Generalized Noise Contrastive Estimation

Miika Pihlaja

Joint work with *Michael Gutmann* and *Aapo Hyvärinen*

University of Helsinki
Dept. of Mathematics and Statistics,
Dept. of Computer Science & HIIT

Motivation - Unnormalized statistical model

- Want to estimate a parameterized model for the data pdf $p_d(\mathbf{x})$ of r.v. X from N_d i.i.d. observations $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{N_d})$
- An unnormalized probabilistic model $p_m^0(\mathbf{x}; \theta)$ is a model for $p_d(\mathbf{x})$ which does not integrate to one for all θ
- It defines a normalized model via

$$p_m(\mathbf{x}; \theta) = \frac{p_m^0(\mathbf{x}; \theta)}{Z(\theta)}, \qquad Z(\theta) = \int p_m^0(\mathbf{x}; \theta) d\mathbf{x}$$

- Computing the value of partition function $Z(\theta)$ is often not feasible. $\Rightarrow$ Want to estimate parameters θ without having to compute $Z(\theta)$
- Applications: Estimating parameters of MRFs, multilayer network models …

Why Maximum Likelihood is problematic

In MLE, partition function cannot be ignored, toy example follows

- Estimate the variance of Gaussian

$$x \sim \mathcal{N}(0, \sigma^2), \qquad p_m(x; \sigma^2) = \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}}}_{Z(\sigma^2)} \exp(-\frac{x^2}{2\sigma^2})$$

- log-likelihood includes the partition function $\Rightarrow Z(\sigma^2)$ must be computed

$$\ell(\sigma^2) = -\left(\Sigma_i x_i^2\right)/(2\sigma^2) - N_d \log Z(\sigma^2)$$

- could we plug it in as another parameter, $c = -\log Z(\sigma^2)$

$$\ell(\sigma^2, c) = -\left(\Sigma_i x_i^2\right)/(2\sigma^2) + N_d c$$

- No, $\ell(\sigma^2, c) \to \infty$ as $c \to \infty$, problem not well defined

Maximum Likelihood as variational problem

- We want to find density f which minimizes $D_{KL}(p_d||f)$

$$\int p_d(x)\log\frac{p_d(x)}{f(x)}\,dx \quad = \quad \int p_d(x)\left(\log p_d(x) - \log f(x)\right)dx$$

- Equivalently we can maximize objective

$$J(f) = \int p_d(x)\log f(x)$$

- Need constraints for f - positive and integrates to 1

$$\begin{aligned} J(f) &= \int p_d(x)\log f(x) + \lambda\left(\int f(x) - 1\right)dx \\ \frac{\delta J}{\delta f} &= \frac{p_d}{f} + \lambda \end{aligned}$$

- Setting the derivative to zero and solving $\lambda = -1$, we find $f = p_d$

Maximum Likelihood as variational problem

- Knowing $\lambda = -1$, we can write the objective simply as

$$J(f) = \int p_d(x) \log f(x) - \int f(x)\, dx$$

- We have transformed the constrained minimization of KL-divergence to an unconstrained optimization problem
- But we still need to compute the second integral
  Introduce auxiliary density $p_n(x)$, use *Importance Sampling*

$$J(f) = \int p_d(x) \log f(x) - \int p_n(x) \frac{f(x)}{p_n(x)}\, dx$$

- *Problem*: ratio $f(x)/p_n(x)$ can have very large values $\Rightarrow$ large variance in estimation

Generalization to a family of estimators

- Replace log and identity by two nonlinear functions $g_1, g_2 : \mathbb{R}_+ \to \mathbb{R}$

$$J(f) = \int p_d(x) \underbrace{\log f(x)}_{g_1\left(\frac{f(x)}{p_n(x)}\right)} - \int p_n(x) \underbrace{\left(\frac{f(x)}{p_n(x)}\right)}_{g_2\left(\frac{f(x)}{p_n(x)}\right)} dx$$

$$J(f) = \int p_d(x) g_1\left(\frac{f(x)}{p_n(x)}\right) - \int p_n(x) g_2\left(\frac{f(x)}{p_n(x)}\right) dx$$

Theorem

If $g_1()$ and $g_2()$ are strictly increasing and fulfill

$$\frac{g_2'(x)}{g_1'(x)} = x,$$

then (under some regularity conditions) $J(f)$ attains it's maximum exactly when $f = p_d$

Bregman divergence view

- Bregman divergence between $p_d(x)$ and $f(x)$ generated by convex function U is defined as

$$D_U[p_d, f] = \int U(p_d(x)) - U(f(x)) - U'(f(x))(p_d(x) - f(x))\,dx$$

- Define a *scaled* Bregman divergence[1]

$$D_U^{p_n}(p_d, f) = \int p_n \left[U\left(\frac{p_d}{p_n}\right) - U\left(\frac{f}{p_n}\right) - U'\left(\frac{f}{p_n}\right)\left(\frac{p_d}{p_n} - \frac{f}{p_n}\right)\right] dx$$

- Denote by V the Fenchel-Legendre conjugate of U, then

$$-D_U^{p_n}(p_d, f) = \int p_d \underbrace{U'\left(\frac{f}{p_n}\right)}_{g_1(\cdot)} - \int p_n \underbrace{V\left(U'\left(\frac{f}{p_n}\right)\right)}_{g_2(\cdot)} dx$$

[1] Stummer & Vajda, arXiv:0911.2784 (2009)

Estimation in practice

- To estimate unnormalized $p_m^0(\mathbf{x};\alpha)$ model and its normalizing constant, we define

$$\log p_m(\mathbf{x};\theta) = \log p_m^0(\mathbf{x};\alpha) + c \quad \text{with} \quad \theta = \{\alpha, c\}$$

- And need to maximize

$$J(\theta) = \int p_d(\mathbf{x}) g_1\left(\frac{p_m(\mathbf{x},\theta)}{p_n(\mathbf{x})}\right) - \int p_n(\mathbf{x}) g_2\left(\frac{p_m(\mathbf{x};\theta)}{p_n(\mathbf{x})}\right) d\mathbf{x}$$

- Compute empirical expectations with samples $(\mathbf{x}_1, \ldots, \mathbf{x}_{N_d})$ from p_d and $(\mathbf{y}_1, \ldots, \mathbf{y}_{N_n})$ from p_n

$$J(\theta) = \frac{1}{N_d}\sum_{i=1}^{N_d} g_1\left(\frac{p_m(\mathbf{x}_i;\theta)}{p_n(\mathbf{x}_i)}\right) - \frac{1}{N_n}\sum_{j=1}^{N_n} g_2\left(\frac{p_m(\mathbf{y}_j;\theta)}{p_n(\mathbf{y}_j)}\right)$$

- Estimate $\hat{\theta}$ by maximizing $J(\theta)$

Estimation in practice

Theorem

Estimator $\hat{\theta}$ is *consistent* and *asymptotically normal*,
$\sqrt{N_d}(\hat{\theta} - \theta^\star) \sim \mathcal{N}(0, \Sigma_g)$

- Family of estimators parameterized by the choice of
  - auxiliary density p_n
  - nonlinearities $g_1()$ and $g_2()$ (fixing one determines the other)
  - size of auxiliary sample N_n and possibly data sample N_d
- We can try to minimize MSE

$$E_d \,||\, \hat{\theta} - \theta^\star||^2 = \mathrm{tr}(\Sigma_g)/N_d$$

by choosing these carefully

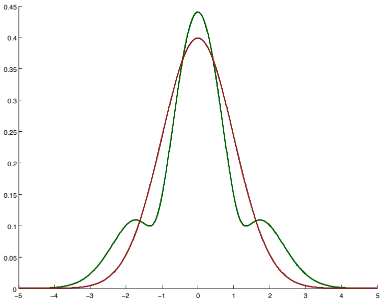
Choice of auxiliary distribution p_n

- We would like $p_n(\mathbf{x})$ to fulfill following properties
  - Easy to sample from
  - Easy to evaluate for any $\mathbf{x}$
  - Give small MSE for the estimator

- For the importance sampling case $g_1(x) = \log x$ and $g_2(x) = x$, we have expression for optimal p_n

$$p_n(\mathbf{x}) \propto ||\mathcal{I}^{-1}\psi(\mathbf{x})||\, p_d(\mathbf{x})$$

where $\psi = \nabla_\theta \log p_m(\mathbf{x};\theta^\star)$ is a score function evaluated at true parameter value and $\mathcal{I}$ is a generalization of Fisher information matrix



- In practice, use e.g. multivariate Gaussian

Choice of nonlinearities $g_1()$ and $g_2()$

Some examples of nonlinearities

- Importance Sampling

$g_1(q)$	$g_2(q)$	Objective $J_g(\theta)$	$\nabla_\theta J_g(\theta)$
$\log q$	q	$E_d \log p_m - E_n \frac{p_m}{p_n}$	$E_d\, \psi - E_n \frac{p_m}{p_n} \psi$

- Noise Contrastive[2]

$\log(\frac{q}{1+q})$	$\log(1+q)$	$E_d \log(\frac{p_m}{p_m+p_n}) + E_n \log(\frac{p_n}{p_m+p_n})$	$E_d \left(\frac{p_n}{p_m+p_n}\right) \psi - E_n \left(\frac{p_m}{p_m+p_n}\right) \psi$

- Inverse Importance Sampling

$-\frac{1}{q}$	$\log q$	$-E_d \frac{p_n}{p_m} - E_n \log p_m$	$E_d \frac{p_n}{p_m} \psi - E_n\, \psi$

- Importance Sampling

$g_1(q)$	$g_2(q)$	Objective $J_g(\theta)$	$\nabla_\theta J_g(\theta)$
$\log q$	q	$E_d \log p_m - E_n \frac{p_m}{p_n}$	$E_d\, \psi - E_n \frac{p_m}{p_n} \psi$

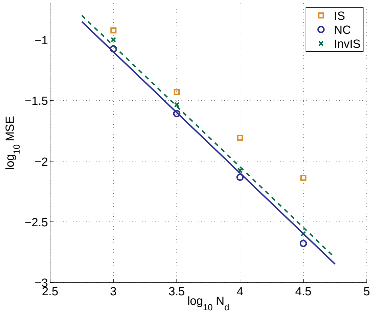
- Noise Contrastive[3]

Estimation of Independent Component Analysis model

- ICA model: $\mathbf{x} = \mathbf{A}\mathbf{s}, \quad \mathbf{B} = \mathbf{A}^{-1}$
- independent Laplacian sources s_i,
  $\mathbf{x} \in \mathbb{R}^4$
  $\dim(\theta) = 17$

$$\log p_d(\mathbf{x}) = -\sum_{i=1}^{4} \sqrt{2}|(\mathbf{b}_i^*)^T \mathbf{x}| - \log 4|\mathbf{A}|$$

$$\log p_m(\mathbf{x};\theta) = -\sum_{i=1}^{4} \sqrt{2}|\mathbf{b}_i^T \mathbf{x}| + c$$

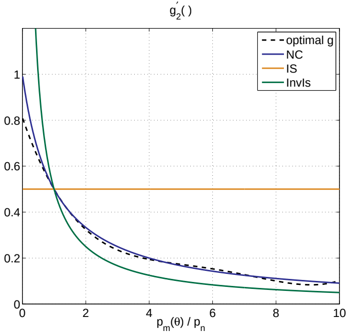


- See [*Gutmann & Hyvärinen, AISTATS 2010*] for simulations with real data and more complex models

Optimal nonlinearities $g_1()$ and $g_2()$ for ICA model

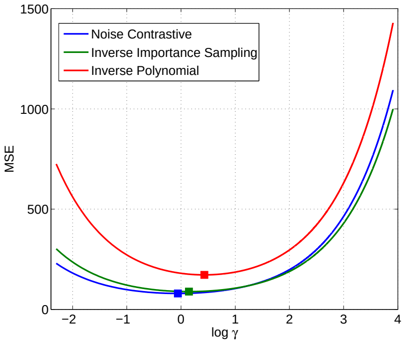
$$J(\theta) = \int p_d \, g_1(p_m/p_n) - \int p_n \, g_2(p_m/p_n),$$

- Using Gaussian noise as p_n, we can numerically optimize $g_2()$
- With super-Gaussian ICA-model and Gaussian noise, $g_1(\cdot)$ and $g_2(\cdot)$ of Noise Contrastive estimation are very close to optimal!



Optimal ratio of data and auxiliary samples

- Can analyze how the estimator behaves when we change the ratio of data and auxiliary sample $\gamma = \frac{N_d}{N_n}$
- We can solve the optimal γ in the ICA model, when $N_{tot} = N_d + N_n$ is kept fixed.



Conclusions

- Maximum Likelihood estimation computationally problematic for unnormalized models
- We propose simple, computationally efficient family of objective functions, including Noise Contrastive Estimation as a special case
- Depends on design parameters: auxiliary density p_n, nonlinearities $g_1()$ and $g_2()$ and ratio of data and auxiliary sample sizes
- For more details [*Pihlaja, Gutmann & Hyvärinen, UAI 2010*; *Gutmann & Hyvärinen, AISTATS 2010*]