fixed sigmoidal non-linearity and considering the synaptic strength fixed point structure in the two-dimensional parameter space defined by the neuron's threshold and gain, we show that this space is carved up into sub- and super-Gaussian-input-finding regimes, possibly with regimes of simultaneous stability of sub- and super-Gaussian sources or regimes of instability of all sources; a single Gaussian source may also be stabilised by the presence of a non-Gaussian source. A neuron's "operating point" (essentially its threshold and gain coupled with its input statistics) therefore critically determines its computational repertoire. Intrinsic plasticity mechanisms induce trajectories in this parameter space but do not fundamentally modify it. Un-

less the trajectories cross critical boundaries in this space, intrinsic plasticity is irrelevant and the neuron's non-linearity may be frozen with identical receptive field refinement dynamics.

1 Introduction

Neurons in early sensory pathways exhibit a wide variety of dynamics on different time scales, including adaptation to changes in input statistics (Kohn, 2007), changes in synaptic strengths leading to receptive field refinement (Katz and Shatz, 1996), changes in intrinsic excitability (Zhang and Linden, 2003), and homeostatic changes (Turrigiano and Nelson, 2004). A complete understanding of the functional properties of, for example, the early visual system would require an understanding of how all these processes interact to perform a transformation of ecologically relevant visual stimuli into dynamic neuronal representations that ultimately subserve an animal's behaviour (Simoncelli and Olshausen, 2001; Carandini et al., 2005).

In a series of papers, Triesch has developed a set of related models of intrinsic plasticity coupled to Hebbian synaptic plasticity. Based on ideas of sparse coding (Olshausen and Field, 1996, 1997; Baddeley et al., 1997; Lennie, 2003), Triesch develops an intrinsic plasticity mechanism that modifies the parameters defining the response or transfer function of a neuron so that its output firing distribution becomes exponential, or as close to exponential as possible, either in a rate-based setting (Triesch, 2007) or a spike-based setting (Savin et al., 2010). An exponential distribution has maximum entropy under the constraint of a fixed mean, and so adapting a neuron's output firing distribution to exponential maximises information transfer and thus develops an efficient neuronal code (cf. Attneave, 1954; Barlow, 1961; Laughlin, 1981; Atick, 1992;

van Hateren, 1992; DeWeese, 1996; Olshausen and Field, 1996; Bell and Sejnowski, 1997; Wainwright, 1999; Brenner et al., 2000; Maravall et al., 2007). When such an intrinsic plasticity mechanism is coupled to Hebbian synaptic plasticity, Triesch finds (amongst other things) that the synaptic strength vector converges on a direction corresponding to heavy-tailed or super-Gaussian sources (Triesch, 2007; Savin et al., 2010). Finding super-Gaussian sources is a classic signature of independent component analysis (ICA; Hyvärinen et al., 2001), an approach that has been employed extensively to find the independent "components" of natural images (Bell and Sejnowski, 1997; Olshausen and Field, 1997; van Hateren, 1998; Simoncelli and Olshausen, 2001; Hyvärinen et al., 2009; Lyu and Simoncelli, 2009).

Being motivated by information-theoretic principles, Triesch only considers a sparse exponential output firing distribution, so the extent to which the finding of heavy-tailed input distributions depends on this particular choice of output firing distribution is unclear. Nor is it clear precisely in what way the conjoint functioning of intrinsic and synaptic plasticity facilitates the discovery of heavy-tailed input distributions. Triesch only examines whether intrinsic plasticity is necessary for successful receptive field development in the case of Földiák bar input (Földiák, 1990), with somewhat ambiguous results. If a neuron has a fixed non-linearity defined by parameters taken from an adapting non-linearity during the process of receptive field refinement, then the fixed non-linearity will also develop an appropriate receptive field; however, if the parameters are taken at the end point of the refinement process, then a cor-

responding fixed non-linearity will not develop an appropriate receptive field (Triesch, 2007; Savin et al., 2010).

Here, by employing Triesch's novel and very stimulating approach as a launching point, we examine in some detail whether and when simultaneous intrinsic plasticity and synaptic plasticity are required for acquiring appropriate fixed points of the synaptic strength vector. We are principally interested in whether and when a single, isolated neuron with a fixed or adapting sigmoidal non-linearity can extract the independent components from its inputs in a manner similar to conventional ICA algorithms. As such, we typically use standard ICA-like inputs, i.e. inputs that are generated from centred, statistically-independent and orthogonally-mixed sources with symmetric probability density functions (PDFs). Such assumptions about the inputs may lack biological or ecological relevance, although centring can easily be achieved via separate "on" and "off" channels for supra- and sub-mean firing, respectively (see, for example, Savin et al., 2010). To overcome these standard criticisms, we will follow Triesch and also consider Földiák bar inputs (Földiák, 1990), since such inputs are neither centred nor white. The results for both types of input are, in fact, qualitatively rather similar. Of course, in focusing on the capacity of an isolated neuron with a fixed or adapting sigmoidal non-linearity to extract independent components from it inputs, we are explicitly ignoring the very real possibility that neurons may serve entirely different computational roles than those considered here.

The structure of the remainder of our paper is as follows. First, in Sec-

tion 2, we consider Triesch's rate-based model (Triesch, 2007), in which the neuron's sigmoidal non-linearity may be characterised by its threshold and gain. We show that intrinsic plasticity in the sparse coding regime leads to the discovery of super-Gaussian inputs generally and not heavy-tailed inputs specifically; if the inputs are all sub-Gaussian, then the strength vector does not converge on any one of these inputs even though some inputs will have heavier tails than others. By considering a sparse but not exponential output firing rate distribution, we show that these super-Gaussian-input-finding dynamics do not require an exponential output firing rate distribution but only sparseness. We then relax sparseness by considering larger mean output firing rates in an exponential output distribution and observe that such a neuron switches from finding super-Gaussian inputs to finding sub-Gaussian inputs. These results suggest that a sigmoidal non-linearity defined by threshold and gain parameters performs ICA in a parameter-dependent manner. In Section 3 we consider the performance of a fixed rather than an adapting non-linearity. We consider an extension of our earlier model of adaptation to input statistics (Elliott et al., 2008) by considering not only adaptation to changing input statistics but also adaptation to changes in synaptic strengths, in order to maintain an approximately invariant output firing rate PDF. For centred, statistically-independent and orthogonally-mixed sources with synaptic strengths normalised on the unit hypersphere, however, the mean and variance of the (standard, linearly summed) total input to a neuron are in fact independent of the neuron's synaptic strengths and thus a neuron's threshold and gain

need not be changed in response to changes in synaptic strengths with such inputs. Via this round-about argument, we reduce to examining the fixed point structure of the synaptic strength vector for a neuron with a fixed non-linearity for whitened, statistically-independent inputs, but examining this structure in the two-dimensional parameter space defined by the neuron's threshold and gain. We consider the stabilities of some specific sub- and super-Gaussian input distributions in this parameter plane before proving that super-Gaussian sources are always stable for any sufficiently large modulus threshold. A large, positive threshold corresponds to a sparse firing regime, but a large, negative threshold corresponds to an "anti-sparse" firing regime. Thus, while sparseness is sufficient to discover super-Gaussian sources, it is not necessary. The fixed point structure for threshold around zero is rather more complicated and idiosyncratic, but it is in this regime that sub-Gaussian inputs may be stable. We observe regimes in which both sub- and super-Gaussian sources are simultaneously stable and regimes in which neither are stable; we also observe regimes in which a Gaussian source may be stabilised by non-Gaussian sources. The conclusion of these considerations is that Triesch's intrinsic plasticity algorithm simply drives the neuron's threshold to large values, in the process making output firing sparse, and this is sufficient to find super-Gaussian sources. However, the neuron's threshold could be fixed at such a large value and it would still find super-Gaussian sources, so intrinsic plasticity is not in fact necessary. Moreover, by forcing the threshold large, Triesch's algorithm misses the sub-Gaussian-input-finding regime. In Section 4, we turn from standard

ICA-like inputs to Földiák bar input (Földiák, 1990), which is not whitened, centred or linearly-mixed. In this case, the threshold and gain in our model do change as synaptic strengths change. We first obtain the stability regions in the two-dimensional response parameter space in which appropriate, single-bar receptive fields are developed for a fixed non-linearity. We may then plot the trajectories of the threshold and gain parameters in this space in both Triesch's and our own model, observing how they change in relation to the single-bar stability regions. Triesch's model pushes these parameters to near criticality, in the sense that they approach the boundary at which (putative) single-bar receptive fields become unstable fixed points of the strength vector. Thus, we explain why the response parameters may be fixed during refinement but not when refinement is complete in Triesch's model. Such criticality is not generic, however, as we demonstrate a simpler system in which the fixed point values of the threshold and gain do not approach critical boundaries in parameter space. Our own model with Földiák bar input does not exhibit such dynamics, and its non-linearity may therefore be frozen at any point along the parameter trajectory, including its terminus. Finally, in Section 5, we discuss these results.

2 Sparseness, Exponential Firing Rates and Heavy-Tailed Distributions

We begin by considering Triesch's model of intrinsic plasticity coupled to Hebbian synaptic plasticity in a single, rate-based neuron (Triesch, 2007). Triesch has extended his work to a spike-based framework including spike-timing-dependent synaptic plasticity (Savin et al., 2010), but the key results are captured in a purely rate-based framework.

2.1 Implementation of Coupled Intrinsic and Synaptic Plasticity

Let the neuron receive n inputs with activities a_i, $i = 1, \ldots, n$, through n synapses of strengths v_i, $i = 1, \ldots, n$. The total input to the neuron is taken for simplicity to be $x = \boldsymbol{v} \cdot \boldsymbol{a}$, where $\boldsymbol{v}$ and $\boldsymbol{a}$ are the vectors of synaptic strengths and input activities, respectively, and the "$\cdot$" denotes the dot product. The PDF of this total input x is denoted by $f_X(x)$. Triesch uses a sigmoidal nonlinearity for the neuron's output firing rate or response function,

$$r(x) = \frac{1}{1 + e^{-(\alpha x + \beta)}}, \tag{2.1}$$

but we prefer the equivalent although somewhat more intuitive parametrisation (Elliott et al., 2008)

$$r(x) = \frac{1}{2} \left[1 + \tanh 2\gamma (x - \theta) \right], \tag{2.2}$$

where θ is the neuron's threshold or total input at semi-saturation, and γ is the gain of the response function at semi-saturation. Eq. (2.2) is identical to Eq. (2.1) when we set $\alpha = 4\gamma$ and $\beta = -4\gamma\,\theta$. We denote the PDF of the neuron's output firing rate by $f_R(r)$. This output PDF depends on both the PDF of the total synaptic input, $f_X(x)$, and the two response parameters γ and θ. Because $r(x)$ is a monotonic function in x, the PDFs $f_R(r)$ and $f_X(x)$ are related through the equation $f_X(r) = \frac{dr(x)}{dx} f_R(r)$.

Triesch implements intrinsic plasticity by adapting the two parameters α and β (or our γ and θ) to bring the output PDF $f_R(r)$ as close as possible, according to some suitable measure, to some target output PDF, which we denote by $g_R(r)$. Because some experimental evidence suggests that cortical neurons exhibit an exponential output firing rate distribution (Baddeley et al., 1997; but see Franco et al. (2007) and Lehky et al. (2011) for evidence for sparse but non-exponential firing rate distributions), and because an exponential distribution has maximum entropy on an unbounded interval, Triesch sets $g_R(r) = \hat{\mu}^{-1} \exp(-r/\hat{\mu})$, where $\hat{\mu}$ is the neuron's desired mean output firing rate.[2] In earlier work, he simply adapts α and β so that the mean and variance of the output firing rate match those of an exponential distribution

[2]Of course, $r(x)$ in Eq. (2.1) is bounded in $[0, 1]$ and so the corresponding maximum entropy distribution is in fact the uniform distribution. However, a uniform distribution on $[0, 1]$ has a fixed mean of $1/2$, while Triesch requires a distribution with an adjustable mean.

with parameter $\hat{\mu}$ (Triesch, 2005a). Later, he adapts α and β by minimising the Kullback-Leibler divergence between $f_R(r)$ and $g_R(r) = \hat{\mu}^{-1} \exp(-r/\hat{\mu})$ (Triesch, 2005b, 2007), defined by

$$\mathcal{D}[f_R \| g_R] = \int dr\, f_R(r)\, \log_e \left[f_R(r)/g_R(r) \right]. \tag{2.3}$$

Implementing gradient descent in $\mathcal{D}[f_R \| g_R]$ by setting

$$\frac{d\alpha}{dt} = -\varepsilon_{\rm ip} \frac{\partial \mathcal{D}}{\partial \alpha}, \tag{2.4}$$

$$\frac{d\beta}{dt} = -\varepsilon_{\rm ip} \frac{\partial \mathcal{D}}{\partial \beta}, \tag{2.5}$$

where $\varepsilon_{\rm ip}$ is a "learning rate" that sets the overall rate of intrinsic plasticity, it is routine to confirm Triesch's results (Triesch, 2007),

$$\frac{\partial \mathcal{D}}{\partial \alpha} = -\left\langle \alpha^{-1} + x \left[1 - 2\, r(x) \right] \right\rangle_X + \hat{\mu}^{-1} \left\langle x\, r(x) \left[1 - r(x) \right] \right\rangle_X, \tag{2.6}$$

$$\frac{\partial \mathcal{D}}{\partial \beta} = -\left\langle 1 - 2\, r(x) \right\rangle_X + \hat{\mu}^{-1} \left\langle r(x) \left[1 - r(x) \right] \right\rangle_X, \tag{2.7}$$

where $\langle \rangle_X$ denotes an average over the distribution of the total input x. For an online or stochastic learning rule, this averaging may be discarded provided that $\varepsilon_{\rm ip}$ is small enough. In Eqs. (2.6) and (2.7) we have kept the contributions from $\int dr f_R(r) \log_e f_R(r)$ (the first term on the right hand sides) and $\int dr f_R(r) \log_e g_R(r)$ (the second term on the right hand sides) separate for clarity as we will consider a non-exponential target output firing rate PDF $g_R(r)$ later. Similar expressions for the adaptation of the parameters γ and θ may also be obtained either by directly recomputing $\partial \mathcal{D}/\partial \gamma$ and $\partial \mathcal{D}/\partial \theta$ or much

more simply by using the chain rule. We obtain

$$\frac{d\gamma}{dt} \equiv -\varepsilon_{\rm ip}\,\frac{\partial\mathcal{D}}{\partial\gamma} = 4\left(\frac{d\alpha}{dt} - \theta\,\frac{d\beta}{dt}\right), \tag{2.8}$$

$$\frac{d\theta}{dt} \equiv -\varepsilon_{\rm ip}\,\frac{\partial\mathcal{D}}{\partial\theta} = -4\,\gamma\,\frac{d\beta}{dt}. \tag{2.9}$$

Finally, Hebbian synaptic plasticity is implemented in the standard way, by writing

$$\frac{d\boldsymbol{v}}{dt} = \varepsilon_{\rm sp}\,\mathbb{P}_{\boldsymbol{v}}\left[\boldsymbol{a}\,r(\boldsymbol{v}\cdot\boldsymbol{a})\right], \tag{2.10}$$

with the strength vector normalised on the unit hypersphere $\boldsymbol{v}\cdot\boldsymbol{v} = 1$. The projection matrix $\mathbb{P}_{\boldsymbol{v}} = \mathbb{I} - \boldsymbol{v}\,\boldsymbol{v}^{\rm T}$, with $\mathbb{I}$ being the $n\times n$ identity matrix and a superscript T denoting the transpose, implements multiplicative synaptic normalisation by projecting any growth of the strength vector off this hypersphere radially back onto it. When the synaptic plasticity learning rate $\varepsilon_{\rm sp}$ is small enough so that large fluctuations are suppressed, Eq. (2.10) may safely be replaced by

$$\frac{d\boldsymbol{v}}{dt} = \varepsilon_{\rm sp}\,\mathbb{P}_{\boldsymbol{v}}\left\langle\boldsymbol{a}\,r(\boldsymbol{v}\cdot\boldsymbol{a})\right\rangle_A, \tag{2.11}$$

where $\langle\rangle_A$ denotes an average over the multivariate PDF $f_A(\boldsymbol{a})$ defining the n inputs' joint activity patterns. The final, stable strength vector is then a solution of $\mathbb{P}_{\boldsymbol{v}}\left\langle\boldsymbol{a}\,r(\boldsymbol{v}\cdot\boldsymbol{a})\right\rangle_A = \boldsymbol{0}$ ($\boldsymbol{0}$ being the zero vector) when the output response parameters γ and θ have stabilised, i.e. when $d\gamma/dt$ and $d\theta/dt$ are also (on average) zero.

2.2 Sparse, Exponential Firing Finds Super-Gaussian Sources

By considering the limit in which intrinsic plasticity is much faster than synaptic plasticity, or $\varepsilon_{\text{ip}} \ll \varepsilon_{\text{sp}}$, Triesch argues that the distribution of $r(x) \equiv r(\boldsymbol{v}\cdot\boldsymbol{a})$ will reach, or be as close as possible to, the target exponential distribution $g_R(r) = \hat{\mu}^{-1}\exp(-r/\hat{\mu})$ before synaptic strengths can change significantly. If the output firing rate is sparse, achieved by setting $\hat{\mu} \ll 1$, then $\langle \boldsymbol{a}\, r(\boldsymbol{v}\cdot\boldsymbol{a})\rangle_A$ will be dominated by those inputs that generate the largest responses $r(x)$. Such responses will arise for the input distributions with the heaviest tails and thus we might expect that the combination of synaptic and intrinsic plasticity will lead to the strength vector $\boldsymbol{v}$ converging on one of these inputs. Triesch supports this argument in simulation, with $\hat{\mu} = 1/10$, by considering two independent inputs, one drawn from a Laplace distribution and the other from a uniform distribution, and showing that his algorithm converges on the Laplace input (Triesch, 2007). Similar results are in fact observed regardless of the size of ε_{ip} relative to ε_{sp} (Triesch, 2007). We shall explain this insensitivity to the relative scales of ε_{ip} and ε_{sp} in Section 3. This convergence to the Laplace input generalises to the case in which the inputs are generated by mixing sources $s_1, \ldots, s_n$ via an orthogonal mixing matrix $\mathbb{M}$ (with $\mathbb{M}^{\mathrm{T}}\mathbb{M} = \mathbb{I}$), so that $\boldsymbol{a} = \mathbb{M}\boldsymbol{s}$, where $\boldsymbol{s}$ is the vector of sources. In this case, the strength vector converges on a row of $\mathbb{M}^{-1}$, which is just the same as a column of $\mathbb{M}$ for orthogonal $\mathbb{M}$. The total input $x = \boldsymbol{v}\cdot\boldsymbol{a} = \boldsymbol{v}^{\mathrm{T}}\mathbb{M}\boldsymbol{s}$ then corresponds to precisely

one source, so that the algorithm converges on the Laplace source when the other sources are uniformly-distributed. With this understood, we will restrict for simplicity to the unmixed case $\mathbb{M} = \mathbb{I}$ in the following.

Does Triesch's algorithm converge on the heaviest-tailed distribution specifically, as competitive dynamics under synaptic normalisation may lead us to suspect, or on any heavy-tailed distribution amongst the inputs more generally? The standard measure of the heaviness of a distribution's tails is its (excess) kurtosis or fourth order cumulant. A Laplace distribution has kurtosis 3 (so super-Gaussian) while a uniform distribution has kurtosis $-6/5$ (so sub-Gaussian). A logistic or sech-squared distribution is also super-Gaussian, but has a smaller kurtosis of $+6/5$ compared to the Laplace distribution's kurtosis of 3. When we implement Triesch's algorithm in the presence of these two distinct super-Gaussian input distributions, we find that it converges on either one of these inputs and not exclusively on the Laplace input with the heavier tail (Fig. 1A–D), although the basin of attraction around the logistic input is smaller than that around the Laplace input. Furthermore, if we consider two differing sub-Gaussian inputs, say the uniform distribution and a binary-valued distribution taking only the values of ± 1 with probabilities $1/2$, which has the smallest possible value of -2 for its kurtosis,[3] then we find that

[3]In order to avoid binary-valued inputs, we could consider two narrow Gaussians centred around -1 and $+1$, suitably scaled and normalised to obtain precisely unit variance, and then draw from one or other of these Gaussians

Triesch's algorithm converges on neither of these input channels (Fig. 1E–F). Of course, sub-Gaussian inputs would not conventionally be regarded as heavy-tailed, but differing sub-Gaussian inputs will have tails of differing heavinesses and the heaviest-tailed inputs should induce the largest neuronal responses at their extremes. Moreover, we might expect that a neuron's gain should compensate for tail heaviness, especially for fixed input mean and variance, by adjusting the neuron's dynamic range to the range of its input distribution so that a more heavily-tailed distribution would induce a smaller gain, while a less heavily-tailed distribution would induce a larger gain (cf. Kvale and Schreiner, 2004). These observations suggest that Triesch's algorithm is not finding the heaviest-tailed distributions in the inputs, as the exponentially-distributed output firing rate argument might suggest, but rather is finding any specifically super-Gaussian distribution in the inputs in general.

FIGURE 1 ABOUT HERE

It is worth noting here for later reference that the final values of the response parameters γ and θ, shown in Fig. 1, are relatively insensitive to the precise structure of the input statistics, i.e. the final values are all similar, regardless of whether the algorithm converges on a Laplace input or a logistic input, or indeed does not converge on a single input channel at all. Indeed, we see that the final values are largely established before the strength vector begins to converge on its final direction. As the strength vector converges on its final

equiprobably, but the result would be identical.

direction, there are some changes in the response parameters, but the changes are relatively small. Since the inputs are centred and whitened, their common means and variances are zero and unity, respectively. These identical lowest-order input statistics explain in part the relative insensitivity of γ and θ to the input statistics. However, we also see from Eqs. (2.6) and (2.7) that the evolution of α and β (and therefore γ and θ) is governed by both input *and* output firing rates. In fact, the evolution of β is governed *only* by output firing rates. Only the evolution of α has a contributing term that depends directly on the input firing rate without accompanying, multiplying factors of the output firing rate. Thus, the exponential output firing rate distribution with a fixed, target mean must also contribute in part to the relative insensitivity of γ and θ to the input statistics. Specifically, the choice of the target output mean $\hat{\mu}$ will certainly strongly influence the final value of the threshold θ.

2.3 Relaxing the Requirement for an Exponential Firing Rate Distribution

To what extent does Triesch's algorithm require that the intrinsic plasticity mechanism generates an exponentially-distributed output firing rate distribution quite specifically as opposed merely to sparse output firing more generally? To address this question, we now consider a non-exponential form for $g_R(r)$ but one that nevertheless encodes the requirement for sparseness of the output firing rate (cf. Franco et al., 2007; Lehky et al., 2011). Perhaps the simplest

choice for $g_R(r)$ to achieve sparseness is

$$g_R(r) = \begin{cases} g_0 & \text{for } 0 \leq r < \hat{\theta} \\ g_1 & \text{for } \hat{\theta} < r \leq 1 \end{cases}, \tag{2.12}$$

where $\hat{\theta} \in (0,1)$ is a threshold that determines the transition point between a higher likelihood, g_0, for smaller r and a lower likelihood, g_1, for larger r. Normalising the PDF and setting its mean to $\hat{\mu}$ require that g_0 and g_1 are given by

$$g_0 = 1 + \frac{1-2\hat{\mu}}{\hat{\theta}}, \tag{2.13}$$

$$g_1 = 1 - \frac{1-2\hat{\mu}}{1-\hat{\theta}}. \tag{2.14}$$

The condition that $g_0 > g_1$ translates into $\hat{\mu} < 1/2$, which is intuitive, while the condition that $g_1 > 0$ translates into $\hat{\theta} < 2\hat{\mu}$. This latter condition can be satisfied simply by setting $\hat{\theta} = \hat{\mu}$ and simultaneously removing a degree of freedom. Although perhaps the simplest choice for $g_R(r)$, the hard threshold creates difficulties in deriving $\partial\mathcal{D}/d\alpha$ and $\partial\mathcal{D}/\partial\beta$. To derive $\partial\mathcal{D}/d\alpha$ and $\partial\mathcal{D}/\partial\beta$ we therefore replace the step function with a sharp sigmoidal non-linearity,

$$g_R(r) = \frac{g_1+g_0}{2} - \frac{g_0-g_1}{2}\tanh\left[\frac{2\hat{\gamma}}{g_0-g_1}\left(r-\hat{\theta}\right)\right], \tag{2.15}$$

where $\hat{\gamma}$ sets the sharpness of the transition around $r = \hat{\theta}$. Normalisation and setting the mean of this distribution to $\hat{\mu}$ requires solving horribly implicit, transcendental equations to obtain g_0 and g_1. However, for $\hat{\gamma} \gg 1$ the solutions in Eqs. (2.13) and (2.14) provide extremely good approximations that serve us very well. Re-deriving the equations for α and β for this different choice of

$g_R(r)$, we obtain

$$\begin{aligned}
\frac{\partial \mathcal{D}}{\partial \alpha} &= -\left\langle \alpha^{-1} + x\,[1 - 2\,r(x)] \right\rangle_X \\
&\quad + \left\langle x\,r(x)\,[1 - r(x)]\,\hat{\gamma}\,\mathrm{sech}^2 \left\{ \frac{2\hat{\gamma}}{g_0 - g_1} \left[r(x) - \hat{\theta} \right] \right\} \right\rangle_X , \quad (2.16) \\
\frac{\partial \mathcal{D}}{\partial \beta} &= -\left\langle 1 - 2\,r(x) \right\rangle_X \\
&\quad + \left\langle r(x)\,[1 - r(x)]\,\hat{\gamma}\,\mathrm{sech}^2 \left\{ \frac{2\hat{\gamma}}{g_0 - g_1} \left[r(x) - \hat{\theta} \right] \right\} \right\rangle_X . \quad (2.17)
\end{aligned}$$

Comparing these to Eqs. (2.6) and (2.7), the second terms on the right hand sides of Eqs. (2.16) and (2.17) have each acquired an additional factor that essentially acts like a Dirac delta function for $\hat{\gamma} \gg 1$.

Implementing this modified form of Triesch's algorithm, we find virtually identical results to those discussed or obtained above with an exponentially-distributed $g_R(r)$ (Fig. 2). The exponentially-distributed output firing rate is therefore irrelevant to the precise details of these results, suggesting that sparse output firing is much more important than the precise shape of the output firing rate distribution for higher firing rates. We remark that for the conservative choice of $\hat{\gamma} = 10$ used to generate Fig. 2, g_0 and g_1 cannot strictly be set according to Eqs. (2.13) and (2.14), because this leads to a poorly normalised target PDF in Eq. (2.15).[4] Nevertheless, we see from Fig. 2C that the final output PDF (which is necessarily normalised correctly because the multivariate input PDF is correctly normalised) is in fact closer to the preferred, step PDF in

[4]We note in passing that Triesch's choice $g_R(r) = \hat{\mu}^{-1} \exp(-r/\hat{\mu})$ is itself not correctly normalised on the actual output interval $r \in (0, 1)$, although the error is small for $\hat{\mu} \ll 1$.

Eq. (2.12) than to the more tractable, target PDF in Eq. (2.15). Regardless of these details, the final output firing rate PDF, however it has been acquired, is sparse but distinctly non-exponentially distributed, and this is all we require to establish the irrelevance of a specifically exponentially-distributed output firing rate distribution to the performance of Triesch's model of intrinsic plasticity.

FIGURE 2 ABOUT HERE

2.4 Relaxing the Sparseness Requirement

To what extent, then, is sparseness, achieved by setting the desired mean output firing rate to $\hat{\mu} = 1/10$, critical to these results? Reverting back for simplicity to the exponential form $g_R(r) = \hat{\mu}^{-1} \exp(-r/\hat{\mu})$ but instead setting $\hat{\mu} = 1/2$ to move the output neuron away from a sparse firing regime, we find essentially opposite results to those above (Fig. 3). Now, super-Gaussian inputs (or sources) are never found. Instead the algorithm only converges on any one of the sub-Gaussian inputs (or sources) that may be present; when sub-Gaussian inputs are not present, the algorithm does not converge on any single input channel. Again, we note that although the final values of the response parameters γ and θ differ from those obtained in the sparse coding regime with $\hat{\mu} = 1/10$ in Fig. 1, with $\hat{\mu} = 1/2$ in Fig. 3 they are still relatively insensitive to which of the sub-Gaussian inputs the algorithm converges or to whether it converges to any single input channel at all. Their final values are still, in this non-sparse regime, close to those established even before the

strength vector has begun to converge on its final direction. Again, this is due at least in part to the use of centred, whitened inputs.

FIGURE 3 ABOUT HERE

In order to examine the dependence of the final values of γ and θ on the target output mean firing rate $\hat{\mu}$, we use Eqs. (2.8) and (2.9) to find the fixed point locations for γ and θ for a given specification of the distribution $f_X(x)$ and choice of $\hat{\mu}$. When the strength vector has converged on an input, the distribution of the total input x to the neuron is precisely the distribution of the input (or source) on which the strength vector has converged. We consider two cases, one with a Laplace input and the other with a binary-valued input, because the corresponding distributions are very different in terms of their higher-order statistical structure, one being super-Gaussian and the other being sub-Gaussian. For the Laplace input, we numerically evaluate the integrals over its PDF and then numerically find the fixed point locations for $d\gamma/dt = 0$ and $d\theta/dt = 0$ for different mean output firing rates $\hat{\mu} \in (0, 1)$. For the binary-valued input, the integrals over its PDF collapse and we find that the fixed point locations are the solutions of the two equations

$$\hat{\mu}\cosh[4\gamma(1-\theta)] - 4\gamma\hat{\mu}\sinh[4\gamma(1-\theta)] = -\hat{\mu} + 2\gamma, \tag{2.18}$$

$$\hat{\mu}\cosh[4\gamma(1+\theta)] - 4\gamma\hat{\mu}\sinh[4\gamma(1+\theta)] = -\hat{\mu} - 2\gamma. \tag{2.19}$$

Although we may obtain explicit solutions for θ in terms of γ, the resulting equations in γ are transcendental and must be solved numerically. The fixed point locations for γ and θ for these two forms of input statistics are shown

in Fig. 4. We see that the corresponding solutions for the gain γ and the corresponding solutions for the threshold θ for each form of input are very similar, despite the inputs' statistics being radically different in terms of the higher-order statistical structure to which standard ICA algorithms are typically exquisitely sensitive. The fixed point locations are therefore at least in part determined by the mean output firing rate $\hat{\mu}$, although the use of non-zero-mean and non-whitened inputs would of course also affect their locations. For smaller (sparser) values of $\hat{\mu}$, the thresholds are higher and the gains are lower, while for larger (less sparse) values of $\hat{\mu}$, the thresholds are lower and the gains are higher. Although the variations in the thresholds as a function of $\hat{\mu}$ are quite large, those in the gains are quite small.

FIGURE 4 ABOUT HERE

3 Maintaining a Neuron's Operating Point by Adapting to Synaptic Strength Changes

Above, we examined Triesch's model of intrinsic plasticity working in concert with synaptic plasticity (Triesch, 2007). Although Triesch reported that his model acts as a heavy-tailed distribution detector, based on the exponentially-distributed output firing rate argument, we saw that the exponential output distribution is, as a matter of fact, irrelevant to his results. Rather, sparseness of output firing appears to be the critical property, and that instead of finding the heaviest-tailed inputs from a set of inputs in the sparse firing regime, the

model actually finds any super-Gaussian input. When we then relaxed the sparse firing requirement by increasing the mean output firing rate, we also found that the model switches to a regime in which it finds sub-Gaussian rather than super-Gaussian inputs. Finally, we observed that the fixed point values of the response parameters γ and θ are relatively insensitive to the input distribution. This is due in part to the use of standard, ICA-like inputs (i.e. centred and whitened), but also because Triesch's algorithm adapts γ and θ to achieve a target output firing rate distribution with a given, specified mean $\hat{\mu}$, and thus defined output firing rate statistics. By adopting a somewhat different perspective, we now shed further light on these results and in the process show that Triesch's model of intrinsic plasticity is itself, somewhat ironically, entirely irrelevant to almost all these results. Almost all these results are a direct consequence of a non-linear response function, $r(x)$, as the partitioning of the dynamics into sub- and super-Gaussian-input-finding, with its suggestion of ICA, might lead us to suspect.

3.1 Adapting θ and γ to the Statistics of a Neuron's Total Input

Consider a scenario in which a neuron is forced into a sparse firing regime but receives only sub-Gaussian inputs, as in Fig. 1E, or contrariwise is forced into a non-sparse firing regime but receives only super-Gaussian inputs, as in Fig. 3E. In these cases, the strength vector will not converge on any one of

its input channels. The neuron then discovers nothing about the regularities in its environment precisely because it has adapted its output statistics in a manner that largely ignores its input statistics. One of the major features of sensory neurons, however, is that they adapt to their input statistics, changing their thresholds and gains dynamically and rapidly as their input statistics change (Barlow and Mollon, 1982; Shapley and Enroth-Cugell, 1984; Meister and Berry, 1999; Kvale and Schreiner, 2004; Zaghloul et al., 2005; Bonin et al., 2006; Dean et al., 2008).

In previous work, we proposed a phenomenological adaptation principle that allows a neuron to maintain an (approximately) invariant output firing rate PDF in the face of changing input statistics (Elliott et al., 2008). We called this invariant output firing rate PDF the neuron's "operating point". For total input x with mean μ and variance σ^2, we showed that for a wide range of simple input PDFs, if a neuron has response function $r(x)$ with threshold θ and gain γ as in Eq. (2.2) [or in fact any similarly parametrised response function depending only on the particular combination $\gamma(x-\theta)$], then setting

$$\theta = \mu + \Theta\,\sigma, \tag{3.1}$$

$$\gamma = \Gamma/\sigma, \tag{3.2}$$

will leave the neuron's output firing PDF invariant. The parameters Γ and Θ are constants intrinsic to the neuron that determine the neuron's preferred operating point. For more general input statistics, these rules will result in only approximate invariance, but we suggested that real neurons, in the face

of intrinsic noise and stochasticity, may not need to maintain precise invariance and that these rules may be good enough for most practical purposes (Elliott et al., 2008).

In that work, we only considered adaptation to changes in the inputs' statistics encoded in the multivariate PDF $f_A(\boldsymbol{a})$, i.e. we essentially ignored the synaptic strengths. This is because adaptation occurs on a time-scale much faster than synaptic plasticity and we were concerned only with a model of adaptation to changes in sensory input. The total input $x = \boldsymbol{v} \cdot \boldsymbol{a}$ that the neuron receives, however, corresponds to these sensory inputs $\boldsymbol{a}$ filtered through the synaptic strengths $\boldsymbol{v}$, and changes in synaptic strengths will of course also modify the statistics of the total input x. We therefore propose as an alternative model of intrinsic plasticity that a neuron should set its threshold and gain according to Eqs. (3.1) and (3.2) where μ and σ change not only as input statistics change but also as synaptic strengths change. If the input statistics are fixed, then changes in μ and σ will directly reflect changes in synaptic strengths, and the neuron will adapt its response function to maintain an (approximately) invariant output PDF in the face of changes in synaptic strength induced by ongoing synaptic plasticity.

Ironically, for whitened and centred inputs with synaptic strengths normalised on the unit hypersphere $\boldsymbol{v} \cdot \boldsymbol{v} = 1$, with adaptation occurring essentially instantaneously compared to the much slower changes in synaptic strengths, μ and σ are independent of the synaptic strengths:

$$\mu \equiv \langle \boldsymbol{v} \cdot \boldsymbol{a} \rangle_A = \boldsymbol{v} \cdot \langle \boldsymbol{a} \rangle_A \equiv 0, \tag{3.3}$$

and

$$\begin{aligned}\sigma^2 &\equiv \left\langle \boldsymbol{v}^T \left(\boldsymbol{a}\boldsymbol{a}^T - \mu^2 \mathbb{I}\right) \boldsymbol{v} \right\rangle_A \\ &= \boldsymbol{v}^T \left\langle \left(\boldsymbol{a}\boldsymbol{a}^T - \mu^2 \mathbb{I}\right) \right\rangle_A \boldsymbol{v} \\ &= \boldsymbol{v}^T \mathbb{C}\, \boldsymbol{v} \equiv 1, \end{aligned} \tag{3.4}$$

since the covariance matrix $\mathbb{C} \equiv \mathbb{I}$ for whitened inputs. Centring and whitening therefore fix the very first and second order input statistics of the total input x to which we propose that a neuron should adapt. Be that as it may, we are forced to consider this white scenario if we are to shed further light on the analysis in Section 2, because the inputs used in that case are zero-mean and white. In Section 4, we will consider a standard problem in which centring and whitening of the inputs is not performed, so that adaptation to synaptic strength changes does occur. Such inputs are biologically much more realistic than the standard, ICA-like, centred and whitened inputs, so our analysis in Section 4 also permits us to consider these more realistic scenarios.

Although $\mu \equiv 0$ and $\sigma \equiv 1$ for whitened inputs, the neuron must nevertheless maintain running estimates of μ and σ in order to set θ and γ appropriately according to Eqs. (3.1) and (3.2). Defining the quadruple $(\mu, \sigma; \Theta, \Gamma)$, which reflects both the relevant total input statistics and the neuron's preferred operating point, it is easy to see that there is an equivalence, in terms of the induced values of θ and μ, between different sets of quadruples:

$$(\mu', \sigma'; \Theta', \Gamma') \longleftrightarrow \left(\mu, \sigma; \frac{\mu' - \mu}{\sigma} + \frac{\sigma'}{\sigma}\Theta', \frac{\sigma}{\sigma'}\Gamma'\right). \tag{3.5}$$

So, if the estimated total input mean μ' and standard deviation σ' at an oper-

ating point defined by Θ' and Γ' are shifted over time to the actual total input mean μ and standard deviation σ, then the neuron's response function $r(x)$ would be unchanged if the neuron's preferred operating point is also shifted over time from Θ' and Γ' to $\Theta = \frac{\mu'-\mu}{\sigma} + \frac{\sigma'}{\sigma}\Theta'$ and $\Gamma = \frac{\sigma}{\sigma'}\Gamma'$. If the response function is unchanged, then of course the fixed points of the synaptic strength vector under the synaptic plasticity rule in Eq. (2.11) would also be unchanged. For whitened and centred inputs, we may therefore also ignore the processes by which the neuron estimates μ and σ and simply examine the fixed point structure induced by the synaptic plasticity rule in Eq. (2.11), for $\mu = 0$ and $\sigma = 1$, as a function of the operating point parameters Θ and Γ. In this white case, we have simply $\gamma = \Gamma$ and $\theta = \Theta$, so that

$$r(x) = \frac{1}{2}\left[1 + \tanh 2\Gamma(x - \Theta)\right]. \tag{3.6}$$

In this section, it thus suffices to examine the dependence of the synaptic strength vector's fixed point locations and stabilities on this fixed response non-linearity as a function of the operating point parameters Θ and Γ. This analysis of course carries over to Triesch's model with $\gamma = \Gamma$ and $\theta = \Theta$ but with intrinsic plasticity switched off.

3.2 Super- and Sub-Gaussian Source Directions are Stable and Unstable, Respectively, for Large $|\Theta|$

For white and independent inputs, so that $f_A(\boldsymbol{a}) = \prod_{i=1}^{n} f_{A_i}(a_i)$ where $f_{A_i}(a_i)$ is the PDF of input i, a simple and standard calculation shows that the n

strength vectors $\boldsymbol{v} = \boldsymbol{e}_i$, $i = 1, \ldots n$, where the jth component of $\boldsymbol{e}_i$ is δ_{ij} (the Kronecker delta function), are all fixed points of Eq. (2.11).[5] This result holds for any response function $r(x)$, not just the sigmoidal response function in Eq. (3.6). If the inputs are generated by orthogonally mixing independent sources, then the fixed points are just the n columns of $\mathbb{M}$. Again, we consider only the $\mathbb{M} = \mathbb{I}$ case for simplicity. To determine the stabilities of these fixed points, we linearise Eq. (2.11) in perturbations around them as usual and find that $\boldsymbol{v} = \boldsymbol{e}_i$ is linearly stable if

$$\langle r'(a_i) - a_i\, r(a_i) \rangle_{A_i} < 0, \tag{3.7}$$

where $r'(x)$ denotes the derivative of $r(x)$ with respect to its argument, and again this result is valid for any form of $r(x)$. Of course, this is a classic result in ICA (Hyvärinen et al., 2001). At these fixed points, the total input x is precisely a_i for some $i \in \{1, \ldots, n\}$ and $f_X(x)$ is precisely $f_A(a_i)$. For notational simplicity, in order to avoid having to specify input i, we will therefore consider the quantity $\langle r'(x) - x\, r(x) \rangle_X$ with the understanding that x and X refer to some particular input and its corresponding distribution. Since

$$r'(x) - x\, r(x) = e^{+x^2/2} \frac{d}{dx} \left[e^{-x^2/2}\, r(x) \right], \tag{3.8}$$

---

[5] There are also a further n, sign-reversed fixed points, $\boldsymbol{v} = -\boldsymbol{e}_i$, $i = 1, \ldots n$, but we ignore this redundancy.

if X has support on the whole real line $\mathbb{R}$, then

$$\begin{aligned}\left\langle r'(x) - x\, r(x)\right\rangle_X &= +\int_{-\infty}^{+\infty} dx f_X(x) e^{+x^2/2} \frac{d}{dx}\left[e^{-x^2/2}\, r(x)\right] \\ &= -\int_{-\infty}^{+\infty} dx\, r(x) e^{-x^2/2} \frac{d}{dx}\left[e^{+x^2/2} f_X(x)\right]. \end{aligned} \quad (3.9)$$

In particular, if X is Gaussian, then $f_X(x) = \exp(-x^2/2)/\sqrt{2\pi}$ and Eq. (3.9) vanishes identically. Again, this is a classic result in ICA: the Gaussian distribution partitions the space of input (or source) distributions into sub- and super-Gaussian distributions (Hyvärinen et al., 2001).

Our task, then, is to examine the dependence of the sign of the quantity $\left\langle r'(x) - x\, r(x)\right\rangle_X$ for the particular choice of $r(x)$ in Eq. (3.6) on the operating point parameters Θ and Γ (or, equivalently, on θ and γ). We write $\Delta_X(\Gamma, \Theta) = \left\langle r'(x) - x\, r(x)\right\rangle_X$ for ease of reference. The constant term $1/2$ in $r(x)$ can be ignored because it drops out of $\Delta_X(\Gamma, \Theta)$ due to the vanishing of its derivative and the centring of the input. Because the distribution $f_X(x)$ for ICA-like inputs is usually assumed to be symmetric around $x = 0$ (so that all odd-order moments vanish), and since tanh is anti-symmetric, we in fact need only consider $\Theta \in [0, \infty)$:

$$\begin{aligned}\Delta_X(\Gamma, +\Theta) &= \frac{1}{2}\int_{-\infty}^{+\infty} dx f_X(x) e^{+x^2/2} \frac{d}{dx}\left\{e^{-x^2/2}\, \tanh\left[2\Gamma(x-\Theta)\right]\right\} \\ &= \frac{1}{2}\int_{-\infty}^{+\infty} dx f_X(x) e^{+x^2/2} \frac{d}{dx}\left\{e^{-x^2/2}\, \tanh\left[2\Gamma(x+\Theta)\right]\right\}, \\ &\equiv \Delta_X(\Gamma, -\Theta), \end{aligned} \quad (3.10)$$

where the second line follows from the change of variable $x \to -x$. The fixed points therefore have the same stabilities for either $+\Theta$ or $-\Theta$.

The simplest scenario to consider is when the input is binary-valued, corresponding to a sub-Gaussian input, and has the advantage of being analytically completely tractable. The details may be found in Appendix A. For Θ around zero, this sub-Gaussian source is stable, while for Θ large enough, it is unstable. For any given value of Γ, we denote the solutions of $\Delta_X(\Theta, \Gamma) = 0$, at which transitions in source stability in the Θ–Γ plane occur, as $\Theta_0(\Gamma)$. In the limit $\Gamma \to \infty$, we find that for the positive solution, $\Theta_0(\Gamma) \to 1$, as shown in Appendix A. Fig. 5 illustrates these results by explicitly plotting $\Delta_X(\Gamma, \Theta)$ against Θ for selected values of Γ (Fig. 5A) and showing the region in the relevant part of the Θ–Γ plane in which this binary-valued, sub-Gaussian input is stable (Fig. 5B, shaded region). As Γ increases from zero, the interval around $\Theta = 0$ in which the input is stable initially decreases before increasing somewhat and then asymptoting to unity.

FIGURE 5 ABOUT HERE

For other input distributions, we must typically resort to numerical methods to obtain the solutions of $\Delta_X(\Theta, \Gamma) = 0$. For "simple" input distributions, such as the Laplace, sech-squared or uniform distributions considered above, the qualitative features of the results for the binary-valued distribution carry over directly, except that the stabilities of sub- and super-Gaussian inputs are reversed in the Θ–Γ plane (Fig. 6). In particular, simple super-Gaussian inputs are unstable for Θ around zero and stable for large Θ, while this is reversed for simple sub-Gaussian inputs. We also see from Fig. 6 that as Γ becomes large,

$\Theta_0(\Gamma)$ always asymptotes to a constant value. For the binary-valued distribution, this value is unity. This asymptotic behaviour is not difficult to understand; the details may be found in Appendix B. Writing $\lim_{\Gamma\to\infty}\Theta_0(\Gamma) = \Theta_0^*$, for the Laplace, logistic and uniform distributions, we find that $\Theta_0^* = 1/\sqrt{2}$, $\Theta_0^* \approx 0.9321$ and $\Theta_0^* = 1$, respectively, in agreement with the asymptotic behaviours observed in Fig. 6.

FIGURE 6 ABOUT HERE

Although we shall see that the trends in the stabilities of simple sub- and super-Gaussian inputs around $\Theta = 0$ are not generic, the trends for large Θ are generic. That is, any super-Gaussian (respectively, sub-Gaussian) input, for Θ large enough, is stable (respectively, unstable). We provide the details of the proof of this result in Appendix C. The key step is to view $r(x)$ on a large enough scale so that it may be approximated as a step function. We find that the stability of a source direction is then dominated by the kurtosis of the source's distribution, with super-Gaussian sources (with positive kurtosis) being stable and sub-Gaussian sources (with negative kurtosis) being unstable.

3.3 Source Stability for Θ Near Zero is Highly Idiosyncratic and Distribution-Dependent

In contrast to the behaviour of $\Delta_X(\Theta, \Gamma)$ for large Θ, its behaviour for Θ around zero observed above for simple distributions is not generic. We can see this by explicitly constructing some rather more complicated sub- and super-

Gaussian input distributions. Writing $f_L(x) = \exp(\sqrt{2}\,|x|)/\sqrt{2}$ and $f_G(x) = \exp\left(-x^2/2\right)/\sqrt{2\pi}$ for the Laplace and Gaussian distributions, respectively, we may, for example, consider the PDFs

$$f_+(x) = \frac{1}{3}\sigma_+ \left[f_L(\sigma_+ x - 1) + f_L(\sigma_+ x) + f_L(\sigma_+ x + 1)\right], \tag{3.11}$$

$$f_-(x) = \frac{1}{4}\sigma_- \left[f_G(\sigma_- x - 6) + f_G(\sigma_- x - 2) + f_G(\sigma_- x + 2) + f_G(\sigma_- x + 6)\right], \tag{3.12}$$

where $\sigma_+ = \sqrt{5/3}$ and $\sigma_- = \sqrt{21}$ ensure that these distributions are normalised correctly, to unit integral, zero mean and unit variance. The PDF $f_+(x)$ corresponds to a super-Gaussian distribution with kurtosis $21/25$, while $f_-(x)$ corresponds to a sub-Gaussian distribution with kurtosis $-544/441$. For these two distributions, Fig. 7 shows graphs of $\Delta_X(\Theta, \Gamma)$ as a function of Θ for particular choices of Γ, and also shows the solutions of $\Delta_X(\Theta, \Gamma) = 0$ in the Θ–Γ plane. Although the tendency for super-Gaussian (respectively, sub-Gaussian) inputs to be unstable (respectively, stable) in the vicinity of $\Theta = 0$ is still observed, the dynamics are now interrupted by "oscillations" of reversing stability for larger values of Γ. We can obtain essentially as many such oscillations as we please by considering sufficiently complicated input distributions. For the particular choices of distributions used in Fig. 7, the behaviour at precisely $\Theta = 0$ does respect that observed for simpler distributions. However, even this is not generic: we can write down super-Gaussian distributions that are stable at $\Theta = 0$ and sub-Gaussian distributions that are unstable there. As Γ increases, the solutions in Θ of $\Delta_X(\Theta, \Gamma) = 0$ undergo bifurcations, giving

rise to new pairs of solutions. If a single pair is created, then this pair must straddle $\Theta = 0$ since solutions for Θ must occur in $\pm\Theta$ pairs.[6] Such a new solution pair will change the stability of the input at $\Theta = 0$ at the corresponding critical value of Γ at which the new pair arose. By the same argument, however, if an even number of pairs is created, then the stability of the input at $\Theta = 0$ will not change. This reasoning explains why the stability of the input at $\Theta = 0$ cannot be generic, in terms of being set entirely by the sign of the input distribution's kurtosis. Finally, for these more complicated distributions, the solutions Θ_0^* for the limit $\Gamma \to \infty$ will in general occur in multiple pairs and not just a single pair. The number of such pairs for this limit will, of course, give the number of bifurcations in the solutions as Γ increases from zero. For $f_+(x)$, these solutions are ± 0.1355, ± 0.5525 and ± 1.2780, while for $f_-(x)$, they are ± 0.3452, ± 0.5767 and ± 1.0955, agreeing with Fig. 7.

FIGURE 7 ABOUT HERE

For Γ large enough, for two inputs with distributions governed by the PDFs in Eqs. (3.11) and (3.12), the values of Θ_0^* enumerated above reveal intervals in Θ in which both inputs are simultaneously stable despite being super- and sub-Gaussian inputs, and intervals in which neither input is stable. Consider starting at large Θ and slowly dialling it down towards zero. For $\Theta > 1.2780$ the super-Gaussian input is stable and the sub-Gaussian input in unstable.

[6] If an odd number of pairs is created, then one of the pairs must straddle zero.

For $1.0955 < \Theta < 1.2780$ neither input is stable, the super-Gaussian input having turned unstable at the upper value while the sub-Gaussian input will become stable only at the lower value. For $0.5767 < \Theta < 1.0955$, the sub-Gaussian input is stable while the super-Gaussian input remains unstable. For $0.5525 < \Theta < 0.5767$, again neither input is stable, the sub-Gaussian input having become unstable at the upper limit while the super-Gaussian input has yet to become stable again. For $0.5525 < \Theta < 0.3452$, the super-Gaussian input is stable again while the sub-Gaussian input remains unstable. For $0.1355 < \Theta < 0.3452$, both inputs are simultaneously stable. Finally, for $0 \leq \Theta < 0.1355$, only the sub-Gaussian input is stable. Such complexity is not unique to the large Γ regime (compare Fig. 7B and Fig. 7D for, say, $\Gamma = 4$), nor is it particular to these more complicated distributions. For example, focusing on the lines corresponding to $\Delta_X(\Theta, \Gamma) = 0$ for the Laplace and uniform distributions in Fig. 6D, we can see that for $\Gamma \lessapprox 0.9$, there is an interval in Θ in which neither input is stable, while for $\Gamma \gtrapprox 0.9$, there is an interval in Θ in which both inputs are stable.

It is instructive to examine how the basins of attraction around fixed points change as we approach a regime in which, say, both a Laplace and a uniform input are stable. Picking the line $\Gamma = 2$ in Fig. 6D, we find that the uniform input becomes stable at $\Theta \approx 0.9769$ while the Laplace input becomes unstable at $\Theta \approx 0.7783$. Writing $\boldsymbol{v} = (\cos\phi, \sin\phi)^{\mathrm{T}}$ for $n = 2$ inputs, the Hebbian

learning rule in Eq. (2.11) becomes

$$\frac{d\phi}{dt} = \varepsilon_{\rm sp} \left\langle (-a_1 \sin\phi + a_2 \cos\phi)\, r\, (a_1 \cos\phi + a_2 \sin\phi) \right\rangle_A , \tag{3.13}$$

where, say, $i = 1$ corresponds to the Laplace input and $i = 2$ to the uniform input; the angle ϕ is the angle between the strength vector and the Laplace input. The basins of attraction around the two inputs' fixed points at $\phi = 0$ and $\phi = \pi/2$ (ignoring the other, sign-reversed fixed points) can be easily visualised by plotting $d\phi/dt$ as a function of ϕ, as shown in Fig. 8. At $\Theta = 1$, the Laplace input is stable and the strength vector will converge on it from any initial direction (except from precisely $\phi = \pi/2$ for the averaged rule). However, incipient bifurcations are apparent, as can be seen by comparing the $\Theta = 1$ curve in Fig. 8A to the $\Theta = 1.5$ curve in Fig. 8B, this latter being shown as a reference point well away from any critical behaviour. At $\Theta = 0.98$, these bifurcations have already occurred, leading to the creation of new, stable fixed points not associated with the two inputs. At these bifurcations, the basin of attraction of the Laplace input suddenly collapses, reducing in size to a relatively small angular range in synaptic strength space. As Θ passes through $\Theta \approx 0.9769$, the uniform input becomes stable. At $\Theta = 0.97$, the fixed points not associated with either of the two inputs are still present, but by $\Theta = 0.95$, they have disappeared, leaving only the stable fixed points associated with the inputs. Thus, over a rather small interval of Θ, the uniform input becomes stable while simultaneously the Laplace input, although remaining stable, loses a large angular range of its basin of attraction. Given a random starting

direction in synaptic strength space, the stable uniform input is therefore much easier to find than the stable Laplace input, in this parameter range.

FIGURE 8 ABOUT HERE

3.4 Stabilising a Single Gaussian Source Direction

As mentioned earlier, $\Delta_X(\Gamma, \Theta)$ vanishes identically for a Gaussian input in virtue of Eq. (3.9). This classic result in ICA owes its origin to the whitening of the inputs, because the whitened multivariate Gaussian distribution is spherically symmetric and thus there is no possibility for the Hebbian learning rule in Eq. (2.11) to break the symmetry between the inputs and converge on any single one of them. Indeed, any direction in synaptic strength space could in principle correspond to an input direction. Even if a single input is Gaussian with all other inputs being non-Gaussian, the Gaussian input would typically drop out entirely from a standard ICA learning rule. For example, under the maximisation of kurtosis or the maximisation of negentropy with a cubic non-linearity as an approximation to negentropy (Hyvärinen et al., 2001), the average learning rule may be written as

$$\frac{d\boldsymbol{v}}{dt} = \varepsilon_{\rm sp}\, k\, \mathbb{P}_{\boldsymbol{v}} \sum_{i=1}^{n} \kappa_i \left(\boldsymbol{v} \cdot \boldsymbol{m}_i\right)^3 \boldsymbol{m}_i, \tag{3.14}$$

where the vectors $\boldsymbol{m}_i$, $i = 1, \ldots, n$, are the n columns of the orthogonal mixing matrix $\mathbb{M}$, κ_i is the kurtosis of source i, and $k = \pm 1$ according to whether the sources are super-Gaussian $(+1)$ or sub-Gaussian (-1); $\mathbb{P}_{\boldsymbol{v}}$ is again the projection operator implementing $\boldsymbol{v} \cdot \boldsymbol{v} = 1$. If all sources are Gaussian, then $d\boldsymbol{v}/dt$

vanishes identically, but if a single source is Gaussian, then its contribution to the right hand side of Eq. (3.14) drops out completely and $\boldsymbol{v}$ will never converge on the corresponding column of $\mathbb{M}$. Nevertheless, for the ICA rule in Eq. (3.14), a single Gaussian source remains a fixed point of the dynamics but is cubicly unstable. Only by reversing the sign of the learning rule and destabilising all other non-Gaussian sources would it be possible to stabilise the single Gaussian source.

This non-convergence of an ICA learning rule to a Gaussian input is not, in fact, the case for Eq. (2.11) with the form of response function, $r(x)$, considered here. Although the stability of a Gaussian input is linearly indeterminate because $\Delta_X(\Gamma, \Theta) \equiv 0$, if we extend the stability analysis out to higher order, then we find that a Gaussian input (or source, with orthogonal mixing) can be a stable fixed point of the strength vector. Consider $n = 2$ inputs for simplicity and again write $\boldsymbol{v} = (\cos\phi, \sin\phi)^{\mathrm{T}}$, with the $\phi = 0$ direction corresponding to a Gaussian input and the orthogonal direction to some other, non-Gaussian input. Expanding the right hand side of the Hebbian learning rule to cubic order around $\phi = 0$, we have that

$$\langle(-a_1 \sin\phi + a_2 \cos\phi)\, r\,(a_1 \cos\phi + a_2 \sin\phi)\rangle_A = \xi_1\phi + \frac{1}{3!}\xi_3\phi^3 + \mathrm{O}(\phi^5), \quad (3.15)$$

where

$$\xi_1 = \left\langle r'(a_1) - a_1 r(a_1)\right\rangle_{A_1}, \quad (3.16)$$

$$\xi_3 = \left\langle a_1 r(a_1) + 3a_1^2\, r'(a_1) - 4r'(a_1) - 6a_1 r''(a_1) + (3+\kappa_2)\, r'''(a_1)\right\rangle_{A_1}, \quad (3.17)$$

where κ_2 is the kurtosis of the non-Gaussian input and $\langle\rangle_{A_1}$ means an average over the Gaussian distribution of the $i = 1$ input. Of course, $\xi_1 \equiv 0$, as expected. Moreover, all the terms except the κ_2 term reduce to zero in ξ_3 over a Gaussian average, and we are left with $\xi_3 = \kappa_2 \, \Xi(\Gamma, \Theta)$, where

$$\Xi(\Gamma, \Theta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} dx \, r(x) \, x(x^2 - 3) \exp\left(-x^2/2\right). \tag{3.18}$$

We again have a symmetry under a change in sign of Θ, so that $\Xi(\Gamma, +\Theta) = \Xi(\Gamma, -\Theta)$, and in the limit of large Γ, we have

$$\lim_{\Gamma \to \infty} \Xi(\Gamma, \Theta) = \frac{1}{\sqrt{2\pi}} \left(\Theta^2 - 1\right) \exp\left(-\Theta^2/2\right). \tag{3.19}$$

The sign of $\Xi(\Gamma, \Theta)$ in conjunction with the sign of the other input's kurtosis determines the stability of the Gaussian input. In Fig. 9 we show the zero contour of $\Xi(\Gamma, \Theta)$ in the Θ–Γ plane, dividing the plane into a region around $\Theta = 0$ in which $\Xi(\Gamma, \Theta) < 0$ and its complement away from $\Theta = 0$ in which $\Xi(\Gamma, \Theta) > 0$. We see this zero contour asymptoting to the line $\Theta = 1$ in the large Γ limit, consistent with Eq. (3.19). For $\Xi(\Gamma, \Theta) < 0$, a super-Gaussian input will stabilise the Gaussian input while a sub-Gaussian input will destabilise it; and *vice versa* for $\Xi(\Gamma, \Theta) > 0$. Roughly speaking, then, if Θ is large enough, a super-Gaussian input will be stable but the Gaussian input will be unstable, while if Θ is small enough, a simple super-Gaussian input may be unstable but the Gaussian input will be stable. This may appear reminiscent of the scenario described above in which the sign of the ICA learning rule must be reversed in order to destabilise the non-Gaussian sources and stabilise the otherwise cubicly-unstable Gaussian source. Here, roughly speaking,

for the regime in which a super-Gaussian input is stable, a single Gaussian source is unstable and *vice versa*; similarly for a sub-Gaussian input. The key difference, however, is the existence of multistable regimes in which both super- and sub-Gaussian inputs may be simultaneously stable, and the highly distribution-dependent behaviour of the inputs' stabilities in the vicinity of $\Theta = 0$. By comparing Fig. 9 to Figs. 6 and 7, we can see that there are regions of parameter space in which, say, a super-Gaussian input and a Gaussian input can be simultaneously stable.

FIGURE 9 ABOUT HERE

3.5 Understanding Coupled Intrinsic and Synaptic Plasticity: Dynamics in the θ–γ Plane

We may now use our results in this section further to illuminate Triesch's model of intrinsic plasticity coupled to Hebbian synaptic plasticity. We may summarise our analysis in this section by stating that for the sigmoidal response non-linearity in Eq. (3.6), the directions in synaptic strength space corresponding to the inputs (or to the underlying sources for orthogonal mixing) are always fixed points of the synaptic strength vector. This is in fact true for any response non-linearity, because of the assumption of centred, whitened, statistically-independent input distributions. However, the stabilities of the inputs depend in a highly sensitive manner on the response non-linearity and therefore, for our choice of $r(x)$ in Eq. (3.6), on the operating point parameters

Γ and Θ. Despite this sensitivity, we have proved one general property: for Θ large enough, super-Gaussian inputs are stable while sub-Gaussian inputs are unstable. For large, positive Θ, the response $r(x)$ will be suppressed for all but the largest inputs that can overcome the neuron's large firing threshold. Large, positive Θ therefore corresponds to a sparse firing regime. But, there is a symmetry between large, positive Θ and large, negative Θ. For large, negative Θ, super-Gaussian inputs are also stable with sub-Gaussian inputs unstable. A large, negative Θ essentially sets the firing threshold so low that the neuron is hyper-excitable. Almost all inputs saturate its output, and only a few very strongly negative inputs can pull the total input well below the neuron's low firing threshold and prevent it from firing strongly. We may refer to this firing scenario as an "anti-sparse" firing regime. Sparseness (or perhaps better, hypo-excitability) in these approaches is therefore not a necessary condition for the stability of super-Gaussian inputs: anti-sparseness, or hyper-excitability, is also a possibility. Since $\gamma \to \Gamma$ and $\theta \to \Theta$ with white inputs, these conclusions of course carry over directly to the neuron with gain γ and threshold θ in the response function in Eq. (2.2). In Section 2, we showed that the exponential output firing rate distribution is not critical to Triesch's results. In this section, we have therefore also shown *en passant* that sparseness of output firing is not critical either. While this is a mathematical and not a biological statement, we note that increased excitability in neurons is commonly observed, for example as a result of homeostatic plasticity (Turrigiano and Nelson, 2004). Furthermore, changes in the excitability of neurons are believed to underlie, in

part, memory allocation processes (Silva et al., 2009). Increased excitability does not necessarily correspond to hyper-excitability, of course. However, how negative Θ must be in order to stabilise a super-Gaussian source via the hyper-excitability discussed here will be highly distribution-dependent, and it may not necessarily be so negative as to be biologically implausible as a scenario. Nevertheless, hyper-excitability to the extent of almost always complete saturation of a neuron's output firing rate is implausible and constitutes a merely mathematical solution without a corresponding biological reality.

How the neuron's threshold θ and gain γ are set is essentially irrelevant: for any given values of these response parameters, the stabilities of the inputs will be uniquely determined by the response function $r(x)$ and the input statistics. Although Triesch uses an intrinsic plasticity mechanism, based on adapting the output firing rate distribution to a sparse exponential distribution (enforcing $\hat{\mu} \ll 1$), all this mechanism achieves is to move the threshold of the neuron to a regime that is "large enough" in the above sense. Furthermore, for the white and centred inputs scenario considered in this section, it is clear that identical outcomes, in terms of finding stable synaptic strength vectors, would be obtained whether or not intrinsic plasticity is operating. If a neuron possesses a fixed threshold and a fixed gain corresponding to those final values that are obtained by a neuron employing intrinsic plasticity, then the stabilities of the former neuron's inputs would be identical to those of the latter's. This explains, furthermore, the insensitivity of Triesch's results to the relative sizes of $\varepsilon_{\mathrm{sp}}$ and $\varepsilon_{\mathrm{ip}}$.

For $\varepsilon_{\text{ip}} \gg \varepsilon_{\text{sp}}$, the neuron moves very quickly to relatively final values of θ and γ. Hebbian synaptic plasticity therefore operates in the presence of an essentially fixed non-linearity, and the fixed points stabilities' are set almost immediately by this non-linearity. Any slight shifts in θ and γ that do occur as the neuron focuses down on a single input are unlikely to move the neuron out of the large θ, super-Gaussian-input-finding regime. For $\varepsilon_{\text{ip}} \sim \varepsilon_{\text{sp}}$, the response non-linearity drifts as the synaptic strengths change. However, the locations of the fixed points never change despite this drifting non-linearity, because of the assumption of whitened, independent inputs. In this regime, it is likely that the strength vector and the response parameters will converge to their final states roughly simultaneously. The scenario in which $\varepsilon_{\text{ip}} \ll \varepsilon_{\text{sp}}$ is, however, much more interesting, from a dynamical point of view. Here, the synaptic strength vector will be able to converge to stable fixed points before the response non-linearity changes much. Therefore, the Hebbian learning rule merely serves to provide a read-out of the stable fixed points as the response non-linearity slowly adapts. In this case, although the final outcome will still be identical to the $\varepsilon_{\text{ip}} \gg \varepsilon_{\text{sp}}$ and $\varepsilon_{\text{ip}} \sim \varepsilon_{\text{sp}}$ cases in terms of the final stabilities of the inputs, the intermediate dynamics could see the synaptic strength vector jumping rapidly between different sets of inputs as the inputs' stabilities change because of slowly changing θ and γ. In particular, the strength vector could jump between sub- and super-Gaussian inputs as θ moves from a small θ regime to a large θ regime even when parameter regimes do not exist in which sub- and super-Gaussian inputs are simultaneously stable.

4 Földiák Bars

In the previous section, we essentially considered how the operating point parameters Γ and Θ (or equivalently γ and θ for fixed first and second order total input statistics μ and σ^2) in the fixed response non-linearity in Eq. (3.6) affect the stabilities of the fixed points corresponding to the neuron's inputs. Unsurprisingly, such a neuron performs ICA in the presence of centred, whitened and statistically-independent inputs (or sources), but its operating point parameters critically determine the stabilities of sub- and super-Gaussian inputs in a manner that is highly sensitive to the details of the inputs' distributions. The assumption of centred and whitened inputs under a synaptic strength vector normalised on the unit hypersphere $\boldsymbol{v} \cdot \boldsymbol{v} = 1$, however, results in the mean and variance of the total input $x = \boldsymbol{v} \cdot \boldsymbol{a}$ reducing to constants, $\mu = 0$ and $\sigma^2 = 1$, and so independent of changes in synaptic strengths. We could not, therefore, examine the role of adaptation to synaptic strength changes. In this section, we relax the assumptions of centring and whitening so that μ and σ^2 become strength-dependent. As synaptic strengths change, a neuron's threshold and gain will also change because of the intrinsic plasticity mechanism proposed in Eqs. (3.1) and (3.2).

Again following Triesch (2007), in order to facilitate further comparison, we consider the Földiák bars problem (Földiák, 1990). The inputs to a single neuron are taken as the activities of an $N \times N$ array of input units forming a simple model of the retina. With Triesch, we set $N = 10$. To construct an

input pattern, each row or column of this retina is set independently of all other rows and columns to be active (unit input) with probability $1/N$ or inactive (zero input) with probability $1 - 1/N$. Activities do not sum, so that the input unit at the intersection of an active row and an active column supplies only unit activity to the neuron. This non-summation at intersections turns the Földiák bars problem into a classic problem in non-linear ICA: while the "source" columns and rows are activated independently, they are not mixed together linearly. Once such an input pattern is generated on the input array, Triesch (2007) also normalises the total, linearly-summed activity over the retina to unity (i.e. $\sum_i a_i = 1$), and we follow. Such normalisation of course constitutes a much more significant non-linearity than non-summation because each column's or row's actual activity level is effectively sensitive to how many other rows and columns are active and to the number of intersections between active rows and columns. The desired outcome of Hebbian synaptic plasticity in this bars problem is that the neuron should develop a receptive field that is tuned to a single row or a single column in the retinal array.

4.1 Receptive Field Development for a Fixed Response Non-Linearity

Before examining the impact of either Triesch's or our own model of intrinsic plasticity on the emergence of an appropriate receptive field structure, we first consider the development of the neuron's receptive field in the presence

of the fixed response non-linearity with fixed threshold θ and fixed gain γ in Eq. (2.2). Specifically, we scan over the θ–γ plane to determine in which region (or regions) single-bar receptive fields are stable fixed points of the strength vector. An analytical or even purely numerical characterisation of the fixed points of this high-dimensional, non-linear problem is hard and we will endeavour to undertake this task elsewhere. For our current purposes, it suffices to run simulations in order to examine the stability of single-bar receptive fields. We consider a sample of points in the θ–γ plane. At each of these points we run 10 simulations in each of which we set the initial strength vector very close to a putative single bar receptive field. For each simulation, we determine whether the strength vector converges precisely on this putative single-bar receptive field or moves away to some other receptive field structure. Because learning rates and convergence times depend on the precise details of the response non-linearity, we increase the learning rate and increase simulation times for larger values of the threshold θ, thereby ensuring that we do not accidentally misclassify single-bar receptive fields as stable merely because the strength vector has barely moved away from them because of very slow learning.

The results are shown in Fig. 10. In the upper right hand corner of the displayed θ–γ plane, shaded in darker grey, single-bar receptive fields are stable fixed points of the synaptic strength vector. There is a minimum value of θ, around $\theta \sim 0.08$, below which putative single-bar receptive fields are unstable for all values of γ. In the thin sliver of parameter space shaded in lighter grey, putative single-bar receptive fields are not stable, but instead multi-bar

receptive fields are. A multi-bar receptive field corresponds to a strength vector tuned to at least two separate bars, of any orientation. In the unshaded region of parameter space, neither single-bar nor multi-bar receptive fields are stable. Instead, the synaptic strength vector converges on a state in which all elements are of equal size, i.e. $\boldsymbol{v} \to \boldsymbol{\omega} \equiv (1, \ldots, 1)^{\mathrm{T}}/N$. In this state, because of the input normalisation $\sum_i a_i = 1$, the neuron's total input $x = \sum_i v_i a_i$ always take the value $1/N$ whenever at least one bar is present on the input array, i.e. the neuron responds with the same output to any (non-zero) number of bars. In fact, $\boldsymbol{v} = \boldsymbol{\omega}$ is a stable fixed point of the strength vector in all displayed regions of the θ–γ plane, not just the unshaded region, as can be confirmed by taking an initial strength vector close to $\boldsymbol{\omega}$ and determining whether it then converges on or moves away from $\boldsymbol{\omega}$. Therefore, in the darker grey region, both $\boldsymbol{\omega}$ and single-bar receptive fields are stable. However, as we move from bottom left to top right of the displayed θ–γ plane, the basin of attraction of the $\boldsymbol{v} = \boldsymbol{\omega}$ stable fixed point shrinks. That is, for simulations based on random initial strength vectors, we are overwhelmingly more likely to find single-bar fixed points than all-equal fixed points in the darker grey region of response parameter space.

FIGURE 10 ABOUT HERE

We note the broad, qualitative similarity between the stable single-bar region in Fig. 10 and the stable, super-Gaussian input regions in Fig. 6, despite the absence of centring and whitening in the former compared to the latter. In

the latter case, there is an invariance under $+\Theta \leftrightarrow -\Theta$, because of the standard assumption that input PDFs are symmetric with all odd-order moments vanishing. However, for Földiák bars, the total input is always non-negative and the mean total input is positive, breaking any possible symmetry $+\theta \leftrightarrow -\theta$. For $\theta < 0$, we find that $\boldsymbol{v} = \boldsymbol{\omega}$ is always stable and putative single-bar receptive fields are always unstable. As a result, we have shown only the $\theta \geq 0$ region.

4.2 Receptive Field Development with Triesch's Model of Intrinsic Plasticity

Having characterised the single-bar stability regions in the θ–γ plane under a fixed response non-linearity, we may now consider Hebbian synaptic plasticity coupled to a mechanism for intrinsic plasticity that causes the non-linearity to adapt. As with our discussion above in Section 3, an adaptive response non-linearity will have no fundamental impact on Hebbian synaptic plasticity unless the stabilities (or indeed existence) of fixed points of the strength vector change as the non-linearity changes.

Under Triesch's model of intrinsic plasticity with $g_R(r) = \hat{\mu}^{-1} \exp(-r/\hat{\mu})$, sparseness of output firing is again imposed by setting $\hat{\mu} \ll 1$. Triesch (2007) typically sets $\hat{\mu} = 1/(2N) = 0.05$ for the Földiák bars problem, although he also considers other values, showing qualitatively very similar results unless the sparseness condition is violated. In Fig. 11, we show the evolution of the response parameters θ and γ in Triesch's model, with $\hat{\mu} = 0.05$, and setting

the initial values to $\theta = 0$ and $\gamma = 1/4$. We also show the development of the neuron's receptive field over time, where each time step corresponds to one presentation of retinal activity. The initial strength vector is set randomly, with each component drawn from the same uniform distribution, and then normalised. We consider slow Hebbian learning, with $\varepsilon_{\text{sp}} = 10^{-4}$, and fast intrinsic plasticity, with $\varepsilon_{\text{ip}} = 10^{-2}$. During the first approximately 0.6×10^6 time steps, the strength vector remains unstructured, but after an initial increase in θ and a large increase in γ (not shown in Fig. 11), the threshold gradually decreases while the gain continues to increase significantly. At around 0.6×10^6 times steps, the gain reaches its maximum and starts falling while the threshold reaches its minimum and starts rising. It is at this point that the receptive field of the neuron begins to refine, converging on a single-bar fixed point. As the receptive field converges, θ and γ stabilise around their final values, although subject to noise.

FIGURE 11 ABOUT HERE

It is highly illuminating to plot the evolution of θ and γ in the θ–γ plane with the stability regions shaded as in Fig. 10. We do this in Fig. 12, for the same mean output firing rate $\hat{\mu} = 0.05$ used in Fig. 11, but also for two other values of $\hat{\mu}$, corresponding to the non-sparse $\hat{\mu} = 0.5$ and the even sparser $\hat{\mu} = 0.005$. The initial value of θ is set to zero in all cases, while the initial gain is set to $\gamma = 1/4$. At time step 10^4, θ and γ have reached the values indicated by the small solid dots in the figure. Notice that their locations are

outside the single-bar stability region. If the non-linearities were frozen in these states, then the strength vectors in each case would converge on $\boldsymbol{\omega}$. Under intrinsic plasticity, however, the (θ, γ) pairs track north-west in the plane, finally reaching the locations indicated by the large, open circles. For $\hat{\mu} = 0.05$, this location corresponds to the minimum in θ and maximum in γ at around 0.6×10^6 times steps seen in Fig. 11; for the other values of $\hat{\mu}$, similar dynamics occur, with minima in θ and maxima in γ being attained. In all cases, at these locations, the synaptic strength vector is still essentially unstructured. For the sparse values of $\hat{\mu}$, the response non-linearity is, however, inside the single-bar stability region, while for the non-sparse value of $\hat{\mu}$, it is inside the multi-bar stability region. For the sparse values, the strength vectors subsequently converge on single-bar fixed points as the (θ, γ) pairs track south-east in the plane, following trajectories that are very similar, but in reverse, to their earlier, north-west trajectories. The reverse trajectories may be distinguished from the earlier trajectories by the presence of greater levels of noise in the latter. From Fig. 11 it is clear that these reverse trajectories require many more time steps than the earlier trajectories to complete. While there are not enough time steps in the north-west trajectories for the strength vector even to begin to move away from its unstructured state, there are enough time steps in the south-east trajectories for convergence to single-bar receptive fields. This is not a result of employing such a low synaptic plasticity learning rate, $\varepsilon_{\text{sp}} = 10^{-4}$. If instead we set $\varepsilon_{\text{sp}} = 10^{-2}$, keeping ε_{ip} unchanged at 10^{-2}, then we observe identical dynamics, except that the (θ, γ) pair does not track so far north-west in the

plane. Convergence on single-bar fixed points always occurs on the reverse trajectories, and these trajectories always terminate in the same location of the θ–γ plane for any given value of $\hat{\mu}$. For sparse values of $\hat{\mu}$, these trajectories in fact always terminate at, or very close to, the boundary separating the single-bar and multi-bar stability regions. For the non-sparse value of $\hat{\mu}$, once the (θ, γ) pair enters the multi-bar stability region, it never leaves it. By taking $\hat{\mu}$ large enough, we could obtain trajectories that never enter the multi-bar or single-bar stability regions. We note the strong dependence of the terminal, fixed point values of θ and γ on the choice of $\hat{\mu}$ and in particular that smaller values of $\hat{\mu}$ induce larger thresholds θ, despite identical input statistics for Földiák bar inputs, confirming that both input *and* output statistics contribute to the final values of θ and γ in Triesch's model.

FIGURE 12 ABOUT HERE

Triesch (2007) reports that intrinsic plasticity is not entirely necessary for developing single-bar receptive fields on the Földiák bars problems, because if θ and γ are frozen at any intermediate values during which single-bar receptive fields begin to emerge (corresponding to the reverse trajectories described above), then such receptive fields will always develop from a random initial strength vector. However, he reports that if instead the non-linearity is frozen with values of θ and γ corresponding to their final values, then single-bar receptive fields do not emerge. In fact, we find that multi-bar rather than single-bar receptive fields develop. Fig. 12 confirms and elaborates on these observations.

For the sparse values of $\hat{\mu}$, the (θ, γ) pairs are always inside the single-bar stability region on the reverse trajectories before final stabilisation. Thus, any fixed non-linearity on these portions of the trajectories will, by definition of the single-bar stability region, lead to the robust development of single-bar receptive fields. However, with the (θ, γ) pairs instead frozen at the terminal values, the response non-linearity is very close to the boundary between multi-bar and single-bar stability. Here, the basin of attraction of single-bar receptive fields is smaller than the basin of attraction of multi-bar receptive fields. For a random initial synaptic strength vector with a fixed non-linearity sitting at or close to this boundary, the synaptic strength vector is much more likely to converge to multi-bar than single-bar receptive fields.

It is intriguing that Triesch's intrinsic plasticity algorithm leads to fixed point or terminal values of θ and γ that appear in some sense to be "critical" in the Földiák bars problem, being at or very close to the boundary partitioning the θ–γ plane into stable and unstable putative single-bar fixed point regions. It is clear from Fig. 12 that this "criticality" is not exact, in the sense that the terminal values do not follow exactly the single-bar stability curve that partitions the plane. In order to understand this feature fully, it would be necessary to derive analytically the equation of the partitioning curve and to derive the equations for the fixed point values of θ and γ as a function of $\hat{\mu}$ from Eqs. (2.8) and (2.9) (coupled with Eq. (2.10) since the synaptic strength vector evolves too) after averaging over the Földiák bar input patterns. Such an undertaking would be formidably hard if not intractable. However, if

this feature is generic, then we would expect such criticality to arise in other, simpler problems, and in particular for those considered in Sections 2 and 3. In Fig. 4 we plotted the fixed point or terminal values of θ and γ as a function $\hat{\mu}$ for a Laplace input, while in Fig. 6A we plotted the stability region in the Θ–Γ plane for a Laplace input. Since the inputs are centred and whitened in that case, we have $\theta \equiv \Theta$ and $\gamma \equiv \Gamma$. Thus, in Fig. 13 we plot the terminal values of θ and γ in the θ–γ plane as a function of $\hat{\mu}$ (taking $\hat{\mu}$ down to 10^{-3} instead of 10^{-2} as in Fig. 4) with the partitioning stability curve also drawn. We clearly see no indication of criticality for this much simpler system. For $\hat{\mu} \gtrapprox 0.16$, a Laplace input is outside the stability region and is thus unstable, while for $\hat{\mu} \lessapprox 0.16$, the terminal values of the (θ, γ) pairs traverse well inside the Laplace stability region, exhibiting no tendency to remain at or near to the boundary region. The criticality exhibited in the Földiák bars problem is therefore not a general feature of Triesch's model of intrinsic plasticity coupled to synaptic plasticity: the final, adapted values of the neuron's threshold and gain do not in general home in on critical regions in parameter space in which strength vector fixed points are close to changing stability (or ceasing to exist).

FIGURE 13 ABOUT HERE

4.3 Receptive Field Development with Adaptation to Changes in Synaptic Strengths

We now consider our own model of intrinsic plasticity, based on adapting to synaptic strength changes in order to keep a neuron's operating point approximately invariant, described by Eqs. (3.1) and (3.2). As the mean μ and variance σ^2 of the total input $x = \boldsymbol{v} \cdot \boldsymbol{a}$ change due to synaptic plasticity (the retinal input statistics do not change because the retinal stimuli are always Földiák bars), the threshold θ and gain γ change according to Eqs. (3.1) and (3.2). Since adaptation is typically a very fast process, in principle μ and σ^2, and thus θ and γ, should change effectively instantaneously compared to the much slower changes in synaptic strength. For simplicity, however, we instead maintain a running average of the statistical quantities of interest so that they change quickly enough compared to synaptic strength changes but not instantaneously. To compute the running average of some time-dependent quantity $h(t)$, we employ a process with a memory governed by a time scale τ_{ad},

$$\langle h(t) \rangle = \frac{1}{\tau_{\mathrm{ad}}} \int_{-\infty}^{t} dt' \, h(t') \exp\left[-\left(t - t'\right) / \tau_{\mathrm{ad}} \right], \tag{4.1}$$

which is equivalent to the differential update rule

$$\frac{d\langle h(t) \rangle}{dt} = \varepsilon_{\mathrm{ad}} \left[h(t) - \langle h(t) \rangle \right], \tag{4.2}$$

where $\varepsilon_{\mathrm{ad}} = 1/\tau_{\mathrm{ad}}$ is the update or learning rate. We set $\varepsilon_{\mathrm{ad}} = 10^{-4}$ in order to obtain good averages without large moment-to-moment fluctuations, and as above set the synaptic plasticity rate to $\varepsilon_{\mathrm{sp}} = 10^{-4}$. Although $\varepsilon_{\mathrm{sp}} = \varepsilon_{\mathrm{ad}}$, we

find that adaptation occurs quickly enough and that it is not usually necessary to set $\varepsilon_{\text{sp}} \ll \varepsilon_{\text{ad}}$.

Unlike Triesch's model of intrinsic plasticity in which only the mean output firing rate parameter $\hat{\mu}$ can be adjusted, we have two degrees of freedom in setting the neuron's preferred operating point via the two parameters Θ and Γ. We could in principle set Θ and Γ so that θ and γ take any desired values given estimates of μ and σ when the neuron's receptive field is in an initially random, unstructured state. However, in order to facilitate comparison of our model of intrinsic plasticity to Triesch's model, we instead set the neuron's operating point so that its initial threshold and gain are close to those in Triesch's model at the north-west termini of the trajectories in Fig. 12, indicated by the large, open circles in that figure. We may then directly compare how the adaptation processes in both models change a neuron's response non-linearity as the neuron refines its receptive field down to single bars or multi-bars.

Such an example is shown in Fig. 14. We set Γ and Θ so that the neuron's initial threshold and gain approximately coincide with the north-west terminus of the $\hat{\mu} = 0.05$ trajectory in Fig. 12, or equivalently with the maximum of γ and minimum of θ at around 0.6×10^6 time steps in Fig. 11. In Fig. 14, we reproduce the evolution of θ and γ from Fig. 11 and show on the same graph the evolution of θ and γ in our model, taking the starting time of the latter to be 0.6×10^6 time steps for direct comparison of the parameters' respective evolutions in both models. We also show the emergence of the single-bar receptive field in our model at the same time steps as those in Triesch's model, except displaced

by the temporal offset of 0.6×10^6 time steps. We see that the threshold in our model does not increase so much as in Triesch's model during receptive field refinement; indeed, it exhibits a maximum before falling somewhat. Likewise, the gain in our model does not decrease so much as in Triesch's model. In Triesch's model, there are larger fluctuations around the final, stable value of θ than around γ. On the contrary, in our model, the fluctuations around the fixed point values of γ are larger than those around θ. This is because γ is a determined by a second-order statistic, σ^2, whereas θ is principally determined by a first-order statistic, μ, and the estimation of second-order statistics in inevitably more noisy than that of first-order statistics. Despite the synaptic plasticity learning rates being set equal in both models, our model's receptive field is slightly less well developed than that of Triesch's at 6.0×10^6 times steps (even accounting for the temporal offset). This is because of the second-order statistic: we can see from the gains in Fig. 14 that the gain in our model converges to its final value (modulo noise) somewhat later than the gain in Triesch's model.

FIGURE 14 ABOUT HERE

Fig. 15 plots the evolution of the (θ, γ) pair in the θ–γ plane for our model as Fig. 12 does for Triesch's model. In Fig. 15, we consider three different operating points approximately coinciding with Triesch's north-west termini for $\hat{\mu} = 0.5$, $\hat{\mu} = 0.05$ (from Fig. 14) and $\hat{\mu} = 0.005$. We notice that the trajectories in the θ–γ plane are significantly different between the two models.

In particular, for the $\hat{\mu} = 0.05$ and $\hat{\mu} = 0.005$ (or equivalent) trajectories, for which stable single-bar receptive fields are formed, in our model the final, fixed point values of θ and γ do not encroach on the boundary separating the single-bar and multi-bar stability regions. The non-linearity may, then, in our model be frozen at any point on these trajectories and single-bar receptive fields will always develop. While Triesch's intrinsic plasticity algorithm pushes θ larger and larger during receptive field refinement in order to generate sparser and sparser output firing, our model demonstrates no such tendency. Rather than adapting the neuron's output firing rate PDF to an exponential distribution, our adaptation mechanism modifies the neuron's threshold and gain in an attempt to maintain an approximately invariant output firing rate PDF despite the changes in synaptic strength that occur while the neuron's receptive field develops and refines. The two computational principles – maintaining an invariant output PDF and adapting towards an exponential output PDF – are radically different in motivation, and each may serve different roles in different contexts. As least in this Földiák bars setting, however, our model's dynamics do not lead to final response non-linearities that reach the stability boundary and do not, therefore, suffer from the risk of catastrophic, destabilising fluctuations that could drive the receptive field from a single-bar to a multi-bar state.

FIGURE 15 ABOUT HERE

5 Discussion

By using Triesch's (2007) model of intrinsic plasticity coupled to synaptic plasticity as a vehicle, we have essentially explored the fixed point structure of the synaptic strength vector of a neuron with a sigmoidal response non-linearity for a variety of different input distributions. In particular, much of our analysis above has examined a fixed rather than an adapting non-linearity. For centred, statistically-independent and orthogonally-mixed sources, the fixed points of the strength vector under any response non-linearity are well-known to be the columns of the mixing matrix itself, or just the directions in synaptic strength space singling out individual inputs in the absence of mixing (Hyvärinen et al., 2001). Moreover, the linear stabilities of these fixed points are always determined by Eq. (3.7), again for any (sufficiently well-behaved) response non-linearity (Hyvärinen et al., 2001). For a sigmoidal non-linearity with fixed threshold θ and gain γ, or operating point parameters Θ and Γ (since $\theta = \Theta$ and $\gamma = \Gamma$ for whitened inputs and with $\boldsymbol{v} \cdot \boldsymbol{v} = 1$), it has therefore sufficed to consider the stabilities in the θ–γ plane of inputs with different probability distributions. Any intrinsic plasticity mechanism that modifies the parameters defining the neuron's response non-linearity induces a trajectory of the neuron through the θ–γ plane. If the neuron does not cross any critical boundaries during this process, then the intrinsic plasticity is irrelevant: the stable fixed points of the strength vector do not change. But if the neuron crosses critical boundaries, then it can move between different input stability regions.

The key general result of our analysis is that for θ (or Θ) of large enough modulus, whitened inputs with super-Gaussian distributions are always stable fixed points of the strength vector, while sub-Gaussian inputs are always unstable. A large, positive threshold corresponds to a hypo-excitable neuron, or a sparse coding regime; while a large, negative threshold corresponds to a hyper-excitable neuron, or an anti-sparse coding regime. By enforcing sparseness with $\hat{\mu} \ll 1$ in an exponential output firing rate distribution, Triesch's model of intrinsic plasticity forces θ to become large and positive, and so the neuron may traverse into the super-Gaussian-input-finding regime. This is not, however, guaranteed. We saw that the final, fixed point values of θ and γ in Triesch's model are relatively insensitive to the input distributions for ICA-like inputs. This is partly because ICA-like inputs have zero mean and unit variance, but also because the choice of $\hat{\mu} \ll 1$ tends to drive θ large. However, precisely what constitutes the large threshold regime is, as we have seen, exquisitely sensitive to the details of the input statistics. Essentially, the maximum value of $\hat{\mu}$ that enforces "sparseness" will itself be highly distribution-dependent, although undoubtedly taking $\hat{\mu}$ very small will suffice for all but the most contrived input distributions. The finding of super-Gaussian inputs in the large modulus θ regime is not a result, however, that requires any specification of the output firing rate PDF. Although large, positive θ does constitute a sparse firing regime, an exponential output firing rate distribution is entirely irrelevant to this regime. Any (non-trivial) output firing rate distribution with an adjustable mean would drive θ large when the mean is taken to be small enough.

It is the setting of $\hat{\mu} \ll 1$ that is, as it were, doing the "heavy lifting" in driving θ large, not the shape of the output distribution *per se*. And just as the exponential output firing rate distribution is irrelevant to the performance of Triesch's intrinsic plasticity mechanism in terms of finding super-Gaussian inputs, so too, in fact, is sparseness. This is because both sparse and anti-sparse output firing regimes induce the stability of super-Gaussian inputs. In other words, a neuron can learn as much about its inputs by the rare absence of activity when it is hyper-excitable as it can by the rare presence of activity when it is hypo-excitable. Although extreme degrees of hyper-excitability may be biologically implausible, from a purely mathematical point of view, sparseness of output firing does not uniquely result in super-Gaussian-input-finding.

An intrinsic plasticity mechanism that drives θ large also risks failing to find sub-Gaussian inputs. Although the response parameter regimes in which sub-Gaussian inputs are stable appear very sensitive to the precise details of the input distributions except for simple, standard sub-Gaussian distributions such as the uniform distribution, nevertheless, if sub-Gaussian inputs are to be stable fixed points of the strength vector, then θ must be away from the large modulus regime and so somewhere around zero. Moreover, super-Gaussian inputs can also be stable for θ closer to zero.

For centred, statistically-independent and orthogonally-mixed sources, a neuron with a sigmoidal response non-linearity essentially does ICA, and the neuron's operating point (its values of Θ and Γ) critically determines its computational repertoire, i.e. which inputs are stable and which are unstable fixed

points of the strength vector. In many respects, perhaps such a result should not be too surprising, given that many results in ICA are valid for any response non-linearity (Hyvärinen et al., 2001). However, in contrast to standard implementations of ICA, a neuron with a sigmoidal response non-linearity can exhibit parameter regimes in which both sub- and super-Gaussian sources are stable; in which neither sub- nor super-Gaussian sources are stable; in which a single, Gaussian source can be extracted without destabilising all other sources. Indeed, such behaviour is very sensitive to the structure of the possibly many sub- or super-Gaussian sources providing input to a neuron, since each particular source carves the Θ–Γ (or θ–γ) plane into its own distinct but contiguous regions of stability and instability. The intersections between such patchworks of stability and instability for different sources will determine those regions in the Θ–Γ plane in which different sources (possibly sub- and super-Gaussian) will be simultaneously stable, and those regions in which no source is stable.

Conventional ICA learning algorithms tend to be extremely non-linear, based on information-theoretic principles that, for example, maximise kurtosis, negentropy or mutual information (Hyvärinen et al., 2001). From a neurobiological perspective, these learning algorithms may be criticised because they may seem difficult to implement in real, neuronal substrates. Furthermore, for any given sign of the learning rule, conventional ICA algorithms stabilise either super-Gaussian sources or sub-Gaussian sources, but not both. If input distributions consist of a mixture of sub- and super-Gaussian inputs, then only one class of input will be stable, with the sign of the learning rule having to be

reversed for the other class to be stable. Such sign changes may appear unnatural: a neuron would have to switch from Hebbian to anti-Hebbian plasticity. To be sure, ICA is a powerful body of cognate techniques, although most of the results in ICA owe their origins to two major, vastly simplifying assumptions: the statistical independence of the sources and their linear (or more specifically, their orthogonal) mixing. These assumptions are so strong that they perform all the work in establishing the fixed point structures of ICA learning rules for any (sufficiently well-behaved) response non-linearity. This is why it should not be surprising that a sigmoidal non-linearity essentially does ICA. But the differences between conventional ICA approaches and ICA as performed by a sigmoidal non-linearity are remarkable. A sigmoidal response non-linearity, while coarse, may nevertheless be a good, first approximation to real neuronal transfer functions, unlike most ICA non-linearities. Furthermore, its threshold and gain parameters are biologically well-motivated and are the targets of adaptation processes in real neurons, while no such equivalents exist in standard ICA non-linearities. It is these very response parameters that, moreover, allow such a neuron to stabilise super-Gaussian sources, sub-Gaussian sources and even a Gaussian source without the sign of the learning rule having to be changed by fiat. It is remarkable that such a simple, two-parameter system can essentially perform ICA but without any of the standard criticisms that can be levelled at standard ICA. Whether the two major assumptions of ICA have any relevance for the inputs to real neurons is unclear, but it is certainly intriguing that real neurons could perform ICA without any of the standard

complications of conventional ICA.

A critical question is to what extent do these results and observations depend on the precise form of the sigmoidal response non-linearity in Eq. (2.2)? In subsequent work, Triesch extended his analysis from a rate-based neuron (Triesch, 2007) to a spiking neuron (Savin et al., 2010), reporting essentially identical results. The spiking probability of the neuron used in this later work consists of a product of a term imposing a refractory period and a term depending on the neuron's membrane potential. This latter term contains three parameters that are targets for intrinsic plasticity, two of them being essentially threshold and gain parameters (relative to the membrane potential) and a third being an overall scale. During intrinsic plasticity, the threshold parameter is driven more negative, making it harder to depolarise the membrane, so taking the neuron into a sparse firing regime. The refractoriness term essentially imposes saturation of the spiking rate. Thus, although mathematically rather more complicated, the transfer function of this neuron will not be too dissimilar from a sigmoidal non-linearity with threshold and gain at the rate-based level. In deriving the super-Gaussian behaviour for large modulus thresholds, the key step was to look at the response function on a large enough scale so that the transition from no response to saturated response could be regarded as approximately a step function. We would therefore expect any reasonably well-behaved saturating non-linearity with parameters that are threshold-like and gain-like to exhibit similar dynamics, and Triesch's more-complicated non-linearity supports this expectation. We would expect any

saturating non-linearity to exhibit a regime of super-Gaussian-input-finding for large enough threshold. Conversely, the dynamics around zero threshold are likely to be highly idiosyncratic, depending both on the precise details of the input statistics as above, and on the precise nature of the non-linearity. It will be interesting to examine these issues more extensively in later work.

We have seen that both Triesch's and our own intrinsic plasticity mechanisms induce trajectories in the θ–γ plane, and that independently of any intrinsic plasticity mechanism, this plane is carved into stable and unstable input regions, both for classic, ICA-like inputs (i.e. whitened, independent and linearly-mixed) and for other types of inputs (e.g., for Földiák bars, which are neither whitened nor linearly-mixed). Triesch's mechanism pushes a neuron into a large θ regime in order to achieve sparseness, but these dynamics are largely insensitive to the precise details of the input statistics, at least for centred, whitened, ICA-like inputs. While our own mechanism is constructed specifically to permit adaptation to changing input statistics (Elliott et al., 2008), including the generalisation proposed here of adapting to changes in a neuron's own synaptic strengths in order to attempt to maintain an approximately invariant output PDF, nevertheless, a neuron's dynamics will be determined by its operating point parameters Θ and Γ, which are set irrespective of input statistics. Although there can be interesting interactions between synaptic plasticity and intrinsic plasticity as a neuron traverses the θ–γ plane, particularly when intrinsic plasticity is much slower than synaptic plasticity, in both models intrinsic plasticity plays second fiddle to the underlying synaptic

strength fixed point structure induced in the θ–γ plane by Hebbian synaptic plasticity. Certainly, achieving an approximately exponential output firing rate PDF or maintaining an approximately invariant output firing rate PDF may have important computational roles in downstream neuronal processing or for coding efficiency. However, such subsequent processing does not feature in either Triesch's or our own analysis. Thus, intrinsic plasticity may appear as something of a distraction from the computational properties of neurons with fixed response non-linearities. Such a view, however, would be unfortunate. Besides exploring the downstream implications of intrinsic plasticity, there are more direct possibilities to evaluate. For example, if $\hat{\mu}$ is set too small in Triesch's model or Θ is set too large in our own model, both types of neuron would fail to converge on inputs when all inputs are sub-Gaussian. A very powerful strategy for a neuron to adopt, therefore, would be to attempt to determine whether its synaptic strength vector has converged on a state that provides information about the statistical regularities in its inputs and if not, then modify its own operating point accordingly. Although the θ–γ plane would still be carved into distinct stability regions by the underlying Hebbian plasticity rule in conjunction with the input statistics, the neuron would then actively search for regions of this plane in which its strength vector could converge on states that are informative. It will be fascinating to pursue this and similar ideas in future work.

Acknowledgements: I thank Paul Adams, State University of New York at

Stony Brook, for critical comments and suggestions.

Appendix A Analysis of $\Delta_X(\Gamma, \Theta)$ for Binary-Valued Inputs

For binary-valued inputs, the integral defining $\Delta_X(\Gamma, \Theta)$ collapses, and we obtain

$$\begin{aligned} \Delta_X(\Gamma, \Theta) &= \frac{1}{2}\Gamma \left[\operatorname{sech}^2 2\Gamma\,(\Theta + 1) + \operatorname{sech}^2 2\Gamma\,(\Theta - 1)\right] \\ &\quad - \frac{1}{4}\left[\tanh 2\Gamma\,(\Theta + 1) - \tanh 2\Gamma\,(\Theta - 1)\right]. \end{aligned} \tag{A.1}$$

For $\Theta = 0$, we then have $\Delta_X(\Gamma, 0) \propto 2\Gamma \operatorname{sech}^2 2\Gamma - \tanh 2\Gamma$, and it is easy to see that $\Delta_X(\Gamma, 0) < 0$ for $\Gamma > 0$. Thus, for Θ around zero, this sub-Gaussian input is stable. Examining the limit of large Θ by writing $\tanh y \sim 1 - 2e^{-2y}$ for large y, we find that

$$\Delta_X(\Gamma, \Theta) \sim e^{-4\Gamma\,\Theta}\,(4\Gamma \cosh 4\Gamma - \sinh 4\Gamma)\,. \tag{A.2}$$

Since the right hand side is positive, this sub-Gaussian input is unstable for Θ large enough. The solutions of $\Delta_X(\Gamma, \Theta) = 0$ give the locations of the transitions between stability and instability in the Θ–Γ plane. We may find the values of Θ as a function of Γ, call them $\Theta_0(\Gamma)$, at which these transitions occur. Writing $\chi = \exp[4\Gamma\,\Theta_0(\Gamma)]$ and $y = \exp(4\Gamma)$, then χ and hence $\Theta_0(\Gamma)$ can be found from the solutions of the quadratic equation $a_2\chi^2 + a_1\chi + a_0 = 0$,

where $a_2 \equiv a_0$ and

$$\begin{aligned} a_1 &= 16\Gamma y^2 + (1 - y^4), \\ a_0 &= y\left[4\Gamma(1 + y^2) + (1 - y^2)\right], \end{aligned} \tag{A.3}$$

or

$$\Theta_0(\Gamma) = \frac{1}{4\Gamma} \log_e \frac{-a_1 \pm \sqrt{(1-y^2)^2\left[(1-y^2)^2 - 64\Gamma^2 y^2\right]}}{2a_0}. \tag{A.4}$$

The positive square root gives the single, $\Theta_0(\Gamma) > 0$ solution. It is easy to see that

$$\lim_{\Gamma \to \infty} \Theta_0(\Gamma) = 1, \tag{A.5}$$

so that in this limit, the transitions in stability occur precisely at $\Theta_0 = \pm 1$.

Appendix B Asymptotic Behaviour of $\Theta_0(\Gamma)$ for Large Γ

In order to obtain the asymptotic behaviour of $\Theta_0(\Gamma)$ for large Γ, we observe that

$$\begin{aligned} \lim_{\Gamma \to \infty} r(x) &= H\left(x - \Theta\right), \\ \lim_{\Gamma \to \infty} r'(x) &= \delta\left(x - \Theta\right), \end{aligned}$$

where $H(x)$ and $\delta(x)$ are, respectively, the Heaviside step and Dirac delta functions. In this limit (assuming interchange of the order of the limit and integration), the integral defining $\Delta_X(\Theta, \Gamma)$ simplifies, giving

$$\lim_{\Gamma \to \infty} \Delta_X(\Theta, \Gamma) = f_X(\Theta) - \int_{\Theta}^{\infty} dx\, x\, f_X(x). \tag{B.1}$$

Writing $\lim_{\Gamma\to\infty}\Theta_0(\Gamma)=\Theta_0^*$, we then have that Θ_0^* is a solution of

$$f_X\left(\Theta_0^*\right)=\int_{\Theta_0^*}^{\infty}dx\,x\,f_X(x). \tag{B.2}$$

Notice that this equation is valid whether X has support on the whole of $\mathbb{R}$ or only on a subset of $\mathbb{R}$, for if Θ_0^* were to fall outside of the support of X, then $f_X(\Theta_0^*)\equiv 0$ while the integral on the right hand side of Eq. (B.2) must be positive (for distributions $f_X(x)$ that are symmetric around $x=0$, as assumed here). The solutions of Eq. (B.2) must occur in pairs with opposite signs, and in general there may be more than precisely one pair of solutions. Usually, we must solve Eq. (B.2) numerically in order to obtain its solutions.

Appendix C Stability of Source Directions for Large Θ

For $\Theta\gg 0$, the interval over which $r(x)$ switches from 0 to 1 is, compared to Θ, small and can be ignored (regardless of the size of Γ: smaller Γ simply means considering even larger Θ). In this large Θ regime, we may then write $r(x)\approx H(x-\Theta)$. Then, from Eq. (3.9),

$$\Delta_X\left(\Theta,\Gamma\right)\approx-\int_{\Theta}^{\infty}dx\,e^{-x^2/2}\frac{d}{dx}\left[e^{+x^2/2}f_X(x)\right]. \tag{C.1}$$

Expanding $f_X(x)$ in its cumulants up to fourth order, we have

$$f_X(x)=\frac{1}{2\pi}\int d\omega\,e^{-i\omega x}\,\exp\left[-\frac{1}{2}\omega^2+\frac{\kappa}{4!}\omega^4-\cdots\right], \tag{C.2}$$

where κ is the fourth order cumulant or (excess) kurtosis. Since Eq. (C.1) is sensitive only to $x\gg 0$ as $\Theta\gg 0$, we need evaluate $f_X(x)$ in Eq. (C.2) only for

large x. In this regime, the integral in Eq. (C.2) is dominated by the behaviour of the integrand around $\omega = 0$. Thus, we may write

$$\begin{aligned} f_X(x) &\approx \frac{1}{2\pi}\int d\omega\, e^{-i\omega x}\, e^{-\omega^2/2}\left(1+\frac{\kappa}{4!}\omega^4\right) \\ &= \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\left[1+\frac{\kappa}{4!}\left(x^4-6x^2+3\right)\right], \end{aligned}$$

for large x. So,

$$\begin{aligned} \Delta_X(\Theta,\Gamma) &\approx -\frac{1}{\sqrt{2\pi}}\int_\Theta^\infty dx\, e^{-x^2/2}\frac{d}{dx}\left[1+\frac{\kappa}{4!}\left(x^4-6x^2+3\right)\right] \\ &\approx -\frac{1}{\sqrt{2\pi}}\frac{\kappa}{3!}\int_\Theta^\infty dx\, x^3\, e^{-x^2/2}. \end{aligned} \tag{C.3}$$

Thus, the kurtosis κ of any input distribution determines the sign of $\Delta_X(\Theta,\Gamma)$ for large Θ and, in particular, for a super-Gaussian distribution, with $\kappa > 0$, the input is stable, while for a sub-Gaussian distribution, with $\kappa < 0$, the input is unstable.

References

Atick, J.J. 1992. Could information theory provide an ecological theory of sensory processing? *Network: Computation in Neural Systems*, **3**, 213–251.

Atteave, F. 1954. Some informational aspects of visual perception. *Psychological Review*, **61**, 183–193.

Baddeley, R., Abbott, L.F., Booth, M.C., Sengpiel, F., Freeman, T., Wakeman, E.A., Rolls, E.T. 1997. Responses of neurons in primary and inferior temporal visual cortices to natural scenes. *Proceedings of the Royal Society of London Series B*, **264**, 1775–1783.

Barlow, H.B. 1961. Possible principles underlying the transformation of sensory messages. *Pages 217–234 of:* Rosenblith, W.A. (ed), *Sensory Communication*. Cambridge, MA: MIT Press.

Barlow, H.B., Mollon, J.D. 1982. *The Senses*. Cambridge: Cambridge University Press.

Bell, A.J., Sejnowski, T.J. 1997. The "independent components" of natural scenes are edge filters. *Vision Research*, **37**, 3327–3338.

Bonin, V., Mante, V., Carandini, M. 2006. The statistical computation underlying contrast gain control. *Journal of Neuroscience*, **26**, 6346–6353.

Brenner, N., Bialek, W., de Ruyter van Steveninck, R.R. 2000. Adaptative rescaling maximizes information transmission. *Neuron*, **26**, 695–702.

Carandini, M., Demb, J.B., Mante, V., Tolhurst, D.J., Dan, Y., Olshausen, B.A., Gallant, J.L., Rust, N.C. 2005. Do we know what the early visual system does? *Journal of Neuroscience*, **25**, 10577–10597.

Dean, I., Robinson, B.L., Harper, N.S., McAlpine, D. 2008. Rapid neural adaptation to sound level statistics. *Journal of Neuroscience*, **28**, 6430–6438.

DeWeese, M. 1996. Optimization principles for the neural code. *Network: Computation in Neural Systems*, **7**, 325–331.

Elliott, T., Kuang, X., Shadbolt, N.R., Zauner, K.-P. 2008. An invariance principle for maintaining the operating point of a neuron. *Network: Computation in Neural Systems*, **19**, 213–235.

Földiák, P. 1990. Forming sparse representations by local anti-Hebbian learning. *Biological Cybernetics*, **64**, 165–170.

Franco, L., Rolls, E.T., Aggelopoulos, N.C., Jerez, J.M. 2007. Neuronal selectivity, population sparseness, and ergodicity in the inferior temporal visual cortex. *Biological Cybernetics*, **96**, 547-560.

Hyvärinen, A., Karhunen, J., E., Oja. 2001. *Independent Component Analysis*. NY: Wiley.

Hyvärinen, A., Hurri, J., P.O., Hoyer. 2009. *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision.* London: Springer.

Katz, L.C., Shatz, C.J. 1996. Synaptic activity and the construction of cortical circuits. *Science*, **274**, 1133–1138.

Kohn, A. 2007. Visual adaptation: Physiology, mechanisms, and functional benefits. *Journal of Neurophysiology*, **97**, 3155–3164.

Kvale, M.N., Schreiner, C.E. 2004. Short-term adaptation of auditory receptive fields to dynamic stimuli. *Journal of Neurophysiology*, **91**, 604–612.

Laughlin, S. 1981. A simple coding procedure enhances a neuron's information capacity. *Zeitschrift fur Naturforschung C*, **36**, 910–912.

Lehky, S.R., Kiani, R., Esteky, H., Tanaka, K. 2011. Statistics of visual responses in primate inferotemporal cortex to object stimuli. *Journal of Neurophysiology*, **106**, 1097–1117.

Lennie, P. 2003. The cost of neural computation. *Current Biology*, **13**, 493–497.

Lyu, S., Simoncelli, E.P. 2009. Nonlinear extraction of independent components of natural images using radial gaussianization. *Neural Computation*, **21**, 1485–1519.

Maravall, M., Petersen, R.S., Fairhall, A.L., Arabzadeh, E., Diamond, M.E. 2007. Shifts in coding properties and maintenance of information transmission during adaptation in barrel cortex. *PLoS Biology*, **5**, 0323–0334.

Meister, M., Berry, M.J. 1999. The neural code of the retina. *Neuron*, **22**, 435–450.

Olshausen, B.A., Field, D.J. 1996. Emergence of simple-cell receptive field properties by learning a sparse code for natural scenes. *Nature*, **381**, 607–609.

Olshausen, B.A., Field, D.J. 1997. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, **37**, 3311–3325.

Savin, C., Joshi, P., Triesch, J. 2010. Independent component analysis in spiking neurons. *PLoS Computational Biology*, **6**, e1000757.

Shapley, R., Enroth-Cugell, C. 1984. Visual adaptation and retinal gain control. *Progress in Retinal Research*, **3**, 263–346.

Silva, A.J., Zhou, Y., Rogerson, T., Shobe, J. Balaji, J. 2009. Molecular and cellular approaches to memory allocation in neural circuits. *Science*, **326**, 391–395.

Simoncelli, E.P., Olshausen, B.A. 2001. Natural image statistics and neural representation. *Annual Review of Neuroscience*, **24**, 1193–1216.

Triesch, J. 2005a. Synergies between intrinsic and synaptic plasticity in individual neurons. *Pages 1417–1424 of:* Saul, L.K., Weiss, Y., Bottou, L. (eds), *Advances in Neural Information Processing Systems 17*. Cambridge, MA: MIT Press.

Triesch, J. 2005b. A gradient rule for the plasticity of a neuron's intrinsic excitability. *Pages 65–70 of:* Duch, W., Kacprzky, J., Oja, E. Zadrozny, *Proceedings of the International Conference on Artificial Neural Networks (ICANN)*. Berlin: Springer.

Triesch, J. 2007. Synergies between intrinsic and synaptic plasticity mechanisms. *Neural Computation*, **19**, 885–909.

Turrigiano, G.G., Nelson, S.B. 2004. Homeostatic plasticity in the developing nervous system. *Nature Reviews Neuroscience*, **5**, 97–107.

van Hateren, J.H. 1992. A theory maximizing sensory information. *Biological Cybernetics*, **68**, 23–29.

van Hateren, J.H. 1998. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings of the Royal Society of London Series B*, **265**, 359–366.

Wainwright, M.J. 1999. Visual adaptation as optimal information transmission. *Vision Research*, **39**, 3960–3974.

Zaghloul, K.A., Boahen, K., Demb, J.B. 2005. Contrast adaptation in subthreshold and spiking responses of mammalian y-type retinal ganglion cells. *Journal of Neuroscience*, **24**, 860–868.

Zhang, W., Linden, D.J. 2003. The other side of the engram: Experience-driven changes in neuronal intrinsic excitability. *Nature Reviews Neuroscience*, **4**, 885–900.

Figure Captions

Figure 1: The evolution of the strength vector and response parameters in Triesch's model of intrinsic plasticity in the presence of two different input distributions for a sparse firing regime, with $\hat{\mu} = 1/10$. A, C and E show the evolution of the angle ϕ between the strength vector and the input indicated by the solid line; the dotted line indicates the other, orthogonal input. B, D and F show the evolution of the threshold θ and gain γ in the simulations shown in A, C and E, respectively. A and B show results for a Laplace and a logistic input, as do C and D; E and F show results for a uniform and a binary-valued input. The model converges on any super-Gaussian input, not specifically the input with the heaviest tail. A and B show the model finding the Laplace input while C and D show it finding the logistic input when both inputs are super-Gaussian. Note that the final values of the threshold and gain are largely set before the strength vector moves towards a specific input, although there are small changes as the strength vector converges on a specific input. E and F show the model not converging on an input in the presence of two sub-Gaussian inputs. The learning rates are set as $\varepsilon_{\text{sp}} = \varepsilon_{\text{ip}} = 10^{-4}$, while initially we set $\theta = 0$ and $\gamma = 1/16$.

Figure 2: The performance of Triesch's model of intrinsic plasticity with a modified, non-exponential output firing rate PDF, given by Eq. (2.15), but still with sparse output firing, determined by $\hat{\mu} = 1/10$. We have set the sharp-

ness to $\hat{\gamma} = 10$. The modified model still finds super-Gaussian sources. The formats of A and B are identical to A and B in Fig. 1. C shows the final output firing rate PDF for final values of the threshold and gain after the model has converged on the Laplace input. For comparison the target PDF given in Eq. (2.15) is shown, as is the step PDF given in Eq. (2.12). $\varepsilon_{\rm sp}$, $\varepsilon_{\rm ip}$ and initial values of θ and γ are set as in Fig. 1.

Figure 3: Triesch's model of intrinsic plasticity in the presence of non-sparse output firing, with $\hat{\mu} = 1/2$. The format of this figure is identical to that in Fig. 1 except that input distributions differ between the respective panels. For $\hat{\mu} = 1/2$, the model converges on any sub-Gaussian input but does not converge on super-Gaussian inputs.

Figure 4: Fixed point solutions for the threshold θ and gain γ in Triesch's model of intrinsic plasticity as a function of the mean output firing rate $\hat{\mu}$ for $g_R(r) = \hat{\mu}^{-1} \exp(-r/\hat{\mu})$, for two differing forms of input statistics. The solutions for θ and γ are obtained from Eqs. (2.8) and (2.9) using either a single Laplace input or a single binary-valued input. The values of γ obtained for these two radically differing forms of inputs (lines labelled "γ") are very similar, as are the values obtained for θ (lines labelled "θ"). The principal factor determining the fixed point values of γ and θ is therefore the mean output firing rate, $\hat{\mu}$, and not the input statistics.

Figure 5: Stability regions, determined by the sign of $\Delta_X(\Gamma, \Theta)$, for a binary-valued input. A plots $\Delta_X(\Gamma, \Theta)$ against Θ for three different values of Γ, as indicated. B plots the sign of $\Delta_X(\Gamma, \Theta)$ in the Θ–Γ plane. The shaded region corresponds to $\Delta_X(\Gamma, \Theta) < 0$ and thus a binary-valued input being stable.

Figure 6: Stability regions, determined by the sign of $\Delta_X(\Gamma, \Theta)$, for other input distributions, as indicated. Shaded regions always correspond to $\Delta_X(\Gamma, \Theta) < 0$ and thus stability of the input. A shows a Laplace input; B a logistic input; C a uniform input. D shows the zero contours on the same graph for the four indicated inputs, in order to facilitate comparison.

Figure 7: The behaviour of $\Delta_X(\Gamma, \Theta)$ in the small Θ region is not generic. A and B show $\Delta_X(\Gamma, \Theta)$ for the super-Gaussian PDF f_+ in Eq. (3.11), while C and D show $\Delta_X(\Gamma, \Theta)$ for the sub-Gaussian PDF f_- in Eq. (3.12). As usual, shaded regions are stable.

Figure 8: $d\phi/dt$ as a function ϕ for a neuron with a Laplace and a uniform input, where ϕ is the angle between the strength vector and the Laplace input. In A, Θ approaches and passes through the point $\Theta \approx 0.9769$ at which the uniform distribution becomes stable. During this process, the basin of attraction around the Laplace input collapses. For reference, B shows a point well away from such critical behaviour, when the Laplace distribution is deeply stable. In both graphs, $\Gamma = 2$.

Figure 9: The sign of the function $\Xi(\Gamma, \Theta)$, which in part determines the stability of a single Gaussian input, in the Θ–Γ plane.

Figure 10: Stability regions in the θ–γ plane for Földiák bars. The vector $\boldsymbol{\omega} = (1, \ldots, 1)^{\mathrm{T}}/N$ is stable everywhere. Single bar fixed points are stable in the darker grey region. Multi-bars fixed points are stable in the thin, lighter grey slither. These stability regions are obtain in simulation as described in the main text.

Figure 11: Evolution of the threshold and gain in Triesch's model of intrinsic plasticity for Földiák bar inputs. Also shown are graphical representations of the strength vector (as an array over the 10×10 array of inputs) at the indicated times points, during the key period between 0.6×10^6 and 6.0×10^6 time steps when the strength vector converges on a single bar. In these graphical representations, a white cell represents an input unit with the largest synaptic strength while a black cell represents an input unit with zero synaptic strength. The learning rates are set as $\varepsilon_{\mathrm{sp}} = 10^{-4}$ and $\varepsilon_{\mathrm{ip}} = 10^{-2}$, while the initial threshold and gain are set as $\theta = 0$ and $\gamma = 1/4$.

Figure 12: The evolution of the threshold and gain from Fig. 11, for $\hat{\mu} = 0.05$, plotted in the θ–γ plane, with the stability regions for the strength vector shaded as in Fig. 10. Also shown are results for $\hat{\mu} = 0.5$ and $\hat{\mu} = 0.005$, as

indicated. The small black dots show the location of the (θ, γ) pair at 10^4 time steps. The large, open circles show the north-west termini of the (θ, γ) trajectories in the θ–γ plane. Subsequent receptive field refinement begins as the (θ, γ) trajectories reverse and head in a south-east, increasing-threshold direction.

Figure 13: The fixed point values of the threshold and gain in the θ–γ plane as a function of $\hat{\mu}$ in Triesch's model of intrinsic plasticity with a single Laplace input. The shaded region shows the area of stability of a Laplace input. The dashed arrow labelled "$\hat{\mu} \downarrow$" indicates the direction of decreasing $\hat{\mu}$ along the line showing these fixed point values.

Figure 14: Evolution of the threshold and gain in our model of adaptation to synaptic strength changes for Földiák bar inputs. Also shown for reference are the threshold and gain in Triesch's model, from Fig. 11, for $\hat{\mu} = 0.05$. The upper gain line and the lower threshold line represent results from our model. We have set the initial values of θ and γ in our model to coincide with the north-west terminus values in Triesch's model for $\hat{\mu} = 0.05$. Since these values are attained at around 0.6×10^6 time steps, we have offset time by this amount in displaying the threshold and gain in our model in order to facilitate more direct comparison. We also show the refinement of the receptive field, similarly displaced, in our model.

Figure 15: The evolution of the threshold and gain in our model of adaptation to synaptic strength changes for Földiák bar inputs plotted in the θ–γ plane, as in Fig. 12. Three trajectories are shown, with different initial values of θ and γ selected to lie inside the large open circles corresponding with the north-west termini of the trajectories in Fig. 12.

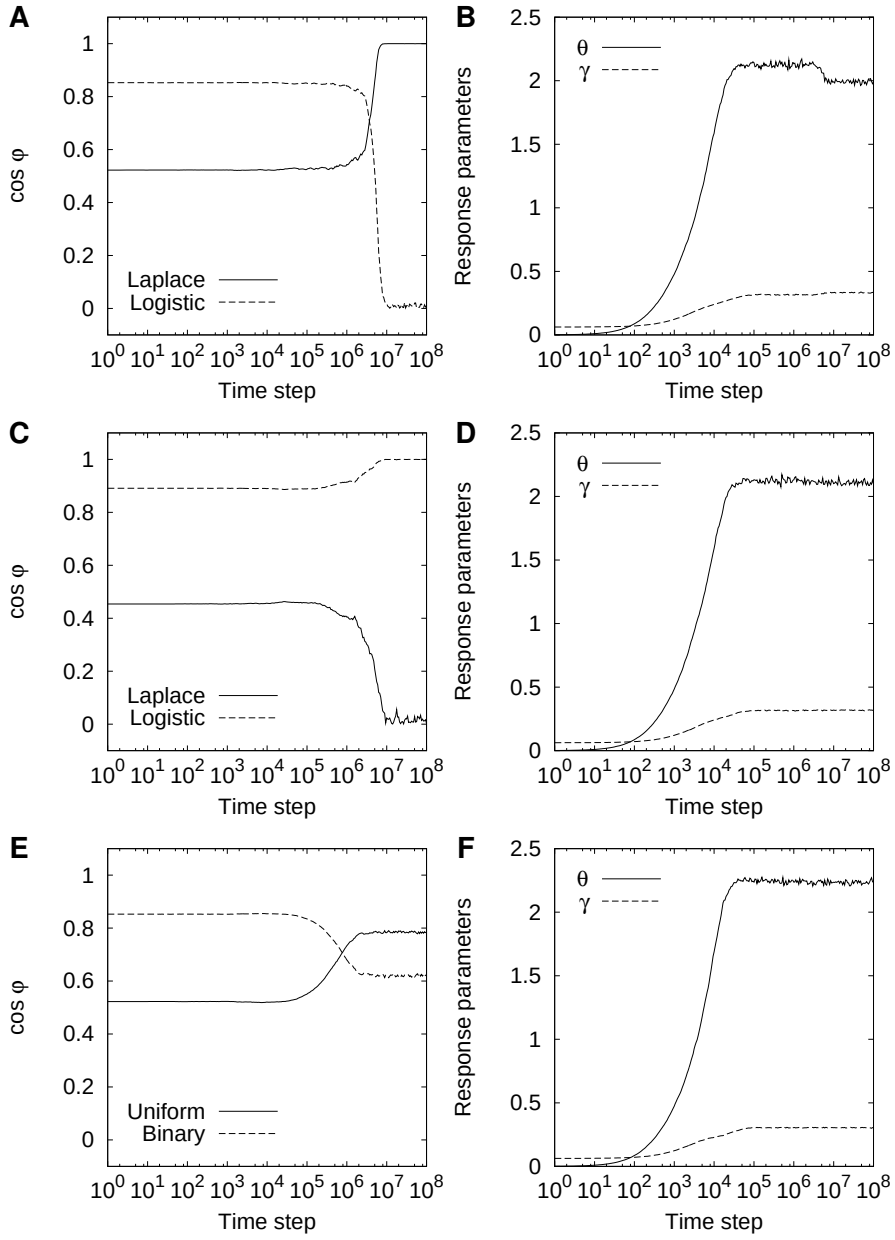

Figure 1. NECO-12-13-2027

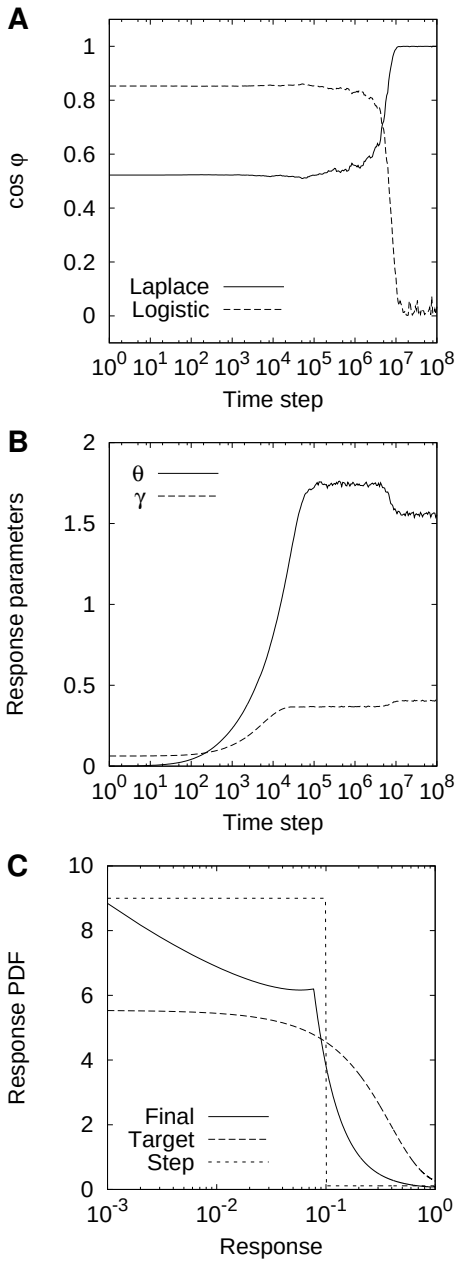

Figure 2. NECO-12-13-2027

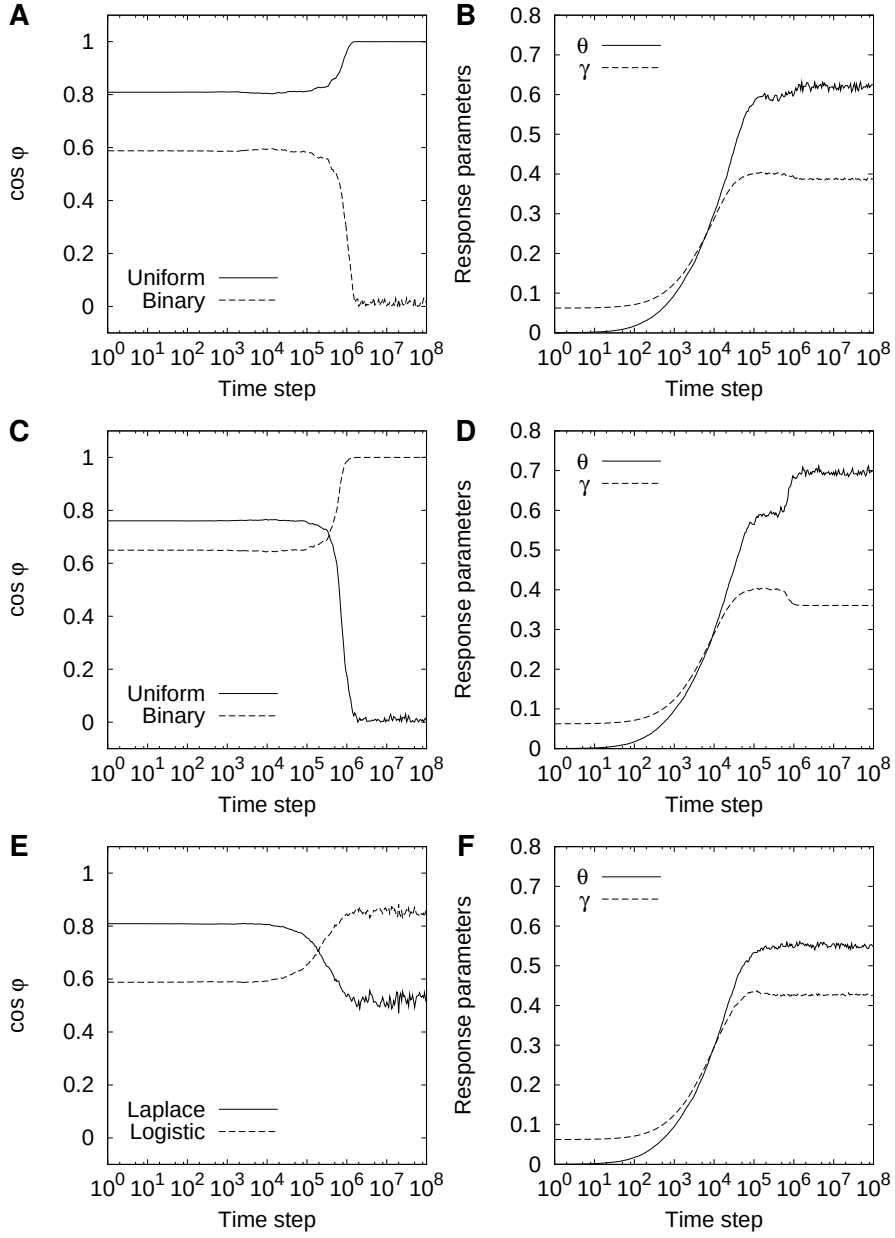

Figure 3. NECO-12-13-2027

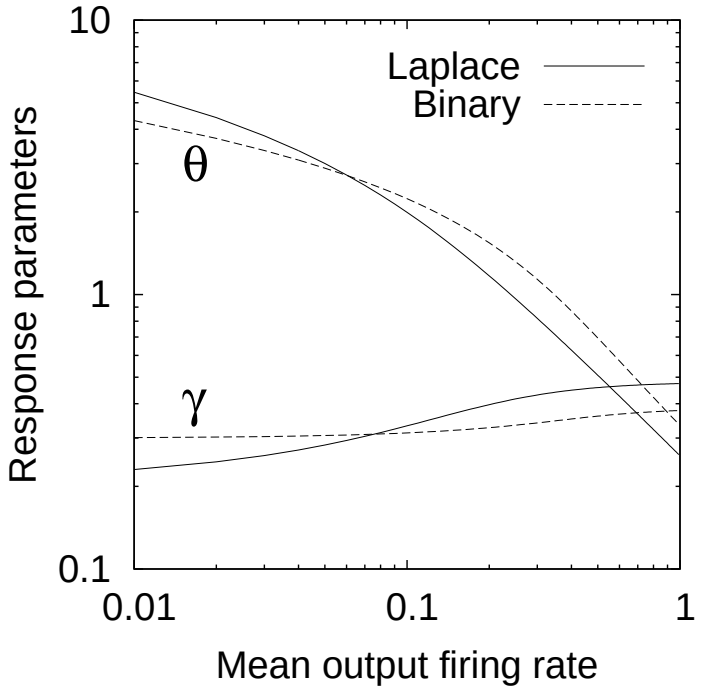

Figure 4. NECO-12-13-2027

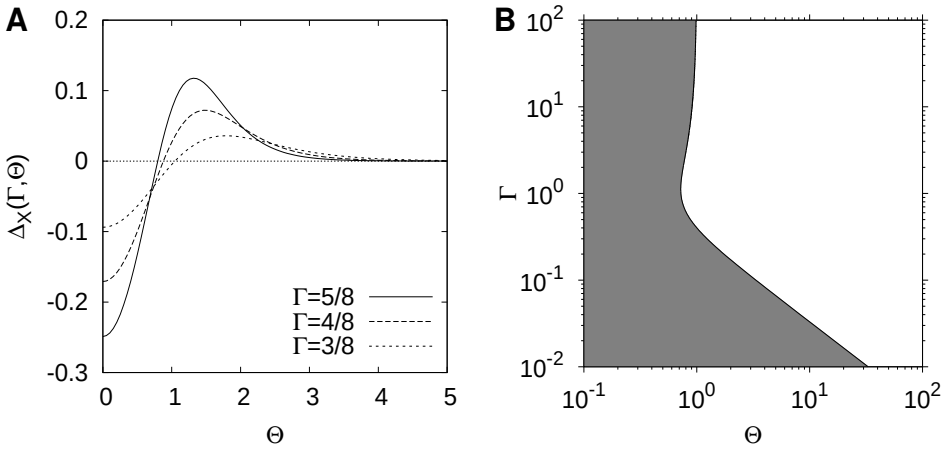

Figure 5. NECO-12-13-2027

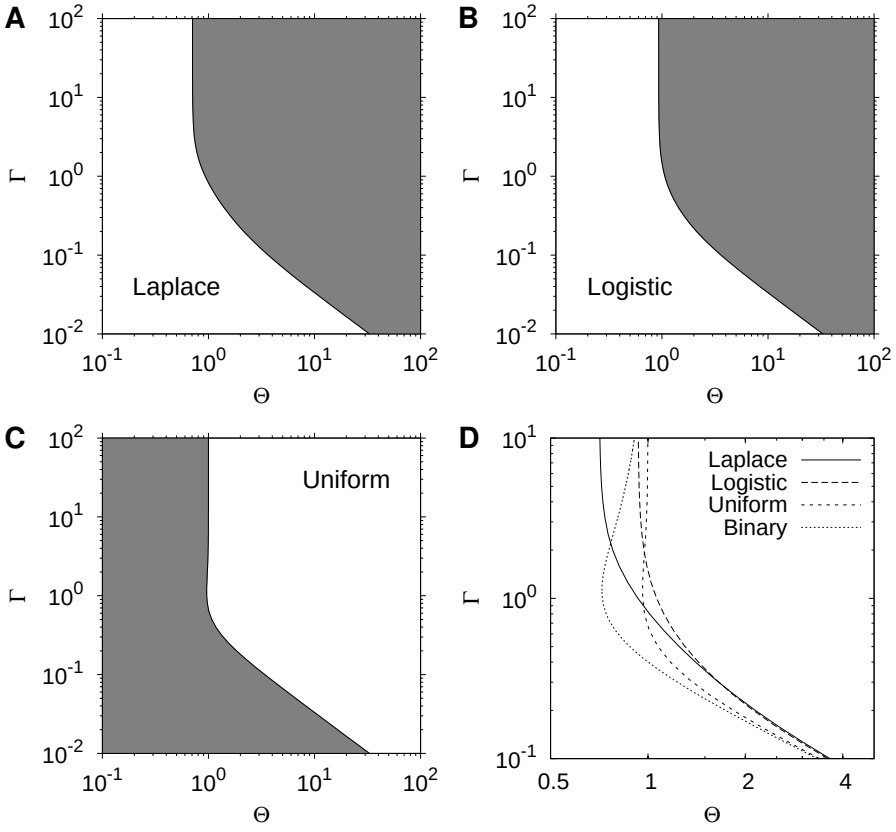

Figure 6. NECO-12-13-2027

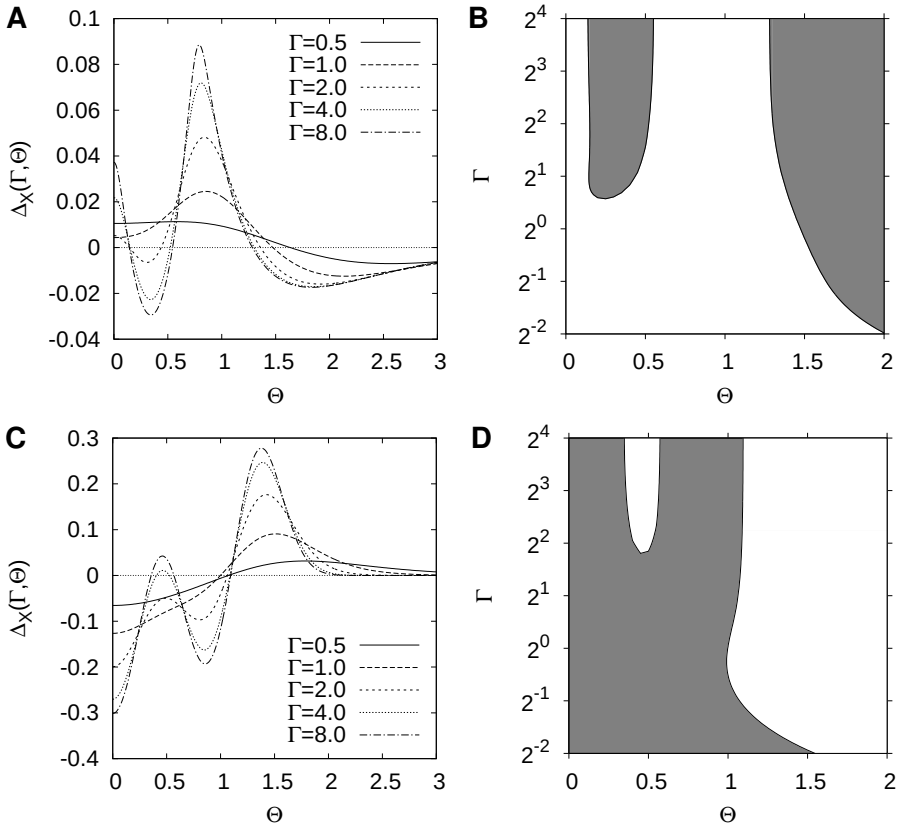

Figure 7. NECO-12-13-2027

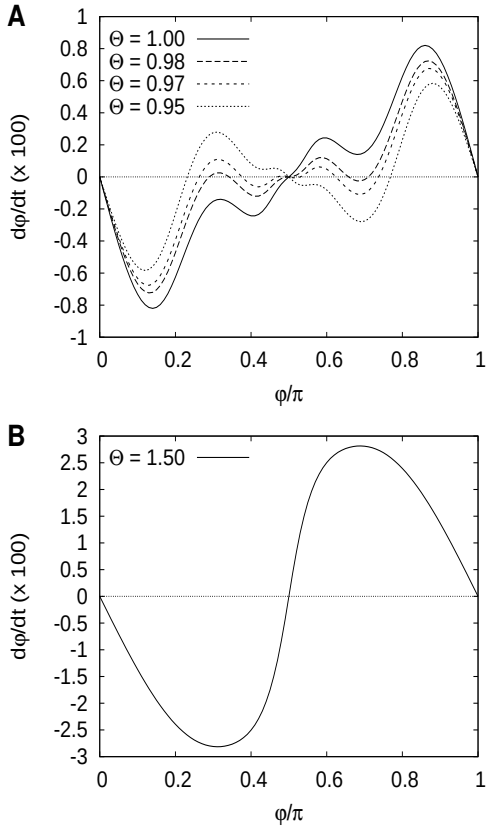

Figure 8. NECO-12-13-2027

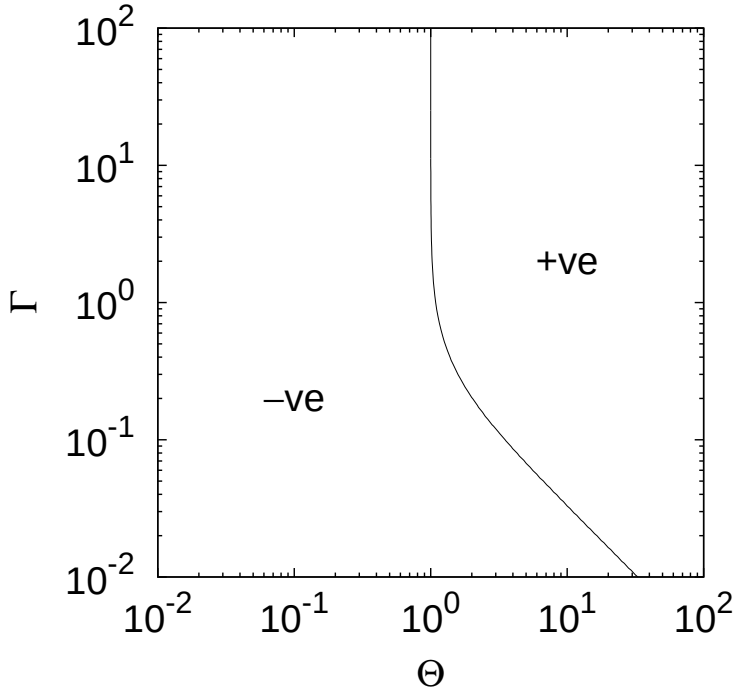

Figure 9. NECO-12-13-2027

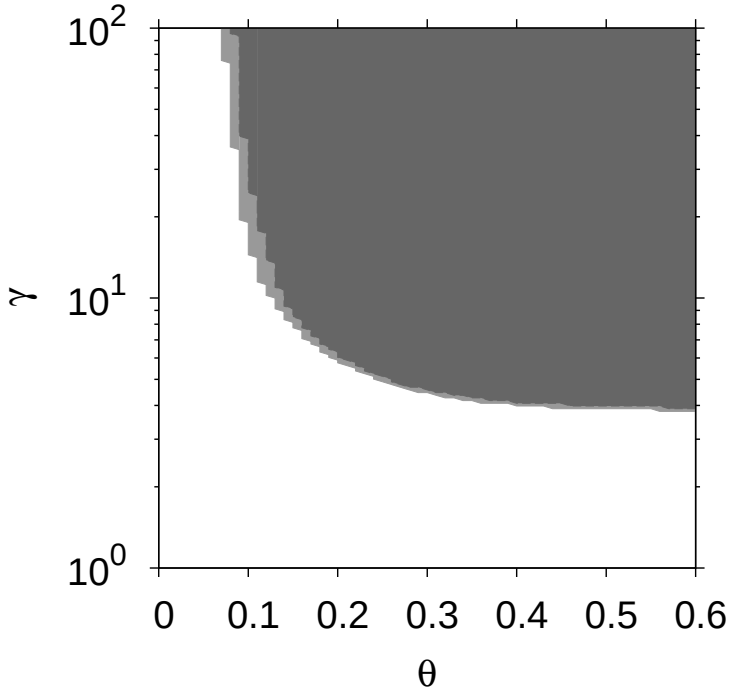

Figure 10. NECO-12-13-2027

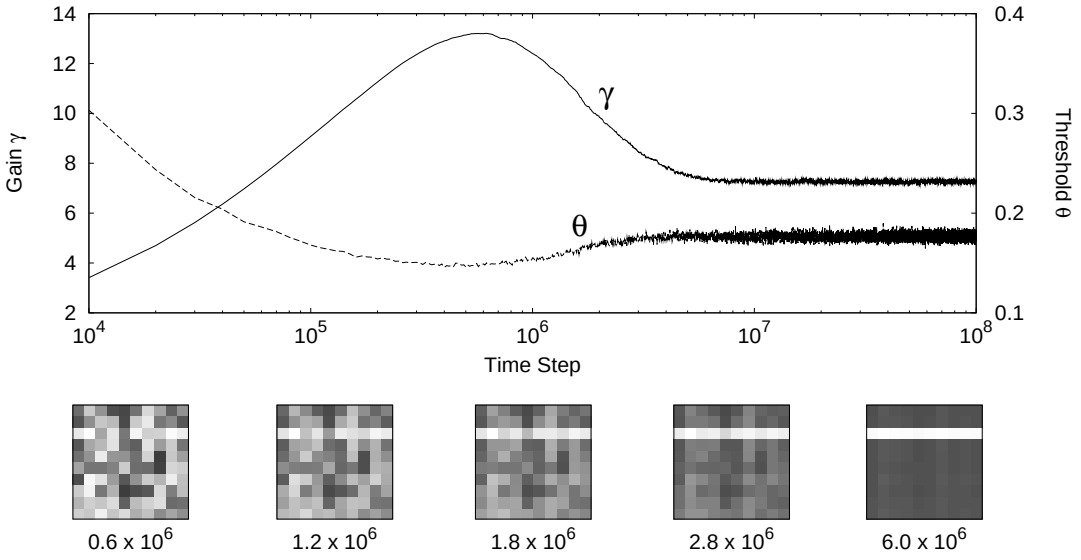

Figure 11. NECO-12-13-2027

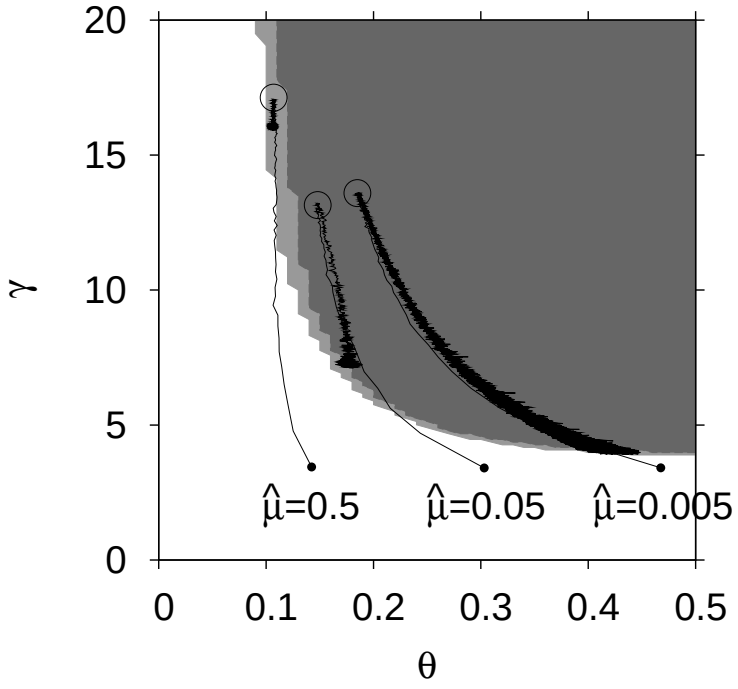

Figure 12. NECO-12-13-2027

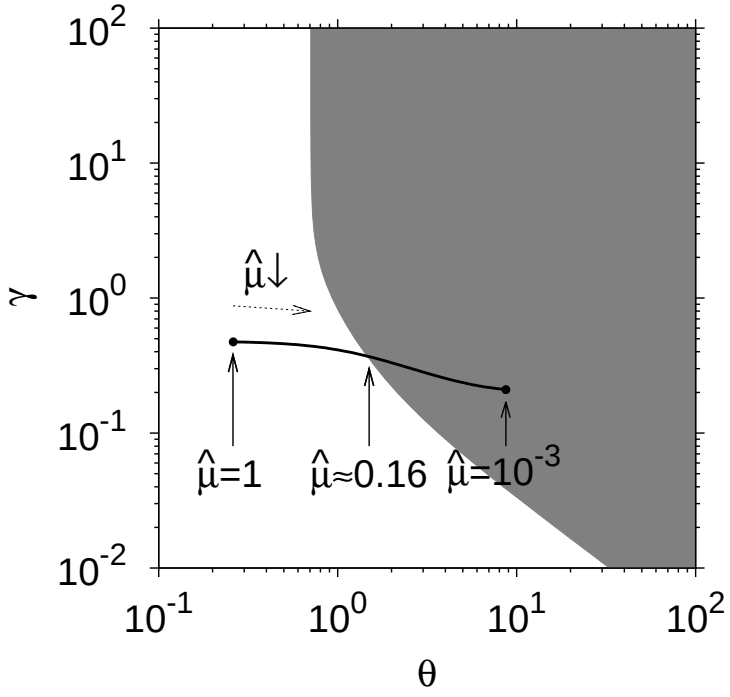

Figure 13. NECO-12-13-2027

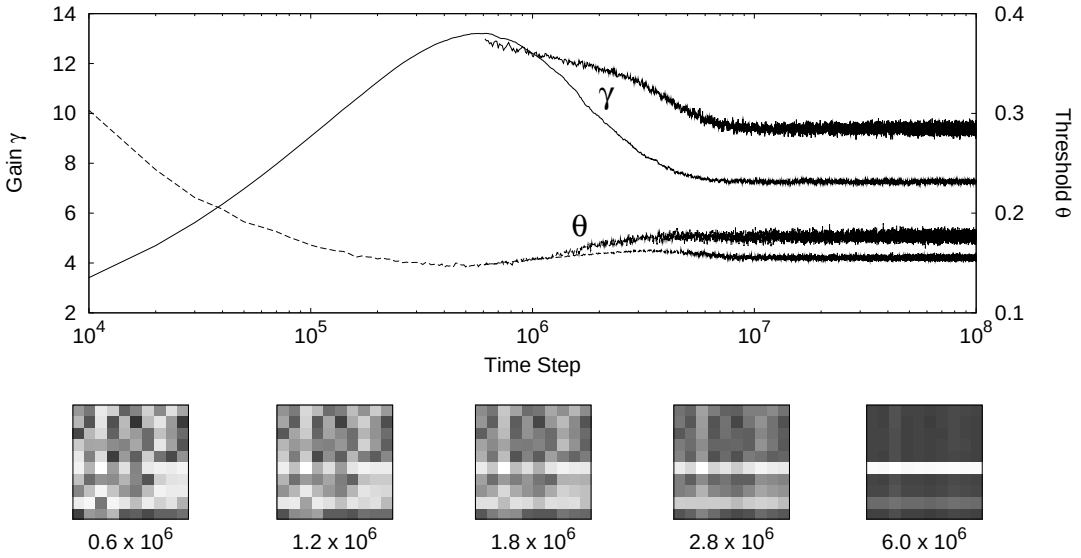

Figure 14. NECO-12-13-2027

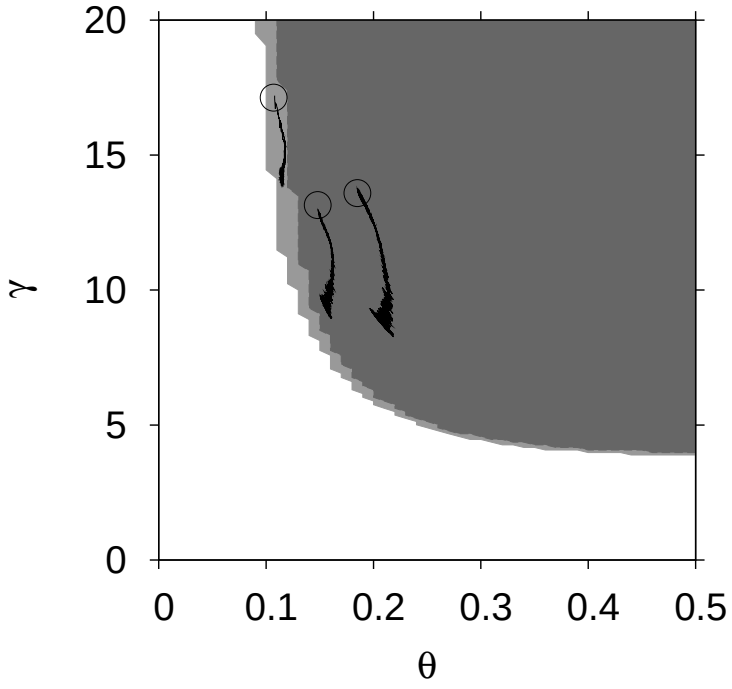

Figure 15. NECO-12-13-2027