Noise-contrastive estimation of unnormalised statistical models

Michael Gutmann

https://sites.google.com/site/michaelgutmann

Institute for Adaptive and Neural Computation
School of Informatics, University of Edinburgh

11th November 2016

Problem statement

- Task: Estimate the parameters $\boldsymbol{\theta}$ of a parametric model $p(.|\boldsymbol{\theta})$ of a d dimensional random vector $\mathbf{x}$
- Given: Data $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ (iid)
- Given: Unnormalized model $\phi(.|\boldsymbol{\theta})$

$$\int_{\boldsymbol{\xi}} \phi(\boldsymbol{\xi}; \boldsymbol{\theta}) \, d\boldsymbol{\xi} = Z(\boldsymbol{\theta}) \neq 1 \qquad p(\mathbf{x}; \boldsymbol{\theta}) = \frac{\phi(\mathbf{x}; \boldsymbol{\theta})}{Z(\boldsymbol{\theta})} \tag{1}$$

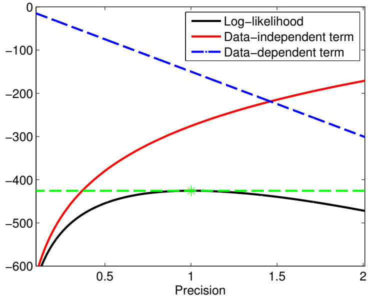
Normalizing partition function $Z(\boldsymbol{\theta})$ not known / computable.

Why does the partition function matter?

- Consider $p(x;\theta) = \frac{\phi(x;\theta)}{Z(\theta)} = \frac{\exp\left(-\theta \frac{x^2}{2}\right)}{\sqrt{2\pi/\theta}}$
- Log-likelihood function for precision $\theta \geq 0$

$$\ell(\theta) = -n \log \sqrt{\frac{2\pi}{\theta}} - \theta \sum_{i=1}^{n} \frac{x_i^2}{2} \tag{2}$$

- Data-dependent (blue) and independent part (red) balance each other.
- If $Z(\boldsymbol{\theta})$ is intractable, $\ell(\boldsymbol{\theta})$ is intractable.



Why is the partition function hard to compute?

$$Z(\boldsymbol{\theta}) = \int_{\boldsymbol{\xi}} \phi(\boldsymbol{\xi}; \boldsymbol{\theta}) \, d\boldsymbol{\xi}$$

- Integrals can generally not be solved in closed form.
- In low dimensions, $Z(\boldsymbol{\theta})$ can be approximated to high accuracy.
- Curse of dimensionality: Solutions feasible in low dimensions become quickly computationally prohibitive as the dimension d increases.

Why are unnormalized models important?

- Unnormalized models are widely used.
- Examples:
  - models of images (Markov random fields)
  - models of text (neural probabilistic language models)
  - models in physics (Ising model)
  - ...
- Advantage: Specifying unnormalized models is often easier than specifying normalized models.
- Disadvantage: Likelihood function is generally intractable.

Program

Noise-contrastive estimation

Intuition and definition

Properties

Bregman divergence to estimate unnormalized models

Framework

Noise-contrastive estimation as member of the framework

Program

Noise-contrastive estimation

Intuition and definition

Properties

Bregman divergence to estimate unnormalized models

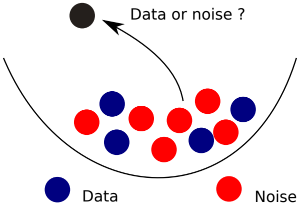
Framework

Noise-contrastive estimation as member of the framework

Intuition behind noise-contrastive estimation

- Formulate the estimation problem as a classification problem: observed data vs. auxiliary “noise” (with known properties)
- Successful classification $\equiv$ learn the differences between the data and the noise
- differences + known noise properties $\Rightarrow$ properties of the data

- Unsupervised learning by supervised learning
- We used (nonlinear) logistic regression for classification

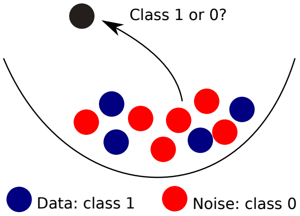


Logistic regression (1/2)

- Let $\mathbf{Y} = (\mathbf{y}_1, \ldots \mathbf{y}_m)$ be a sample from a random variable $\mathbf{y}$ with known (auxiliary) distribution $p_{\mathbf{y}}$.
- Introduce labels and form regression function:

$$P(C = 1|\mathbf{u}; \boldsymbol{\theta}) = \frac{1}{1 + G(\mathbf{u}; \boldsymbol{\theta})} \qquad G(\mathbf{u}; \boldsymbol{\theta}) \geq 0 \qquad (3)$$

- Determine the parameters $\boldsymbol{\theta}$ such that $P(C = 1|\mathbf{u}; \boldsymbol{\theta})$ is
  - large for most $\mathbf{x}_i$
  - small for most $\mathbf{y}_i$.



Logistic regression (2/2)

- Maximize (rescaled) conditional log-likelihood using the labeled data $\{(\mathbf{x}_1, 1), \ldots, (\mathbf{x}_n, 1), (\mathbf{y}_1, 0), \ldots, (\mathbf{y}_m, 0)\}$,

$$J_n^{\mathrm{NCE}}(\boldsymbol{\theta}) = \frac{1}{n} \left(\sum_{i=1}^{n} \log P(C = 1 | \mathbf{x}_i; \boldsymbol{\theta}) + \sum_{i=1}^{m} \log \left[P(C = 0 | \mathbf{y}_i; \boldsymbol{\theta}) \right] \right)$$

- For large sample sizes n and m, $\hat{\boldsymbol{\theta}}$ satisfying

$$G(\mathbf{u}; \hat{\boldsymbol{\theta}}) = \frac{m}{n} \frac{p_{\mathbf{y}}(\mathbf{u})}{p_{\mathbf{x}}(\mathbf{u})} \tag{4}$$

is maximizing $J_n^{\mathrm{NCE}}(\boldsymbol{\theta})$. Without any normalization constraints.

proof

Noise-contrastive estimation

(Gutmann and Hyvärinen, 2010; 2012)

- Assume unnormalized model $\phi(.|\boldsymbol{\theta})$ is parametrized such that its scale can vary freely.

$$\boldsymbol{\theta} \rightarrow (\boldsymbol{\theta}; c) \qquad\qquad \phi(\mathbf{u}; \boldsymbol{\theta}) \rightarrow \exp(c)\phi(\mathbf{u}; \boldsymbol{\theta}) \tag{5}$$

- Noise-contrastive estimation:
  1. Choose $p_{\mathbf{y}}$
  2. Generate auxiliary data $\mathbf{Y}$
  3. Estimate $\boldsymbol{\theta}$ via logistic regression with

$$G(\mathbf{u}; \boldsymbol{\theta}) = \frac{m}{n} \frac{p_{\mathbf{y}}(\mathbf{u})}{\phi(\mathbf{u}; \boldsymbol{\theta})}. \tag{6}$$

Noise-contrastive estimation

(Gutmann and Hyvärinen, 2010; 2012)

- Assume unnormalized model $\phi(.|\boldsymbol{\theta})$ is parametrized such that its scale can vary freely.

$$\boldsymbol{\theta} \rightarrow (\boldsymbol{\theta}; c) \qquad\qquad \phi(\mathbf{u}; \boldsymbol{\theta}) \rightarrow \exp(c)\phi(\mathbf{u}; \boldsymbol{\theta}) \tag{5}$$

- Noise-contrastive estimation:
  1. Choose $p_{\mathbf{y}}$
  2. Generate auxiliary data $\mathbf{Y}$
  3. Estimate $\boldsymbol{\theta}$ via logistic regression with

$$G(\mathbf{u}; \boldsymbol{\theta}) = \frac{m}{n} \frac{p_{\mathbf{y}}(\mathbf{u})}{\phi(\mathbf{u}; \boldsymbol{\theta})}. \tag{6}$$

- $G(\mathbf{u}; \boldsymbol{\theta}) \rightarrow \frac{m}{n} \frac{p_{\mathbf{y}}(\mathbf{u})}{p_{\mathbf{x}}(\mathbf{u})} \quad \Rightarrow \quad \phi(\mathbf{u}; \boldsymbol{\theta}) \rightarrow p_{\mathbf{x}}(\mathbf{u})$

Example

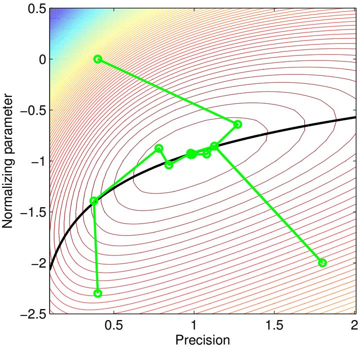
- Unnormalized Gaussian:

$$\phi(u;\boldsymbol{\theta}) = \exp(\theta_2)\exp\left(-\theta_1 \frac{u^2}{2}\right), \quad \theta_1 > 0,\ \theta_2 \in \mathbb{R}, \qquad (7)$$

- Parameters: θ_1 (precision), $\theta_2 \equiv c$ (scaling parameter)

Contour plot of $J_n^{\mathrm{NCE}}(\boldsymbol{\theta})$:

- Gaussian noise with $\nu = m/n = 10$
- True precision $\theta_1^\star = 1$
- Black: normalized models
  Green: optimization paths



Statistical properties

(Gutmann and Hyvärinen, 2012)

- Assume $p_{\mathbf{x}} = p(.|\boldsymbol{\theta}^\star)$
- Consistency: As n increases,

$$\hat{\boldsymbol{\theta}}_n = \text{argmax}_{\boldsymbol{\theta}} J_n^{\text{NCE}}(\boldsymbol{\theta}), \tag{8}$$

  converges in probability to $\boldsymbol{\theta}^\star$.
- Efficiency: As $\nu = m/n$ increases, for any valid choice of $p_{\mathbf{y}}$, noise-contrastive estimation tends to "perform as well" as MLE (it is asymptotically Fisher efficient).

Validating the statistical properties with toy data

- Let the data follow the ICA model $\mathbf{x} = \mathbf{A}\mathbf{s}$ with 4 sources.

$$\log p(\mathbf{x}; \boldsymbol{\theta}^\star) = -\sum_{i=1}^{4} \sqrt{2}|\mathbf{b}_i^\star \mathbf{x}| + c^\star \tag{9}$$

with $c^\star = \log |\det \mathbf{B}^\star| - \frac{4}{2}\log 2$ and $\mathbf{B}^\star = \mathbf{A}^{-1}$.

- To validate the method, estimate the unnormalized model

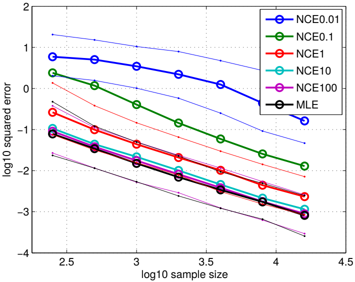
$$\log \phi(\mathbf{x}; \boldsymbol{\theta}) = -\sum_{i=1}^{4} \sqrt{2}|\mathbf{b}_i \mathbf{x}| + c \tag{10}$$

with parameters $\boldsymbol{\theta} = (\mathbf{b}_1, \ldots, \mathbf{b}_4, c)$.

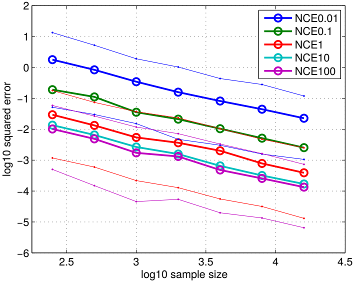
- Contrastive noise $p_\mathbf{y}$: Gaussian with the same covariance as the data.

Validating the statistical properties with toy data

- Results for 500 estimation problems with random **A**, for $\nu \in \{0.01, 0.1, 1, 10, 100\}$.
- MLE results: with properly normalized model



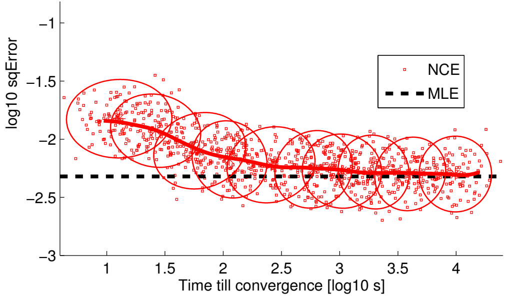
(a) Mixing matrix



(b) Normalizing constant

Computational aspects

- The estimation accuracy improves as m increases.
- Trade-off between computational and statistical performance.
- Example: ICA model as before but with 10 sources. $n = 8000$, $\nu \in \{1, 2, 5, 10, 20, 50, 100, 200, 400, 1000\}$.
  Performance for 100 random estimation problems:



Computational aspects

How good is the trade-off? Compare with

1. MLE where partition function is evaluated with importance sampling. Maximization of

$$J_{\mathrm{IS}}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \log \phi(\mathbf{x}_i; \boldsymbol{\theta}) - \log \left(\frac{1}{m} \sum_{i=1}^{m} \frac{\phi(\mathbf{y}_i; \boldsymbol{\theta})}{p_{\mathbf{y}}(\mathbf{y}_i)} \right) \qquad (11)$$

2. Score matching: minimization of

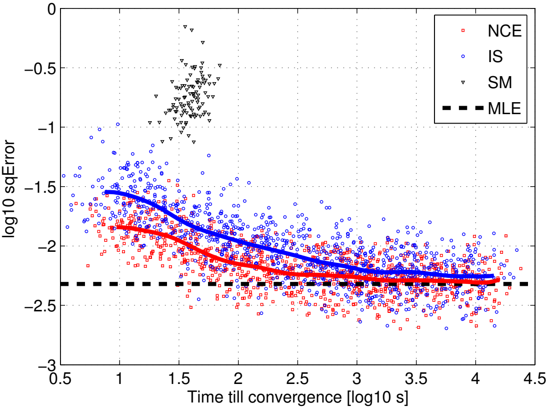
$$J_{\mathrm{SM}}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{10} \frac{1}{2}\Psi_j^2(\mathbf{x}_i; \boldsymbol{\theta}) + \Psi_j'(\mathbf{x}_i; \boldsymbol{\theta}) \qquad (12)$$

with $\Psi_j(\mathbf{x}; \boldsymbol{\theta}) = \frac{\partial \log \phi(\mathbf{x}; \boldsymbol{\theta})}{\partial x_j}$ (here: smoothing needed!)

(see Gutmann and Hyvärinen, 2012, for more comparisons)

Computational aspects

- NCE is less sensitive to the mismatch of data and noise distribution than importance sampling.
- Score matching does not perform well if the data distribution is not sufficiently smooth.



Application examples

- Models of text: e.g. Mnih and Teh, 2012, *A fast and simple algorithm for training neural probabilistic language models*
- Models of images: e.g. Gutmann and Hyvärinen, 2013, *A three-layer model of natural image statistics*
- Machine translation: e.g. Zoph et al, 2016, *Simple, fast noise-contrastive estimation for large RNN vocabularies*
- Product recommendation: e.g. Tschiatschek et al, 2016, *Learning probabilistic submodular diversity models via noise contrastive estimation*

Program

Noise-contrastive estimation

Intuition and definition

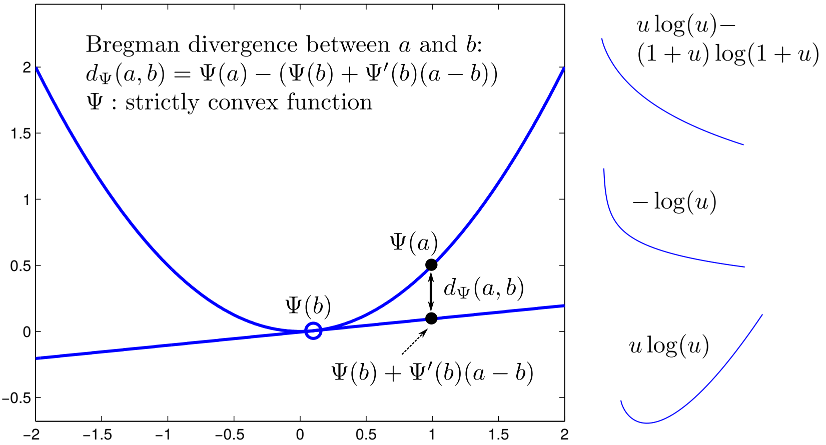
Properties

Bregman divergence to estimate unnormalized models

Framework

Noise-contrastive estimation as member of the framework

Bregman divergence between two vectors a and b



$$d_\Psi(a,b) = 0 \Leftrightarrow a = b \qquad d_\Psi(a,b) > 0 \text{ if } a \neq b$$

Bregman divergence between two functions f and g

- Compute $d_\Psi(f(\mathbf{u}), g(\mathbf{u}))$ for all $\mathbf{u}$ in their domain; take weighted average

$$\tilde{d}_\Psi(f, g) = \int d_\Psi(f(\mathbf{u}), g(\mathbf{u}))\mathrm{d}\mu(\mathbf{u}) \tag{13}$$

$$= \int \Psi(f) - \left[\Psi(g) + \Psi'(g)(f - g)\right]\mathrm{d}\mu \tag{14}$$

- Zero iff $f = g$ (a.e.); no normalization condition on f or g
- Fix f, omit terms not depending on g,

$$J(g) = \int \left[-\Psi(g) + \Psi'(g)g - \Psi'(g)f\right]\mathrm{d}\mu \tag{15}$$

Estimation of unnormalized models

$$J(g) = \int \left[-\Psi(g) + \Psi'(g)g - \Psi'(g)f \right] \mathrm{d}\mu$$

- Idea: Choose f, g, and μ so that we obtain a computable cost function for consistent estimation of unnormalized models.
- Choose $f = T(p_{\mathbf{x}})$ and $g = T(\phi)$ such that

$$f = g \Rightarrow p_{\mathbf{x}} = \phi \tag{16}$$

  Examples:
  - $f = p_{\mathbf{x}}$, $g = \phi$
  - $f = \frac{p_{\mathbf{x}}}{\nu p_{\mathbf{y}}}$, $g = \frac{\phi}{\nu p_{\mathbf{y}}}$
  - ...
- Choose μ such that the integral can either be computed in closed form or approximated as sample average.

(Gutmann and Hirayama, 2011)

Estimation of unnormalized models

(Gutmann and Hirayama, 2011)

- Several estimation methods for unnormalized models are part of the framework
  - Noise-contrastive estimation
  - Poisson-transform (Barthelmé and Chopin, 2015)
  - Score matching (Hyvärinen, 2005)
  - Pseudo-likelihood (Besag, 1975)
  - ...
- Noise-contrastive estimation:

$$\Psi(u) = u \log u - (1+u)\log(1+u) \quad (17)$$

$$f(\mathbf{u}) = \frac{\nu p_{\mathbf{y}}(\mathbf{u})}{p_{\mathbf{x}}(\mathbf{u})} \qquad \mathrm{d}\mu(\mathbf{u}) = p_{\mathbf{x}}(\mathbf{u})\mathrm{d}\mathbf{u} \quad (18)$$

proof

Conclusions

- Point estimation for parametric models with intractable partition functions (unnormalized models)
- Noise contrastive estimation
  - Estimate the model by learning to classify between data and noise
  - Consistent estimator, has MLE as limit
  - Applicable to large-scale problems
- Bregman divergence as general framework to estimate unnormalized models.

Appendix

Maximizer of the NCE objective function

Noise-contrastive estimation as member of the Bregman framework

Maximizer of the NCE objective function

Noise-contrastive estimation as member of the Bregman framework

Proof of Equation (4)

For large sample sizes n and m, $\hat{\boldsymbol{\theta}}$ satisfying

$$G(\mathbf{u}; \hat{\boldsymbol{\theta}}) = \frac{m}{n} \frac{p_{\mathbf{y}}(\mathbf{u})}{p_{\mathbf{x}}(\mathbf{u})}$$

is maximizing $J_n^{\mathrm{NCE}}(\boldsymbol{\theta})$,

$$J_n^{\mathrm{NCE}}(\boldsymbol{\theta}) = \frac{1}{n} \left(\sum_{i=1}^{n} \log P(C = 1 | \mathbf{x}_i; \boldsymbol{\theta}) + \sum_{i=1}^{m} \log \left[P(C = 0 | \mathbf{y}_i; \boldsymbol{\theta}) \right] \right)$$

without any normalization constraints.

Proof of Equation (4)

$$J_n^{\mathrm{NCE}}(\boldsymbol{\theta}) = \frac{1}{n}\left(\sum_{i=1}^{n} \log P(C=1|\mathbf{x}_i;\boldsymbol{\theta}) + \sum_{i=1}^{m} \log\left[P(C=0|\mathbf{y}_i;\boldsymbol{\theta})\right]\right)$$
$$= \frac{1}{n}\sum_{t=1}^{n} \log P(C=1|\mathbf{x}_i;\boldsymbol{\theta}) + \frac{m}{n}\frac{1}{m}\sum_{t=1}^{m} \log\left[P(C=0|\mathbf{y}_i;\boldsymbol{\theta})\right]$$

Fix the ratio $m/n = \nu$ and let $n \to \infty$ and $m \to \infty$. By law of large numbers, J_n^{NCE} converges to J^{NCE},

$$J^{\mathrm{NCE}}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x}}\left(\log P(C=1|\mathbf{x};\boldsymbol{\theta})\right) + \nu\mathbb{E}_{\mathbf{y}}\left(\log P(C=0|\mathbf{y};\boldsymbol{\theta})\right) \quad (19)$$

With $P(C=1|\mathbf{x};\boldsymbol{\theta}) = \frac{1}{1+G(\mathbf{x};\boldsymbol{\theta})}$ and $P(C=0|\mathbf{y};\boldsymbol{\theta}) = \frac{G(\mathbf{y};\boldsymbol{\theta})}{1+G(\mathbf{y};\boldsymbol{\theta})}$ we have

$$J^{\mathrm{NCE}}(\boldsymbol{\theta}) = -\mathbb{E}_{\mathbf{x}}\log(1+G(\mathbf{x};\boldsymbol{\theta})) + \nu\mathbb{E}_{\mathbf{y}}\log G(\mathbf{y};\boldsymbol{\theta}) - \nu\mathbb{E}_{\mathbf{y}}\log\left(1+G(\mathbf{y};\boldsymbol{\theta})\right) \quad (20)$$

Consider the objective $J^{\mathrm{NCE}}(\boldsymbol{\theta})$ as a function of $H = \log G$ rather than $\boldsymbol{\theta}$,

$$
\begin{aligned}
\mathcal{J}^{\mathrm{NCE}}(H) &= -\mathbb{E}_{\mathbf{x}} \log(1 + \exp H(\mathbf{x})) + \nu \mathbb{E}_{\mathbf{y}} H(\mathbf{y}) - \nu \mathbb{E}_{\mathbf{y}} \log\left(1 + \exp H(\mathbf{y})\right) \\
&= -\int p_{\mathbf{x}}(\boldsymbol{\xi}) \log(1 + \exp H(\boldsymbol{\xi})) d\boldsymbol{\xi} + \nu \int p_{\mathbf{y}}(\boldsymbol{\xi}) H(\boldsymbol{\xi}) d\boldsymbol{\xi} \\
&\quad - \nu \int p_{\mathbf{y}}(\boldsymbol{\xi}) \log(1 + \exp H(\boldsymbol{\xi})) d\boldsymbol{\xi} \\
&= -\int (p_{\mathbf{x}}(\boldsymbol{\xi}) + \nu p_{\mathbf{y}}(\boldsymbol{\xi})) \log(1 + \exp H(\boldsymbol{\xi})) d\boldsymbol{\xi} + \\
&\quad \nu \int p_{\mathbf{y}}(\boldsymbol{\xi}) H(\boldsymbol{\xi}) d\boldsymbol{\xi}
\end{aligned}
$$

We now expand $\mathcal{J}^{\mathrm{NCE}}(H + \epsilon q)$ around H for an arbitrary function q and a small scalar ϵ.

With

$$\log(1+\exp\left[H(\boldsymbol{\xi})+\epsilon q(\boldsymbol{\xi})\right]) = \log(1+\exp H(\boldsymbol{\xi})) + \frac{\epsilon q(\boldsymbol{\xi})}{1+\exp(-H(\boldsymbol{\xi}))} + \frac{\epsilon^2}{2}\frac{q(\boldsymbol{\xi})}{1+\exp(-H(\boldsymbol{\xi}))}\frac{q(\boldsymbol{\xi})}{1+\exp(H(\boldsymbol{\xi}))} + O(\epsilon^3)$$

we have

$$\begin{aligned}\mathcal{J}^{\mathrm{NCE}}(H+\epsilon q) = & -\int (p_{\mathbf{x}}(\boldsymbol{\xi})+\nu p_{\mathbf{y}}(\boldsymbol{\xi}))\log(1+\exp H(\boldsymbol{\xi}))d\boldsymbol{\xi} \\ & -\epsilon\int \frac{p_{\mathbf{x}}(\boldsymbol{\xi})+\nu p_{\mathbf{y}}(\boldsymbol{\xi})}{1+\exp(-H(\boldsymbol{\xi}))}q(\boldsymbol{\xi})d\boldsymbol{\xi} \\ & -\frac{\epsilon^2}{2}\int \frac{p_{\mathbf{x}}(\boldsymbol{\xi})+\nu p_{\mathbf{y}}(\boldsymbol{\xi})}{1+\exp(-H(\boldsymbol{\xi}))}\frac{q(\boldsymbol{\xi})^2}{1+\exp(H(\boldsymbol{\xi}))}d\boldsymbol{\xi} \\ & +\nu\int p_{\mathbf{y}}(\boldsymbol{\xi})H(\boldsymbol{\xi})d\boldsymbol{\xi} + \epsilon\nu\int p_{\mathbf{y}}(\boldsymbol{\xi})q(\boldsymbol{\xi})d\boldsymbol{\xi} + O(\epsilon^3)\end{aligned}$$

Collecting terms gives:

$$\mathcal{J}^{\rm NCE}(H+\epsilon q) = \mathcal{J}^{\rm NCE}(H) + \epsilon \int \left(\nu p_{\mathbf{y}}(\boldsymbol{\xi}) - \frac{p_{\mathbf{x}}(\boldsymbol{\xi}) + \nu p_{\mathbf{y}}(\boldsymbol{\xi})}{1+\exp(-H(\boldsymbol{\xi}))} \right) q(\boldsymbol{\xi}) d\boldsymbol{\xi}$$
$$- \frac{\epsilon^2}{2} \int \frac{p_{\mathbf{x}}(\boldsymbol{\xi}) + \nu p_{\mathbf{y}}(\boldsymbol{\xi})}{1+\exp(-H(\boldsymbol{\xi}))} \frac{q(\boldsymbol{\xi})^2}{1+\exp(H(\boldsymbol{\xi}))} d\boldsymbol{\xi} + O(\epsilon^3)$$

The second-order term is negative for all (non-trivial) q and H.
The first-order term is zero for all q if and only if

$$\nu p_{\mathbf{y}}(\boldsymbol{\xi}) = \frac{p_{\mathbf{x}}(\boldsymbol{\xi}) + \nu p_{\mathbf{y}}(\boldsymbol{\xi})}{1+\exp(-H^*(\boldsymbol{\xi}))}$$
$$\nu p_{\mathbf{y}}(\boldsymbol{\xi}) + \nu p_{\mathbf{y}}(\boldsymbol{\xi}) \exp(-H^*(\boldsymbol{\xi})) = p_{\mathbf{x}}(\boldsymbol{\xi}) + \nu p_{\mathbf{y}}(\boldsymbol{\xi})$$
$$\exp(-H^*(\boldsymbol{\xi})) = \frac{p_{\mathbf{x}}(\boldsymbol{\xi})}{\nu p_{\mathbf{y}}(\boldsymbol{\xi})}$$

which shows that $\hat{\boldsymbol{\theta}}$ such that $G(\boldsymbol{\xi};\hat{\boldsymbol{\theta}}) = \exp(H^*(\boldsymbol{\xi})) = \nu \frac{p_{\mathbf{y}}}{p_{\mathbf{x}}}$ is maximizing $J^{\rm NCE}(\boldsymbol{\theta})$.

back

Appendix

Maximizer of the NCE objective function

Noise-contrastive estimation as member of the Bregman framework

Proof

In noise-contrastive estimation, we maximize

$$J_n^{\mathrm{NCE}}(\boldsymbol{\theta}) = \frac{1}{n}\left(\sum_{i=1}^{n} \log P(C=1|\mathbf{x}_i;\boldsymbol{\theta}) + \sum_{i=1}^{m} \log\left[P(C=0|\mathbf{y}_i;\boldsymbol{\theta})\right]\right)$$

Sample version of

$$J^{\mathrm{NCE}}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x}}\left(\log P(C=1|\mathbf{x};\boldsymbol{\theta})\right) + \nu\mathbb{E}_{\mathbf{y}}\left(\log P(C=0|\mathbf{y};\boldsymbol{\theta})\right)$$

With

$$P(C=1|\mathbf{u};\boldsymbol{\theta}) = \frac{1}{1+G(\mathbf{u};\boldsymbol{\theta})} \quad P(C=0|\mathbf{u};\boldsymbol{\theta}) = \frac{1}{1+1/G(\mathbf{u};\boldsymbol{\theta})}$$

$$J^{\mathrm{NCE}}(\boldsymbol{\theta}) = -\mathbb{E}_{\mathbf{x}}\log(1+G(\mathbf{x};\boldsymbol{\theta})) - \nu\mathbb{E}_{\mathbf{y}}\log(1+1/G(\mathbf{y};\boldsymbol{\theta})) \quad (21)$$

where $G(\mathbf{u};\boldsymbol{\theta}) = \frac{\nu p_{\mathbf{y}}(\mathbf{u})}{\phi(\mathbf{u};\boldsymbol{\theta})}$.

The general cost function in the Bregman framework is

$$J(g) = \int \left[-\Psi(g) + \Psi'(g)g - \Psi'(g)f \right] \mathrm{d}\mu \tag{22}$$

With

$$\Psi(g) = g\log(g) - (1+g)\log(1+g) \tag{23}$$
$$\Psi'(g) = \log(g) - \log(1+g) \tag{24}$$

we have

$$\begin{aligned} J(g) = \int \Big[& -g\log(g) + (1+g)\log(1+g) \\ & + \log(g)g - \log(1+g)g \\ & - \log(g)f + \log(1+g)f \Big] \mathrm{d}\mu \end{aligned} \tag{25}$$

$$J(g) = \int \left[\log(1+g) - \log(g) f + \log(1+g) f \right] \mathrm{d}\mu \tag{26}$$

$$= \int \left[\log(1+g) + \log(1+1/g) f \right] \mathrm{d}\mu \tag{27}$$

With

$$f(\mathbf{u}) = \frac{\nu p_{\mathbf{y}}(\mathbf{u})}{p_{\mathbf{x}}(\mathbf{u})} \qquad g(\mathbf{u}) = G(\mathbf{u}; \boldsymbol{\theta}) \qquad \mathrm{d}\mu(\mathbf{u}) = p_{\mathbf{x}}(\mathbf{u}) \mathrm{d}\mathbf{u} \tag{28}$$

we have

$$J(G(.;\boldsymbol{\theta})) = \int p_{\mathbf{x}}(\mathbf{u}) \log(1 + G(\mathbf{u};\boldsymbol{\theta})) \mathrm{d}\mathbf{u}$$

$$+ \nu p_{\mathbf{y}}(\mathbf{u}) \log(1 + 1/G(\mathbf{u};\boldsymbol{\theta})) \mathrm{d}\mathbf{u} \tag{29}$$

$$= - J^{\mathrm{NCE}}(\boldsymbol{\theta}) \tag{30}$$

back