On One Variational Problem of Information Theory

Roman V. Belavkin[1]*

School of Engineering and Information Sciences
Middlesex University, London NW4 4BT, UK

Abstract. This short article discusses an important variational problem that first appeared in statistical physics and information theory. First, some necessary background material for solving the conditional extrema problems is outlined. Then, a simple example is used to introduce the problem and show its solution. The importance of the problem is illustrated by numerous application of its solution in various fields.

1 Unconditional Extremum

Many optimisation problems are formulated as problems of maximisation of some real function $f : Y \to \mathbb{R}$ (objective function, utility). A solution, if exists, is an element $\bar{y} \in Y$ such that $f(y) \leq f(\bar{y})$ for all y. If y are elements of a linear space Y and f is differentiable on the neighbourhood of $\bar{y}$, then it is necessary that f has zero derivative $f'(\bar{y}) = 0$ or gradient $\nabla f(\bar{y}) = 0$ at $\bar{y}$, if Y is a multidimensional linear space. Geometrically, this condition simply states that f has constant tangent hyperplane at $\bar{y}$ (i.e. a constant function $c(y) = f(\bar{y})$). The sufficient condition should guarantee that the graph of f lays below this hyperplane. In particular, if f is a closed, concave function, then the condition $\nabla f(\bar{y}) = 0$ is both necessary and sufficient for f to attain its absolute maximum value at $\bar{y}$.

Remark 1. If Y is a finite-dimensional linear space, then the gradient of $f(y) = f(y_1, \dots, y_m)$ is a vector of partial derivatives:

$$\nabla f(y) = \Big(\frac{\partial f(y)}{\partial y_1}, \dots, \frac{\partial f(y)}{\partial y_m}\Big)$$

The gradient defines a linear function $\langle \nabla f(y), \cdot \rangle : Y \to \mathbb{R}$ on Y as follows:

$$\langle \nabla f(y), z \rangle := \frac{\partial f(y)}{\partial y_1} z_1 + \cdots + \frac{\partial f(y)}{\partial y_m} z_m = \sum_{i=1}^{m} \frac{\partial f(y)}{\partial y_i} z_i$$

If f is weakly differentiable, then this function coincides with the *directional derivative* $f'(y, z)$, a derivative of function f at point y in the direction of z:

$$f'(y, z) = \lim_{t \downarrow 0} \frac{f(y + tz) - f(y)}{t}$$

* This work was supported in part by EPSRC grant EP/H031936/1.

The derivative $f'(y,z) = \langle \nabla f(y), z \rangle$ gives the best linear approximation of f at some point z:

$$f(z) \approx f(y) + \langle \nabla f(y), z - y \rangle$$

The right hand side of the above expression defines the tangent hyperplane of f at point y. This tangent hyperplane is constant if and only if $\nabla f(y) = 0$, that is $f'(y,z) = 0$ in all directions z. These facts are also true when Y is infinite-dimensional. In this case, $\nabla f(y)$ can be called a Gâteaux differential or a variation of functional f.

Remark 2. If f is not differentiable at $\bar{y}$, then the concepts of a gradient ∇f and tangent hyperplane are replaced by the concepts of subgradients (or supgradients) ∂f and the sets of supporting hyperplanes. Subdifferential calculus is well-developed and widely used in non-smooth analysis. In particular, the fact that f achieves its extreme value at point $\bar{y}$ implies that the set of supporting hyperplanes of f at $\bar{y}$ includes constant hyperplane (the only difference from the smooth case is that there can be other, non-constant supporting hyperplanes). This translates into condition that the subdifferential (or supdifferential) set (the set of all sub- or supgradients) includes zero vector:

$$\partial f(\bar{y}) \ni 0$$

For more information, see for example [1, 2].

Remark 3. If f is not concave, but a proper upper semicontinuous function (i.e. $f(y) < \infty$ and the set $\{y : \lambda \leq f(y)\}$ is closed for each λ), then the existence of $\bar{y}$ maximising f on some closed and bounded subset of Y is still guaranteed. The solution $\bar{y}$ is then given by the necessary and sufficient condition $\nabla f^{**}(\bar{y}) = 0$, where f^{**} is concave closure of f (i.e. biconjugate in the concave sense).

2 Conditional Exremum and Method of Lagrange Multipliers

Often, the optimisation problems are formulated with additional conditions (constraints) expressed by equalities $g_i(y) = \lambda_i$ or inequalities $g_i(y) \leq \lambda_i$ using other functions $g_i : Y \to \mathbb{R}$, $i \in [1, \ldots, m]$. In set comprehension notation this problem is written as:

$$\bar{f}(\lambda) := \sup\{f(y) : g_i(y) \leq \lambda_i\}$$

This is the problem of a conditional extremum, because the constraints define the subset $C := \{y : g_i(y) \leq \lambda_i\} \subseteq Y$ of *feasible* solutions, and therefore the optimal solution $\bar{y} \in C$ of the above problem with constraints is generally 'worse' than that of the unconstrained problem. It is clear, however, that solutions of unconstrained problems are not very useful if they are unfeasible, and therefore taking the constraints into account is absolutely crucial in problem formulation.

Solutions to the conditional extremum problem are often found using the method of Lagrange multipliers. First, one introduces a suitable Lagrange function $K : Y \times \mathbb{R}^m \to \mathbb{R}$, where $\mathbb{R}^m$ is a real linear space for the Lagrange multipliers $(\alpha_1, \ldots, \alpha_m) \in \mathbb{R}^m$. The number of the multipliers (and the number m of dimensions of $\mathbb{R}^m$) equals to the number of constraints $g_i(y) \leq \lambda_i$. The Lagrange function is chosen so that $K(y, 0) = f(y)$, such as the following linear combination of functions f and g_i:

$$K(y, \alpha) = f(y) + \sum_{i=1}^{m} \alpha_i [\lambda_i - g_i(y)]$$

Here α_i are the Lagrange multipliers related to the constraints λ_i, and the value $\alpha = (\alpha_1, \ldots, \alpha_m) = 0$ corresponds to the unconstrained optimisation. The necessary condition of conditional extremum is $\nabla K(\bar{y}, \alpha) = 0$, and if the Lagrange function is concave in each argument, then this condition is also sufficient (e.g. this is the case when f is concave, g_i are convex and $\alpha \geq 0$).

Let us consider the case of one constraint $g(y) \leq \lambda$ and one Lagrange multiplier. The condition $\nabla K(\bar{y}, \alpha) = 0$ corresponds to the following partial conditions:

$$\begin{aligned} \nabla_y K(\bar{y}, \alpha) &= \nabla f(\bar{y}) - \alpha \nabla g(\bar{y}) = 0 \\ \nabla_\alpha K(\bar{y}, \alpha) &= \lambda - g(\bar{y}) = 0 \end{aligned}$$

Here, $\nabla_y K(y, \alpha)$ denotes partial gradient of $K(y, \alpha)$ in subspace Y of $Y \times \mathbb{R}$, and $\nabla_\alpha K(y, \alpha)$ is just partial derivative of $K(y, \alpha)$ over $\alpha \in \mathbb{R}$. The first condition states that the optimal solutions $\bar{y}$ are such that the gradients of functions f and g are proportional $\nabla f(\bar{y}) = \alpha \nabla g(\bar{y})$. In fact, the solutions are a one parameter family $\bar{y} = \bar{y}(\alpha)$, and the parameter is related to the constraint $\alpha = \alpha(\lambda)$. The latter is obtained by inverting the second condition $g(\bar{y}(\alpha)) = \lambda$. Another useful observation is that α is the derivative of function $\bar{f}(\lambda) := \sup\{f(y) : g(y) \leq \lambda\}$. Indeed, $\bar{f}(\lambda) = K(\bar{y}, \alpha) = f(\bar{y}) + \alpha[\lambda - g(\bar{y})]$, and therefore:

$$\bar{f}'(\lambda) = \frac{\partial(f(\bar{y}) + \alpha[\lambda - g(\bar{y})])}{\partial \lambda} = \alpha$$

Because $\bar{f}(\lambda)$ is a non-decreasing function, it follows that $\alpha \geq 0$.

3 Fundamental Variational Problem of Information Theory and Statistical Physics

Having equipped ourselves with the background for solving conditional extremum problems, let us now consider the following example. First, we shall see it simply as an illustration of the method of Lagrange multipliers. Then we shall discuss the importance of this example.

Example 1 (Exponential family). Let f be a linear function of $y \in \mathbb{R}^m$:

$$f(y) = \sum_{i=1}^{m} x_i y_i \,,$$

where $x \in \mathbb{R}^m$ is a fixed vector. Thus, f is both concave and convex. Let g be defined as follows

$$g(y) = \begin{cases} \sum_{i=1}^{m} (\ln y_i - 1) y_i \,, & \text{if } y > 0 \\ 0\,, & \text{if } y = 0 \\ +\infty\,, & \text{if } y < 0 \end{cases}$$

This function is convex. Thus, the corresponding Lagrange function is concave

$$K(y, \alpha) = \sum_{i=1}^{m} x_i y_i + \alpha[\lambda - \sum_{i=1}^{m} (\ln y_i - 1) y_i]$$

and conditions $\nabla f(\bar{y}) = \alpha \nabla g(\bar{y})$, $g(\bar{y}) = \lambda$, $\alpha \geq 0$ are both necessary and sufficient for the conditional extremum. The gradients of functions f and g are:

$$\nabla f(y) = x\,, \qquad \nabla g(y) = \ln y$$

Therefore, $x = \alpha \ln \bar{y}$, and the solutions are defined by the following relations:

$$\bar{y} = e^{x/\alpha}\,, \qquad \sum_{i=1}^{m} (\alpha^{-1} x_i - 1) e^{x_i/\alpha} = \lambda$$

The first relation defines the optimal vector $\bar{y}$ as a function of fixed vector x and Lagrange multiplier α, which is determined by x and constraint λ using the second relation.

Example 1 is more than a simple illustration of optimisation problem with constraints — it plays a very important role in probability theory, thermodynamics, statistical mechanics and information theory. To see this, let us normalise the optimal vector $\bar{y}$:

$$\bar{p}(x_i) = \frac{\bar{y}_i}{\sum_{i=1}^{m} y_i} = \frac{e^{x_i/\alpha}}{\sum_{i=1}^{m} e^{x_i/\alpha}} \tag{1}$$

The function above is the Boltzmann (or Gibbs) distribution — a member of exponential family of probability distributions. In fact, the one parameter exponential family can be obtained as a solution of optimisation problem $\sup\{f(y) : g(y) \leq \lambda\}$ in Example 1 with slight modification of function $g(y)$ and including the normalisation condition $\sum y_i = 1$.

The objective function $f(p) = \sum x_i p_i$, evaluated at normalised positive vector p, corresponds to the expected value $\mathbb{E}_p\{x\}$ of random variable x. Function $g(p) = \sum p_i \ln p_i - \sum p_i = -H(p) - 1$, where $H(p) = -\sum p_i \ln p_i$ is the entropy of distribution p. Thus, the problem in Example 1 can be considered as the problem of maximisation of expected value $\mathbb{E}_p\{x\} = \sum x_i p_i$ over all probability distributions p satisfying the constraint on entropy $H(p) \geq -\lambda - 1$.

The problem of maximisation of the expected value of random variable x arises in numerous applications. Thus, in economics x may represent a utility or profit, in game theory x is a payoff, and in cybernetics x is a reward function. Equivalently, in estimation problems $-x$ may represent cost function to be minimised. In physics, $-f(p) = -\sum x_i p_i$ represents internal energy to be minimised.

Entropy is an important measure of uncertainty, and the maximum entropy principle in thermodynamics or statistical mechanics can be equivalently represented as minimisation of internal energy $-f(p)$ (i.e. maximisation of $f(p)$) with constraints on entropy $H(p) \geq -\lambda - 1$ (i.e. $g(p) \leq \lambda$). The fact that Boltzmann distribution (1) is the solution of this problem is well-known in physics. In thermodynamics, the Lagrange multiplier α corresponds to temperature, and the expressions are often written using its inverse $\beta = \alpha^{-1}$.

Negative entropy is closely related to information. Thus, the constraint $g(p) = -H(p) - 1 \leq \lambda$ can be interpreted as a particular kind of information constraint. Variational problems with information constraints were used in information theory to determine the maximum channel capacity [3]. A similar problem was used to study the utility of Shannon information [4]. The solutions of all these problems can be explained as slight variations of Example 1.

These problems have deep relation to combinatorial optimisation, machine learning, neural networks and cognitive science. Indeed, the Boltzmann distribution (1) with variable temperature α is used in the simulated annealing [5]. In machine learning, it is used to control the trade-off between exploration and exploitation [6]. In artificial neural networks, this distribution arises in Boltzmann machines [7]. In cognitive science, it is used in some cognitive architectures for conflict resolution that can simulate the ‘soft-max’ properties of human choice strategies [8].

This relation between optimisation problems with information constraints and learning and adaptive systems has been investigated recently in the area of information dynamics [9, 10]. A theory has been developed in quite a general setting. However, many results can be understood by using the simple problem of Example 1 as an illustration, and for this reason it plays fundamental role in the theory.

References

1. Rockafellar, R.T.: Conjugate Duality and Optimization. Volume 16 of CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics, PA (1974)
2. Tikhomirov, V.M.: Convex Analysis. In: Analysis II. Volume 14 of Encyclopedia of Mathematical Sciences. Springer-Verlag (1990) 1–92
3. Shannon, C.E.: A mathematical theory of communication. Bell System Technical Journal **27** (July and October 1948) 379–423 and 623–656

4. Stratonovich, R.L.: On value of information. Izvestiya of USSR Academy of Sciences, Technical Cybernetics **5** (1965) 3–12 In Russian.
5. Kirkpatrick, S., Gelatt, C.D., Vecchi, J.M.P.: Optimization by simulated annealing. Science **220**(4598) (May 1983) 671–680
6. Kaelbling, L.P.: Learning in Embedded Systems. The MIT Press, Cambridge, MA (1993)
7. Hinton, G.E., Sejnowski, T.J., Ackley, D.H.: Boltzmann machines: Constraint satisfaction networks that learn. Technical Report 119, Carnegie–Mellon University (1984)
8. Anderson, J.R.: The adaptive character of thought. Lawrence Erlbaum, Hillsdale, NJ (1990)
9. Belavkin, R.V.: Bounds of optimal learning. In: 2009 IEEE International Symposium on Adaptive Dynamic Programming and Reinforcement Learning, Nashville, TN, USA, IEEE (2009) 199–204
10. Belavkin, R.V.: Information trajectory of optimal learning. In Hirsch, M.J., Pardalos, P.M., Murphey, R., eds.: Dynamics of Information Systems: Theory and Applications. Volume 40 of Springer Optimization and Its Applications Series. Springer (2010)