Mapping the Energy Landscape of Non-Convex Optimization Problems

Maira Pavlovskaia[1], Kewei Tu[2], and Song-Chun Zhu[1]

[1] Department of Statistics, University of California, Los Angeles, 8125 Math Science Bldg, Los Angeles, CA 90095, USA,
{mariapavl,sczhu}@ucla.edu
[2] School of Information Science and Technology, ShanghaiTech University, No. 8 Building, 319 Yueyang Road, Shanghai 200031, China
tukw@shanghaitech.edu.cn

Abstract. An *energy landscape map* (ELM) characterizes and visualizes an energy function with a tree structure, in which each leaf node represents a local minimum and each non-leaf node represents the barrier between adjacent energy basins. We demonstrate the utility of ELMs in analyzing non-convex energy minimization problems with two case studies: clustering with Gaussian mixture models and learning mixtures of Bernoulli templates from images. By plotting the ELMs, we are able to visualize the impact of different problem settings on the energy landscape as well as to examine and compare the behaviors of different learning algorithms on the ELMs.

1 Introduction

In many computer vision, pattern recognition and learning problems, the energy function to be optimized is highly non-convex. A large body of work has been devoted to designing algorithms that are capable of efficiently finding a good local optimum in the non-convex energy landscape. On the other hand, much less work has been done in analyzing the properties of such non-convex energy landscapes.

In this paper, inspired by the success of visualizing the landscapes of Ising and Spin-glass models by [2] and [14], we compute *Energy Landscape Maps* (ELMs) in the high-dimensional hypothesis spaces for a few model learning problems in computer vision and pattern recognition — learning mixtures of Gaussian and learning mixtures of Bernoulli templates. An ELM is a tree structure in which each leaf node represents a local minimum whose energy determines the y-axis position of the leaf node; each non-leaf node represents the energy barrier between local minima. Figure 1 shows an example energy function and the corresponding ELM. The ELM of an energy landscape reveals important characteristics of the landscape, including

- the number of local minima and their energy levels;
- the energy barriers between adjacent local minima; and

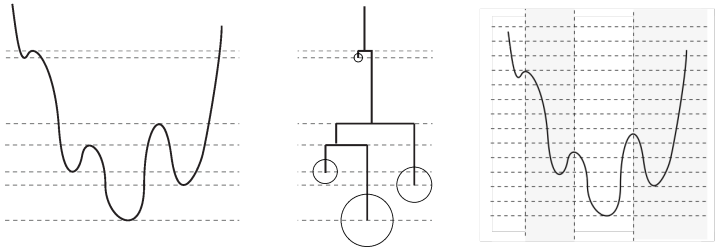


Fig. 1. (Left) An energy function. (Middle) Its corresponding ELM. The y-axis of the ELM is the energy level. Each leaf node is a local minimum and the leaf nodes are connected at the energy barrier between their energy basins. The probability mass or volume of an energy basin is indicated by the size of the circle around the leaf node. (Right) Partition of the spaces into bins according to basins and energy levels.

– the probability mass and volume of each local minimum.

Such information can be very useful in analyzing the intrinsic complexity of the optimization problems (for either inference or learning tasks), analyzing the effects of various conditions on the complexity, and visualizing the behavior of different optimization algorithms (i.e. how they move in the landscape).

ELMs can be efficiently constructed by running a MCMC algorithm that features a dynamic reweighting scheme allowing the sampler to cross energy barriers and efficiently traverse the entire space. In the literature, Becker and Karplus [2] presents the first work for visualizing multidimensional energy landscapes for the spin-glass model. Liang [6, 7] generalizes the Wang-Landau algorithm [13] for random walks in the state space. Zhou [14] uses the generalized Wang-Landau algorithm to plot the landscape for Ising model with hundreds of local minima and proposes an effective way for estimating the energy barriers. In contrast to the above work that compute the landscapes in "state" spaces for inference problems, our work is focused on the landscapes in "hypothesis" spaces (the sets of all models) for statistical learning problems. We modify the previous MCMC algorithm to handle several new issues that arise in plotting ELMs of continuous hypothesis spaces.

2 ELM construction in hypothesis spaces

Let $\mathcal{H}$ be a hypothesis space for a learning problem and let $E(x)$ be the energy of a hypothesis $x \in \mathcal{H}$. For example, in a n-component mixture of Gaussian clustering problem, given a training dataset, a posterior probability $\pi(x)$ is defined and x includes the model parameters such as the means and variances of the n unknown Gaussians; the landscape is defined by energy function $E(x) = -\log \pi(x)$. For simplicity, we bound $\mathcal{H}$ by limiting x to a finite range calculated from the input data points.

As Figure 1 (right) shows, the finite hypothesis space is partitioned into energy basins D_i and each basin is further partitioned into energy intervals $[u_{j+1}, u_j)$. Thus the space $\mathcal{H}$ is divided into bins $D_{i,j}$

$$D_{i,j} = \{x : x \in D_i,\ E(x) \in [u_{j+1}, u_j)\}. \tag{1}$$

Let $\phi(x)$ be the index mapping x to the bin index (i, j), and $\beta_{ij} = \pi(D_{ij})$ the probability mass of bin $D_{i,j}$. Our goal is to design an MCMC algorithm with equal probability visiting all bins, i.e. its state at time t follows a new equalized probability,

$$x_t \sim \pi^+(x) \propto \frac{\pi(x)}{\beta_{\phi(x)}}. \tag{2}$$

The generalized Wang-Landau algorithm estimates β_{ij} by γ_{ij} using stochastic gradient. The algorithm goes as follows:

1. Initialize a sample $x_0 \in \mathcal{H}$ and the bin weights γ^0_{ij} for the bins $D_{i,j}$. Repeat step 2-6:
2. At step t, sample $y \sim Q(x_t, y)$ from some proposal distribution Q.
3. Perform steepest descent initialized with y to find the energy basin that y belongs to. Let $\phi(y)$ be the index of the bin containing y.
4. Accept proposal y with probability $\alpha(x_t, y)$:

$$\alpha(x_t, y) = \min\left(1, \frac{Q(y, x_t)}{Q(x_t, y)} \frac{\pi(y)}{\pi(x_t)} \frac{\gamma^t_{\phi(x_t)}}{\gamma^t_{\phi(y)}}\right). \tag{3}$$

5. If the proposal is accepted, increase the weight $\gamma^{t+1}_{\phi(y)} = \gamma^t_{\phi(y)} * f$ for some constant $f > 1$.
6. If x_t and y belong to different basins D_k and D_l, then perform ridge descent to update the estimated upper-bound of the energy barrier between the two basins. In ridge descent we search for a local minimum along the ridge between the two basins, by starting with $a_0 = x_t, b_0 = y$ and iterating to find (a_t, b_t):

$$a_t = \text{argmin}_a \{E(a) : a \in \text{Neighborhood}(b_{t-1}) \cap D_k\}$$
$$b_t = \text{argmin}_b \{E(b) : b \in \text{Neighborhood}(a_t) \cap D_l\}$$

   until $b_{t-1} = b_t$. The neighborhood of a sample is defined as the subspace surrounding the sample with its size controlled by an adaptive radius.
7. After the algorithm converges, construct the ELM based on the energy of the basins that have been discovered and the estimated energy barriers between them. We check the convergence of the algorithm using the multivariate extension of the Gelman and Rubin criterion [5].

Figure 2 illustrates the Markov chain produced by the algorithm. Note that the modified acceptance probability in eqn.(3) will reject sample y if the Markov chain has visited bin $\phi(y)$ many times, forcing the sampler to move into less explored space.

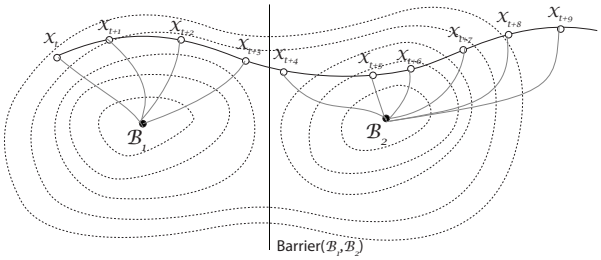


Fig. 2. Sequential MCMC samples $x_t, x_{t+1}, \ldots, x_{t+9}$. For each sample, we perform gradient descent to determine which energy basin the sample belongs to. If two sequential samples fall into different basins (x_{t+3} and x_{t+4} in this example), we estimate or update the upper-bound of the energy barrier between their respective basins (B_1 and B_2 in this example).

Unlike in previous work that samples from discrete state spaces, several new issues arise in plotting the ELMs of continuous hypothesis spaces. For example, many of the basins in the hypothesis space have a flat bottom which may result in a large number of false local minima, and thus we merge local minima identified by gradient descent based on the following criteria: (1) the distance between two local minima is smaller than a constant ϵ; or (2) there is no barrier along the straight line between two local minima. Besides, there may be constraints between parameters (e.g., a probability vector should lie on the surface of a unit simplex), and thus we may need to run our algorithm on a manifold. More details of our algorithm can be found in [8].

3 ELMs of Gaussian Mixture Models

An n-component Gaussian Mixture Model (GMM) is a weighted mixture of n Gaussians. The energy function of data clustering using GMM is the negative log of the posterior, given by $E(x) = -\log P(x|z_i : i = 1 \ldots m) - \log P(x)$ for m input data examples $\{z_i\}$. We use a Dirichlet prior on the weights of the model and the Normal-inverse-Wishart prior on the means and variances of the model components.

3.1 Experiments on Synthetic Data

We synthesize a 2-dimensional, 3-component GMM, draw m samples from it, and run our algorithm to plot the ELM. We want to analyze how the separability c affects the energy landscape. The separability of the GMM represents the overlap between separate components of the model and is defined as $c = \min\left(\frac{||\mu_i - \mu_j||}{\sqrt{n}\max(\sigma_1,\sigma_2)}\right)$ [4]. We also look at the effect of partial supervision on the energy landscape by assigning ground truth labels to a fraction of the samples.

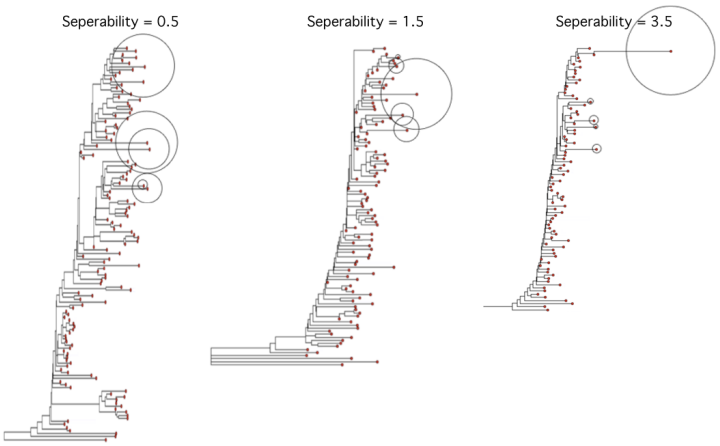


Fig. 3. ELMs for 100 samples drawn from GMMs with low, medium and high separability ($c = 0.5, 1.5, 3.5$). The relative probability mass of the energy basins corresponding to the 5 lowest-energy minima are indicated by circle size around the local minima.

Comparing Different Ground-truth Models Figure 3 shows some of the ELMs with the separability being $\{0.5, 1.5, 3.5\}$ for $m = 100$ samples. The energy landscape becomes increasingly simple (containing fewer local minima) as the separability increases. The landscape for the high separability ($c = 3.5$) case has relatively small energy barriers between the high-energy local minima and a pronounced low-energy global minimum. Conversely, the landscape for the low separability has a structure with high energy barriers between local minima and multiple local minima with similar energy to the global minimum. This indicates that the complexity of learning the GMM model should increase as the separability decreases, as we would expect.

The probability mass of the 5 energy basins corresponding to the lowest-energy local minima are shown in Figures 3 by the circles (similarly we can also show the volume of each basin). The ratio of the mass of the lowest energy basin to the mass of the remaining energy basins increases with separability. This is also consistent with the intuition that high-separability landscapes have lower complexity, as it is more likely that the global optimal solution can be found by gradient descent from a randomly sampled starting point.

We examine the affects of partial supervision by assigning ground truth labels (i.e. which Gaussian cluster a point belongs to) to a portion of the data samples. Figure 4 shows the ELMs of a synthesized GMM (dimension = 2, number of components = 3, separability $c = 1.0$, number of samples = 100) with $\{0\%, 5\%, 10\%, 50\%, 90\%, 100\%\}$ labelled data points. Figure 5 shows the number of local minima in the ELM for the labeling of $1, \ldots, 100$ samples. This shows a significant decrease in landscape complexity for the first 10 labels, and diminishing returns from supervised input after the initial 10%.

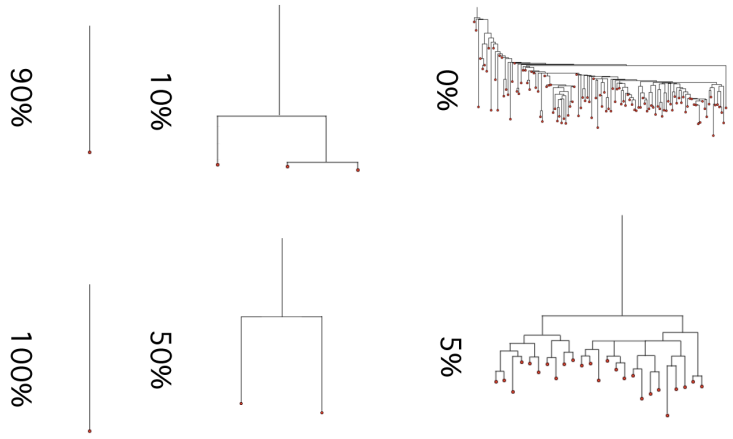


Fig. 4. ELMs of synthesized GMMs (separability $c = 1.0$, nSamples $= 100$) with $\{0\%, 5\%, 10\%, 50\%, 90\%, 100\%\}$ labelled data points.

Behavior of Learning Algorithms: EM, K-mean and SW-Cut Expectation-maximization (EM) is one of the most popular algorithms for learning a GMM from data. K-means is another popular learning algorithm of GMM which can be seen as a degraded variant of EM with hard assignments in the E-step and the assumption of identical spherical Gaussian components. The Swedsen-Wang Cut (SW-cut) algorithm [1] is a generalization of the Swendson-Wang method [11] to arbitrary probabilities. It is a MCMC method that has much faster convergence rates than classic Markov Chain Monte Carlo methods such as the Gibbs sampler in cases when model states are strongly coupled (such as the Ising-Potts model) [9].

For each synthetic dataset, we ran the three algorithms for 200 times and found the energy basins of the ELM that the learned models belong to. Hence we obtain a histogram of the learned models on the leaf nodes of the ELM for each learning algorithm as shown in Figure 6–7.

Figures 6 and 7 show a comparison of the EM, K-mean, and SW-cut algorithms for $n = 100$ samples drawn from low ($c = 0.5$) and high ($c = 3.5$) separability GMMs. The SW-cut algorithm performs best in each situation, always converging to the global optimal solution. In the low separability case, the K-mean algorithm converges to one of the seven local minima, with a higher probability of converging to those with lower energy. The EM algorithm almost always finds the global minimum and thus outperforms K-mean. This can be explained by the fact that K-mean is a degraded variant of EM with extra assumptions that may not hold. However, in the high separability case, the K-mean

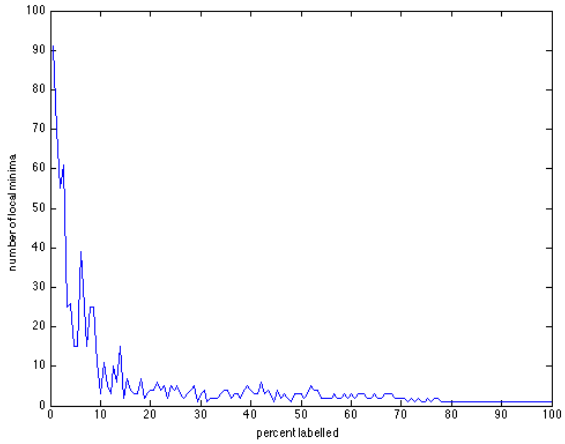
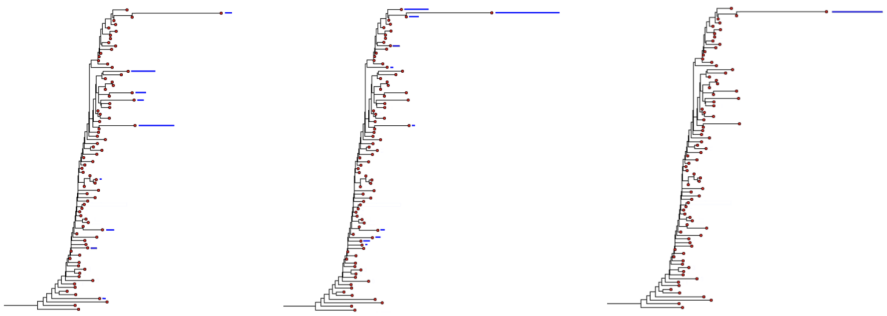


Fig. 5. Number of local minima versus the percentage of labelled data points for a GMM with separability $c = 1.0$.

(a) EM (b) k-means (c) SW-cut

Fig. 6. Low separability $c = 0.5$: histogram of EM, k-means, and SW-cut algorithm results on the ELM.



(a) EM (b) k-means (c) SW-cut

Fig. 7. High separability $c = 3.5$: histogram of EM, k-means, and SW-cut algorithm results on the ELM.

algorithm converges to the true model the majority of the time, while the EM almost always converges to a local minimum with higher energy than the true model. This can be explained by a recent theoretical result showing that the objective function of hard-EM (with k-means as a special case) is the summation of the standard energy function of GMM with an inductive bias in favor of high-separability models [12, 10].

3.2 Experiments on Real Data

We ran our algorithm to plot the ELM for the well-known Iris data set from the UCI repository [3]. The data set contains 150 points in 4 dimensions and can be modeled as a 3-components 4-dimensional GMM. The three components each represent a type of iris plant and the true component labels are known. The points corresponding to the first component are linearly separable from the others, but the points corresponding to the remaining two components are not linearly separable.

Figure 8 shows the ELM of the Iris dataset. We visualize the local minima by plotting the ellipsoids of the covariance matrices centered at the means of each component in 2 of the 4 dimensions.

The 6 lowest energy local minima are shown on the right and the 6 highest energy local minima are shown on the left. The high energy local minima are less accurate models than the low energy local minima. The local minima (E) (B) and (D) have the first component split into two and the remaining two (non-separable) components merged into one. The local minima (A) and (F) have significant overlap between the 2nd and 3rd components and (C) has the components overlapping completely. The low-energy local minima (G-L) all have the same 1st components and slightly different positions of the 2nd and 3rd components.

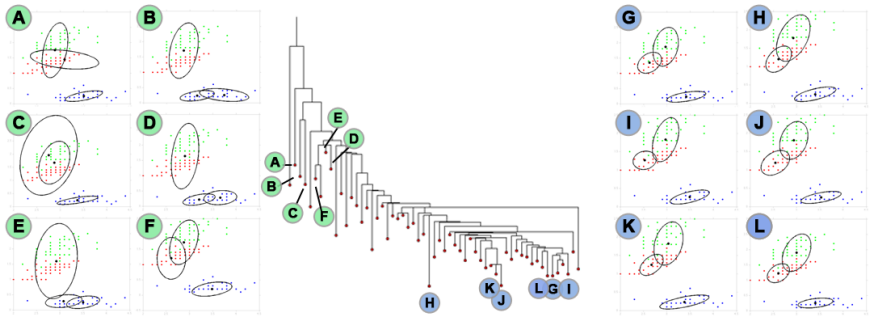


Fig. 8. ELM of the Iris dataset and corresponding local minima.

4 Learning Mixtures of Bernoulli Templates

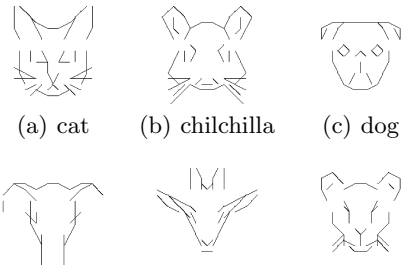
An object image can be converted to a dense edge map or a sparse sketch map using Gabor filters. We can quantize the edges/sketches into finite locations and orientations, and thus each input image is transformed to a binary vector. A Bernoulli template $P \in \{0,1\}^n$ is an n-dimensional binary vector. A sample x is generated from P with independent Bernoulli noise: the i-th coordinate x_i is equal to P_i with a fixed probability p and equal to $1 - P_i$ with probability $1-p$. An K-component Mixture of Bernoulli Templates (MBT) B is a weighted mixture of K Bernoulli templates defined by the set of templates $\{P_i\}$ and weights $\{w_i | w_i \in [0,1]\}$ for $i \in \{0,\ldots,K\}$ with $\sum w_i = 1$. Samples s_j are drawn from B by first sampling a component P_i from the discrete distribution of weights $\{w_i\}$, then sampling from the template P_i as outlined above. We wish to compute the energy landscape map of the space of MBTs with a fixed noise level p. The energy function that we use is the negative log of the posterior, given by $E(B) = -\log P(B|z_i : i = 1 \ldots M)$ for M samples $\{z_i\}$. The probability of a sample z_i given a MBT is defined as:

$$P(z_i|B) = \sum_{i=1}^{m} w_i p^{\sum_{j=1}^{n} I(z_i(j)=P_i(j))} (1-p)^{\sum_{j=1}^{n} I(z_i(j)\neq P_i(j))},$$

where $P_i(j)$ is the j-th component of the i-th Bernoulli template in B, and $z_i(j)$ is the j-th component of the i-th sample. When constructing the ELMs, we discretize the hypothesis space by allowing the weights to take values $w_i \in \{0, 0.1, \ldots, 1.0\}$.

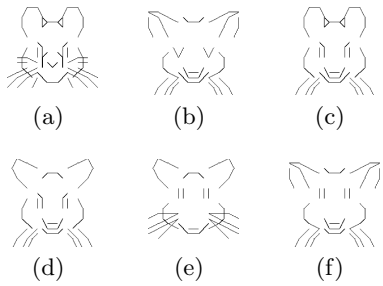
4.1 Experiment on synthetic data

We synthesized Bernouilli templates which represent animal faces as show in Figure 9. Each animal face is a 9x9 grid with each cell containing up to 3 sketches. The dictionary of sketches contains 18 elements, each of which is a straight line



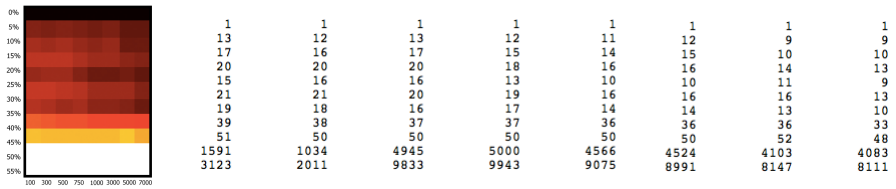
(a) cat (b) chilchilla (c) dog (d) elephant (e) goat (f) lion

Fig. 9. Animal face templates - low overlap



(a) (b) (c) (d) (e) (f)

Fig. 10. Mouse face templates - high overlap



(a) landscape map (b) number of local minima

Fig. 11. The number of local minima in the energy landscape of learning MBT with varying values of noise level p and number of samples.

connecting the endpoints or midpoints of the cell edges. The Bernouilli template can therefore be represented as a $18 \times 9 \times 9$ dimensional binary vector. There are 10 animals in total, so we have a Bernoulli mixture model with the number of component $M = 10$.

We construct the energy landscape maps of the Bernouilli mixture model for varying numbers of samples $n = 100, 300, \ldots, 7000$ and varying noise level $p = 0, 0.05, \ldots, 0.5, 0.55$. The number of local minima in each energy landscape is tabulated in Figure 11 (b) and drawn as a heat map in Figure 11 (a). As expected, the number of local minima increases as the noise level p increases, and decreases as the number of samples decreases. In particular, with no noise, the landscape is convex and with noise $p > 0.45$, there are too many local minima and the algorithm does not converge.

We repeat the same experiment using variants of a mouse face as shown in Figure 10. We swap out components of the mouse face (the eyes, ears, whiskers, nose, mouth, head top and head sides) for three different variants. We thereby generate 20 Bernouilli templates which have relatively high degrees of overlap. We generate the ELMs of various MBTs containing three of the 20 templates with noise level $p = 0$. In each MBT, the three templates have different degrees of overlap. Hence we plot the number of local minima in the ELMs versus the degree of overlap as show in Figure 12. As expected, the number of local minima

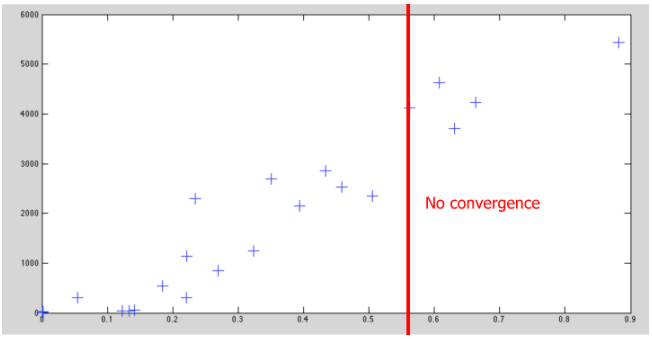


Fig. 12. Number of local minima found for varying degrees of overlap in the Bernoulli templates.

increases with the degree of overlap, and there are too many local minima for the chains to converge past overlap $c = 0.5$.

4.2 Experiment on real data

We perform the Bernouilli templates experiment on a set of real images of animal faces. We binarize the images by extracting the prominent sketches on a 9x9 grid. Eight Gabor filters with eight different orientations centered in the centers and corners of each cell are applied to the image. The filters with a strong response above a fixed threshold correspond to edges detected in the figure; these are mapped to the dictionary of 18 elements. Thus each animal face is represented as a $18 \times 9 \times 9$ dimensional binary vector. The Gabor filter responses on animal face pictures are shown in Figure 13. The binarized animal faces are shown in Figure 14.

We chose 3 different animal types – deer, dog and cat, with an equal number of images chosen from each category (Figure 15). The binarized versions of these can be modeled as a mixture of 3 Bernouilli templates - each template corresponding to one animal face type.

The ELM is shown in Figure 16 along with the Bernouilli templates corresponding to three local minima separated by large energy barriers. We make two observations: 1. the templates corresponding to each animal type are clearly identifiable, and therefore the algorithm has converged on reasonable local minima. 2. The animal faces have differing orientations across the local minima (the deer face on in the left-most local minimum is rotated and tilted to the right and the dog face in the same local minimum is rotated and lilted to the left), which explains the energy barriers between them.

Figure 17 shows a comparison of the SW-cut, k-means, and EM algorithm performance as a histogram on the ELM of animal face Bernouilli Mixture Model. The histogram is obtained by running each algorithm 200 times with a random initialization, then finding the closest local minimum in the ELM to the output of the algorithm. The counts of the closest local minima are then displayed as a

Fig. 13. Animal face images and corresponding binary sketches indicates the existence of a Gabor filter response above a fixed threshold.

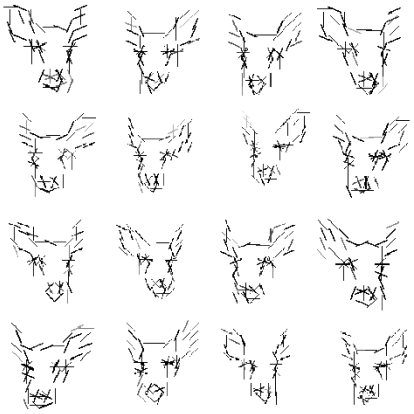


Fig. 14. Deer face sketches binarized from real images.

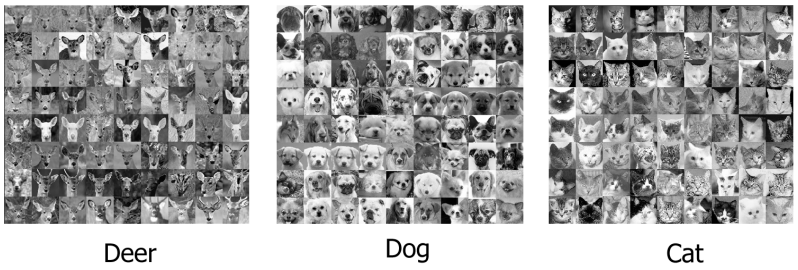


Fig. 15. Animal face images

bar plot next to each local minimum. It can be seen that SW-cut always finds the global minimum, while k-means performs the worst probably because of the high degree of overlap between the sketches of the three types of animal faces.

5 Conclusion

We present a method for computing the energy landscape maps (ELMs) in hypothesis spaces and thus visualize for the first time the non-convex energy minimization problems in computer vision, pattern recognition and statistical learning. We demonstrate the methods in two cases: clustering with Gaussian mixture models in low dimensional space, and learning mixtures of Bernoulli templates from images in very high dimensional space. By plotting the ELMs, we have shown how different problem settings, such as separability and levels of supervision, impact the complexity of the energy landscape. We have also examined

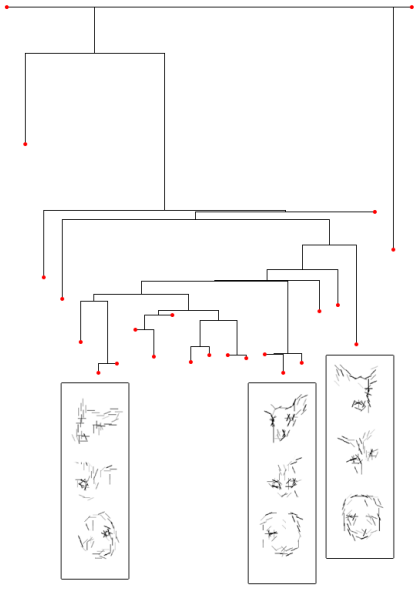
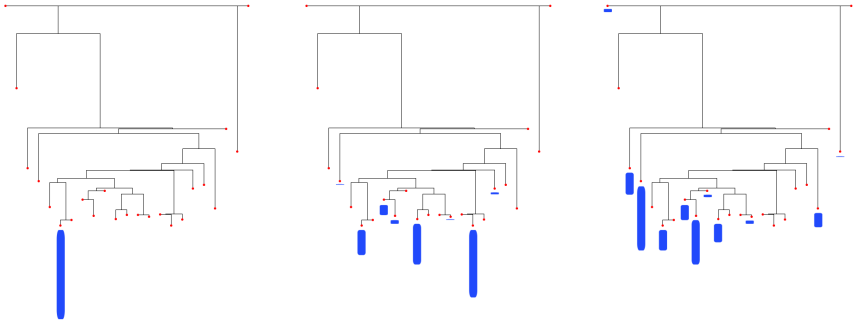


Fig. 16. ELM of three animal faces (dog, cat, and deer). We show the Bernouilli templates corresponding to three local minima with large energy barriers.



(a) SW-cut (b) EM (c) k-means

Fig. 17. Comparison of SW-cut, k-means, and EM algorithm performance on the ELM of animal face Bernouilli Mixture Model.

the behaviors of different learning algorithms in the ELMs. More experimental results and analysis can be found in our technical report [8].

Acknowledgments. The authors thank Dr. Qing Zhou for his tutorial on the algorithm and many helpful suggestions, thank Drs. Yingnian Wu and Adrian Barbu for their discussions, and acknowledge the support of a DARPA MSEE project FA 8650-11-1-7149.

References

1. Barbu, A., Zhu, S.C.: Generalizing swendsen-wang to sampling arbitrary posterior probabilities. IEEE Transactions on Pattern Analysis and Machine Intelligence 27, 1239–1253 (2005)
2. Becker, O.M., Karplus, M.: The topology of multidimensional potential energy surfaces: Theory and application to peptide structure and kinetics. The Journal of Chemical Physics 106(4), 1495–1517 (1997)
3. Blake, C.L., Merz, C.J.: UCI repository of machine learning databases (1998)
4. Dasgupta, S., Schulman, L.J.: A two-round variant of em for gaussian mixtures. In: Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence. pp. 152–159. UAI'00, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2000)
5. Gelman, A., Rubin, D.B.: Inference from Iterative Simulation Using Multiple Sequences. Statistical Science 7(4), 457–472 (1992)
6. Liang, F.: Generalized wang-landau algorithm for monte carlo computation. Journal of the American Statistical Association 100, 1311–1327 (2005)
7. Liang, F.: A generalized wang-landau algorithm for monte carlo computation. JASA. Journal of the American Statistical Association 100(472), 1311–1327 (2005)
8. Pavlovskaia, M., Tu, K., Zhu, S.C.: Mapping energy landscapes of non-convex learning problems. arXiv preprint arXiv:1410.0576 (2014)
9. Potts, R.B.: Some Generalized Order-Disorder Transformation. In: Transformations, Proceedings of the Cambridge Philosophical Society. vol. 48, pp. 106–109 (1952)
10. Samdani, R., Chang, M.W., Roth, D.: Unified expectation maximization. In: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 688–698. Association for Computational Linguistics (2012)
11. Swendsen, R.H., Wang, J.S.: Nonuniversal critical dynamics in Monte Carlo simulations. Physical Review Letters 58(2), 86–88 (Jan 1987)
12. Tu, K., Honavar, V.: Unambiguity regularization for unsupervised learning of probabilistic grammars. In: Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing and Natural Language Learning (EMNLP-CoNLL 2012) (2012)
13. Wang, F., Landaul, D.: Efficient multi-range random-walk algorithm to calculate the density of states. Phys. Rev. Lett. 86, 2050–2053 (2001)
14. Zhou, Q.: Random walk over basins of attraction to construct ising energy landscapes. Phys. Rev. Lett. 106, 180602 (May 2011)