

Differential Priors for Elastic Nets

Miguel Á. Carreira-Perpiñán¹, Peter Dayan², and Geoffrey J. Goodhill³

¹ Dept. of Computer Science & Electrical Eng., OGI, Oregon Health & Science University
miguel@cse.ogi.edu

² Gatsby Computational Neuroscience Unit
dayan@gatsby.ucl.ac.uk

³ Queensland Brain Institute and Dept. of Mathematics, University of Queensland
goodhill@uq.edu.au

Abstract. The elastic net and related algorithms, such as generative topographic mapping, are key methods for discretized dimension-reduction problems. At their heart are priors that specify the expected topological and geometric properties of the maps. However, up to now, only a very small subset of possible priors has been considered. Here we study a much more general family originating from discrete, high-order derivative operators. We show theoretically that the form of the discrete approximation to the derivative used has a crucial influence on the resulting map. Using a new and more powerful iterative elastic net algorithm, we confirm these results empirically, and illustrate how different priors affect the form of simulated ocular dominance columns.

1 Introduction

The elastic net was originally introduced as a continuous optimisation method for the traveling salesman problem (TSP) [1]. The basic idea is to represent a tour in continuous city space by a collection of linked centroids (the elastic net) that optimizes a trade-off between matching the given cities and keeping the sum of squared link lengths small. This trade-off is biased towards short nets at first and is slowly reversed until only the matching term matters – a deterministic annealing algorithm. From a dimension reduction perspective [2], the elastic net is a probabilistic model (Gaussian mixture) with a specific type of prior (the sum of squared link lengths) that models a high-dimensional city space in terms of a low-dimensional, discretized space. It is thus related to latent-variable models such as the generative topographic mapping [3], and can be seen as a probabilistic self-organising map (SOM) [4]. The elastic net has also been applied to modeling ocular dominance and orientation maps [5, 6], by interpreting the net centroids as the preferred stimuli of points in a discretized primary visual cortex and using the training set of “cities” as a sample of the continuous stimulus space. Biologically, the learning rule can be interpreted in terms of Hebbian learning and the prior in terms of an intracortical connectivity function. In this context, Dayan [7] suggested the use of more general quadratic forms for the intracortical connectivity prior. Here we define the generalized probabilistic model and analyse by theory and simulations the effect of different discrete priors on the structure of the resulting cortical maps.

2 The Generalized Elastic Net

Given a collection of centroids $\{\mathbf{y}_m\}_{m=1}^M \subset \mathbb{R}^D$ and a scale parameter $\sigma \in \mathbb{R}^+$, consider a Gaussian-mixture density $p(\mathbf{x}) = \sum_{m=1}^M \frac{1}{M} p(\mathbf{x}|m)$ with $\mathbf{x}|m \sim \mathcal{N}(\mathbf{y}_m, \sigma^2 \mathbf{I}_D)$. Expressing the centroids as a $D \times M$ matrix $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_M)$, define a smoothing, or neighborhood-preserving, prior on the centroids $p(\mathbf{Y}; \beta) \propto \exp(-\frac{\beta}{2} \text{tr}(\mathbf{Y}^T \mathbf{Y} \mathbf{S}))$ where β is a regularisation hyperparameter and \mathbf{S} is a (semi)positive definite $M \times M$ matrix. This prior can also be seen as a Gaussian process prior, with matrix \mathbf{S} the inverse of the Gaussian process covariance matrix. Without the prior, the centroids could be permuted at will with no change in the model, since the variable m is just an index. The prior can be used to convey the topological (dimension and shape) and geometric (e.g. curvature) structure of a manifold implicitly defined by the centroids – i.e., as if the centroids resulted from discretizing a continuous latent variable model with a uniform density in a latent space of dimension L (the “cortex”), a nonparametric mapping from latent to data space (defined by the centroids \mathbf{Y}) and an isotropic Gaussian noise model in data space of dimension D (the “stimuli space”) [8].

Given a training set expressed as a $D \times N$ matrix $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)$, we are interested in deterministic annealing algorithms that minimize the energy function

$$E(\mathbf{Y}, \sigma) = -\sigma \sum_{n=1}^N \log \sum_{m=1}^M e^{-\frac{1}{2} \|\frac{\mathbf{x}_n - \mathbf{y}_m}{\sigma}\|^2} + \frac{\beta}{2} \text{tr}(\mathbf{Y}^T \mathbf{Y} \mathbf{S}) \quad (1)$$

over \mathbf{Y} for fixed σ , starting with a large σ and tracking the minimum to a small value of σ . E is derived from the log posterior of the full model. We call the first term the *fitness term*, arising from the Gaussian mixture $p(\mathbf{X}|\mathbf{Y}, \sigma)$, and the second term the *tension term*, arising from the prior $p(\mathbf{Y})$. The multiplication of the fitness term by σ downweights the fitness term with respect to the tension term as σ decreases. We do this because (1) one can find good solutions to combinatorial optimisation problems such as the TSP (which require $\sigma \rightarrow 0$); and (2) if considered as a dynamical system for a continuous latent space, the evolution of the net as a function of σ and the iteration index models the temporal evolution of cortical maps [5, 6]. We also investigate the behavior of the model for a larger range of β values than has previously been possible, thanks to the new algorithm we introduce in Sect. 5.

To apply the elastic net to practical problems such as the TSP or cortical map modeling, \mathbf{S} should incorporate some knowledge of the problem being modeled. \mathbf{S} specifies the expected topological and geometric structure of the solutions. Here, we fix the topology, i.e., the neighborhood relationship in the net, to be one or two dimensions, with open or periodic boundary conditions, and focus on the way that \mathbf{S} represents priors on curvature. The prior has a crucial impact on the final solutions. For instance, in the application of the elastic net to ocular dominance, it helps determine the (empirically testable) width of the ocular dominance stripes. The next section studies this in an idealized, continuum limit; Sect. 4 considers the unexpectedly dramatic effect of discretization; Sect. 5 presents the new elastic net algorithm which allows us to study a much wider range of conditions; and Sect. 6 uses this algorithm to verify and extend our theoretical results. In Sect. 3–4 we assume an $M \times 1$ vector $\mathbf{y} \stackrel{\text{def}}{=} \mathbf{Y}^T$, since the tension term separates in a sum of D terms, one per dimension in \mathbb{R}^D (stimulus space).

3 Regularisation of a Continuous Elastic Net

In a 1D continuous setting for which $\mathbf{y} = (y_m)$ becomes a real function of a real variable $y = y(t)$, consider a tension term of the smoothness functional type:

$$\frac{\beta}{2} \int_{-\infty}^{\infty} (\mathcal{D}y(t))^2 dt = \frac{\beta}{2} \int_{-\infty}^{\infty} ((s * y)(t))^2 dt. \quad (2)$$

\mathcal{D} is a differential operator and s a kernel, e.g. for the p th-order derivative $\mathcal{D}_p = \frac{d^p}{dt^p}$ and $s = \frac{d^p \delta}{dt^p}$ where δ is the delta function. Such operators characterize the metric properties of y , such as its curvature. Note that a basis of the nullspace of \mathcal{D}_p is $\{1, t, \dots, t^{p-1}\}$. When the fitness term is also quadratic, such as $\int (y - g)^2 dt$ for fixed g , regularisation problems like this can be approached from the point of view of function approximation in a Hilbert space – e.g. of the functions y having derivatives up to order p that are square-integrable, i.e., with seminorm $\int \|\cdot\|^2$. This is the case for spline regression. In our case, the fitness term is not quadratic but results from Gaussian-mixture density estimation; however, we can still gain insight in the Fourier domain. By applying Parseval's theorem to the continuous tension term (2) with p th-order derivative \mathcal{D}_p and calling \hat{y} the Fourier transform of y , we can see that the tension energy is the same in both domains:

$$\frac{\beta}{2} \int_{-\infty}^{\infty} \left(\frac{d^p y}{dt^p} \right)^2 dt = \frac{\beta}{2} \int_{-\infty}^{\infty} |(i2\pi k)^p \hat{y}(k)|^2 dk = \frac{\beta}{2} \int_{-\infty}^{\infty} (2\pi k)^{2p} |\hat{y}(k)|^2 dk$$

since the Fourier transform of $\frac{d^p y}{dt^p}$ is $(i2\pi k)^p \hat{y}(k)$. This means that \mathcal{D}_p is acting as a high-pass filter whose cutoff frequency increases monotonically with p ; see Fig. 1(A). Therefore, high-frequency functions will incur a high penalty and the minima of the energy will likely have low frequencies – subject to the effect of the fitness term.

Regularization is therefore straightforward in the continuous case. How this extends to the discrete case is the topic of the rest of the paper.

4 Discrete Nets: Construction of \mathbf{S} , Analysis of the Tension Term

We consider $\mathbf{S} = \mathbf{D}^T \mathbf{D}$ where \mathbf{D} is obtained from a translationally invariant (convolution) filter that we represent via a *stencil*. The stencil approximates a derivative via a finite-difference scheme, so \mathbf{S} has a zero eigenvalue associated with the eigenvector of ones. This implies that \mathbf{S} will not be positive definite so the prior will be improper, and that the tension term will be invariant to rigid motions of the net (note the fitness term is invariant to permutations of \mathbf{Y} but not to rigid motions of it). We thus have a differential prior $\text{tr}(\mathbf{Y}^T \mathbf{Y} \mathbf{S}) = \|\mathbf{D} \mathbf{Y}^T\|^2$ in terms of the Frobenius norm.

In 1D, we write a stencil as $\zeta = (\dots, \zeta_{-2}, \zeta_{-1}, \zeta_0, \zeta_1, \zeta_2, \dots)$, so that the rows of the \mathbf{D} matrix result from successively shifting ζ and padding with zeros at the ends. We have $\mathbf{D}^T \mathbf{y} = \zeta * \mathbf{y}$. If M is large and ζ has only a few nonzero coefficients, \mathbf{D} will be sparse. For example, for $M = 7$ and periodic boundary conditions (b.c.):

$$\zeta = (a \ b \ c \ d \ e) \quad \mathbf{D} = \begin{pmatrix} c & d & e & 0 & 0 & a & b \\ b & c & d & e & 0 & 0 & a \\ a & b & c & d & e & 0 & 0 \\ 0 & a & b & c & d & e & 0 \\ 0 & 0 & a & b & c & d & e \\ e & 0 & 0 & a & b & c & d \\ d & e & 0 & 0 & a & b & c \end{pmatrix} \quad \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_7 \end{pmatrix}.$$

With periodic b.c., \mathbf{D} is a *circulant* matrix. When other types of b.c. are used, e.g. for nets with non-rectangular shapes or holes, \mathbf{D} will be in general a *quasi-Toeplitz* matrix ($d_{mn} = d_{n-m}$ except at a few elements). In elastic nets of two or more dimensions, \mathbf{D} is circulant or quasi-Toeplitz by blocks. The original elastic net model [1, 2, 5] is obtained by using a stencil $(0, -1, 1)$; in 1D the tension term simplifies to the sum of squared lengths $\frac{\beta}{2} \sum_m \|\mathbf{y}_{m+1} - \mathbf{y}_m\|^2$.

Compared with the continuous case, the discrete case has an extra degree of freedom since the p th derivative can be represented by many different stencils differing in truncation error. If the goal is accuracy at a reasonable computation cost (as in the numerical solution of PDEs), then one seeks stencils having few nonzero coefficients and high-order truncation error. For example, both $(0, -1, 1)$ (forward-difference) and $(-\frac{1}{2}, 0, \frac{1}{2})$ (central-difference) approximate the first derivative, but with linear and quadratic error, respectively, so the latter is preferred. Surprisingly, in the elastic net (where accuracy is not the goal) it turns out that stencils of the same order can result in nets of completely different characteristics. In particular, the highest frequency components can fail to be penalized, giving rise to a “sawtooth” solution (Fig. 3), which is physically unreasonable. Here, we give a characterisation of discrete differential stencils. The basic idea is that we can understand the tension term in terms of the eigenspace of \mathbf{S} , and that the latter coincides for periodic b.c. with the Fourier spectrum of the stencil, i.e., the eigenvectors are (co)sinewaves in a bounded, discrete frequency interval.

For periodic b.c., \mathbf{D} and \mathbf{S} are circulant. The mathematical analysis of 1D nets is straightforward; we state the main results without proof (full details appear in [9]). Many of them carry over to the L -dimensional case. Assume the net has M centroids (where M is even, for simplicity).

The eigenvectors of \mathbf{S} are discrete plane waves $v_{mn} = \cos(2\pi \frac{m}{M}n)$ and $w_{mn} = \sin(2\pi \frac{m}{M}n)$. Unlike the continuous case, the discrete frequency m is upper bounded by $\frac{M}{2}$. This highest frequency corresponds to a sawtooth wave $(1, -1, 1, -1, \dots, 1, -1)^T$, which plays a significant role with certain stencils. \mathbf{S} has $\frac{M}{2}$ distinct eigenvalues ν_m (since $\nu_m = \nu_{M-m}$ for $m = 1, \dots, M-1$) which are real nonnegative. The power spectrum of the stencil ζ is equal to the eigenspectrum of \mathbf{S} , i.e., $|\hat{\zeta}_\kappa|^2 = \nu_\kappa$ where $\hat{\zeta}_\kappa = \sum_{m=0}^{M-1} \zeta_m e^{-i2\pi \frac{\kappa}{M}m}$ is the discrete Fourier transform of ζ , $\nu_0 = 0$ is associated with the constant eigenvector $\mathbf{v}_0 = (1, \dots, 1)^T$; and $\nu_{\frac{M}{2}}$ is associated with the sawtooth eigenvector $\mathbf{v}_{\frac{M}{2}} = (1, -1, 1, -1, \dots, 1, -1)^T$.

By decomposing the net in the eigenvector basis of \mathbf{S} as a superposition of plane waves, it becomes clear that frequency m contributes a tension-term penalty proportional to ν_m . Thus, given a stencil, its power spectrum gives the penalty for each frequency of the net. The constant eigenvector of ones is the sampled version of a constant net and incurs zero penalty since $\nu_0 = 0$ for a differential operator. However, since the rest of eigenvalues are nonzero in general, eigenvectors of the form $v_{mn} = n^m$ for fixed $m \in \mathbb{Z}^+$, corresponding to a monomial t^m , are not nullified by \mathbf{D} in the circulant case (but they can be nullified in the nonperiodic b.c. case).

A stencil ζ has zero power at the sawtooth frequency iff $\sum_{m \text{ even}} \zeta_m = \sum_{m \text{ odd}} \zeta_m = 0$. We call *sawtooth stencil* a stencil satisfying this condition. The convolution of any stencil with a sawtooth stencil is also sawtooth. For sawtooth stencils, the highest frequency incurs no penalty in the tension term, just as the zero-frequency wave (the con-

stant net) does – unlike the continuous case, where a wave of frequency m has an average penalty proportional to m^{2p} .

Applying p times a first-order stencil ζ results in a p th-order stencil with matrix \mathbf{S}^p . The family of ζ consists of all such stencils for $p \geq 1$. We give specific results for two families that are particularly important; see Fig. 1(B,C).

Forward-difference family. This is defined by the first-order forward-difference stencil $\zeta = (0, -1, 1)$. The p th-order derivative stencil has eigenvalues $\nu_m = (2 \sin(\pi \frac{m}{M}))^{2p}$. The stencils are not sawtooth. The first four stencils are:

p	Finite difference scheme	Error term	Stencil
1	$y'_m \approx \frac{y_{m+1} - y_m}{h}$	$-y''(\xi) \frac{h}{2}$	$(0, \underline{1}, 1)$
2	$y''_m \approx \frac{y_{m+2} - 2y_{m+1} + y_m}{h^2}$	$-y'''(\xi)h$	$(0, 0, 1, \underline{2}, 1)$
3	$y'''_m \approx \frac{y_{m+3} - 3y_{m+2} + 3y_{m+1} - y_m}{h^3}$	$-y^{iv}(\xi) \frac{3h}{2}$	$(0, 0, 0, \underline{1}, 3, \underline{3}, 1)$
4	$y^{iv}_m \approx \frac{y_{m+4} - 4y_{m+3} + 6y_{m+2} - 4y_{m+1} + y_m}{h^4}$	$-y^v(\xi)2h$	$(0, 0, 0, 0, 1, \underline{4}, 6, \underline{4}, 1)$

Figure 1(B) shows that the forward-difference family forms a progression with p similar to that of the continuous case where the curves slope up more slowly for larger p . Even though the nullspace is strictly that of the constant wave, since the only null eigenvalue is ν_0 , as p increases there are ever more near-zero eigenvalues for the low frequencies. Thus, in this family low frequencies are practically not penalized for high p .

Central-difference family. This is defined by the first-order central-difference stencil $\zeta = (-\frac{1}{2}, 0, \frac{1}{2})$. The p th-order derivative stencil has eigenvalues $\nu_m = \sin^{2p}(2\pi \frac{m}{M})$. All stencils are sawtooth. The stencil of order p can be obtained from the forward-difference stencil of order p by intercalating 0 every two components and dividing by 2^p . Fig. 1(C) shows that this family also has a progression with decreasing slopes at low frequencies, but since every one of its stencils is a sawtooth stencil, both the low and high frequencies are not penalized. As M grows, more frequencies are allowed and the net approaches the continuum limit. However, for the central-difference family the sawtooth frequency remains unpenalized.

5 Annealing Algorithms for Parameter Estimation

Existing algorithms for optimizing the elastic net energy function include gradient descent [1] and matrix iteration methods (Gauss-Seidel in [5]). However, for large β , the step size required for convergence of gradient descent becomes very small, and the matrix iteration methods may occasionally diverge. We therefore developed a new, more powerful, iteration scheme. For constant σ , we look for stationary points:

$$\frac{\partial E}{\partial \mathbf{Y}} = -\frac{1}{\sigma}(\mathbf{XW} - \mathbf{YG}) + \beta \mathbf{Y} \left(\frac{\mathbf{S} + \mathbf{S}^T}{2} \right) = \mathbf{0} \implies \mathbf{YA} = \mathbf{XW} \quad (3)$$

with weight matrix $\mathbf{W} = (w_{nm})$ and invertible diagonal matrix $\mathbf{G} = \text{diag}(g_m)$

$$w_{nm} \stackrel{\text{def}}{=} \frac{e^{-\frac{1}{2} \left\| \frac{x_n - y_m}{\sigma} \right\|^2}}{\sum_{m'=1}^M e^{-\frac{1}{2} \left\| \frac{x_n - y_{m'}}{\sigma} \right\|^2}} \quad g_m \stackrel{\text{def}}{=} \sum_{n=1}^N w_{nm} \quad \mathbf{A} \stackrel{\text{def}}{=} \mathbf{G} + \sigma \beta \left(\frac{\mathbf{S} + \mathbf{S}^T}{2} \right).$$

The weight w_{nm} is also the responsibility $p(m|\mathbf{x}_n)$ of centroid $\boldsymbol{\mu}_m$ for generating point \mathbf{x}_n , and so g_m is the total responsibility of centroid $\boldsymbol{\mu}_m$. The matrix \mathbf{XW} is then a list of average centroids. Since \mathbf{S} will typically be sparse (with a banded or block-banded structure), we can solve the system $\mathbf{YA} = \mathbf{XW}$ efficiently and robustly without inverting \mathbf{A} by computing (with prior reordering if desired) the Cholesky factorisation $\mathbf{A} = \mathbf{LL}^T$, where \mathbf{L} is lower triangular with nonnegative diagonal elements. \mathbf{Y} is then obtained by solving two triangular systems by Gaussian elimination.

The Cholesky factorisation [10] is stable for symmetric semipositive definite matrices and terminates in $\mathcal{O}(\frac{1}{3}M^3)$ for dense \mathbf{A} but much faster for sparse \mathbf{A} , since \mathbf{L} preserves the banded structure. Naturally, since \mathbf{W} and \mathbf{G} depend on \mathbf{Y} , the method should be iterated. The algorithm can be seen as an EM algorithm, so it converges with linear order. Thus, although compared with conventional gradient descent, it requires slightly more computation time per step (which is always dominated by the computation of \mathbf{W}), it generally requires fewer iterations to converge. This and its reliability render it preferable.

6 Simulation Results for Cortical Map Modeling

In cortical map modeling [5, 6], the nets first develop uniform retinotopy, and then at a bifurcation of the energy, ocular dominance (OD) and orientation (OR) maps emerge. There, the maps are waves of a specific frequency κ^* for a range of β values. This frequency κ^* results from a trade-off between the fitness term pushing for high frequencies and the tension term penalizing frequency m proportionally to $\beta\nu_m$. The analysis of the tension term alone indicates that the frequency κ^* should decrease with β and increase with the stencil order p , given the progression of the eigenvalue curves towards passing higher frequencies with increasing p . This is confirmed by our simulations (Fig. 2; only OD is shown). In [11], we show that these differential priors correspond to Mexican-hat lateral interactions with p oscillations (excitatory vs inhibitory); and that for $p > 1$ the geometric relations between the OD and OR maps (e.g. intersection angles) do not match those of the biological maps that have been observed so far.

The elastic nets resulting from the central-difference family very often contain sawtooth patterns. Such sawtooth patterns may take all the net or part of it, and can appear superimposed on a lower-frequency wave for some values of σ (see Fig. 3). One can also understand why this happens by noting that the tension term decouples in two, one for the even centroids and the other for the odd centroids. Other well-known stencils are also sawtooth, for example in 2D the quadratic-error Laplacian $\nabla_{\times}^2 = \frac{1}{2} \begin{pmatrix} 1 & 0 & 1 \\ 0 & -4 & 0 \\ 1 & 0 & 1 \end{pmatrix}$.

7 Discussion

The behavior of the generalized elastic net is determined by the joint effect of the fitness and tension terms of its energy function. Our separate analysis of the tension term provides insight into the structure of the minima of the energy and makes explicit the discrepancies between the continuous and the discrete formulations of the net. This analysis paves the way for further work in the generalized elastic net, such as the study

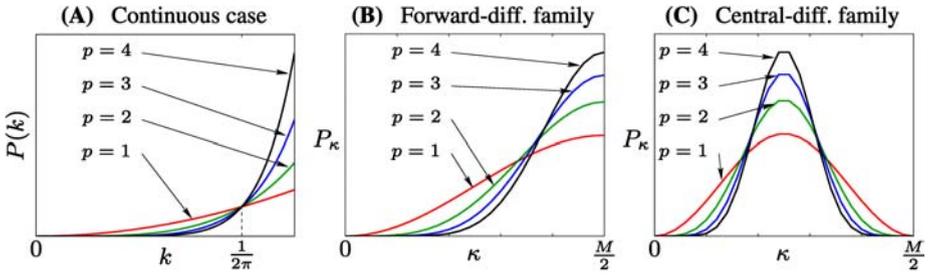


Fig. 1. 1D power spectrum associated with the derivative of order p of a continuous function, $P(k) = (2\pi k)^{2p}$, $k \in \mathbb{R}^+$; and for discrete nets with stencils (normalized by total power) from the forward- and central-difference families, $P_\kappa \stackrel{\text{def}}{=} |\hat{\zeta}_\kappa|^2$, $\kappa = 0, \dots, M/2$

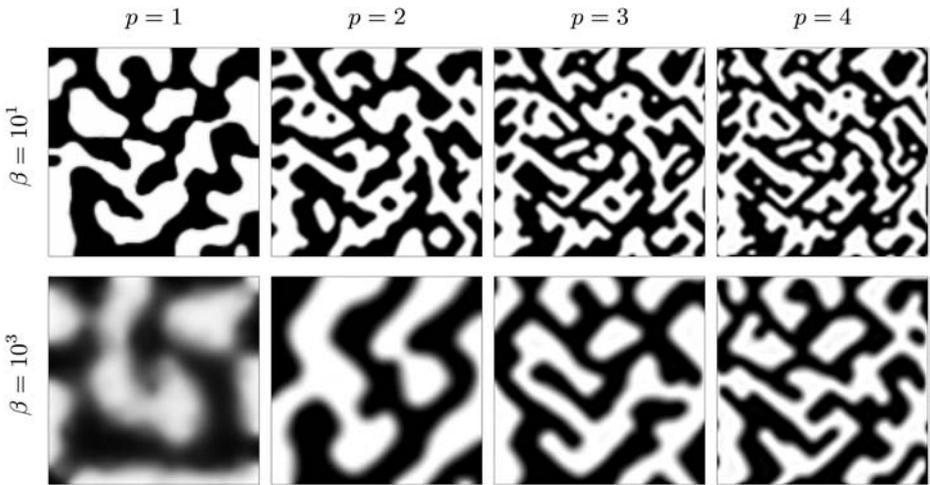


Fig. 2. Selected subset of ocular dominance map simulations with a 2D net with the forward-difference family and open b.c. Except for p and β , all other parameters and the initial conditions are the same. The stripes are narrower for low β and high p

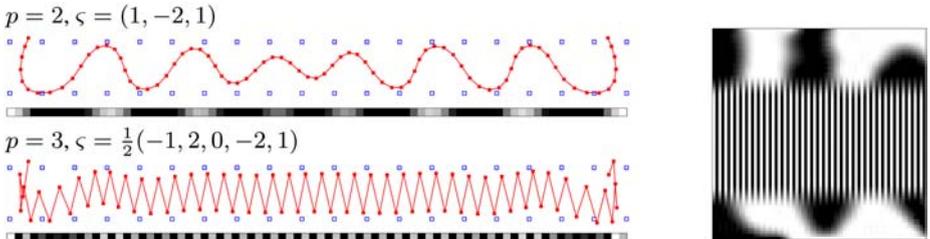


Fig. 3. Sawtooth stencils. *Left:* 1D nets in a stimulus space of (retinotopy, ocular dominance), open b.c., with a non-sawtooth stencil above and a sawtooth one below. *Right:* 2D sawtooths in ocular dominance with 3rd-order stencil $\varsigma = \frac{1}{2}(-1, 2, 0, -2, 1)$; cf. Fig. 2

of other stencil types, nonperiodic b.c., and the analysis of bifurcations from the Hessian of E and the character of the emergent nets. It also applies to other problems for which a discretized version of a quadratic smoothness regularizer is used, e.g. in computer vision, image processing or inverse problems. The model can also be used for unsupervised learning. We can extract MAP estimates for both \mathbf{Y} and σ (perhaps with a prior on σ) with an EM algorithm, again based on Cholesky factorization; or perform Bayesian inference on β to attain good generalisation to unseen data (e.g. [12]).

Can differential priors be used with the SOM and GTM? This is tricky for the SOM, since it is not defined through an objective function. However, it may be possible to introduce them in the learning rule (similarly to momentum terms). Since GTM defines a continuous mapping, differential priors can, in principle, be used exactly. However, this is practically cumbersome, so that they may be more conveniently approximated by discrete stencils as here.

Difference schemes have been used extensively in the numerical solution of differential equations [13]. In this context, the concerns regarding the choice of difference scheme are: accuracy (whether the truncation error is of high order, which will mean a faster convergence), stability (whether the approximate solution converges to the true one as the step size tends to zero) and sparsity (whether the scheme has few nonzero coefficients, for computational efficiency). In our context of differential priors with a discrete net, accuracy and stability are not so relevant, while whether the stencil is for instance sawtooth is crucial.

References

1. Durbin, R., Willshaw, D.: An analogue approach to the traveling salesman problem using an elastic net method. *Nature* **326** (1987) 689–691
2. Durbin, R., Szeliski, R., Yuille, A.: An analysis of the elastic net approach to the traveling salesman problem. *Neural Computation* **1** (1989) 348–358
3. Bishop, C.M., Svensén, M., Williams, C.K.I.: GTM: The generative topographic mapping. *Neural Computation* **10** (1998) 215–234
4. Kohonen, T.K.: *Self-Organizing Maps*. Springer-Verlag (1995)
5. Durbin, R., Mitchison, G.: A dimension reduction framework for understanding cortical maps. *Nature* **343** (1990) 644–647
6. Goodhill, G.J., Willshaw, D.J.: Application of the elastic net algorithm to the formation of ocular dominance stripes. *Network: Computation in Neural Systems* **1** (1990) 41–59
7. Dayan, P.: Arbitrary elastic topologies and ocular dominance. *Neural Computation* **5** (1993) 392–401
8. Carreira-Perpiñán, M.Á.: *Continuous Latent Variable Models for Dimensionality Reduction and Sequential Data Reconstruction*. PhD thesis, University of Sheffield, UK (2001)
9. Carreira-Perpiñán, M.Á., Goodhill, G.J.: Generalized elastic nets. In revision (2005)
10. Golub, G.H., van Loan, C.F.: *Matrix Computations*. Johns Hopkins University Press (1996)
11. Carreira-Perpiñán, M.Á., Goodhill, G.J.: Influence of lateral connections on the structure of cortical maps. *J. Neurophysiol.* **92** (2004) 2947–2959
12. Utsugi, A.: Hyperparameter selection for self-organizing maps. *Neural Computation* **9** (1997) 623–635
13. Samarskiĭ, A.A.: *The Theory of Difference Schemes*. Marcel Dekker (2001)