Learning Doubly Intractable Latent Variable Models via Score Matching

Eszter Vértes, Maneesh Sahani
Gatsby Unit, University College London

Background

- Latent variable models are powerful tools for learning about the underlying structure of a dataset in an unsupervised setting.
- Learning is intractable in most complex (e.g. non-Gaussian) models.

Double intractability:
1. The posterior distribution is intractable, i.e. we cannot compute the normalizer for the latent variables: $Z(\theta) = \int p(x, z) dz$
2. For some latent variable models the joint distribution is only available up to proportionality:
   $$p(x, z) = \frac{1}{Z(\theta)} \tilde{p}(x, z), \text{ where } Z(\theta) = \int \tilde{p}(x, z) dx dz$$
- Variational algorithms are infeasible since we do not have access to the normalized log-joint.

Score matching (SM)

- The exact SM objective for jointly exponential family models:
  $$J(\theta) = \sum_x \sum_z \frac{1}{2} \| \theta^T \partial_z S(x, z) \|_2^2 + \langle (\theta^T \partial_z S(x, z) ) \rangle_d - \langle \partial_z \theta^T E_\theta(x, z) \rangle_d$$
- We can propagate derivatives wrt $\theta$ into the expectations without knowing the normaliser of $p(z|x)$ or $p(x, z)$ by using the property of exp. family:
  $$\partial_\theta \log p(z|x) = S(x, z) - \langle S(x, z) \rangle_{z|x}$$
- The posterior $p(z|x)$ appears in the resulting gradient $\nabla_\theta J(\theta)$ only in terms of its expectations.
- We approximate these integrals using a Hamiltonian Monte Carlo sampler (Hoffman et al., 2011)

Score matching for latent variable models

- For energy based models of the form: $p(x, z) \propto \exp(-E_\theta(x, z))$
  - The score function can be expressed as an expectation:
    $$\partial_\theta \log p_\theta(x) = \int p(z|x)(-\partial_z E_\theta(x, z)) dz$$
  - The score matching objective can be rewritten (Swersky et al., 2011):
    $$J(\theta) = \sum_x \sum_z \frac{1}{2} \langle \partial_z E_\theta(x, z) \rangle_{z|x}^2 + \langle (\partial_z E_\theta(x, z) ) \rangle_{z|x} - \langle \partial_\theta \theta^T E_\theta(x, z) \rangle_{z|x}$$

Exponential family

- Jointly exponential family model:
  $$p(x, z) = \exp(\theta^T S(x, z) - A(\theta))$$
  where $\theta$: natural parameter vector, $S(x, z)$: sufficient statistic
  - Useful property: $\nabla_\theta A(\theta) = \langle S(x, z) \rangle_{x,z}$

SM for doubly intractable models

- Estimation of non-normalized statistical models by score matching.

Experiments

- Rectified latent Gaussian model defined as:
  $$p(z) \propto \mathcal{N}(z(0, \Sigma)) \prod_i \theta(z_i)$$
  $$p(x|z) = \mathcal{N}(Wz, \sigma^2 I)$$
  - Sufficient statistics: $S(x, z) = \text{vec} [x^T x, x^T z, z^T z]$.
  - In general, the normalizer for the joint model cannot be computed analytically.
  - $z \in \mathbb{R}^d, x \in \mathbb{R}^d$, we learn $\Sigma, W, \sigma$

Contours of learned and true densities

Total variation distance

- Empirical distance between two densities:
  $$\delta(P, Q) = \sup_x |P(x) - Q(x)|$$
  - Computed between pairs of data sets generated from the true and learned models (green) and between two data sets coming from the true model (blue)

Summary

- Score matching can be applied to doubly intractable jointly exponential family models
- SM allows for learning flexible latent variable models with arbitrary sufficient statistics
- No need for fixed form approximations of the posterior distribution
- In contrast to the Boltzmann machine learning rule or contrastive divergence, Monte Carlo simulation is only required for sampling from the posterior, not from the joint distribution

References


This work was funded by the Gatsby Charitable Foundation.
Contact: eszter@gatsby.ucl.ac.uk