An introduction to Variational calculus in Machine Learning

Anders Meng

February 2004

1 Introduction

The intention of this note is not to give a full understanding of calculus of variations since this area are simply to big, however the note is meant as an appetizer. Classical variational methods concerns the field of finding the extremum of an integral depending on an unknown function and its derivatives. Methods as the finite element method, used widely in many software packages for solving partial differential equations is using a variational approach as well as e.g. maximum entropy estimation [6].

Another intuitively example which is derived in many textbooks on calculus of variations; consider you have a line integral in euclidian space between two points a and b. To minimize the line integral (functional) with respect to the functions describing the path, one finds that a linear function minimizes the line-integral. This is of no surprise, since we are working in an euclidian space, however, if the integral is not as easy to interpret, calculus of variations comes in handy in the more general case.

This little note will mainly concentrate on a specific example, namely the *Variational EM algorithm for incomplete data*.

2 A practical example of calculus of variations

To determine the shortest distance between to given points, A and B positioned in a two dimensional Euclidean space (see figure (1) page (2)), calculus of variations will be applied as to determine the functional form of the solution. The problem can be illustrated as minimizing the line-integral given by

$$I = \int_A^B 1 ds(x, y). \tag{1}$$

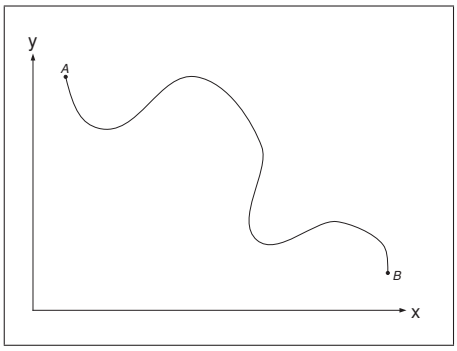

Figure 1: *A line between the points A and B*

The above integral can be rewritten by a simple (and a bit heuristic) observation.

$$\begin{aligned} ds &= (dx^2 + dy^2)^{\frac{1}{2}} \\ &= dx\left[1 + \left(\frac{dy}{dx}\right)^2\right]^{\frac{1}{2}} \\ \frac{ds}{dx} &= \left[1 + \left(\frac{dy}{dx}\right)^2\right]^{\frac{1}{2}}. \end{aligned} \tag{2}$$

From this observation the line integral can be written as

$$I(x, y, y') = \int_A^B \left[1 + \left(\frac{dy}{dx}\right)^2\right]^{\frac{1}{2}} dx. \tag{3}$$

The line integral given in equation (3) is also known as a functional, since it depends on x, y, y', where $y' = \frac{dy}{dx}$. Before performing the minimization, two important rules in variational calculus needs attention

Rules of the game

- *Functional derivative*

$$\frac{\delta f(x)}{\delta f(x')} = \delta(x - x') \tag{4}$$

where

$$\delta(x) = \begin{cases} 1 & \text{for x} = 0 \\ 0 & \text{otherwise} \end{cases}$$

- *Communitative*

$$\frac{\delta}{\delta f(x')}\frac{\partial f(x)}{\partial x} = \frac{\partial}{\partial x}\frac{\delta f(x)}{\delta f(x')} \tag{5}$$

- The *Chain Rule* known from function differentiation applies.

The function $f(x)$ should be a smooth function, which usually is the case.

It is now possible to determine the type of functions which minimizes the functional given in equation (3) page (2)

$$\frac{\delta I(x, y, y')}{\delta y(x_p)} = \frac{\delta}{\delta y(x_p)} \int_A^B \left[1 + \left(\frac{dy}{dx}\right)^2\right]^{\frac{1}{2}} dx. \tag{6}$$

where we now define $y_p \equiv y(x_p)$ to simplify the notation. Using "the rules of the game", it is possible to differentiate the expression inside the integral to give

$$\begin{aligned} \frac{\delta I(x, y, y')}{\delta y(x_p)} &= \int_A^B \frac{1}{2}(1 + y'^2)^{-\frac{1}{2}} 2y' \frac{\delta}{\delta y_p}\frac{dy}{dx} dx \\ &= \int_A^B \frac{y'}{(1 + y'^2)^{\frac{1}{2}}}\frac{d}{dx}\frac{\delta y(x)}{\delta y(x_p)} dx \\ &= \int_A^B \underbrace{\frac{y'}{(1 + y'^2)^{\frac{1}{2}}}}_{g}\underbrace{\frac{d}{dx}\delta(x - x_p)}_{f} dx. \end{aligned} \tag{7}$$

The chain rule in integration :

$$\int f(x)g(x)dx = F(x)g(x) - \int F(x)g'(x)dx \tag{8}$$

can be applied to the expression above, which results in the following equations

$$\frac{\delta I(x,y,y')}{\delta y(x_p)} = \left[\delta(x-x_p)\frac{y'}{(1+y'^2)^{\frac{1}{2}}}\right]_A^B - \int_A^B \delta(x-x_p)\frac{y''}{(1+y'^2)^{\frac{3}{2}}}dx. \tag{9}$$

The last step serves an explanation. Remember that

$$g(x) = \frac{y'}{(1+y'^2)^{\frac{1}{2}}}. \tag{10}$$

The derivative of this w.r.t. x calculates to (try it out)

$$g'(x) = \frac{y''}{(1+y'^2)^{\frac{3}{2}}} \tag{11}$$

Now assuming that on the boundaries, A and B, the first part of the expression given in equation (9) disappears, so $\delta(A-x_p)=0$ and $\delta(B-x_p)=0$, then the expression simplifies to (assuming that $\int_A^B \delta(x-x_p)dx = 1$)

$$\frac{\delta I(x,y,y')}{\delta y(x_p)} = -\frac{y_p''}{(1+y_p'^2)^{\frac{3}{2}}} \tag{12}$$

Setting this expression to zero and determine the function form of $y(x_p)$:

$$\frac{y_p''}{(1+y_p'^2)^{\frac{3}{2}}} = 0. \tag{13}$$

A solution to the differential equation $y_p'' = 0$ is $y_p(x) = ax + b$; the family of straight lines, which intuitively makes sense. However in cases where the space is not Euclidean, the functions which minimized the integral expression or functional may not be that easy to determine.

In the above case, to guarantee that the found solution is a minimum the functional have to be differentiated once more with respect to $y_p(x)$ to make sure that the found solution is a minimum.

The above example was partly inspired from [1], where other more delicate examples might be found, however the derivation procedure in [1] is not the same as shown here.

The next section will introduce variational calculus in machine learning.

3 Calculus of variations in Machine Learning

The practical example which will be investigated is the problem of lower bounding the marginal likelihood using a variational approach. Dempster et al. [4] proposed the EM-algorithm for this purpose, but in this note a variational EM - algorithm is derived in accordance with [5]. Let $\mathbf{y}$ denote the observed variables, $\mathbf{x}$ denote the latent variables and $\boldsymbol{\theta}$ denote the parameters. The log - likelihood of a data-set $\mathbf{y}$ can be lower bounded by introducing any distribution over the latent variables and the input parameters as long as this distribution have support where $p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}, m)$ does, and using the trick of Jensen's inequality.

$$\begin{aligned} \ln p(\mathbf{y}|m) = \ln \int p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}|m) d\mathbf{x} d\boldsymbol{\theta} &= \ln \int q(\mathbf{x}, \boldsymbol{\theta}) \frac{p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}|m)}{q(\mathbf{x}, \boldsymbol{\theta})} d\mathbf{x} d\boldsymbol{\theta} \\ &\geq \int q(\mathbf{x}, \boldsymbol{\theta}) \ln \frac{p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}|m)}{q(\mathbf{x}, \boldsymbol{\theta})} d\mathbf{x} d\boldsymbol{\theta} \end{aligned} \tag{14}$$

The expression can be rewritten using Bayes rule to show that maximization of the above expression w.r.t. $q(\mathbf{x}, \boldsymbol{\theta})$ actually corresponds to minimizing the Kullback-Leibler divergence between $q(\mathbf{x}, \boldsymbol{\theta})$ and $p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}, m)$, see e.g. [5].

If we maximize the expression w.r.t. $q(\mathbf{x}, \boldsymbol{\theta})$ we would find that $q(\mathbf{x}, \boldsymbol{\theta}) = p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}, m)$, which is the true posterior distribution[1]. This however is not the way to go since we would still have to determine a normalizing constant, namely the marginal likelihood. Another way is to factorize the probability into $q(\mathbf{x}, \boldsymbol{\theta}) \approx q_x(\mathbf{x}) q_\theta(\boldsymbol{\theta})$, in which equation (14) can be written as

$$\ln p(\mathbf{y}|m) \geq \int q_x(\mathbf{x}) q_\theta(\boldsymbol{\theta}) \ln \frac{p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}|m)}{q_x(\mathbf{x}) q_\theta(\boldsymbol{\theta})} d\mathbf{x} d\boldsymbol{\theta} = \mathbf{F}_m(q_x(\mathbf{x}), q_\theta(\boldsymbol{\theta}), \mathbf{y}) \tag{15}$$

which further can be split into parts depending on $q_x(\mathbf{x})$ and $q_\theta(\boldsymbol{\theta})$. See below

$$\begin{aligned} \mathbf{F}_m(q_x(\mathbf{x}), q_\theta(\boldsymbol{\theta}), \mathbf{y}) &= \int q_x(\mathbf{x}) q_\theta(\boldsymbol{\theta}) \ln \frac{p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}|m)}{q_x(\mathbf{x}) q_\theta(\boldsymbol{\theta})} d\mathbf{x} d\boldsymbol{\theta} \\ &= \int q_x(\mathbf{x}) q_\theta(\boldsymbol{\theta}) \left(\ln \frac{p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta}, m)}{q_x(\mathbf{x})} + \ln \frac{p(\boldsymbol{\theta}|m)}{q_\theta(\boldsymbol{\theta})} \right) d\mathbf{x} d\boldsymbol{\theta} \\ &= \int q_x(\mathbf{x}) q_\theta(\boldsymbol{\theta}) \ln \frac{p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta}, m)}{q_x(\mathbf{x})} d\mathbf{x} d\boldsymbol{\theta} + \int q_\theta(\boldsymbol{\theta}) \ln \frac{p(\boldsymbol{\theta}|m)}{q_\theta(\boldsymbol{\theta})} d\boldsymbol{\theta} \end{aligned}$$

where $\mathbf{F}_m(q_x(\mathbf{x}), q_\theta(\boldsymbol{\theta}), \mathbf{y})$ is our functional. The big problem, is to determine the optimal form of the distributions $q_x(\mathbf{x})$ and $q_\theta(\boldsymbol{\theta})$. In order to do this, we calculate the derivative of the functional with respect to the two "free" functions / distributions, and determine iterative updates of the distributions. It will now be shown how the update-formulas are calculated using calculus of variation. The results of the calculations can be found in [5].

[1] You can see this by inserting $q(\mathbf{x}, \boldsymbol{\theta}) = p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}, m)$ into eq. (14), which turns the inequality into a equality

The short hand notation : $q_x \equiv q_{\mathbf{x}}(\mathbf{x})$ and $q_\theta \equiv q_\theta(\boldsymbol{\theta})$ is used in the following derivation. Another thing that should be stressed is the following relation;

Assume we have the functional $G(f_x(x), f_y(y))$, then the following relation applies :

$$\begin{aligned} G(f_x(x), f_y(y)) &= \int\int f_x(x) f_y(y) \ln p(x,y) dx dy \\ \frac{\delta G(f_x(x), f_y(y))}{\delta f_x(x')} &= \int\int \delta(x - x') f_y(y) ln(p(x,y)) dx dy \\ &= \int f_y(y) ln(p(x', y)) dy. \end{aligned} \tag{16}$$

This will be used in the following when performing the differentiation with respect to $q_x(\mathbf{x})$. Also a Lagrange-multiplier have been added, as to ensure that we find a probability distribution.

$$\begin{aligned} F_m(q_x, q_\theta, \mathbf{y}) &= \int q_x q_\theta \ln \frac{p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta}, m)}{q_x} d\mathbf{x} d\boldsymbol{\theta} + \int q_\theta \ln \frac{p(\boldsymbol{\theta}|m)}{q_\theta} d\boldsymbol{\theta} + \lambda \left(1 - \int q_x(\mathbf{x}) d\mathbf{x}\right) \\ \frac{\delta F_m(\cdot)}{\delta q_x} &= \int \left[q_\theta \ln \frac{p(\mathbf{y}, \mathbf{x}'|\boldsymbol{\theta}, m)}{q_{x'}} - q_{x'} q_\theta \frac{q_{x'}}{p(\mathbf{y}, \mathbf{x}'|\boldsymbol{\theta}, m)} q_{x'}^{-2} p(\mathbf{y}, \mathbf{x}'|\boldsymbol{\theta}, m) \right] d\boldsymbol{\theta} - \lambda \\ &= \int \left[q_\theta \ln p(\mathbf{y}, \mathbf{x}'|\boldsymbol{\theta}, m) - q_\theta \ln q_{x'} - q_\theta \right] d\boldsymbol{\theta} - \lambda \\ &= \int q_\theta \ln p(\mathbf{y}, \mathbf{x}'|\boldsymbol{\theta}, m) d\boldsymbol{\theta} - \ln q_{x'} - 1 - \lambda = 0 \end{aligned} \tag{17}$$

Isolating with respect to $q_{x'}(\mathbf{x}')$ gives the following relation.

$$\begin{aligned} q_{x'}(\mathbf{x}') &= \exp\left\{ \int q_\theta(\boldsymbol{\theta}) \ \ln p(\mathbf{y}, \mathbf{x}'|\boldsymbol{\theta}, m) d\boldsymbol{\theta} - Z_x \right\} \\ &= \exp\left\{ < \ln p(\mathbf{y}, \mathbf{x}'|\boldsymbol{\theta}, m) >_{\boldsymbol{\theta}} - Z_x \right\} \end{aligned} \tag{18}$$

where Z_x is a normalization constant. The above method can be used again, now just determining the "derivative" w.r.t. $q_\theta(\boldsymbol{\theta})$. The calculation of the derivative is quite similar however, below, one can see how it is calculated :

$$\begin{aligned} F_m(q_x, q_\theta, \mathbf{y}) &= \int q_x q_\theta \ln \frac{p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta}, m)}{q_x} d\mathbf{x} d\boldsymbol{\theta} + \int q_\theta \ln \frac{p(\boldsymbol{\theta}|m)}{q_\theta} d\boldsymbol{\theta} + \lambda \left(1 - \int q_\theta(\boldsymbol{\theta}) d\boldsymbol{\theta}\right) \\ \frac{\delta F_m(\cdot)}{\delta q_{\theta'}} &= \int q_x \ln \frac{p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta'}, m)}{q_x} d\mathbf{x} + \ln \frac{p(\boldsymbol{\theta'}|m)}{q_{\theta'}} - 1 - \lambda \\ &= \int q_x \ln p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta'}, m) d\mathbf{x} + \ln \frac{p(\boldsymbol{\theta'}|m)}{q_{\theta'}} - C - \lambda \\ &= \int q_x \ln p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta'}, m) d\mathbf{x} + \ln p(\boldsymbol{\theta'}|m) - \ln q_{\theta'} - C - \lambda = 0 \end{aligned} \tag{19}$$

Isolating with respect to $q_{\theta'}(\boldsymbol{\theta}')$ gives the following relation.

$$\begin{aligned} q_{\theta'}(\boldsymbol{\theta}') &= p(\boldsymbol{\theta}'|m) \exp\left\{ \int q_x \ln p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta}', m) d\mathbf{x} - Z_\theta \right\} \\ &= p(\boldsymbol{\theta}'|m) \exp\left\{ < \ln p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta}', m) >_{\mathbf{x}} - Z_\theta \right\} \end{aligned} \tag{20}$$

Where Z_θ is a normalization constant. The two equations we have just determined is coupled equations. One way to solve these equations is to iterate these until convergence (fixed point iteration). So denoting a iteration number as superscript t the equations to iterate look like

$$q_x^{(t+1)}(\mathbf{x}) \propto \exp\left[\int \ln(p(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta}, m) q_\theta^{(t)}(\boldsymbol{\theta}) d\boldsymbol{\theta} \right] \tag{21}$$

$$q_\theta^{(t+1)}(\boldsymbol{\theta}) \propto p(\boldsymbol{\theta}|m) \exp\left[\int \ln(p(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta}, m) q_x^{(t+1)}(\mathbf{x}) d\mathbf{x} \right] \tag{22}$$

where the normalization constants have been avoided in accordance with [5]. The lower bound is increased at each iteration.

In the above derivations the principles of calculation of variations what used. Due to the form of the update-formulas there is a special family of models which comes in handy, these are the so called : *Conjugate-Exponential Models*, but will not be further discussed. However the Phd-thesis by Matthew Beal, goes into much more details concerning the CE-models [3].

References

[1] Charles Fox, *An introduction to the Calculus of Variations*. Dover Publications, Inc. 1987.

[2] C. Bishop, *Neural Networks for Pattern Recognition*. Oxford: Clarendon Press, 1995.

[3] M.J. Beal, *Variational Algorithms for Approximate Bayesian Inference*, PhD. Thesis, Gatsby Computational Neuroscience Unit, University College ,London, 2003. (281 pages).

[4] A.P. Dempster, N.M. Laird and D.B. Rubin, *Maximum Likelihood from Incomplete Data via the EM algorithm*. Journal of the Royal Statistical Society Series B, vol. 39, no. 1, pp. 1–38, Nov. 1977.

[5] Matthew J. Beal and Zoubin Ghahramani, "The Variational Bayesian EM Algorithm for Imcomplete data: With applications to Scoring graphical Model Structure", *In Bayesian Statistics 7, Oxford University Press,* 2003

[6] T. Jaakkola, "Tutorial on variational approximation methods", *In Advanced mean field methods: theory and practice. MIT Press,* 2000, citeseer.nj.nec.com/jaakkola00tutorial.html.