

Notes on K-L divergence and MaxEnt learning

Robert Staubs

`rstaubs@linguist.umass.edu`

October 13, 2014

1 What is this?

Solving with numerical optimizers is greatly aided if we have an explicit, known gradient. MaxEnt offers us (relatively) easy answers in likelihood maximization and K-L divergence minimization.

Here I include some notes on how these gradients are derived, as well as comments on their interpretation and implementation. These notes serve both as a tool for a successor to work such as HGR, as well as a codification of things I have in scattered notes.

Among those contents are notes on the calculation of Hessians (second derivatives) for MaxEnt. These are of potential use in optimization, but have not before been a part of HGR. Hessian calculations will be added to this document when I can typeset them.

Please let me know if you have comments, questions, corrections, clarifications, etc.

2 Definitions

Let X be the set of inputs, with members x .

Let Y_x be the sets of outputs, with members y . (I will abbreviate these.)

Let $z \subset y$ denote the hidden structures z compatible with the output y . Z_x is the set of hidden structures available in the tableau for x .

Let $w_{(i)}$ indicate the i th weight, $v_{(i)}$ the i th element of a violation vector (etc.)

N_x is the MaxEnt normalization for a tableau with input x .

p and q are the predicted MaxEnt distribution and the empirical distribution, respectively.

3 Recurring gradients

$$\frac{\partial}{\partial w_{(i)}} N_x = \frac{\partial}{\partial w_{(i)}} \sum_{y' \in Y_x} \sum_{z' \subset y'} e^{w^T v_{xz'}} \quad \text{def.} \quad (1)$$

$$= \sum_{y' \in Y_x} \sum_{z' \subset y'} v_{xz'(i)} e^{w^T v_{xz'}} \quad \text{chain rule} \quad (2)$$

$$= \sum_{z \in Z_x} v_{xz(i)} e^{w^T v_{xz}} \quad (3)$$

$$\frac{\partial}{\partial w_{(i)}} p(y|x) = \frac{\partial}{\partial w_{(i)}} \sum_{z \subset y} p(y, z|x) \quad \text{hidden struc. def.} \quad (4)$$

$$= \frac{\partial}{\partial w_{(i)}} \sum_{z \subset y} \frac{e^{w^T v_{xz}}}{N_x} \quad \text{MaxEnt def.} \quad (5)$$

$$= \sum_{z \subset y} \frac{(v_{xz(i)} e^{w^T v_{xz}})(N_x) - (e^{w^T v_{xz}})(\frac{\partial N_x}{\partial w_{(i)}})}{N_x^2} \quad \text{quotient rule} \quad (6)$$

$$= \sum_{z \subset y} \frac{(v_{xz(i)} e^{w^T v_{xz}})(N_x) - (e^{w^T v_{xz}})(\sum_{z' \in Z_x} v_{xz'(i)} e^{w^T v_{xz'}})}{N_x^2} \quad \text{see above} \quad (7)$$

$$= \sum_{z \subset y} p(y, z|x) \left(v_{xz(i)} - \sum_{z' \in Z_x} p(y, z'|x) v_{xz'(i)} \right) \quad \text{MaxEnt defs.} \quad (8)$$

$$= \sum_{z \subset y} p(y, z|x) (v_{xz(i)} - E[v_{x(i)}]) \quad \text{def. exp.} \quad (9)$$

$$= \sum_{z \subset y} (p(y, z|x) v_{xz(i)}) - p(y|x) E[v_{x(i)}] \quad (10)$$

$$\frac{\partial}{\partial w_{(i)}} \log p(y|x) = \frac{1}{p(y|x)} \frac{\partial}{\partial w_{(i)}} p(y|x) \quad \text{chain rule} \quad (11)$$

$$= \frac{1}{p(y|x)} \left(\sum_{z \subset y} (p(y, z|x) v_{xz(i)}) - p(y|x) E[v_{x(i)}] \right) \quad \text{above} \quad (12)$$

$$= \left(\sum_{z \subset y} \frac{p(y, z|x)}{p(y|x)} v_{xz(i)} \right) - E[v_{x(i)}] \quad (13)$$

$$= \left(\sum_{z \subset y} p(z|x, y) v_{xz(i)} \right) - E[v_{x(i)}] \quad \text{def. cond. prob.} \quad (14)$$

$$= E[v_{x(i)}|y] - E[v_{x(i)}] \quad \text{def. cond. exp.} \quad (15)$$

4 Gradients for Kullback-Leibler divergences

K-L divergence is not symmetric: $D(p||q) \neq D(q||p)$, in general. We have been using $D(q||p)$ up til now in HGR. This is fairly typical, using the divergence which places the true values on the left.

Computing the gradient is largely a matter of plugging in what we have from above:

$$D(q||p) = \sum_{x \in X} \sum_{y \in Y_x} q(y|x) \log \frac{q(y|x)}{p(y|x)} \quad \text{def.} \quad (16)$$

$$\frac{\partial D}{\partial w_{(i)}} = \frac{\partial}{\partial w_{(i)}} \sum_{x \in X} \sum_{y \in Y_x} q(y|x) (\log q(y|x) - \log p(y|x)) \quad \text{log properties} \quad (17)$$

$$= - \sum_{x \in X} \sum_{y \in Y_x} q(y|x) \left(\frac{\partial}{\partial w_{(i)}} \log p(y|x) \right) \quad q \text{ constant w.r.t. } w \quad (18)$$

$$= - \sum_{x \in X} \sum_{y \in Y_x} q(y|x) [E[v_{x(i)}|y] - E[v_{x(i)}]] \quad \text{see above} \quad (19)$$

$E[v_{x(i)}]$ is the expected amount of violation of the i th constraint, under the predicted distribution. Computing it therefore involves computing the distribution over *full* structures for a tableau and weighting the violations. These are then summed. This is one-liner if done in matrix math, as it probably should be.

$E[v_{x(i)}|y]$ is a similar expectation, but taken only over a certain output. To compute this, the

distribution over full structures compatible with a given output is computed and used to weight violations. The one-liner is similar here, but it has to be embedded in some logic that subdivides the data into sub-tableaux for each output.

q is the empirical distribution, and therefore involves no novel calculation.

In the maximum likelihood case, there is only a single winner in each tableau. The K-L gradient thus reduces to the following, where y_x^* is the target output for the input x .

$$\frac{\partial D}{\partial w_{(i)}} = - \sum_{x \in X} \sum_{y \in Y_x} q(y|x) [E[v_{x(i)}|y] - E[v_{x(i)}]] \quad \text{above} \quad (20)$$

$$= - \sum_{x \in X} \sum_{y=y_x^*} [E[v_{x(i)}|y] - E[v_{x(i)}]] \quad \text{only one winner} \quad (21)$$

$$(22)$$

When there is no hidden structure, it is instead the conditional expectation that simplifies:

$$\frac{\partial D}{\partial w_{(i)}} = - \sum_{x \in X} \sum_{y \in Y_x} q(y|x) [E[v_{x(i)}|y] - E[v_{x(i)}]] \quad \text{above} \quad (23)$$

$$= - \sum_{x \in X} \sum_{y \in Y_x} [v_{xy(i)} - E[v_{x(i)}]] \quad \text{one full structure per output} \quad (24)$$

These combine trivially in the case where there is a single, fully specified target output for every input:

$$\frac{\partial D}{\partial w_{(i)}} = - \sum_{x \in X} \sum_{y=y_x^*} [v_{xy(i)} - E[v_{x(i)}]] \quad \text{one full structure, one winner} \quad (25)$$

This is all that is needed to implement K-L as found in HGR. It might be that someone would want the other direction on K-L. It is here:

$$D(p||q) = \sum_{x \in X} \sum_{y \in Y_x} p(y|x) \log \frac{p(y|x)}{q(y|x)} \quad \text{def.} \quad (26)$$

$$\frac{\partial D}{\partial w_{(i)}} = \frac{\partial}{\partial w_{(i)}} \sum_{x \in X} \sum_{y \in Y_x} p(y|x) (\log p(y|x) - \log q(y|x)) \quad \text{logs} \quad (27)$$

$$= \sum_{x \in X} \sum_{y \in Y_x} \left(\frac{\partial}{\partial w_{(i)}} p(y|x) \right) (\log p(y|x) - \log q(y|x)) \quad (28)$$

$$+ p(y|x) \left(\frac{\partial}{\partial w_{(i)}} (\log p(y|x) - \log q(y|x)) \right) \quad \text{prod}$$

$$= \sum_{x \in X} \sum_{y \in Y_x} \left(\sum_{z \subset y} (p(y, z|x) v_{xz(i)}) - p(y|x) E[v_{x(i)}] \right) (\log p(y|x) - \log q(y|x)) \quad (29)$$

$$+ p(y|x) (E[v_{x(i)}|y] - E[v_{x(i)}]) \quad \text{see above}$$

$$= \sum_{x \in X} \sum_{y \in Y_x} p(y|x) (E[v_{x(i)}|y] - E[v_{x(i)}]) (\log p(y|x) - \log q(y|x)) \quad (30)$$

$$+ p(y|x) (E[v_{x(i)}|y] - E[v_{x(i)}]) \quad \text{cond exp}$$

$$= \sum_{x \in X} \sum_{y \in Y_x} p(y|x) (E[v_{x(i)}|y] - E[v_{x(i)}]) (\log p(y|x) - \log q(y|x) + 1)$$

The core expectation comparison is the same as before, but it is somewhat obscured. Note that within this is the p - q divergence— p multiplied by the log difference between p and q . A form reflecting this seems to more obfuscate than clarify, however.

N.B. I have not numerically checked the final form here for hidden structure, though I have checked it for overt structure. I advise asking me or checking the result numerically if you implement this.