The dual of constrained KL-Divergence is the MLE of the log-linear model

Dingquan Wang

February 6, 2016

Given a distribution of interest p. We are looking for an estimation q that approaches p by minimizing the KL-Divergence with constraints:

$$q* = \arg\max_{q \in \mathcal{Q}} KL(q||p) = \arg_{q \in \mathcal{Q}} \max E_q[\log \frac{q(X)}{p(X)}] \tag{1}$$

$$\text{s.t.} \quad E_q[f(X)] - E_p[f(X)] \leq \xi; ||\xi||_\beta < \epsilon \tag{2}$$

$\mathcal{Q}$ is a distribution family. $f(X)$ is a measurement vector of X that we are interested. The general goal is while minimizing $q \in \mathcal{Q}$ approaching the true distribution we force some quantities agree with some observation in expectation. Consider the Lagrangian:

$$\max_{\lambda \geq 0, \alpha \geq 0} \min_{q(X), \xi} L(q(X), \epsilon, \lambda, \alpha, \gamma) \tag{3}$$

where:

$$L(q(X), \epsilon, \lambda, \alpha, \gamma) = KL(q||p) + \lambda \cdot (E_q[f(X)] - E_p[f(X)] - \xi) \tag{4}$$

$$+ \alpha \cdot (||\xi||_\beta - \epsilon) + \gamma \cdot (\int_X q(X) - 1) \tag{5}$$

In order to compute the dual of this Lagrangian, we first represent:

$$\alpha||\xi||_\beta = \max \xi \cdot \eta \;\; \text{s.t.} ||\eta||_\beta \leq \alpha \tag{6}$$

This results in a variational Lagrangian:

$$\max_{\lambda \geq 0, \alpha \geq 0} \max_{||\eta||_\beta \leq \alpha} \min_{q(X), \xi} L(q(X), \epsilon, \lambda, \alpha, \gamma) \tag{7}$$

with $L(q(X), \epsilon, \lambda, \alpha, \gamma)$ defined as:

$$L(q(X), \epsilon, \lambda, \alpha, \gamma) = E_q[\log \frac{q(X)}{p(X)}] + \lambda \cdot (E_q[f_(X)] - E_p[f_(X)] - \xi) \tag{8}$$

$$+ \xi\eta - \alpha\epsilon + \gamma \cdot (\int_X q(X) - 1) \tag{9}$$

$$\frac{\partial L(q(X), \epsilon, \lambda, \alpha, \gamma)}{\partial q(X)} = \log q(X) - \log p(X) + 1 + \lambda \cdot f(X) + \gamma = 0 \tag{10}$$

$$\tag{11}$$

$$\frac{\partial L(q(X), \epsilon, \lambda, \alpha, \gamma)}{\partial \xi_i} = \eta_i - \lambda_i \rightarrow \eta = \lambda \tag{12}$$

$$\tag{13}$$

Plugging $q(Y)$, $\eta = \lambda$ in $L(q(X), \epsilon, \lambda, \alpha, \gamma)$ and taking the derivative with respect to γ

$$\frac{\partial L(\lambda, \alpha, \gamma)}{\partial \gamma} = \int_X \frac{p(X) \exp(-\lambda \cdot f(X))}{e \exp(\gamma)} - 1 = 0 \tag{14}$$

$$\rightarrow \gamma = \log(\frac{\int_X p(X) \exp(-\lambda \cdot f(X))}{e}) \tag{15}$$

plug γ into (10)

$$\log q(X) = \log p(X) - 1 - \lambda \cdot f(X) - \log(\frac{\int_X p(X) \exp(-\lambda \cdot f(X))}{e}) \tag{16}$$

$$q(X) = \exp(\log p(X)) \exp(-1) \exp(-\lambda \cdot f(X)) \exp(-\log(\frac{\int_X p(X) \exp(-\lambda \cdot f(X))}{e})) \tag{17}$$

$$= \frac{p(X) \exp(-\lambda \cdot f(X))}{\int_X p(X) \exp(-\lambda \cdot f(X))} \tag{18}$$

$$= \frac{p(X) \exp(-\lambda \cdot f(X))}{Z_\lambda} \tag{19}$$

where $Z_\lambda = \int_X p(X) \exp(-\lambda \cdot f(X))$. Plugging γ and $q(X)$ into $L(q(X), \epsilon, \lambda, \alpha, \gamma)$.

$$L(\lambda, \alpha) = E_q[\log \frac{q(X)}{p(X)}] + \lambda \cdot (E_q[f_(X)] - E_p[f_(X)] - \xi) \tag{20}$$

$$+ \xi\lambda - \alpha\epsilon + \gamma \cdot (\int_X q(X) - 1) \tag{21}$$

$$L(\lambda, \alpha) = E_q[\log \frac{q(X)}{p(X)}] + \lambda \cdot (E_q[f_(X)] - E_p[f_(X)]) - \alpha\epsilon \tag{22}$$

$$= \int_X q(X) \log \frac{q(X)}{p(X)} + \lambda \cdot E_q[f_(X)] - \lambda \cdot (E_p[f_(X)] - \alpha\epsilon \tag{23}$$

$$= \int_X \frac{p(X) \exp(-\lambda \cdot f(X))}{Z_\lambda} \log \frac{\exp(-\lambda \cdot f(X))}{Z_\lambda} + \lambda \cdot E_q[f_(X)] - \lambda \cdot (E_p[f_(X)] - \alpha\epsilon \tag{24}$$

$$= \int_X \frac{p(X) \exp(-\lambda \cdot f(X))}{Z_\lambda} \cdot -\lambda \cdot f(X) - log Z_\lambda \int_X \frac{p(X) \exp(-\lambda \cdot f(X))}{Z_\lambda} \tag{25}$$

$$+ \lambda \cdot E_q[f_(X)] - \lambda \cdot (E_p[f_(X)] - \alpha\epsilon \tag{26}$$

$$= -\lambda \cdot E_q[f_(X)] - log Z_\lambda + \lambda \cdot E_q[f_(X)] - \lambda \cdot (E_p[f_(X)] - \alpha\epsilon \tag{27}$$

$$= -\log Z_\lambda - \lambda \cdot E_p[f_(X)] - \alpha\epsilon \tag{28}$$

The new objective is:

$$\max_{\lambda \geq 0, \alpha \geq 0} L(\lambda, \alpha) = -\log(Z_\lambda) - \lambda \cdot E_p[f_(X)] - \alpha\epsilon \text{ s.t. } ||\lambda||_\beta^* \leq \alpha \tag{29}$$

We can analytically see that the optimum of this objective with respect to α is $\alpha = ||\lambda||_{\beta}^{*}$ and placing this in $L(\lambda, \alpha)$ we get the dual objective:

$$\max_{\lambda \geq 0} L(\lambda) = -\log(Z_{\lambda}) - \lambda \cdot E_{p}[f_{(}X)] - \epsilon ||\lambda||_{\beta}^{*} \tag{30}$$

This is the same objective as the MLE of the log-linear model. as desired.