Learning in energy based models via score matching

REDWOOD CENTER for Theoretical Neuroscience

Jascha Sohl-Dickstein and Bruno Olshausen

Redwood Center, University of California, Berkeley

Main Points

- The neural representation of sensory input has in many cases been shown to correspond to intrinsic statistical structure in the world. Building probabilistic models of the world can thus strongly inform our understanding of its representation in the brain.
- Many potentially powerful probabilistic models in neuroscience and machine learning go unused because intractability of their partition function makes learning impossible
- This problem can be sidestepped by replacing the full partition function integral with an approximate integral over patches around each of the observed data points.
- This approximation can be viewed as a formalization of the philosophy espoused in Contrastive Divergence
- The objective function learned in this fashion is the Score Matching objective function, recently proposed by Hyvärinen [3]
- Score Matching allows fast learning in some previously highly challenging cases
- Score Matching promises to allow simpler learning of heterogeneous and hierarchical models

Derivation

- Finding the Maximum Likelihood parameter estimate of a probabilistic model

$$q(\mathbf{x}) = \frac{e^{-E(\mathbf{x};\theta)}}{Z(\theta)} \tag{1}$$

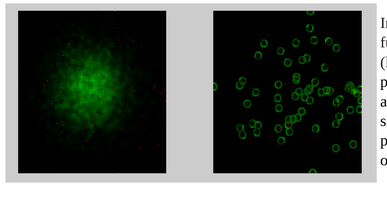
  is in general intractable since it requires solving the partition function integral $Z(\theta) = \int e^{-E(\mathbf{x};\theta)} d\mathbf{x}$

- Maximum Likelihood parameter estimation can be accomplished by performing gradient descent using the following gradient

$$\left\langle \frac{\partial E(\mathbf{x};\theta)}{\partial \theta} \right\rangle_{p(\mathbf{x})} - \left\langle \frac{\partial E(\mathbf{x},\theta)}{\partial \theta} \right\rangle_{q(\mathbf{x})} \tag{2}$$

  where the second term stems from the partition function in equation 1.

- Contrastive Divergence [1] suggests that the distribution $q(\mathbf{x})$ in the second term be replaced by the first step in a Markov Chain Monte Carlo chain initiated at $p(\mathbf{x})$.
- In a more precisely defined fashion, one might instead imagine replacing the model distribution with a model distribution built out of small patches of the model around each data point



In score matching, the full model distribution (left pane, green) is replaced (right pane) with a model distribution consisting of cutouts in hyperspheres around each of the data points (red)

- This can be done by first Taylor expanding the change in $\frac{\partial E(\mathbf{x};\theta)}{\partial \theta}$ around each of the data points

$$\frac{\partial E(\mathbf{x};\theta)}{\partial \theta} - \frac{\partial E(\mathbf{x}+\mathbf{r};\theta)}{\partial \theta} \approx -\nabla_x \frac{\partial E(\mathbf{x}+\mathbf{r};\theta)}{\partial \theta} \cdot \mathbf{r} \tag{3}$$

  and then approximating the average value of that Taylor expansion in a hypersphere around the data points using its divergence

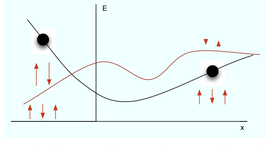
$$\begin{aligned} \left\langle \frac{\partial E(\mathbf{x};\theta)}{\partial \theta} \right\rangle_{p(\mathbf{x})} - \left\langle \frac{\partial E(\mathbf{x},\theta)}{\partial \theta} \right\rangle_{q(\mathbf{x})} &\propto -\left\langle \frac{1}{q(x)} \nabla_x \cdot (q(x) \nabla_x f(x;\theta)) \right\rangle_{p(x)} \\ &\propto \left\langle \nabla_x E(x;\theta) \cdot \nabla_x \frac{\partial E(x;\theta)}{\partial \theta} - \nabla_x \cdot \nabla_x \frac{\partial E(x;\theta)}{\partial \theta} \right\rangle_{p(x)} \end{aligned} \tag{4}$$

- This approximate gradient can be integrated to produce an alternative Maximum Likelihood objective function which is guaranteed (subject to certain reasonable constraints) to be at a global minimum when model and world probability distributions agree:

$$\hat{\theta} = \arg\min_{\theta} \left\langle \frac{1}{2} \nabla_x E(x;\theta) \cdot \nabla_x E(x;\theta) - \nabla_x \cdot \nabla_x E(x;\theta) \right\rangle_{p(x)} \tag{5}$$

Intuition and Practicalities

- In the objective function above each data point is trying to pull the model around itself such that it lies at the bottom of a well in the energy landscape (first derivative of E as close to 0 as possible, second derivative as large as possible). The net result of all the data points greedily doing this is to pull the model distribution onto the data distribution.

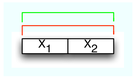


$E(\mathbf{x}) = -\log p(\mathbf{x})$ for the *(black)* data and *(red)* model distributions. The pull applied to the model by each of the *(black balls)* data points is indicated by the red arrows, first derivative term above the line, second derivative term below.

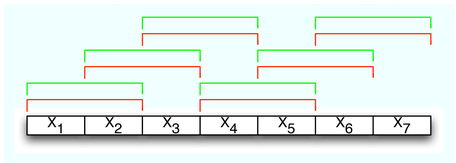
- This Score Matching derivation can be thought of as replacing the model distribution with a distribution whose local structure comes from the model and whose global structure comes from the data. This is a process nearly diametrically opposed to mean field theory techniques.
- In general, Score Matching will take a time $\propto$ [number of model parameters] $\cdot$ [dimension of data] $\cdot$ [number of samples] per learning step.
- Learning is more effectively done via line searches than straightforward gradient ascent, as the 3rd derivatives in the objective function's gradient can vary wildly in magnitude over many functions' parameter spaces.
- Score Matching is equivalent to contrastive divergence with infinitesimal steps and momentum-less Langevin dynamics (as noted by Hyvärinen [2]). This correspondence occurs because, unlike in Metropolis-Hastings Monte Carlo, Langevin dynamics constrain a system to evolve in time in an unbiased fashion

Application to Fields of Experts

- Product of Experts: $E(\mathbf{x}) = E_1(\mathbf{x}) + E_2(\mathbf{x})$
  All of the experts occur above a single image patch



- Field of Experts: $E(\mathbf{x}) = E_1(\{x_1, x_2\}) + E_2(\{x_1, x_2\}) + E_1(\{x_2, x_3\}) + E_2(\{x_2, x_3\})$...
  All of the experts occur above every image patch

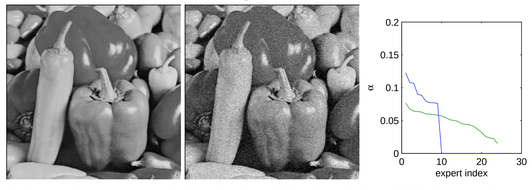


- Field of Experts with student-t test experts, as in Roth and Black [4]

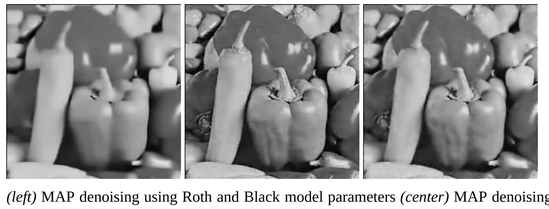
$$q(\mathbf{x}) = \frac{1}{Z} \prod_k \prod_i e^{-E_i(\mathbf{x_k})} \tag{6}$$

$$E_i(\mathbf{x}) = \log\left(1 + \sum_j \lambda_i^{\mathbf{T}} \mathbf{x}\right)^{\alpha_i} \tag{7}$$

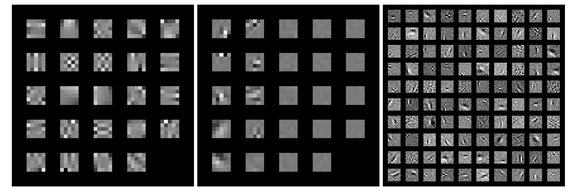
- Score Matching was used to learn the ML parameters for the Field of Experts model. Field of Experts model parameters have previously been learned via Contrastive Divergence
- Receptive fields learned by Score Matching are more biologically consistent (Gabor-like) than those learned via Contrastive Divergence, and the Score Matching model performs marginally better at denoising without inclusion of an ad-hoc adjustment to the noise variance during reconstruction



(left) Sample image provided by Roth and Black with their FOE demo code *(center)* Sample image with additive gaussian noise ($\sigma_{noise} = 0.28\sigma_{image}$) *(right)* Learned α_i values (sorted) for Contrastive Divergence learning (green) and Score Matching learning (blue)



(left) MAP denoising using Roth and Black model parameters *(center)* MAP denoising using Roth and Black model parameters and a fudge factor of $\sigma_{MAP} = 5.25\sigma_{noise}$ applied to the noise during reconstruction *(right)* MAP denoising using Score Matching model parameters

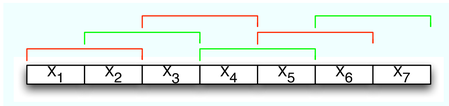


(left) 5x5 receptive fields learned via Contrastive Divergence *(center)* 5x5 receptive fields learned via Score Matching *(right)* Random sample of 100 receptive fields learned using Score Matching and a 10x10 by 3x overcomplete Tapestry of Experts model (see below)

The Future

- Tapestry of Experts (spatially heterogeneous units)
  Experts repeat in a regular pattern above the image, but do not overlap with shifted versions of themselves.

$$E(\mathbf{x}) = E_r(\{x_1, x_2\}) + E_g(\{x_2, x_3\}) + E_r(\{x_3, x_4\}) + E_g(\{x_4, x_5\})...$$



  (current work with a Tapestry of Experts approach produces performance equivalent to the Field of Experts but requires a lower degree of overcompleteness)
- Functionally heterogeneous units,
  $E_i \in \left\{ \alpha \log\left(1 + (\lambda^{\mathbf{T}}\mathbf{x})^2\right), \alpha\left(1 + |\lambda^{\mathbf{T}}\mathbf{x}|\right)^{\beta}, \sigma\left(\lambda^{\mathbf{T}}\mathbf{x} + \alpha\right), ... \right\}$
- Continuous, deterministic, energy based hierarchical models
  - Derivatives of energies can be easily propagated up and down the hierarchy via the chain rule
  - In the case of experts consisting of linear transformations followed by pointwise nonlinearities, the data and first derivative of the energy function can instead be transformed upwards together and chain rule propagation is unnecessary

References

[1] M. A. Carreira-Perpignan, G. E. Hinton. On Contrastive Divergence Learning. Artificial Intelligence and Statistics, 2005

[2] A. Hyvärinen. Connections between score matching, contrastive divergence, and pseudolikelihood for continuous-valued variables. Submitted manuscript.

[3] A. Hyvärinen. Estimation of non-normalized statistical models using score matching. Journal of Machine Learning Research, 2005.

[4] S. Roth, M. Black. Fields of Experts: A Framework for Learning Image Priors. IEEE Computer Vision and Pattern Recognition, 2005

Acknowledgments

This work was supported by NGA grant MCA 015894-UCB and NSF grant IIS-06-25223.