MERL – A MITSUBISHI ELECTRIC RESEARCH LABORATORY
http://www.merl.com

Pattern discovery via entropy minimization

Matthew Brand

TR-98-21 October 1998

Abstract

We propose a framework for learning hidden-variable models by optimizing entropies, in which entropy minimization, posterior maximization, and free energy minimization are all equivalent. Solutions for the maximum *a posteriori* (MAP) estimator yield powerful learning algorithms that combine all the charms of expectation-maximization and deterministic annealing. Contained as special cases are the methods of maximum entropy, maximum likelihood, and a new method, maximum structure. We focus on the maximum structure case, in which entropy minimization maximizes the amount of evidence supporting each parameter while minimizing uncertainty in the sufficient statistics and cross-entropy between the model and the data. In iterative estimation, the MAP estimator gradually extinguishes excess parameters, sculpting a model structure that reflects hidden structures in the data. These models are highly resistant to over-fitting and have the particular virtue of being easy to interpret, often yielding insights into the hidden causes that generate the data.

Publication History:–
1. 8mar98 first circulated.
2. 6oct98 accepted to Uncertainty'99, Society of Artificial Intelligence and Statistics
3. 29oct98 final version uploaded for publication by Morgan Kaufmann.

Pattern discovery via entropy minimization

Matthew Brand*
MERL
Cambridge, MA 02139, USA
8mar98 revised 29oct98.

Abstract

We propose a framework for learning hidden-variable models by optimizing entropies, in which entropy minimization, posterior maximization, and free energy minimization are all equivalent. Solutions for the maximum *a posteriori* (MAP) estimator yield powerful learning algorithms that combine all the charms of expectation-maximization and deterministic annealing. Contained as special cases are the methods of maximum entropy, maximum likelihood, and a new method, maximum structure. We focus on the maximum structure case, in which entropy minimization maximizes the amount of evidence supporting each parameter while minimizing uncertainty in the sufficient statistics and cross-entropy between the model and the data. In iterative estimation, the MAP estimator gradually extinguishes excess parameters, sculpting a model structure that reflects hidden structures in the data. These models are highly resistant to over-fitting and have the particular virtue of being easy to interpret, often yielding insights into the hidden causes that generate the data.

1 Motivation

In pattern discovery we seek a model that reflects the structure of the data, which we hope in turn reflects the structure of the generating process. The unspoken premise is that the universe is constructed out of relatively small processes; in order to produce an object (dataset) larger than itself, a small process must have some kind of repetition structure, e.g., loops. Therefore a dataset is an "unrolled" record of the process' internal structure. If we can find a compact model of the data, we feel increasingly confident that 1) we can interpret the model and in doing so learn about the process, and 2) predictions made from the model will be consistent with new samples taken from the process. We now know that #2 is well-founded: Recent theorems in algorithmic complexity assure us that of all possible models (e.g., of all possible Turing machines) the one yielding the smallest two part encoding—model plus data relative to the the model—is the best strategy for hypothesis identification [Vitanyi and Li, 1996] and almost always the best strategy for prediction [Vitányi and Li, 1997].

Sadly, the function which yields the most compact model is not computable. Furthermore, we only know how to search efficiently for accurate models (disregarding compactness) in quite restricted spaces, e.g., for locally optimal parameterizations of stochastic models whose kind and structure we fix in advance. In this paper we broaden that scope by estimating both model structure and parameters in a manner that reveals salient structures in the data. Fortuitously, our framework also yields a fast hill-climbing procedure for finding nearly global optima.

2 Setting

We are interested in learning (point estimating) a single model that is highly predictive and whose structure reflects the structure of the generating process. We must address three important limits: finite data, finite time, and finite precision. In particular, we want to extract as much *essential* structure as possible from a dataset without modeling any of its *accidental* structure (e.g., noise and sampling artifacts); we want to do so without any wasted or speculative computation (e.g., models discarded by a selection process); and we want to maximize the information content of all parameters (e.g., bits of evidence supporting each parameter) while being realistic about the information capacity of digitally represented numbers. We begin in a Bayesian setting, but in §6 we shall show that our framework has its clearest interpretation when learning is considered a problem of minimizing entropies.

*Email: brand@merl.com or brand@media.mit.edu.

We begin with a set of observations $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ and a hypothesis class of hidden-variable models whose likelihoods are (approximated in practice by) products of densities drawn from the flat-exponential family, e.g., those having minimal sufficient statistics and parameter vectors of equal dimension[1]. We use hidden variables because we assume that the data has latent structure and/or is incomplete. The vector $\boldsymbol{\theta} = \{\theta_1, \cdots, \theta_N\}$ parameterizes a model and specifies its structure via a sparse encoding. Starting from a random over-parameterized model $\boldsymbol{\theta}$ we seek an optimal embedded model $\boldsymbol{\theta}^*$ that maximizes the posterior given by Bayes' rule: $\boldsymbol{\theta}^* = \operatorname{argmax}_{\boldsymbol{\theta}} P(\boldsymbol{\theta}|\boldsymbol{X}) \propto P(\boldsymbol{X}|\boldsymbol{\theta})P(\boldsymbol{\theta})$, where the likelihood $P(\boldsymbol{X}|\boldsymbol{\theta})$ measures accuracy in modeling the data and the prior $P(\boldsymbol{\theta})$ measures consistency with our background knowledge.

3 Maximum structure priors

What is our background knowledge? Let us assert that the vector of model parameters $\boldsymbol{\theta}$ is itself taken from a random variable $\boldsymbol{\Theta}$ which generates all possible processes. We know that on average, these processes are not infinitely unpredictable—otherwise learning would be impossible! Therefore our background knowledge is that, on average, learning will succeed. More formally, we say that the expected entropy of processes from $\boldsymbol{\Theta}$ is finite. We write this prior knowledge as $\xi\colon\ E_{\boldsymbol{\Theta}}[H(\boldsymbol{\theta})] = h$, where $H(\boldsymbol{\theta})$ is an entropy measure assessed on the model specified by $\boldsymbol{\theta}$, $E_{\boldsymbol{\Theta}}$ is the expectation with regard to $\boldsymbol{\Theta}$, and h is some finite value. The classic maximum-entropy method [Jaynes, 1982, §2] allows us to derive the mathematical form of a distribution from knowledge ξ about its expectations via Euler-Lagrange equations, yielding

$$P(\boldsymbol{\theta}|\xi) \propto \exp[-\lambda H(\boldsymbol{\theta})] \tag{1}$$

where the Lagrange multiplier λ depends on h and is unknown. We shall find meaningful interpretations for several values of λ, but we shall concentrate on the assumption that $\lambda = 1$. This we shall call the entropic prior

$$P_e(\boldsymbol{\theta}) \stackrel{\text{def}}{\propto} e^{-H(\boldsymbol{\theta})} \tag{2}$$

We assume ξ henceforth and drop it from notation. We note in passing that by making a similar assertion about the expected perplexity ($e^{H(\boldsymbol{\theta})}$) and assuming a measure on λ, it is also possible to integrate out the Lagrange multiplier and arrive directly at eqn. 2.

We call the reader's attention to two properties that derive from the definition of entropy: 1) $P_e(\cdot)$ is a bias for compact models having less ambiguity, more determinism, and therefore more structure. 2) $P_e(\cdot)$ is invariant to reparameterizations of the model, because the entropy is defined in terms of the model's joint and/or factored distributions.

We turn now from inferential principles to optimization problems. Models with hidden variables can be very difficult to fit and notoriously prone to over-fitting. Eqn. 2 is a remarkable form which we shall exploit to simultaneously estimate structure and parameters of complex probability models—a mixed combinatorial (structure) and continuous (parameter) optimization problem. We shall develop MAP estimators such that in expectation maximization (EM), the prior drives weakly supported parameters toward extinction, allowing them to be removed from the model without loss of posterior probability. The resulting models can be quite good; however, both hidden-variable and combinatorial optimization problems have notoriously rough energy surfaces and EM is merely a local hill-climber. Before presenting estimators in §5 we shall develop a technique that uses the entropic prior to improve the quality of local optima found by EM, and which finds a quasi-global optimum in the limiting case.

4 Prior balancing

We now introduce a generalization of the posterior, by rewriting Bayes' rule with a manipulation of the prior:

$$\tilde{P}(\boldsymbol{\theta}|\boldsymbol{X}, T, T_0) \stackrel{\text{def}}{\propto} P(\boldsymbol{X}|\boldsymbol{\theta})P(\boldsymbol{\theta})^{T_0 - T}\delta(T) \tag{3}$$

where T is normally positive and $\delta(T) \in (0, 1]$ is a driving term that monotonically increases as T approaches zero (e.g., $\delta(T) \propto e^{-T^2/2}$). Varying T balances the prior against the likelihood, which is useful in iterative parameter estimation because it allows $\boldsymbol{\theta}$ to get into the right neighborhood with respect to one constraint before attempting to satisfy the other. Of course, to obtain a proper probability we are obliged to make $Z = T_0 - T$ converge to a meaningful value (e.g., 1) which can be done by following the gradient $\Delta T \propto \frac{\partial}{\partial T} \log \tilde{P}(\boldsymbol{\theta}|\boldsymbol{X}, T, T_0)$ or by iteratively tracking the MAP estimate of $\hat{T} = \operatorname{argmax}_T \tilde{P}(\boldsymbol{\theta}|\boldsymbol{X}, T, T_0)$. Using the shorthand $H = -\log P(\boldsymbol{\theta})$, the gradient and MAP estimate for the Gaussian driving function are

$$\Delta T \propto H - T \ ;\ \hat{T} = H \tag{4}$$

T's decay rate can be controlled with a variance term on δ, or by choice of a different driving function. E.g., $\delta(T)$ may be one-sided (e.g., $\delta(T) = e^{T - e^T}$; $\Delta T \propto H + 1 - e^T$; $\hat{T} = \log(1 + H)$) or a barrier term (e.g., $\delta(T) = e^{-T^{3/2}}$; $\Delta T \propto H - \sqrt{T}$; $\hat{T} = H^2$). If T_0 is not known, the gap from equilibrium T to the desired Z can be bridged with a heuristic schedule.

We introduce this as an optimization technique; it is unclear that there is any useful interpretation as an inference

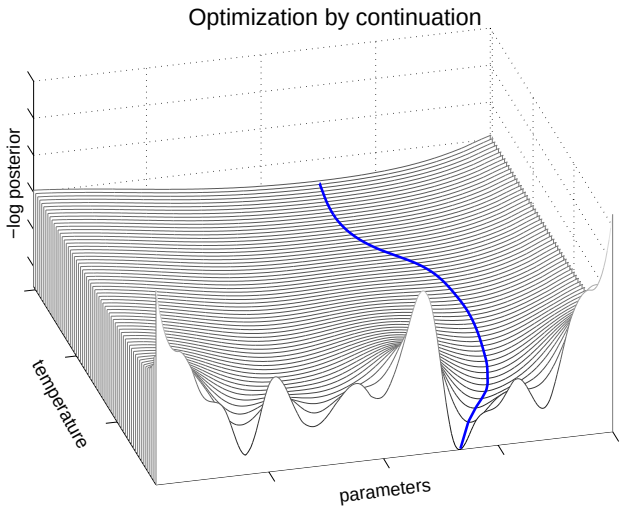
[1] These are the maximum entropy distributions given the assertion that a random variable can be characterized by its expectations

principle. However, prior balancing does take on a physical meaning with the entropic prior. Taking the negated logarithm of eqn. 3 with the entropic prior $P_e(\cdot)$, we obtain

$$\begin{aligned} &-\log \tilde{P}_e(\boldsymbol{\theta}|\boldsymbol{X},T,T_0) \\ &= \quad \tilde{F} \;=^{+}\; E-(T-T_0)H(\boldsymbol{\theta})-\log\delta(T) \\ &\geq^{+} \quad F \;=\; E-TH \end{aligned} \tag{5}$$

where $E=-\log P(\boldsymbol{X}|\boldsymbol{\theta})$ is the error or energy cost of a given parameterization, and T is understood as temperature. The notation $=^{+}$ indicates equality assuming additive constants (normalizing terms) that have been omitted. $\tilde{F}$ is an upper bound on the Helmholtz free energy equation of statistical physics, with equality at $\delta(T)=1,\ T_0=0$. Maximizing the modified posterior thus minimizes the free energy, which is analogous to finding an equilibrium configuration in a complex model whose different parts compete to explain the data. In models with factorizable likelihoods and priors, the prior on each independent parameter can be balanced separately, allowing different parts of the model to mature at different rates.

Four interesting cases immediately fall out of $\tilde{F}$: **1)** Iteratively re-estimating $\boldsymbol{\theta}$ while $T\to 0$ gives deterministic annealing (DA) [Rose et al., 1990; Miller et al., 1996; Hofmann et al., 1998], a pseudo-global optimizer for pock-marked energy surfaces. DA belongs to a family of continuation techniques that convolve (smooth) a energy surface to make it globally convex, then track the optimum as gradual deconvolution reintroduces surface texture[2]. The following illustration shows how tracking a local minimum leads to the global minimum of the function in the forefront:



During DA, entropy at high temperatures keeps the system from prematurely committing to nearby local optima and forces it to explore the energy surface's large-scale structure. In a hidden-variable model, this is equivalent to maximizing almost equally w.r.t. all possible hypotheses within the model (e.g., all possible paths through a hidden Markov model), then concentrating on the most promising hypotheses as the temperature declines. In §5 we'll introduce MAP estimators that fold DA into EM at no additional computational cost. Note that our modified posterior gives a useful amendment to DA—an automatic annealing schedule that tracks the quality (inversely, entropy) of the model via the gradients or MAP estimates w.r.t. T.

The remaining cases of interest arise out of different terminal temperatures: **2)** At $T-T_0=1$ we obtain a maximum-entropy solution. **3)** At $T-T_0=0$ we obtain the maximum-likelihood (ML) solution[3]. **4)** At $-Z=T-T_0=-1$ we obtain the maximum structure solution corresponding the the entropic prior in eqn. 2. Conveniently, we have derived a single MAP estimator for all four cases.

5 MAP estimators

We obtain MAP parameter estimators by solving for maxima of the log of the balanced posterior,

$$\hat{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta}} \left[\log P(\boldsymbol{X}|\boldsymbol{\theta}) - ZH(\boldsymbol{\theta})\right] \tag{6}$$

Although this can lead to systems of transcendental equations, we have obtained solutions for most simple distributions and, by subadditivity principles, for all models composed thereof.

For the multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{K}$ the entropy is $\frac{1}{2}\log((2\pi)^d e|\boldsymbol{K}|)$. The entropic prior is thus $P_e(\boldsymbol{\mu},\boldsymbol{K}) \propto |\boldsymbol{K}|^{-1/2}$, which is uniform in $\boldsymbol{\mu}$ and inversely proportional to the volume of $\boldsymbol{K}$. To find the entropic MAP estimator we set the gradient of the log-posterior to zero. Without loss of generality we assume a zero mean to simplify the derivation:

$$0 = \frac{d}{d\boldsymbol{K}} \log\left(\left[\prod_n^N \mathcal{N}(x_n;\boldsymbol{K})\right] P_e(\boldsymbol{K})^Z\right) \tag{7}$$

$$= \frac{d}{d\boldsymbol{K}}\left[\sum_n^N \left[-\tfrac{1}{2}\boldsymbol{x}_n^\top \boldsymbol{K}^{-1}\boldsymbol{x}_n - \tfrac{1}{2}\log(2\pi)^d|\boldsymbol{K}|\right]\right] - Z\frac{d}{d\boldsymbol{K}}H(\mathcal{N}(\boldsymbol{K})) \tag{8}$$

$$= \frac{d}{2d\boldsymbol{K}}\left[\sum_n^N \left[-\boldsymbol{x}_n^\top \boldsymbol{K}^{-1}\boldsymbol{x}_n - \log|\boldsymbol{K}|\right] - Z\log|\boldsymbol{K}| - C\right] \tag{9}$$

$$= \tfrac{1}{2}\sum_n^N 2\boldsymbol{K}^{-1}\boldsymbol{x}_n\boldsymbol{x}_n^\top\boldsymbol{K}^{-1} - \boldsymbol{K}^{-1}\boldsymbol{x}_n\boldsymbol{x}_n^\top\boldsymbol{K}^{-1} \circ \boldsymbol{I}$$

[2] A statistical analogue is robust M-estimation [Huber, 1981] with a shrinking scale parameter [Li, 1996].

[3] We thank the anonymous reviewers for pointing out work by [Ukeda and Nagano, 1995], who developed this special case directly via maximum entropy considerations, albeit without estimators or annealing schedule.

$$-\tfrac{1}{2}\sum_{n}^{N+Z} 2\boldsymbol{K}^{-1} - \boldsymbol{K}^{-1}\circ\boldsymbol{I} \tag{10}$$

$$= \sum_{n}^{N} \boldsymbol{K}^{-1}\boldsymbol{x}_n\boldsymbol{x}_n^\top\boldsymbol{K}^{-1} - \tfrac{N+Z}{N}\boldsymbol{K}^{-1}$$

$$-\tfrac{1}{2}\sum_{n}^{N} \left(\boldsymbol{K}^{-1}\boldsymbol{x}_n\boldsymbol{x}_n^\top\boldsymbol{K}^{-1} - \tfrac{N+Z}{N}\boldsymbol{K}^{-1}\right)\circ\boldsymbol{I} \tag{11}$$

Left- and right-multiplying by $\boldsymbol{K}$ reveals the entropic MAP estimator for covariances, which is essentially an $N+Z$ normalization of the scatter of N samples:

$$\hat{\boldsymbol{K}} = \frac{\sum_n^N \boldsymbol{x}_n\boldsymbol{x}_n^\top}{N+Z} \tag{12}$$

Note that the maximum structure (Z=1) MAP estimator is the best mean squared-error estimate, while the maximum-entropy (Z=-1) MAP estimator is the best unbiased estimate. Similarly normalized estimators give the scale parameters of related distributions, e.g., exponential, Laplace, and gamma.

A multinomial distribution has entropy $H(\boldsymbol{\theta}) = -\sum_i \theta_i \log\theta_i$, and entropic prior $P_e(\boldsymbol{\theta}) \propto \boldsymbol{\theta}^{\boldsymbol{\theta}} = \prod_i \theta_i^{\theta_i}$. For the MAP estimator given a vector $\boldsymbol{\omega}$ of evidence for each alternative, we set the derivative of the log-posterior to zero, using a Lagrange multiplier to ensure $\sum_i \theta_i = 1$,

$$0 = \frac{\partial}{\partial\theta_i}\left(\log\prod_i \theta_i^{\omega_i+Z\theta_i} + \lambda\sum_i\theta_i\right) \tag{13}$$

$$= \frac{\omega_i}{\theta_i} + Z\log\theta_i + Z + \lambda \tag{14}$$

We solve this a system of simultaneous transcendental equations for θ_i using the Lambert W function [Corless et al., 1996], an inverse mapping satisfying $W(y)e^{W(y)} = y$ and therefore $\log W(y) + W(y) = \log y$. Setting $y = e^x$ and working backwards towards eqn. 14,

$$0 = -W(e^x) - \log W(e^x) + x \tag{15}$$

$$= \frac{-1}{1/W(e^x)} - \log W(e^x) + x + \log q - \log q \tag{16}$$

$$= \frac{-q}{q/W(e^x)} + \log q/W(e^x) + x - \log q \tag{17}$$

Setting $x = 1+\lambda/Z+\log q$ and $q = -\omega_i/Z$, eqn. 17 simplifies to $Z\times$eqn. 14:

$$\begin{array}{lll} 0= \frac{\omega_i/Z}{-(\omega_i/Z)/W(-\omega_i e^{1+\lambda/Z}/Z)} & +\log\frac{-\omega_i/Z}{W(-\omega_i e^{1+\lambda/Z}/Z)} & +1+\lambda/Z \\ =\omega_i/(Z\theta_i) & +\log\theta_i & +1+\lambda/Z \end{array} \tag{18}$$

which implies that

$$\hat{\theta}_i = \frac{-\omega_i/Z}{W(-\omega_i e^{1+\lambda/Z}/Z)} \tag{19}$$

where eqn. 14 and eqn. 19 form a set of fix-point equations for λ that typically converge in 2-5 iterations. This derivation generalizes that given in [Brand, 1997a]. Details on computing W are given in [Brand, 1997b].

Differentiating eqn. 14 again we find that the posterior is concave in any parameter provided $\omega_i > Z\theta_i$, which is always true except for near-degenerate cases when $Z > \max_i \omega_i$ (e.g., there is almost no evidence[4]).

Note that we have assumed that the likelihood and the prior are factorizable, which is fortunately the case for many popular hidden-variable models.

6 Minimization of entropies

For flat exponential distributions, the principle of maximum likelihood is equivalent to minimizing cross-entropy (directed Kullback-Leibler distance) between the data's sufficient statistics and the distribution's estimated parameters. Our framework identifies three entropies associated with modeling, and the MAP estimator minimizes their sum. This is most clearly seen by considering the negated log-posterior of a multinomial parameter in the maximum structure ($Z=1$) case:

$$-\log P(\boldsymbol{\omega}|\boldsymbol{\theta})e^{-H(\boldsymbol{\theta})} = -\log\prod_i \theta_i^{\omega_i+\theta_i} \tag{20}$$

$$= -\sum_i(\omega_i+\theta_i)\log\theta_i \tag{21}$$

$$= -\sum_i(\omega_i\log\theta_i + \theta_i\log\theta_i - \omega_i\log\omega_i + \omega_i\log\omega_i) \tag{22}$$

$$= -\sum_i \omega_i\log\omega_i + \sum_i\omega_i\log\frac{\omega_i}{\theta_i} - \sum_i\theta_i\log\theta_i \tag{23}$$

$$= H(\boldsymbol{\omega}) + D(\boldsymbol{\omega}||\boldsymbol{\theta}) + H(\boldsymbol{\theta}) \tag{24}$$

Each of these entropies has a useful interpretation: $H(\boldsymbol{\omega})$ measures uncertainty in the expected sufficient statistics, and is linearly related to their coding length. The cross-entropy $D(\boldsymbol{\omega}||\boldsymbol{\theta})$ measures error in the model's fit to the data, and is linearly related to the coding costs of aspects of the data not captured by the model. (Ideally, this will dwindle to purely random noise.) Together, these terms give the expected coding length of the data relative to the model. $H(\boldsymbol{\theta})$ measures uncertainty in the model's component distributions, as well as its coding length. In short, MAP estimation reduces all entropies and relative entropies between the model and the data's sufficient statistics. In practice, it does so by extinguishing parts of the model that are ill-matched to the structure of the data, and therefore weakly supported. More evidence concentrates on the surviving parameters (equivalently

[4] In these cases we may either use slightly more elaborate methods to find the global optimum or simply accept a local optimum and ignore its typically negligible effect on the joint distribution.

$H(\boldsymbol{\omega})$ is minimized), and their information content is thereby maximized.

6.1 Variations

One may introduce other entropies into eqn. 24, perhaps to symmetrize the cross entropy (equiv. KL divergence)

$$H(\boldsymbol{\omega}) + D(\boldsymbol{\omega}||\boldsymbol{\theta}) + \underline{D(\boldsymbol{\theta}||\boldsymbol{\omega})} + H(\boldsymbol{\theta}). \quad (25)$$

One might want to conserve information contained in previous estimates

$$H(\boldsymbol{\omega}) + D(\boldsymbol{\omega}||\boldsymbol{\theta}) + \underline{D(\boldsymbol{\theta}||\boldsymbol{\theta}_{\text{old}})} + H(\boldsymbol{\theta}), \quad (26)$$

or stay close to (or far from) a reference model $\boldsymbol{\theta}^*$

$$H(\boldsymbol{\omega}) + D(\boldsymbol{\omega}||\boldsymbol{\theta}) \pm \underline{D(\boldsymbol{\theta}||\boldsymbol{\theta}^*)} + H(\boldsymbol{\theta}). \quad (27)$$

Happily, small variations on the MAP formulæ given above can accommodate any combination of these constraints.

For simplicity of exposition we have identified $H(\boldsymbol{\theta})$ with the entropy of the model's component distributions, leading to an extremely efficient form of description length minimization. One may also choose other entropy measures for $H(\boldsymbol{\theta})$. For example, one may prefer to minimize the joint entropy

$$\begin{aligned} H(\boldsymbol{X},\boldsymbol{\theta}) &= H(\boldsymbol{X}) + H(\boldsymbol{\theta}|\boldsymbol{X}) \\ &= c + \textstyle\sum_j p_j \sum_i \theta_{i|j} \log \theta_{i|j} \end{aligned} \quad (28)$$

where $H(\boldsymbol{X}) = c$ is fixed and p_j are the probabilities of hidden variable values given the data. These are estimated from the data during the E-step of EM, and in some cases can be computed from the parameters (e.g., in Markov models, p_j = the stationary probability of state j). This case leads to another small variation on the MAP estimator, etc., but the result is only approximate, because the entropy is minimized w.r.t. p_j calculated from the previous rather than the current parameter estimates. Nonetheless, we find in practice that this case is very well behaved, especially in deterministic annealing, where p_j changes quite gradually.

7 Trimming

One problem with entropy minimization on digital computers is that numerical error often intrudes before parameters are fully extinguished. Eqn. 14 obliges us to add numbers to their logarithms, which becomes troublesome when parameter values are not near 1. Fortunately, the prior allows us to identify excess parameters that can be trimmed from a model without loss of posterior probability, such that

$$\tilde{P}_e(\boldsymbol{\theta}\backslash\theta_i|\boldsymbol{X},Z) \geq \tilde{P}_e(\boldsymbol{\theta}|\boldsymbol{X},Z) \quad (29)$$

Expanding via Bayes' rule, taking logarithms, and rearranging, we obtain

$$Z[H(\boldsymbol{\theta})\text{–}H(\boldsymbol{\theta}\backslash\theta_i)] \geq \log P(\boldsymbol{X}|\boldsymbol{\theta})\text{–}\log P(\boldsymbol{X}|\boldsymbol{\theta}\backslash\theta_i) \quad (30)$$

Operationally, a parameter is trimmed by setting it to a target value which makes the model simpler or easier to interpret. Typically the target will be the minimal entropy extinction value, e.g., setting a multinomial parameter $\theta_i \leftarrow 0$, but it may simply be a usefully interpreted value, e.g., setting a variance parameter $k_{ii} \leftarrow 1$ (equivalently $\log k_{ii} \leftarrow 0$). When zeroing, eqn. 30 can be approximated via differentials, yielding

$$Z\frac{\partial H(\boldsymbol{\theta})}{\partial \theta_i}\theta_i \;>\; \theta_i \frac{\partial \log P(\boldsymbol{X}|\boldsymbol{\theta})}{\partial \theta_i} \quad (31)$$

One may observe from this that a parameter can be trimmed when varying it increases the entropy faster than the log-likelihood. The sides of eqns. 30 and 31 can be mixed and matched to obtain mathematically convenient forms. For the multinomial trim test, we set l.h.s. eqn. 30 ($=-Z\theta_i \log \theta_i$) against r.h.s. eqn. 31 to obtain:

$$\theta_i < \exp\left[-\frac{1}{Z}\frac{\partial \log P(\boldsymbol{X}|\boldsymbol{\theta})}{\partial \theta_i}\right] \quad (32)$$

The gradient of the log-likelihood is normally computed for re-estimation, so these trimming tests are very cheap.

Solving eqn. 30 via W gives the variance trimming criteria:

$$k_{ii} \;\geq\; -\hat{k}_{ii}/W_0(-\hat{k}_{ii}e^{-\hat{k}_{ii}}) \text{ iff } \hat{\mathrm{k}}_{\mathrm{ii}} > 1 \quad (33)$$

$$k_{ii} \;\leq\; -\hat{k}_{ii}/W_{-1}(-\hat{k}_{ii}e^{-\hat{k}_{ii}}) \text{ iff } \hat{\mathrm{k}}_{\mathrm{ii}} < 1 \quad (34)$$

There is also a trimming test for covariances ($k_{ij} \leftarrow 0$), but it is beyond the scope of this paper.

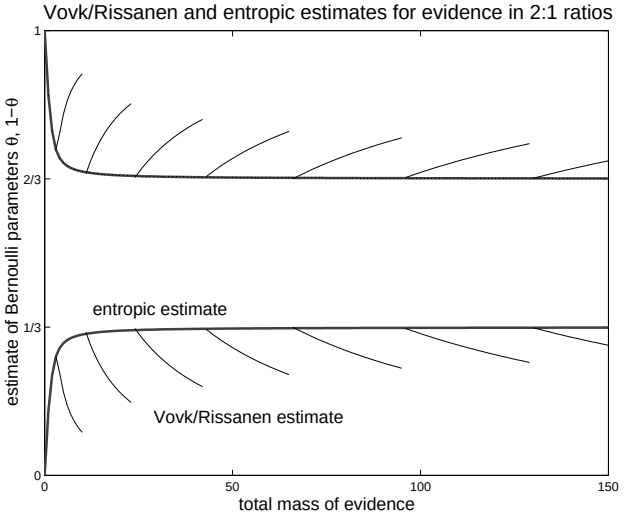
Trimming is a local heuristic that increases but does not maximize the posterior. It makes sense in iterative estimation of models containing hidden variables because it can radically sparsify the model, which in turn reduces ambiguity in the data's expected sufficient statistics. In short, it accelerates learning. In our experiments we have used trimming mainly to speed parameter extinction that would otherwise happen gradually via MAP estimation. As noted above, this becomes particularly desirable when parameters are near extinction and floating-point calculations introduce numerical round-off and underflow errors. In addition, trimming near convergence can "bump" the model out of the local local probability maximum and into a parameter subspace of simpler geometry, thus enabling further training.

Not only do parameter extinction and trimming protect against over-fitting, but by sculpting the model to fit the data, they often reveal a simple machine that explains the data (e.g., a finite-state machine from an HMM; a circuit from a NN). We shall give many examples in §9.

8 Related ideas

Model selection criteria such as MML [Wallace and Boulton, 1968; Wallace and Freeman, 1987], AIC

[Aikake, 1973], MDL [Rissanen, 1978], BIC [Schwartz, 1978], stochastic complexity [Rissanen, 1989] have a rich literature in model selection, but usually cannot be leveraged into priors, let alone estimators, because they force separate treatments of model structure and parameter values. For example, BIC and MDL are asymptotic approximations of the marginal likelihood $p_m(\boldsymbol{X}|S) = \int p(\boldsymbol{X}|\boldsymbol{\theta}, S)d\boldsymbol{\theta}$, assuming structure S and typically ignoring any prior $p(\boldsymbol{\theta}|S)$. Selection criteria are typically formulated at a level of granularity (e.g., number of parameters) that prevents their use in continuous sample spaces. The entropic prior bears some resemblance to all of these selection criteria, but provides a unified treatment of structure and parameter values. Stochastic complexity occupies an interesting middle ground because it has natural interpretations at a bit-level of granularity, leading to approximate estimators and occasionally yielding a point estimator for discrete sample spaces. E.g., [Vovk, 1995] derived a binomial parameter estimator and showed that it minimizes stochastic complexity and, remarkably, that it is MDL in the strong sense of [Vitányi and Li, 1997], with probability close to 1. Unfortunately, MDL based on discretization of continuous values has some severe consequences; the figure below contrasts the discontinuous Vovk estimator with equivalent entropic MAP estimator.



Computer scientists have tried to get more leverage out of model selection criteria by proposing search heuristics that reduce the squandered computation that dominates generate-and-test "structure-learning" algorithms. [Ikeda, 1993] and [Stolcke and Omohundro, 1994] presented search algorithms for HMM structure that generate rival models by splitting or merging states that are likely to reduce description length. Whereas run-times were reported in days, entropic estimation will induce HMM topologies from similar datasets in a matter of minutes. [Friedman, 1998] advertised a "Structural EM" algorithm for Bayesian networks, but the M-step is approximate and there is a penalized generate-and-test search inside the EM loop. We would propose entropy minimization algorithm for Bayesian networks in which parameter extinction makes subgraphs deterministic and thus trimmable. Unfortunately the E-step for dense graphs can be computationally prohibitive, which makes starting with over-complete structure problematic.

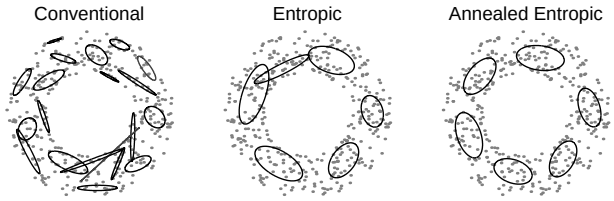
Most exponential density forms have conjugate priors that can be used to obtain estimators with some description-length *reducing* properties. For example, for multinomial parameters one might use Dirichlet priors with exponents $\alpha_i < 1$. Of course, one must choose values for these parameters *a priori*; this is equivalent to inventing extra observations before any actual samples have been taken, a potentially dangerous business.

9 Examples

We have used entropic estimation to obtain low entropy models in a variety of model classes. These have been tested on benchmark datasets as well as real-time data obtained directly from computer vision and speech analysis systems. The resulting models have consistently been smaller, more discriminative, better generalizing, and more predictive than conventionally (e.g., ML) estimated models, except when both are under-fit. Typically the resulting model is interpretable, often providing insight into the causal structure of the process that generated the data. Here we give a variety of examples

9.1 Mixture models

A mixture model was fitted to an ring of Cartesian samples taken from a uniform distribution over the polar coordinates $r \in [1, 2]$; $\theta \in [0, 2\pi]$. The figures show the model estimated conventionally, entropically with trimming, and entropically with trimming and deterministic annealing. Initial conditions were identical for all three cases. Ellipses indicate iso-density contours for the Gaussian components.



In the first case, over-fitting has resulted in a model of the accidental properties of the data (e.g., the clumpiness of the sample). In the second case, trimming has removed excess components and the resulting model looks much more like the essential structure of the data, but a large under-sampled region near the top still affects the model. In the third case, DA circumvents this local optimum and finds a model which generalizes even better.

9.2 Radial basis function networks (RBFNs)

Vowel classification: We obtained the British English vowel recognition dataset from the CMU Neural-Bench Archive. Each example of a vowel is characterized by LPC coefficients from two time-frames. This dataset has been treated several times using perceptrons, neural networks, Kanerva models, radial basis function networks, and non-parametric methods (nearest neighbor). The last yielded the best classification of the test set (56%), while RBFNs peaked at 53% with 528 Gaussian basis functions [Robinson, 1989]. We combined entropically estimated mixture models of each vowel's training data into an RBFN and obtained 58.7% correct classification on the test set. Entropic estimation found mixtures of 1-3 components per vowel, resulting in an RBFN of only 22 basis functions.

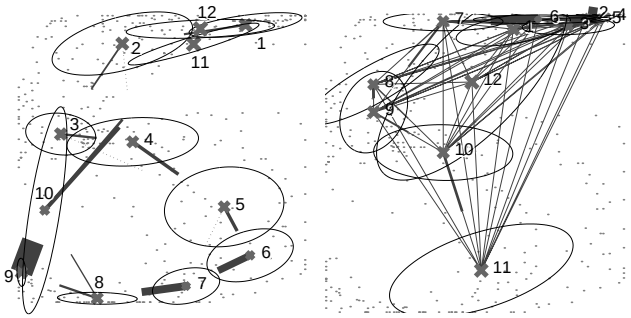
Home value prediction: We obtained the publicly available Boston home value database from StatLib and set out to predict the median home value from 13 other potentially relevant attributes. The features are numeric- or boolean-valued and have different (arbitrary) dynamic ranges; we normalized each dimension to unit variance before mixture modeling. In each trial, all models were identically initialized and trained on half of the examples, randomly selected, then used to predict home values for the remaining test examples. We use the performance of the ML model as a baseline in each trial; our measure of performance is how much an entropically estimated model reduced the mean squared-error of predictions on the test set. The following table shows mean improvement in RBFN prediction accuracy over maximum-likelihood models, as a function of # of RBFs at initialization.

#RBFs	entropic MAP	...+trimming	...+annealing
5	† 0.0047	† 0.0040	* 0.0100
10	† -0.0066	† 0.0014	* 0.0032
20	† -0.0781	0.0207	0.0581
40	0.0078	0.3986	0.4065
80	* 0.0959	2.7565	3.5145

All values are at a $p < 10^{-5}$ level of statistical significance except for those marked with a '†' ($p > 0.1$) or a '*' ($p < 0.01$). The improvement is quite significant given a large initialization (recall that these scores are in terms of unit-variance prices). This can be attributed to entropic estimation's resistance to over-fitting. The advantage disappears or becomes statistically insignificant when the initializations are too small, but notice that even in these under-fitted models, annealing is finding superior local optima.

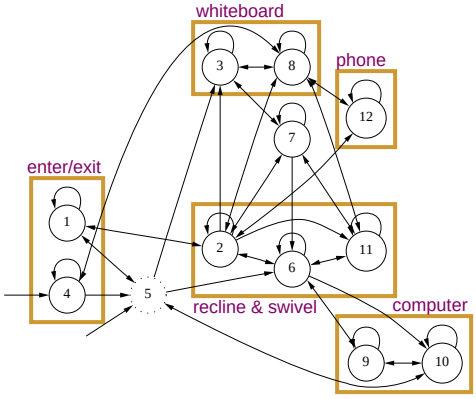
9.3 Hidden Markov models (HMMs)

Handwriting analysis and classification: We obtained handwriting samples by 10 writers from the UNIPEN archive. The diagrams below show entropic and maximum-likelihood models of the pen-strokes for the digit "5," estimated from pen-position data taken at 5msec intervals from 10 different individuals via an electro-magnetic resonance sensing tablet. Initial conditions were identical.



Ellipses indicate show iso-density contours for each state; $\times$s and arcs indicate state dwell and transition probabilities, respectively, by their thicknesses. Entropic estimation induces an interpretable automaton that captures essential structure and timing of the pen-strokes, as well as variations in their ordering between writers. 50 of the original 80 dynamical parameters were trimmed. Estimation without the entropic prior results in a wholly opaque model, in which none of the original dynamical parameters were trimmed. Over 10 trials with different parts of the database held out, entropic models yielded 96% correct classification; ML models yielded 93%.

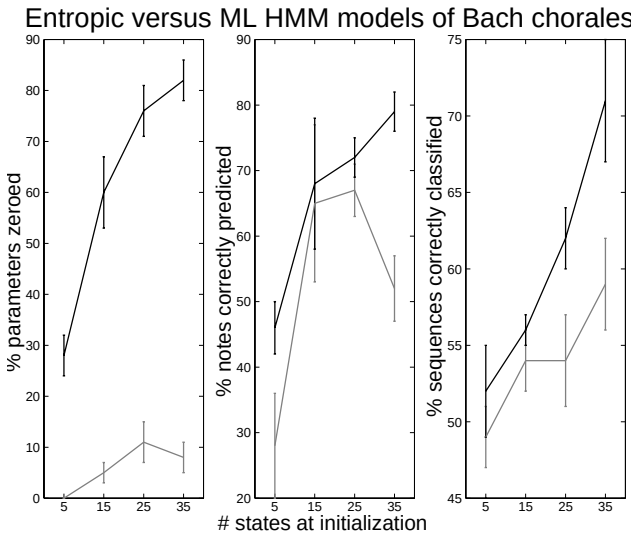
Video analysis: In this example a gigabyte of video randomly taken from an office setting was filtered to extract motion vectors and then modeled entropically with a hidden Markov model [Brand, 1997a]. The resulting HMM state machine is compact enough to read:



Roughly 2/3 of the transitions were trimmed. We have taken the liberty of labeling states by using forward-backward analysis to find their support in new video. The states mapped nicely to typical work activities of the office occupant. An HMM conventionally estimated from the same initial conditions is fully connected and thus too bushy to profitably illustrate or interpret.

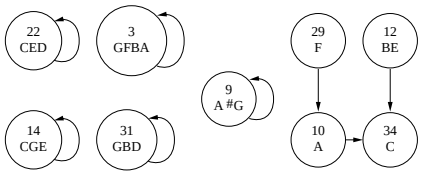
Prediction in Bach chorales: We obtained a dataset of melodic lines from 100 of J.S. Bach's 371 surviving

chorales from the UCI repository [Merz and Murphy, 1998], and transposed all into the key of C. We compared entropically and conventionally estimated HMMs in prediction and classification tasks, training from identical random initial conditions and trying a variety of different initial state-counts. We trained with 90 chorales and testing with the remaining 10. In ten trials, all chorales were rotated into the test set. The results neatly chart the sparsification, classification, and prediction superiority of entropically estimated HMMs.



Lines indicate mean performance over 10 trials; error bars are 2 standard deviations. It is clear that despite substantial loss of parameters to sparsification, the entropically estimated HMMs were, on average, better predictors of notes. (Each test sequence was truncated to a random length and the HMMs were used to predict the first missing note.) They also were better at discriminating between test chorales and temporally reversed test chorales—challenging because Bach famously employed melodic reversal as a compositional device. With larger models, parameter-trimming became state-trimming: An average of 1.6 states were "pinched off" the 35-state models when all incoming transitions were deleted.

While the conventionally estimated HMMs were wholly uninterpretable, in the entropically estimated HMMs one can discern several basic musical structures. Here is a sampling of high-probability states and subgraphs of interest from an entropically estimated 35-state HMM. (Tones output by each state are listed in order of probability. and extraneous arcs have been removed for clarity.)

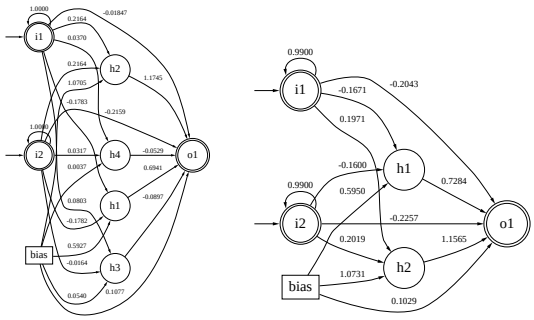


These include self-transitioning states that output only tonic (C-E-G) or dominant (G-B-D) triads, lower- or upper-register diatonic tones (C-D-E or F-G-A-B), and mordents (A-♯G-A). We also found chordal state sequences (F-A-C) and states that lead to the tonic (C) via the mediant (E) or the leading tone (B). Forward-backward analysis of the chorales confirms that these subgraphs are indeed "explaining" these musical structures.

Facial animation: In this case we learned a dynamical model of two covarying signals—acoustic features in the voice and visual features on the face—from video. Inspection revealed that the model organized facial configuration space into prototypes that strongly resemble visemes. Although technically a recognition model, it was so sparse and near-deterministic that it can be used to generate realistic facial motions to accompany a new voice track, e.g., near-photorealistic face syncing. Moreover, the empirical the entropy rate of the model indicates that it is propagating context effects an average of 135msec forward and backward in time—consistent with vocal co-articulation effects—but some context is propagating over 300msec—probably facial co-articulation effects, which typically occur at longer time-scales due to less agile facial tissue. Surprisingly, the model was able to predict upper facial motion even more accurately than motion around the mouth, perhaps by exploiting prosodic information in the acoustic signal.

9.4 Recurrent neural networks (RNNs)

In this rather speculative example, an entropic prior was designed for the weights of a recurrent neural network by assuming that Θ generates neural networks with weights from a Gaussian distribution with functionally related mean and variance. This yields a modified back-propagation rule and a trim test. The figures show a semi-recurrent neural network (a DAG with self-connections for memory; activation propagates one link per cycle) trained to compute XOR using conventional back-prop with weight decay (left) and entropic back-prop with trimming (right). Initial conditions were identical.



Note that the entropically estimated topology is an amalgam of the two minimal feed-forward XOR circuits.

10 Open questions

Our framework is agnostic with regard to two pressing questions: Firstly, what criteria should motivate our choice of an entropy measure $H(\boldsymbol{\theta})$? As discussed in §6, different choices lead to MDL and entropy minimax. In many cases they are identical. Both are consistent with our results, and one can often be used as upper-bound for the other when analytic forms are unavailable. Secondly, when should we remove trimmable parameters? Always? At or near convergence? During DA? Various combinations of entropic training, trimming, and annealing have intriguing physical analogues, including the metallurgical processes of casting, tempering, and bluing.

The entropic prior can pose severe challenges to integration-based Bayesian methods such as model averaging. Because integrating over all hidden-variable models in almost always intractable, Bayesians typically fall back to the marginal likelihoods of a single model structure as an approximation. For some classes of models, marginal likelihoods are more predictive than posterior densities, but in general there is no reason to believe that they are always a good approximation (e.g., HMMs) or even that the model structure having the greatest marginal likelihood shares much probability mass with the posterior mode. Where integration is desirable but infeasible, we propose falling back to the weighted responses of a population of entropically estimated models. Of course, depending on one's choice of $H(\boldsymbol{\theta})$, this might require solving the integral which normalizes the prior—a problem that remains outstanding.

It is worth inquiring whether this framework can be extended to purely discrete optimization problems. Our preliminary results are encouraging: It is possible to generalize the multinomial MAP estimator to compute doubly stochastic matrices. This allows us to entertain graph-theoretic problems whose solutions can be formulated as permutation matrices (zeros but for a single 1 in each row and column). These problems can be "softened" to a doubly stochastic matrix (a weighted super-imposition of all possible solutions) at some high temperature, then "hardened" to a single quasi-optimal solution as entropy is algorithmically minimized. Initial experiments with traveling salesman problems have been quite promising.

11 Summary

We have develops an efficient method for finding compact hidden-variable probability models via entropy minimization. These models are highly predictive and often *interpretable* as theories that identify relationships between hidden causes and observed effects. The main results are: 1) An entropic prior that favors small, unambiguous, and maximally structured models. 2) A prior-balancing manipulation of Bayes' rule that allows one to gradually introduce or remove constraints in the course of iterative re-estimation. When combined with #1 this yields a posterior whose negative logarithm equals or upper-bounds the Helmholtz free energy of the model. This posterior contains as special cases the methods of maximum entropy, maximum likelihood, and our new method, maximum structure. 3) MAP estimators such that entropy optimization and deterministic annealing can be performed wholly within EM. In the maximum structure case, the MAP estimator smoothly extinguishes excess parameters, thereby simplifying the model and preventing over-fitting. 4) Trimming tests that identify excess parameters whose removal will *increase* the posterior. These accelerate learning. The combined result is a class of fast and exact hill-climbing algorithms that mix continuous and combinatoric optimization and evade sub-optimal equilibria.

Acknowledgments

We thank the reviewers for clarifying questions and for making interesting connections to the statistics literature.

References

Aikake, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. and Csàki, F., editors, *Proc, 2nd International Symp. on Inference Theory*, pages 267–281. Akadémiai Kiadó.

Brand, M. (1997a). Learning concise models of human activity from ambient video. Technical Report 97-25, Mitsubishi Electric Research Labs.

Brand, M. (1997b). Structure discovery in conditional probability models via an entropic prior and parameter extinction. *Neural Computation (accepted 8/98)*.

Corless, R. M., Gonnet, G. H., Hare, D. E. G., Jeffrey, D. J., and Knuth, D. E. (1996). On the Lambert W function. *Advances in Computational Mathematics*, 5:329–359.

Friedman, N. (1998). The bayesian structural EM algorithm. In *Proc. 14th Conf. on Uncertainty in Artificial Intelligence*.

Hofmann, T., Puzicha, J., and Buhmann, J. M. (1998). Unsupervised texture segmentation in a deterministic annealing framework. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*.

Huber, P. (1981). *Robust Statistics*. Wiley and Sons.

Ikeda, S. (1993). Construction of phoneme models — Model search of hidden Markov models. In *International Workshop on Intelligent Signal Processing and Communication Systems*, Sendai.

Jaynes, E. T. (1982). *Papers on probability, statistics, and statistical mechanics*, chapter on Brandeis Lectures (1963), pages 39–76. Kluwer Academic.

Li, S. Z. (1996). Robustizing robust M estimation using deterministic annealing. *Pattern Recognition*, 29(1):159–166.

Merz, C. and Murphy, P. (1998). UCI repository of machine learning databases.

Miller, D., Rao, A., Rose, K., and Gersho, A. (1996). A global optimization technique for statistical classifier design. *IEEE Transactions on Signal Processing*, 4:3108–3121.

Rissanen, J. (1978). Modeling by the shortest data description. *Automatica*, 14:465–471.

Rissanen, J. (1989). *Stochastic Complexity and Statistical Inquiry*. World Scientific.

Robinson, A. J. (1989). *Dynamic Error Propagation Networks*. PhD thesis, Cambridge University Engineering Department. See www.boltz.cs.cmu.edu/benchmarks/vowel.html, updated 1997.

Rose, K., Gurewitz, E., and Fox, G. (1990). A deterministic annealing approach to clustering. *Pattern Recognition Letters*, 11(9):589–594.

Schwartz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6:461–464.

Stolcke, A. and Omohundro, S. (1994). Best-first model merging for hidden Markov model induction. Technical Report TR-94-003, International Computer Science Institute, 1947 Center St., Berkeley, CA, 94704, USA.

Ukeda and Nagano (1995). Deterministic annealing in EM. In Tesauro, G., Touretzky, D. S., and Leen, T., editors, *Advances in Neural Information Processing Systems*, volume 7. MIT Press.

Vitanyi, P. and Li, M. (1996). Ideal MDL and its relation to Bayesianism. In *ISIS: Information, Statistics and Induction in Science*, pages 282–291. World Scientific, Singapore.

Vitányi, P. and Li, M. (1997). Minimum description length induction, Bayesianism, and Kolmogorov complexity. *Submitted to IEEE Transations on Information Theory*. Revised version 17dec97 available at http://www.cwi.nl/~paulv/selection.html.

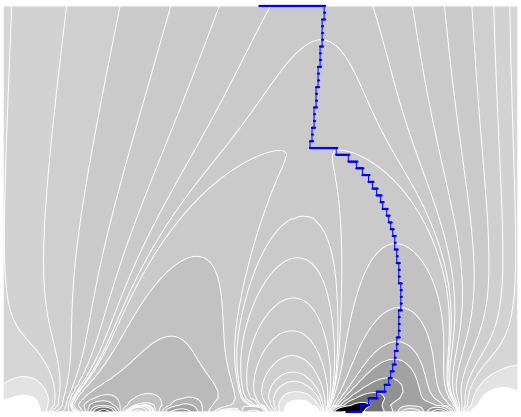
Vovk, V. G. (1995). Minimum description length estimators under the optimal coding scheme. In Vitányi, P., editor, *Proceedings, Computational Learning Theory / Europe*, pages 237–251. Springer-Verlag.

Wallace, C. and Boulton, D. (1968). An information measure for classification. *Computing Journal*, 11:185–195.

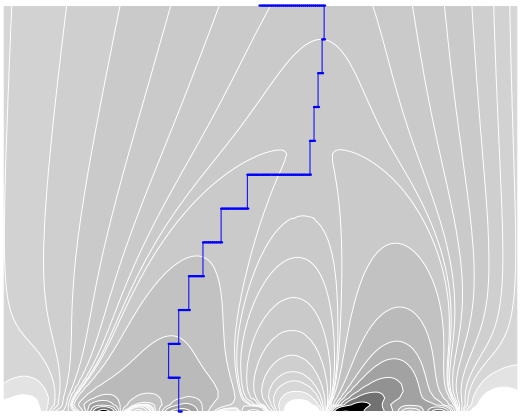
Wallace, C. and Freeman, P. (1987). Estimation and inference by compact coding. *Journal of the Royal Statistical Society, Series B*, 49:240–251.

A Deterministic annealing paths

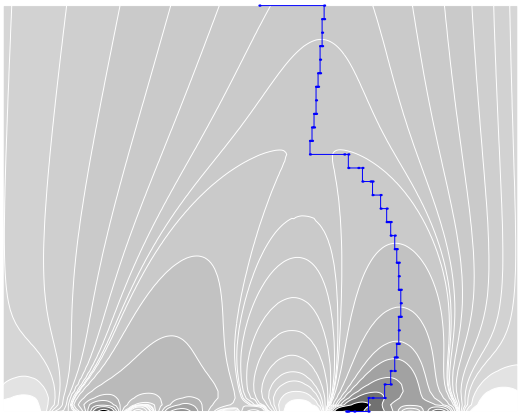
Here we illustrate different annealing methods associated with the continuation shown in §4. The figures below show the iso-contours of that surface, and possible paths to the minimum. The goal is to travel the correct annealing path while computing as few points along it as possible. In this regard gradient descent is profligate:



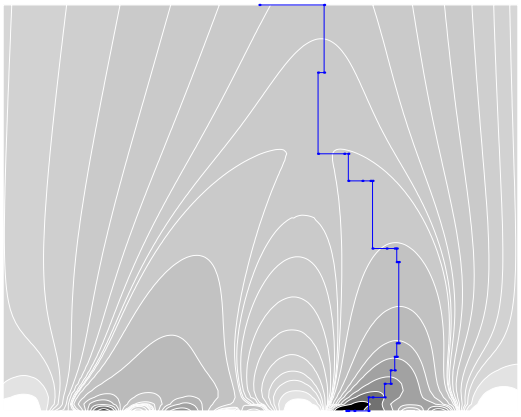
because it finds the local minimum at each temperature by moving in fixed increments, then lowers the temperature according to a fixed schedule. This can lead to disaster if the temperature declines too fast:



Expectation maximization is usually a much more efficient way to get into the neighborhood of the local optimum:



and we exploit the same machinery to calculate safe jumps in temperature:



In practice we find it sufficient just to move *towards* the local optimum at each temperature, interleaving parameter and temperature re-estimation to obtain very fast trajectories to the optimum.