What makes a good model of natural images?

Yair Weiss[1,2]
[1] Hebrew University of Jerusalem
yweiss@cs.huji.ac.il

William T. Freeman[2]
[2] MIT CSAIL
billf@mit.edu

Abstract

Many low-level vision algorithms assume a prior probability over images, and there has been great interest in trying to learn this prior from examples. Since images are very non Gaussian, high dimensional, continuous signals, learning their distribution presents a tremendous computational challenge. Perhaps the most successful recent algorithm is the Fields of Experts (FOE) [20] model which has shown impressive performance by modeling image statistics with a product of potentials defined on filter outputs. However, as in previous models of images based on filter outputs [30], calculating the probability of an image given the model requires evaluating an intractable partition function. This makes learning very slow (requires Monte-Carlo sampling at every step) and makes it virtually impossible to compare the likelihood of two different models. Given this computational difficulty, it is hard to say whether nonintuitive features learned by such models represent a true property of natural images or an artifact of the approximations used during learning.

In this paper we present (1) tractable lower and upper bounds on the partition function of models based on filter outputs and (2) efficient learning algorithms that do not require any sampling. Our results are based on recent results in machine learning that deal with Gaussian potentials. We extend these results to non-Gaussian potentials and derive a novel, basis rotation algorithm for approximating the maximum likelihood filters. Our results allow us to (1) rigorously compare the likelihood of different models and (2) calculate high likelihood models of natural image statistics in a matter of minutes. Applying our results to previous models shows that the nonintuitive features are not *an artifact of the learning process but rather are capturing robust properties of natural images.*

1. Introduction

Significant progress in low-level vision has been achieved by algorithms that are based on energy minimization. Typically, the algorithm's output is calculated by min-
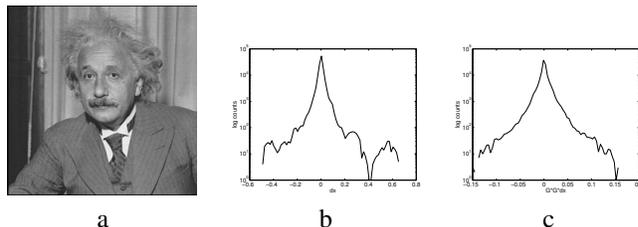
Figure 1. **a.** A natural image. **b-c.** Log histogram of derivatives at different scales. Natural images have characteristic, heavy-tailed, non-Gaussian distributions.
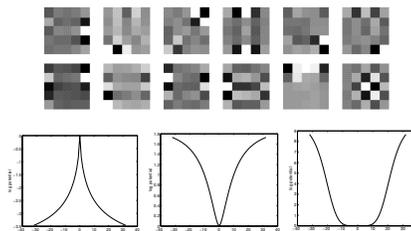
Figure 2. Non-intuitive results from previous models. **Top:** The filters learned by the FOE algorithm [19]. Note that they look nothing like derivative filters. **Bottom:** the potentials learned by the Zhu and Mumford algorithm on natural images [30]. For derivatives at the finest scale (left), the potential is qualitatively similar to the log histogram. But at coarser scales (middle and right) the potential is flipped, favoring many large filter responses.

imizing an energy function that is the sum of two terms: a data fidelity term which measure the likelihood of the input image given the output and a prior term which encodes prior assumptions about the output . Examples of tasks that have been tackled using this approach include optical flow estimation [20, 2], stereo vision [3, 5] and image segmentation. An important subclass of these problems is when the output is itself a "natural image". This includes problems such as transparency analysis [13], removal of camera blur [6] image denoising and image inpainting [19].

For low-level vision tasks where the output is a natural image, the prior should capture some knowledge about the space of natural images. This space is obviously a tiny fraction of the space of $N \times N$ matrices, but how can we

1

characterize it? Some of the earliest energy-based methods used a quadratic smoothness assumption (e.g. [8]). Thus the energy was simply the sum of squared local derivative operators. This corresponds to a Gaussian prior on images and would be most appropriate if the distribution of natural images were indeed Gaussian. Unfortunately, images are very non Gaussian. Figure 1 illustrates a well known property of natural images. When derivative-like filters are applied to natural images the distribution of the filter output is highly non Gaussian - it is peaked at zero and has heavy tails [16, 21, 18]. This property is remarkably robust and holds for a wide range of natural scenes. Similar, non Gaussian marginals are also obtained for optical flow and stereo [19, 9]. Thus a Gaussian prior is not appropriate and more recent algorithms typically assume a robust, non-quadratic energy on local derivatives (e.g. [2, 5]). Ideally, one would like to learn the functional form from a training data. Also, one would like to know whether basing the energy functions on local derivatives is the best thing to do.

In a seminal paper [30] Zhu and Mumford showed how to address both questions using the principle of maximum likelihood. Denoting by x an image, they defined the probability of an image by means of an energy function that depends on the output of linear filters w_k applied to the image:

$$\Pr(x; \{w_k, \Psi_k\}) = \frac{1}{Z(\{w_k, \Psi_k\})} e^{-\sum_{i,k} E_k(w_{ik}^T x)} \quad (1)$$

$$= \frac{1}{Z(\{w_k, \Psi_k\})} \prod_{i,k} \Psi_k(w_{ik}^T x) \quad (2)$$

where i is an index over image pixels and k is an index over linear filters. $w_{ik}^T x$ is the result of applying the linear filter w_k to image x at location i. The *partition function* $Z(\{w_k, \Psi_k\})$ is an explicit normalization constant and is defined by:

$$Z(\{w_k, \Psi_k\}) = \int_x \prod_{i,k} \Psi_k(w_{ik}^T x) dx \quad (3)$$

For an arbitrary energy function the partition function is intractable since it requires integrating over all possible images. If an image has N^2 pixels and we discretize it to have 256 possible gray levels, exact calculation would require summing over 256^{N^2} possible images.

Note that equation 2 contains as special cases some well known priors used in low-level vision. If the filters are just horizontal and vertical derivatives and the energy functions are quadratic, this gives the classical smoothness assumptions. If the filters are horizontal and vertical derivatives while the energy functions are robust norms, this gives the more modern robust smoothness assumptions. Zhu and Mumford proposed learning both the set of filters w_k and the corresponding energies E_k from data by maximizing the likelihood of the training set. Specifically, they assumed the

filters were chosen from a discrete set of oriented derivative-like filters while the energy functions could be arbitrarily shaped. Their findings, illustrated in figure 2 were very nonintuitive. For derivatives at the finest scale, the learned potentials were qualitatively similar to the log histograms and peaked at zero. But at the coarser scales the potentials were *inverted* — they had a minimum at zero, even though the log histograms have a maximum at zero. This inversion effect is more pronounced the coarser the filters.

Roth and Black [19] introduced the Fields of Experts (FOE) model which assumes a parametric, student T distribution for the potentials, but allows the filters to be arbitrary. Again, they used the principle of maximum likelihood to find the optimal filters. Their learned filters (shown in figure 2) do not resemble derivative filters at all. Nevertheless, they showed that using the learned filters gave far superior performance compared to simple derivative filters on a range of image-restoration problems. In [20] they extended this work to optical flow estimation, and again showed that using learned filters improved performance versus hand-chosen filters such as derivative filters.

Despite the progress made by using maximum likelihood to learn energy functions for low-level vision, two significant problems remain. The first problem is that performing the learning is excruciatingly slow. In both [30, 19], learning is performed using gradient ascent - by following the gradient of the log likelihood in equation 2. This gradient includes the gradient of the partition function which is intractable. Zhu and Mumford used Gibbs sampling in order to estimate the gradient of the partition function, and noted that it could take many sweeps of sampling to converge to a suitable gradient. Since sampling needs to be performed before each gradient descent step, learning is extremely slow even when faster sampling techniques are used [29, 28]. Roth and Black used an approximate sampling method called "contrastive divergence" [7]. Even with this approximation , they noted that learning is very slow.

A second problem with existing approaches based on maximum likelihood is that it is extremely difficult to actually compare the likelihood for two competing models. Again, this is due to the intractable partition function in equation 2. Even if we wait long enough for a sampling algorithm to give us fair samples from the model, calculating the partition function from a finite number of samples is a difficult problem [14]. Thus, we have no way of currently saying whether the nonintuitive findings illustrated in figure 2 represent a local minimum of the optimization, or whether they really give higher likelihood to natural images.

In this paper, we build on recent results in machine learning with the closely related *product of experts* model. This model is similar to equation 2 but every linear filter is ap-

plied at only one location:

$$\Pr(x; \{w_k, \Psi_k\}) = \frac{1}{Z(\{w_k, \Psi_k\}} e^{-\sum_k E_k(w_k^T x)} \quad (4)$$

$$= \frac{1}{Z(\{w_k, \Psi_k\})} \prod_k \Psi_k(w_k^T x) \quad (5)$$

For this model it has been shown that when the potentials are Gaussians, the optimal filters are the *minor components* - the principal components of the data with minimal eigenvalue [27]. Thus training in this case is extremely fast and requires just one eigenvector computation on the training set. It has also been shown that in the *undercomplete* case - when the number of filters is smaller than the dimensionality of x, the partition function can be calculated exactly using the singular value decomposition of the vectors w_k. Unfortunately, neither of these results is directly applicable to the case we are interested in - as mentioned earlier images are highly non-Gaussian so Gaussian potentials are not appropriate. Furthermore, the translation invariance of images would suggest that our prior also needs to be translation invariant as in the FOE model. If we have K filters in the FOE model, then the model is K times overcomplete.

It would thus be desirable to obtain (1) a fast algorithm for learning good filters and (2) a way to calculate the partition function in the overcomplete, non-Gaussian case. In this paper we provide both. We derive tractable lower and upper bounds on the partition function based on the Fourier transform of the filters $\{w_k\}$. We also show how to calculate high likelihood filters using iterated PCA *with no sampling required*. Applying our results to previous models shows that the nonintuitive features are *not* an artifact of the learning process but rather are capturing robust properties of natural images.

2. Analysis

We start by reviewing the results of Williams and Agakov [27] for Gaussian potentials. We will use a slightly different derivation that extends more easily to the non Gaussian case. Suppose we define a probability distribution over images using a single, linear filter. In this case, the energy of an image x is given by:

$$E_{Gaussian}(x; w) = (w^T x)^2 + \epsilon \|x\|^2 \quad (6)$$

where the $\epsilon \|x\|^2$ term is there to make sure that $e^{-E(x)}$ is normalizable - otherwise all directions orthogonal to w are completely unconstrained.

The log likelihood of a dataset $\{x_i\}$ is given (up to a term that is independent of w):

$$\ln \Pr(\{x_i\}; w) = -\sum_i \left((w^T x_i)^2 - \ln Z(w) \right) \quad (7)$$

What vector w will give the maximum of the log likelihood ? As pointed out by [12] a good vector should give low energy to the training images (so that $(w^T x_i)^2$ is minimal) but also give high energy to all other images (so that $Z(w)$ is minimized). In the absence of the $Z(w)$ term we could always choose $w = 0$. But what happens if we constrain the norm of w by $w^T w = 1$? For the Gaussian case, we can explicitly calculate $\ln Z(w) = -\ln \det(ww^T + \epsilon I)$ and this can be shown to be constant for any w that satisfies $w^T w = 1$. In fact, the following observation proves this is true for arbitrary energy functions:

Observation 1: Let $E(w^T x)$ be an arbitrary function of $w^T x$. Define $Z(w) = \int_x e^{-E(w^T x) + \epsilon x^T x} dx$. Then $Z(w) = Z(v)$ for any unit norm vectors w, v.

The proof follows from the fact that we can always choose an orthogonal transformation A so that $Aw = v$. The result then follows from a change of integration variables.

Corollary 1: If we restrict ourselves to unit norm vectors w and Gaussian potentials, then the maximum likelihood w is the *minor* component of the data - the eigenvector of the covariance matrix with minimal eigenvalue.

Proof: Since $Z(w)$ is constant for any unit norm vector, the MLE w is given by maximizing the first term:

$$w^* = \arg \min_{w: w^T w = 1} \sum_i (w^T x_i)^2 \quad (8)$$

and this is the minor component of the data.

Observation 1 and Corollary 1 have analogues in the case of K orthogonal vectors.

Observation 2: Let $E(w^T x)$ be an arbitrary function of $w^T x$. Define $Z(w) = \int_x e^{-\sum_k E(w_k^T x) + \epsilon x^T x} dx$. Then $Z(w) = Z(v)$ for any set of K orthonormal vectors $\{w_i\}, \{v_i\}$.

Corollary 2: If we restrict ourselves to orthoronormal set of K vectors w, then the maximum likelihood linear filters w_k for Gaussian potentials are the K *minor* components of the data.

To summarize, in the Gaussian case we can find the optimal orthonormal filters w_k using an extremely simple algorithm - just calculating the minor components of the data. It can also be shown [27] that the restriction to orthonormal vectors is not necessary - if we remove that restriction and solve for the optimal vectors w_k without any costraints we still recover the minor components. The only difference will be that each minor component v_k is rescaled according to its eigenvalue $w_k = v_k \frac{1}{\epsilon + \lambda_k}$. Finally, the result also holds when we restrict our linear filters to be compact in the image. For example, for computational reasons, we may want our filters to be no larger than $M \times M$. In this case, the optimal filters are simply the minor components of the set of $M \times M$ patches of natural images.
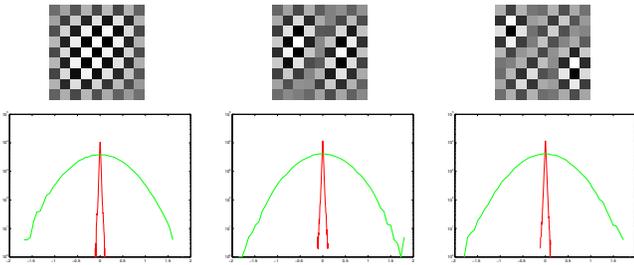
Figure 3. Top: The minor components of 9×9 patches taken from the Einstein image. Bottom: Log histograms of the filter outputs on the Einstein image (red) and a white noise image (green). Even though the minor components look nothing like derivative filters, they give low energy to natural images but high energy to all other images.

Gaussian FOE   Gaussian POE

a   b   c   d

Figure 4. **a.-b** The filter that maximizes the likelihood of the Einstein image in a Gaussian Fields of Experts model along with its power spectrum. **c.-d** The filter that maximizes the likelihood of the Einstein image in a Gaussian Products of Experts model along with its power spectrum.

Figure 3 show the K minor components of a set of natural image patches (of size 9×9). They look nothing like derivative filters and at first glance seem highly counterintuitive. Nevertheless these are the maximum likelihood filters to use with Gaussian potentials when building a model of natural images. This is simply because they rarely fire on natural images (since they minimize $w^T x$) but fire frequently on all possible images (since they are constrained to be unit norm). Figure 3b shows the histogram of the filter output on a natural image (red) versus a white noise image (green).

What about non Gaussian potentials ? Using observation 2, we have:

Corollary 3: If we restrict ourselves to orthoronormal set of K vectors w, then the maximum likelihood linear filters w_k^* for an arbitray energy function $E(w^T x)$ are the K orthogonal vectors that minimize:

$$\{w_k^*\} = \arg \min_{\{w_k\} \in ortho} \sum_i \sum_k E(w_k^T x_i) \qquad (9)$$

Although the minimum in equation 9 cannot be calculated by a simple eigenvector calculation, note that it does not require any sampling or evaluation of Z. In section 3 we show an efficient EM algorithm for performing the minimization for a large class of energy functions.

The fact that the optimal filters can be found without sampling or partition function evaluation for undercomplete models was pointed out by Welling et al. [26]. They used a slightly different probabilistic model - rather than adding $\epsilon \|x\|^2$ to the energy function, they only assumed a Gaussian distribution on directions orthogonal to w_k. This makes it difficult to directly extend their result to the overcomplete case. But the basic idea behind corollaries 1 - 3 is very simple and holds for overcomplete representations as well - if we can find a set for which the partition function is constant, the optimal filters in that set can be found using constrained optimization and without sampling. The challenge is to find a constraint set upon which the partition function is constant.
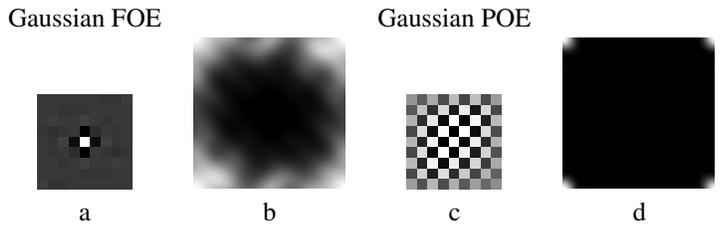
2.1. Gaussian Fields of Experts

A **Gaussian Field of Experts** (GFOE) prior is of the form:

$$\Pr(x; \{w_k\}) = \frac{1}{Z_{GFOE}(\{w_k\})} \prod_{i,k} e^{-(w_{ik}^T x)^2} \qquad (10)$$

Since this is a jointly Gaussian pdf, its partition function is simply:

$$-\ln Z_{GFOE}(\{w_k\}) = \ln \det \left(\sum_{i,k} w_{ik} w_{ik}^T \right) \qquad (11)$$

Since the log likelihood is the energy minus $\ln Z$, given two filter sets that fire equally on the training set, the log likelihood will favor the set that maximize the log determinant in equation 11. But what filters are these ? We can get better intuition by looking at the filters in the frequency domain.

Observation 4: Let $W_k(\omega)$ be the Fourier transform of filter w_k. Then:

$$-\ln Z_{GFOE}(\{w_k\}) = \sum_\omega \ln \left(\sum_k |W_k(\omega)|^2 \right) \qquad (12)$$

Proof: This follows from the fact that w_{ik} is simply the shift of w to pixel i and so applying all w_{ik} to an image is equivalent to *convolving* the image with k filters . This makes the determinant on the right hand side of equation 11 to be the determinant of $\epsilon I + A^T A$ where $A = [A_1; A_2; \cdots A_K]$ and each A_k is a convolution matrix. It can then be shown that $A^T A$ is also a convolution with a filter whose Fourier coefficients are the sum of the squares of the individual filters (e.g. [25]). The eigenvalues of convolution matrices are simply the Fourier transform of the corresponding filter.

4

From observation 4 it is clear that the log determinant can always be increased by increasing the norm of the filters. This is similar to the case of undercomplete products of experts, where we saw that constraining the filters to be unit norm was enough to make the partition function constant. This, however, is no longer the case in GFOE.

Corollary 4: Suppose we constrain all filters to have unit norm. The the negative log partition function $-\ln Z_{GFOE}$ is maximized when the filters satisfy the *tiling constraint:*.

$$\sum_k |W_k(\omega)|^2 = c \qquad (13)$$

In other words, the summed energy of all filters in a given frequency should be a constant (independent of frequency).

Proof: This follows from adding Lagrange multipliers to the equation for $-\ln Z_{GFOE}$ and differentiating.

The tiling constraint is well studied in filter design and signal processing. It is equivalent to the requirement that the set of filters form a *tight frame* or be *self inverting* [22]. It means that recovering the original signal from the k convolved signals is trivial - we simply convolve each filtered signal with a flipped version of the same filter and sum (see [22] for more details). It is interesting that this same constraint comes up in the case of maximum likelihood estimation.

The tiling constraint has a simple interpertation in the case of a single filter w - it simply requires that w be orthogonal to all its translates (the tiling constraint means that the convolution of w with itself is the delta function). One such filter is the delta function. Another example is a filter that is simply white noise - for large filter size, this filter will become orthogonal to its translates. When there are multiple filters, however, none of them needs to be orthogonal to its translates to satisfy the tiling constraint. A simple example are the pair of filters $[1, 1]$ and $[1, -1]$ which can be shown to tile together. This is because the convolution of one filter with itself cancels out the convolution of the second filter with itself everywhere but the origin.

Combining the form of $\log Z$ with the energy (which involves the energy of convolving x with the filters) gives a simple equation for the optimal filters in the Fourier domain.

Observation 5: Let $W_k(\omega)$ be the Fourier transform of filter w_k and $X(\omega)$ the Fourier transform of x, then the maximum likelihood filters for a GFOE model satisfy:

$$\sum_k |W_k(\omega)|^2 = \frac{1}{|X(\omega)|^2} \qquad (14)$$

When there is a single filter $K = 1$, a filter satisfying equation 14 is called a *whitening filter* [1]. Figure 4 shows a 9×9 whitening filter for the einstein image along with its power spectrum. This is the optimal filter for $K = 1$ in

a Gaussian Fields of Experts model. For comparison, figure 4 also shows the minor component for the same image along with its power spectrum. This is the optimal filter for $K = 1$ in a Gaussian product of experts model. Whereas both filters are predominantly high frequency, the Fields of Experts partition function favors filters that approximately tile, and hence the filter is much broader in frequency (and more localized in space).

2.2. Gaussian Scale Mixture Fields of Experts

As mentioned in the introduction, Gaussian potentials are not well suited to models of natural images. It turns out, however, that many of the potentials used in low-level vision are well fit by a *Gaussian Scale Mixture* (GSM) [18]. These are potentials that are a mixture of zero mean Gaussians:

$$\Psi(x) \propto \sum_{j=1}^{J} \frac{\pi_j}{\sigma_j} e^{-\frac{x^2}{2\sigma_j^2}} \qquad (15)$$

Explicit GSM priors were used in [18],[6],[13]. Also, the student T distribution used in the FOE model can be shown to be a GSM [17]. Essentially the only requirement is that the potential be monotonically decreasing away from zero and symmetric at zero. Thus the potentials found by Zhu and Mumford for the finest scale (see figure 2) are GSM priors but those at the coarser scales are not.

We now define a Gaussian Scale Mixture Fields of Experts (GSM FOE) model to be a model of the form:

$$\Pr(x; \{w_k\}) = \frac{1}{Z_{GSM}(\{w_k\})} \prod_{i,k} \Psi(w_{ik}^T x) \qquad (16)$$

where Ψ is a GSM potential.

Since each potential is a mixture of Gaussians the GSM-FOE can also be seen as a mixture of an exponentially large number of Gaussians (every time we multiply a mixture of J Gaussians we obtain a new mixture of J^2 Gaussians). Thus the partition function Z_{GSMFOE} can be expressed analytically as a sum of determinants. Unfortunately, the number of determinants is the sum is exponentially large, so that exact evaluation is intractable. We now present, however, tractable, upper and lower bounds on the partition function.

Theorem 1: Let Z_{GSM} be the partition function of a GSM FOE model with filters $\{w_k\}$ and GSM potential defined by $\{\pi_j, \sigma_j\}_{j=1}^{J}$. We assume that the GSM standard deviations are ordered in increasing magnitude $\sigma_1 \leq \sigma_2 \cdots \leq \sigma_J$. Let Z_{GFOE} be the partition function of Gaussian FOE model with the same filters but scaled by $\frac{1}{\sigma_J}$. Then $\ln Z_{GSM}$ can be bounded above and below by $\ln Z_{GFOE}$ plus constants that do not depend on the filters:

$$\ln Z_{GFOE} + Ma \leq \ln Z_{GSM} \leq \ln Z_{GFOE} + Mb \qquad (17)$$

Filter	Lower	Upper
Laplacian	-230	-135
δ	214	309
9×9 edge	51	146
MCA	44	139

a   b
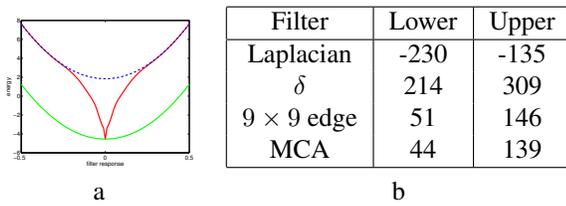
Figure 5. **a.** An illustration of the energy bound lemma. A GSM energy (red) is upper and lower bounded by a quadratic. **b.** Upper and lower bounds on the negative log likelihood for different filters using a GSM prior. Note that there is no overlap between the bounds of the Laplacian and the other filters. Thus the Laplacian provably gives the highest likelihood among this set of filters.

with:

$$a = \ln \frac{\pi_J}{\sigma_J} \tag{18}$$

$$b = \ln \left(\sum_j \frac{\pi_j}{\sigma_j} \right) \tag{19}$$

and M is the number of pixels times the number of filters.

The proof is based on the following lemma.

Energy Bound Lemma: Let $\Psi(x)$ be a GSM potential defined by $\{\pi_j, \sigma_j\}_{j=1}^J$. Let $E(x)$ be the energy of that potential $E(x) = \ln \Psi(x)$. Then:

$$\frac{x^2}{2\sigma_J^2} - \ln \left(\sum_j \frac{\pi_j}{\sigma_j} \right) \leq E(x) \leq \frac{x^2}{2\sigma_J^2} - \ln \frac{\pi_J}{\sigma_J} \tag{20}$$

Proof of Lemma: Since $E(x)$ is the negative log probability of a mixture of Gaussians it has the following, variational, interpertation (e.g. [4, 15]):

$$E(x) = \min_q \sum_j q_j \left(\frac{1}{2\sigma_j^2} x^2 - \ln \frac{\pi_j}{\sigma_j} + \ln q_j \right) \tag{21}$$

Where the minimum is with respect to vectors q that are positive and sum to one. An upper bound is immediately obtained by choosing a particular q that is zero everywhere but the last component. A lower bound is obtained by allowing two different vectors q, one minimizes the first term and the other minimizes the second two terms in equation 21.

Since the energy bound lemma holds for any x, exponentiating all sides of the energy bound and then integrating gives the desired bounds on the partition function.

The tightness of the bounds will depend on various parameters. Obviously, if the GSM is close to a single Gaussian the lower and upper bounds coincide. The table in figure 5 shows the lower and upper bounds of the log likelihood of the Einstein image for several filters using a specific GSM FOE model (the GSM parameters were fit to the histogram in figure 1). As can be seen, even when the upper and

a   b   c
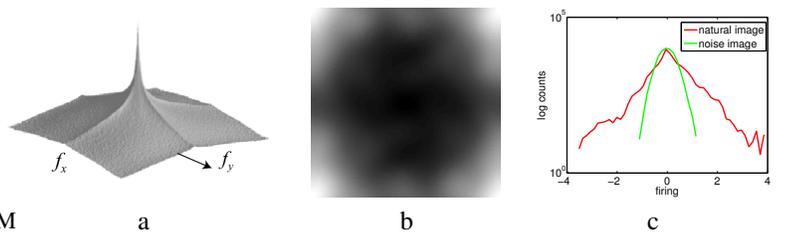
Figure 6. Explanation of non-intuitive findings in previous models. **a.** Natural images have a power spectrum that is mostly concentrated at low spatial frequencies and falls off as $1/f^2$ (figure replotted from [23]). **b.** The Roth and Black filters have a spectrum concentrated at the high spatial frequencies so that they fire rarely on natural images **c.** A coarse scale derivative filter, actually fires more frequently on natural images compared to random images with the same intensity range. Hence the Zhu and Mumford model learned an inverted potential for the coarse derivatives (see figure 2).

lower bounds do not coincide, they still allow us to rigorously prefer some filters over others. In particular, the 3×3 Laplacian, which is close to the optimal filter for a Gaussian model (see figure 4a) is the best filter in this set when a GSM potential is used. Interestingly, Zhu and Mumford also found the 3×3 Laplacian to be the best single filter to use, in their study [30].

2.3. Summary - what makes a good model of natural images ?

To summarize, when we use GSM priors in models based on filter outputs, we seek filters that fire *rarely on natural images* but *frequently on all other images*. But how can we characterize filters that fire rarely on natural images ? One well-known property property of natural images is that their image power spectra tend to fall off with increasing spatial frequency (e.g. [24]). Typically this is modelled by assuming that power falls off as $1/f^2$. This is a remarkably consistent property - figure 6a shows the mean power of 6000 natural scenes (replotted from [23]) which obeys a power law with the exponent 2.02. Since maximum likelihood seeks filters that fire rarely on natural images, this means that good filters should have most of their power concentrated at the high spatial frequencies. Figure 6b shows the average power spectrum of the Roth and Black filters. Although the filters look relatively random, their power spectrum is *not* uniformly spread out, but rather concentrated at the high frequencies.

The $1/f$ amplitude spectrum property also explains the inverted potentials found by Zhu and Mumford for coarse derivatives. In their case, the class of "all other images" was restricted to have the same range in intensities as natural images (i.e. all signals considered had intensity values between 0 and 255). When this restriction is combined with the $1/f$ property, this means that coarse derivatives (which

6

are primarily low spatial frequency) actually fire *more* on natural images than on white noise (compare figure 6c with figure 3). Thus a strong response from a coarse derivative filter on a signal x makes it more likely that x is indeed a natural image, which is precisely what the inverted potentials are imposing.

3. Basis Rotation Algorithm

From the upper and lower bounds on $\ln Z_{GSM}$ we can immediately obtain a lower bound on the log likelihood. As in many variational approaches to machine learning [11] we can then run an optimization algorithm such as gradient ascent to increase the lower bound.

An even simpler strategy is to restrict the search to a set of filters w for which $\ln Z$ is constant, and just find filters in that set that fire rarely on natural images. We can easily define such a set, by considering all possible *rotations* of a single basis set of filters b_k. That is, if we denote by B a matrix whose kth column is b_k and R is any $K \times K$ orthogonal matrix then $\ln Z_{GFOE}(B) = \ln Z_{GFOE}(RB)$ (this follows directly from equation 12).

In order to learn an orthogonal matrix R such that the columns of $W = RB$ minimize the energy on the training set we use a variant of the EM algorithm. We learn the columns of R one by one, where each column r_k is restricted to be unit norm and orthogonal to the previous columns.

We take the training images and divide them into $L \times L$ patches. Let $\{y(t)\}$ denote these training patches.

E step:

$$q_j(t) \propto \frac{\pi_j}{\sigma_j} e^{-\frac{1}{2\sigma_j^2}(w^T y(t))^2} \tag{22}$$

M step:

$$r = eig \min B^T \left(\sum_{t,j} \frac{q_j(t)}{\sigma_j^2} y(t) y^T(t) \right) B \tag{23}$$

$$w = Br \tag{24}$$

It is easy to show that the GSM energy on the training set never increases at every iteration, and the unit norm constraint on r is satisfied.

After finding k columns of r we require that r_{k+1} be orthogonal to the previous columns by building a matrix L whose columns are all orthogonal to $r_1 \cdots r_k$. the M step is then modified by replacing B with LB. The idea of basis rotation has also proven powerful in the context of complete ICA algorithms [10].

4. Experiments

We used the algorithms described above to learn filters for a FOE prior over natural images. We used the same
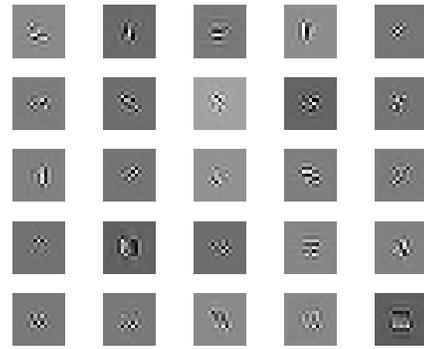
Figure 7. Filters found using the basis rotation algorithm by only considering filter sets that have the exact same mean power spectrum as the Roth and Black filters. These filters give higher upper and lower bounds on the likelihood of the training set compared to the Roth and Black filters, and they exhibit more structure.

Input	FOE (5×5)	Basis Rotation

| | PSNR:30.65 dB | PSNR:31.48 dB |
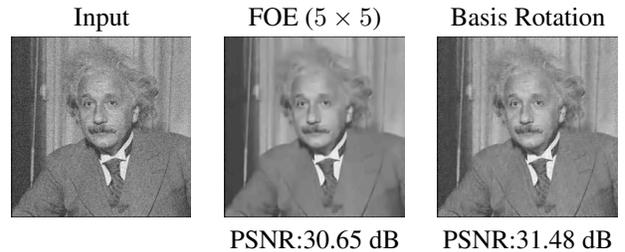
Figure 8. Comparing denoising results with the Roth and Black 5×5 filters and the 15×15 filters learned using the EM algorithm. The larger filters can pick up faint lines and edges (e.g. the texture pattern on the vest, the hair) while the 5×5 filters tend to oversmooth.

training set used by Roth and Black [19] – a subset of the Berkeley segmentation database. For training 5×5 filters, we used the same set of patches used by Roth and Black, while for training larger filters, we sampled $65,000$ 15×15 patches from the training set. In all the experiments reported here, the GSM prior was fixed and had the shape shown in figure 5.

In a first set of experiments, we trained 5×5 and 3×3 filters using conjugate gradient ascent on the approximate log likelihood. We found that learning was quite rapid - less than half an hour to train 24 5×5 filters. The filters found were different between different runs of the learning, but always had the same characteristics as the Roth and Black filters shown in figure 2. They were predominantly high frequencies but otherwise unstructured.

In a second set of experiments, we trained 15×15 filters using the basis rotation algorithm. As a basis set, we either used shifted versions of the whitening filter or shifted versions of a filter whose power spectrum equals the mean power spectrum of the Roth and Black 5×5 filters (i.e. the basis filter was the inverse Fourier transform of the power spectrum showed in figure 6). The results were essentially equivalent when using either basis set. Figure 7 shows the

learned filters using the Roth and Black basis set. By construction, these filters have the same mean power spectrum as the Roth and Black filters, so in terms of the second-order statistics of natural scenes, they are equivalent. But the higher order statistics are quite different - the basis rotation filters are more structured and elongated, oriented receptive fields are consistently found. This is consistent with previous results on learning receptive fields [16, 1].

Since we are searching in a space in which the bound on the partition function is constant, we can compare the approximate likelihoods of different filter sets by simply measuring which filter set fires more rarely on natural images. We indeed find that the larger, more structured filters, have higher upper and lower bounds on the log likelihood compared to the unstructured Roth and Black filters or compared to taking translates of a single whitening filter. However, the differences can be quite small (e.g. the Roth and Black filters achieve 88% of the minimal energy, while the whitening filter by itself achieves 98%).

In a final set of experiments, we compared the denoising performance of the different learned filters. Given an input image y we used conjugate gradient descent to minimize $J(x) = -\ln \Pr(x; w) + \frac{1}{2\sigma^2}\|x - y\|^2$ where σ^2 is the variance of the observation noise and $\Pr(x; w)$ is the GSM FOE prior (equation 16). Figure 8 shows results on the Einstein image (note that this image was not part of the training set). The 15×15 filters tend to preserve weak edges and lines (e.g. the texture patterns on the vest and on the hair) while the 5×5 filters tend to oversmooth. Figure 8 also gives the peak signal to noise ratio (PSNR [19]) which is better for the 15×15 result (although which result looks better may be a matter of taste). The result in figure 8b uses the FOE filters, but it is *not* using the Roth and Black denoising algorithm (which includes a regularization constant tuned on a training set, and filter-specific student T potentials). Their full algorithm gives PSNR 31.72 but the perceptual difference remains - the 5×5 filters consistently oversmooth faint lines and edges which the 15×15 filters can recover.

5. Discussion

Despite much progress in understanding natural image statistics, it has been difficult to translate this knowledge into working machine vision algorithms. For example, the prior learned by Zhu and Mumford, published over 10 years ago, has not been widely adopted in machine vision. The more recent FOE model is somewhat more widely used, but far less than would be expected given the importance of using a good prior in many machine vision applications. Two barriers to adoption have been (1) the huge computational burden to learn them and (2) the nonintuitive features or potentials that have been learned.

In this paper we have addressed both of these problems. We derived a rigorous upper and lower bound on the log

partition function of models based on filter outputs and suggested a novel basis rotation algorithm to learn high likelihood filters. Our analysis indicates that good filters to use with a GSM prior are ones that fire rarely on natural images while frequently on all other images. When this is combined with the $1/f$ property of natural image amplitude spectra, it explains why the Roth and Black filters are predominantly high frequency and why inverted potentials are obtained for coarse derivative filters. We also showed that using our basis rotation algorithm it is possible to learn even better filters in a matter of minutes.

References

[1] A. J. Bell and T. J. Sejnowski. Edges are the independent components of natural scenes. In *NIPS*, pages 831–837, 1996. 5, 8

[2] M. J. Black and P. Anandan. The robust estimation of multiple motions: affine and piecewise smooth fields. Technical Report spl-93-092, Xerox PARC, 1993. 1, 2

[3] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. In *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, 1999. 1

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. B*, 39:1–38, 1977. 6

[5] P. Felzenszwalb and D. Huttenlocher. Efficient belief propagation for early vision. In *Proceedings of IEEE CVPR*, pages 261–268, 2004. 1, 2

[6] R. Fergus, B. Singh, A. Hertzmann, S. T. Roweis, and W. T. Freeman. Removing camera shake from a single photograph. *ACM Trans. Graph.*, 25(3):787–794, 2006. 1, 5

[7] G. E. Hinton and Y. W. Teh. Discovering multiple constraints that are frequently approximately satisfied. In *Proceedings of Uncertainty in Artificial Intelligence (UAI-2001)*, 2001. 7

[8] B. K. P. Horn and B. G. Schunck. Determining optical flow. *Artif. Intell.*, 17(1–3):185–203, August 1981. 2

[9] J. Huang, A. B. Lee, and D. Mumford. Statistics of range images. In *CVPR*, pages 1324–1331, 2000. 2

[10] A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13(4-5):411–430, 2000. 7

[11] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. An introduction to variational methods for graphical models. In M. Jordan, editor, *Learning in Graphical Models*. MIT Press, 1998. 7

[12] Y. LeCun and F. Huang. Loss functions for discriminative training of energy-based models. In *Proc. of the 10-th International Workshop on Artificial Intelligence and Statistics (AIStats'05)*, 2005. 3

[13] A. Levin and Y. Weiss. User assisted separation of reflections from a single image using a sparsity prior. In *Proc. of the European Conference on Computer Vision (ECCV)*, 2004. 1, 5

[14] R. Neal. *Bayesian Learning for Neural Networks*. Springer-Verlag, 1996. 2

[15] R. Neal and G. Hinton. A new view of the EM algorithm that justifies incremental and other variants. *Biometrika*, 1993. submitted. 6

[16] B. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–608, 1996. 2, 8

[17] S. Osindero, M. Welling, and G. E. Hinton. Topographic product models applied to natural scene statistics. *Neural Computation*, 18(2):381–414, 2006. 5

[18] J. Portilla, V. Strela, M. Wainwright, and E. P. Simoncelli. Image denoising using scale mixtures of gaussians in the wavelet domain. *IEEE Trans Image Processing*, 12(11):1338–1351, 2003. 2, 5

[19] S. Roth and M. J. Black. Fields of experts: A framework for learning image priors. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2005. 1, 2, 7, 8

[20] S. Roth and M. J. Black. On the spatial statistics of optical flow. In *ICCV*, pages 42–49, 2005. 1, 2

[21] E. Simoncelli. Statistical models for images:compression restoration and synthesis. In *Proc Asilomar Conference on Signals, Systems and Computers*, pages 673–678, 1997. 2

[22] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger. Shiftable multiscale transforms. *IEEE Transactions on Information Theory*, 38(2):587–607, 1992. 5

[23] A. Torralba and A. Oliva. Statistics of natural image categories. *Network: Computation in Neural Systems*, 14:391–412, 2003. 6

[24] A. van der Schaaf and J. van Hateren. Modelling the power spectra of natural images: statistics and information. *Vision Research*, 36(17):2759–70, 1996. 6

[25] Y. Weiss. Deriving intrinsic images from image sequences. In *Proc. Intl. Conf. Computer Vision*, pages 68–75. 2001. 4

[26] M. Welling, R. S. Zemel, and G. E. Hinton. Efficient parametric projection pursuit density estimation. In *UAI*, pages 575–582, 2003. 4

[27] C. K. I. Williams and F. V. Agakov. Products of gaussians and probabilistic minor component analysis. *Neural Computation*, 14(5):1169–1182, 2002. 3

[28] S. Zhu and X. Liu. Learning in gibbsian fields: How fast and how accurate can it be? *IEEE Trans on PAMI*, 2002. 2

[29] S. C. Zhu, X. W. Liu, and Y. N. Wu. Exploring texture enesmbles by efficient markov chain monte carlo — toward a trichromacy theory of texture. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(6):554–569, 2000. 2

[30] S. C. Zhu and D. Mumford. Prior learning and gibbs reaction-diffusion. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(11):1236–1250, 1997. 1, 2, 6