On determination of function extrema with MaxEnt formulation

Ravi Sankar Saripalli

September 16, 2019

Description

This is an attempt to find extrema of a function with MaxEnt formulation. The usual way to find extrema of a function $f(x)$ is to find x where function derivative $f'(x)$ vanishes. An alternate approach is to use probability distribution function $p(x)$ as an independant parameteric function and seek to maximize the expected value of $f(x)$ based on the $p(x)$ with the condition that entropy (as defined by Shanon) of the probability distribution $p(x)$ is maximized, thus ensuring that the distribution function derived is least biased.

The MaxEnt Lagrangian

$$\mathcal{L}(p,\lambda) = \int f(x)p(x)dx + T\int p(x)ln(p(x))dx + \lambda\left\{\left(\int p(x)dx\right) - 1\right\} \tag{1}$$

The first term in the Lagrangian is the expected value of the function corresponding to the PDF $p(x)$, the second term corresponds to the entropy of the PDF scaled by an arbitray constant T and the last term corresponds to the equality constraint that integral of PDF is one. The multiplier λ is the Lagrangian parameter.
Noting that the Lagrangian is function of $p(x)$ and λ, the extrema of the Lagrangian is obtained by requiring the functional derivative with respect $p(x)$ and the Lagrangian parameter λ are zero.
Although this enables us to determine $p(x)$ and λ corresponding to the extrema, we need additional criterion to determine if the extrema is either minimum or maximum. Sign of the second derivative in functional space similar to simple variables can be used to ascertain if the stationary point corresponds to minima or maxima.

The Euler-Lagrange equation

Consider the following functional (function of functions)

$$F\left(f', f, x\right) = \int \phi\left(f'(x), f(x), x\right) dx \tag{2}$$

Requiring that functional derivative of F with respect to f is zero at extrema (note f is a function not a variable, hence the name functional derivative) following Eauler-Lagrange equation can be derived.

$$\frac{\partial \phi}{\partial y} - \frac{d}{dx}\left(\frac{\partial \phi}{\partial y'}\right) = 0 \tag{3}$$

Derivation of the above equation that is easy to follow is here

Finding Stationary Point of Lagrangian in Eq 1.

Using the above Euler-Lagrange equation, the stationary conditions for the Lagrangian in equation 1, can be derived from functional derivative with respective p and the derivative with respective λ as follows.

$$f(x) + T\{1 + ln(p(x))\} + \lambda = 0 \tag{4}$$

Setting derivative with respect to λ to zero yields the contraint that PDF integral should be 1.

$$\int p(x)\, dx - 1 = 0 \tag{5}$$

Rearranging eqn.4

$$p(x) = e^{-(1+\lambda/T)}\, e^{-f(x)/T} \tag{6}$$

Combining eqns. 5 and 6 the value of lambda can be expressed as follows

$$\lambda = T\left\{ln \int e^{-f(x)/T} dx - 1\right\} \tag{7}$$

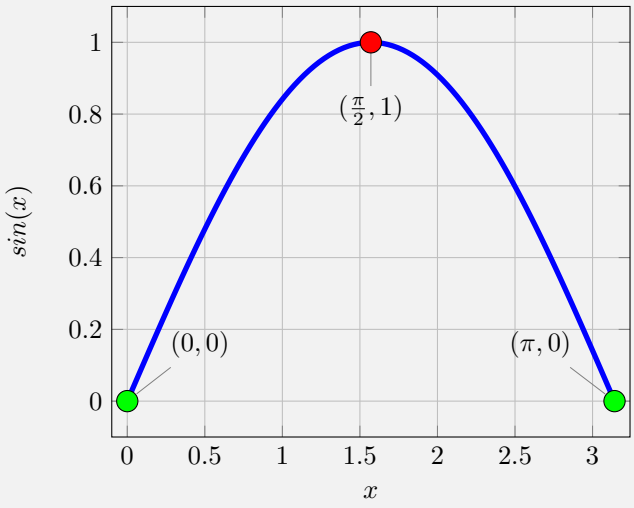
The question to ask is how do we now proceed to get min or max of $f(x)$ given that we have it in analytical form (eg. sin(x)).. What will be the p(x) as T approaches zero. That is the intent of the annealing process as we converge on p(x).

Explorations with sinx

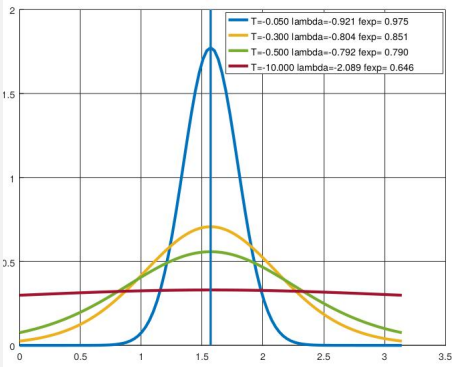
Assuming that we are interested in finding extrema of sin(x) in the interval $[0, \pi]$, one can proceed as follows.

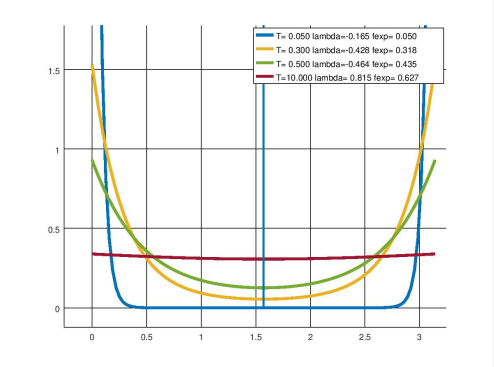
1. For a given T, calculate λ by evaluating the integral numerically
2. With $T and \lambda$ known, the PDF $p(x)$ now is well defined (eq.6). The expected value of $f(x)$ can be determined by numerical integration of $\int p(x) f(x) dx$

$sin(x)$



Change in pdf with T when $T \leq 0$ and $T \geq 0$

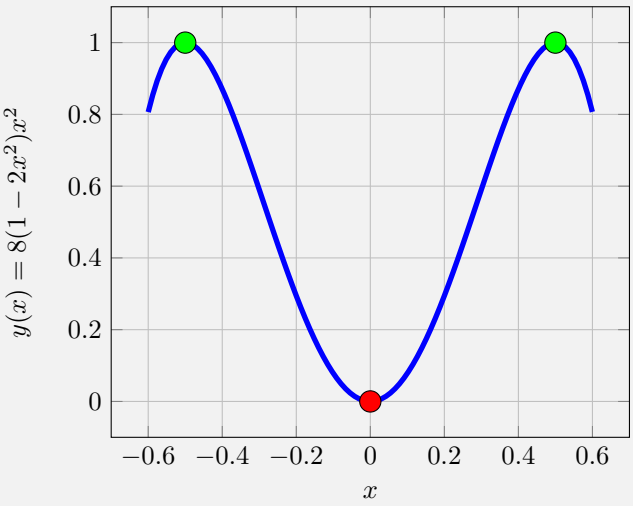




From the above figures, it is clear that for $T \geq 0$ as its magnitude approaches zero, the pdf spread decreases and tends to peak around the true maxima. On the otherhand, when $T \geq 0$, the pdf peaks appear near the end points where the function has minimal value in the interval of interest.

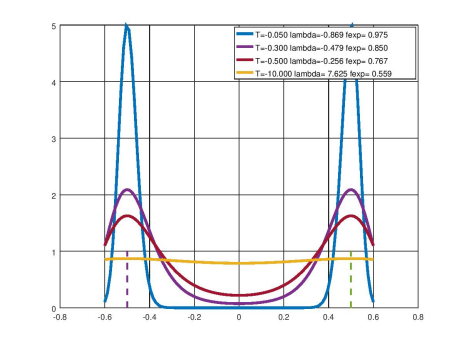
Let us now explore a different function that has two peaks, and trough as shown in the following figure.

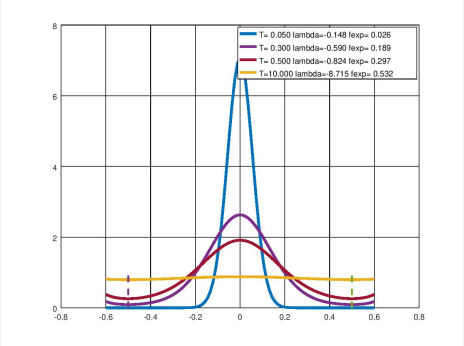
$y(x) = 8(1-2x^2)x^2 - 0.6 < x < 0.6$



Using the above function, the effect of T on PDF corresponding to the stationary point is determined as described earlier.

Effect of T on PDF at stationary condition





Once again, when T value is approaching zero from positive side, the PDF corresponding to extrema condition peaks around the true minima at $x = 0$ with corresponding expected value of function approaching 0. When T approaches zero from the negative side, the PDF peaks around $x = -0.5, and 0.5$ with expected value of function approaching 1.

It therefore appears that when T approaches zero from negative side PDF determined from stationary

condition corresponds to maxima of the function, while it approaches zero from positive side, the PDF peaks around the minima of the function.

Functional Derivatives

In this section the derivation of functional derivatives is provided. While there is no shortage of articles on this topic, thesis work by Peter J. Oliver from University of Minnesota. provides lucid explanation. Let $\eta(x)$ be an arbitrary function which vanishes to zero at the boundaries, and ϵ is an arbitrarily small constant. For the functional F defined in equaiton 2, we denote its gradient with respect to $f(x)$ with ∇F_f, and $\eta(x)$ is the variation of function $f(x)$, then

$$\begin{aligned} &\langle \nabla F_f \, , \eta \rangle = \frac{d}{d\epsilon} F\left(f + \epsilon\eta\right)\Big|_{\epsilon=0} \\ &\text{where} \\ &\langle \nabla F_f, \eta \rangle = \int \nabla F_f \eta(x) dx \end{aligned} \tag{8}$$

$$\begin{aligned} \langle \nabla F_f \, , \eta \rangle &= \frac{d}{d\epsilon} \int \phi\left(fe, f' + \epsilon\eta', x\right) dx \Big|_{\epsilon=0} \\ &= \int \left\{ \eta \frac{\partial \phi\left(fe, f' + \epsilon\eta', x\right)}{\partial\left(f + \epsilon\eta\right)} + \eta' \frac{\partial \phi\left(fe, f' + \epsilon\eta', x\right)}{\partial\left(f' + \epsilon\eta'\right)} \right\} dx \Big|_{\epsilon=0} \\ &= \int \left\{ \eta \frac{\partial \phi\left(f, f', x\right)}{\partial f} + \eta' \frac{\partial \phi\left(f, f', x\right)}{\partial f'} \right\} dx \\ &= \int \eta \frac{\partial \phi\left(f, f', x\right)}{\partial f} dx + \left[\eta \frac{\partial \phi\left(f, f', x\right)}{\partial f} \right]_a^{\overset{0}{b}} - \int \eta \frac{d}{dx} \left\{ \frac{\partial \phi\left(f, f', x\right)}{\partial f'} \right\} dx \\ &= \int \eta \left[\frac{\partial \phi\left(f, f', x\right)}{\partial f} - \frac{d}{dx} \left\{ \frac{\partial \phi\left(f, f', x\right)}{\partial f'} \right\} \right] dx \end{aligned}$$

Since $\eta(x)$ is arbitray function

$$\nabla F_f = \frac{\partial \phi\left(f, f', x\right)}{\partial f} - \frac{d}{dx} \left\{ \frac{\partial \phi\left(f, f', x\right)}{\partial f'} \right\} \tag{9}$$

Similarly the second variation of f can be defined and evaluated as follows. It should be noted that unlike, the first variation, it is not possible to eliminate $\eta\prime$ through integration by parts.

$$\begin{aligned} Q(f, \eta) &= \frac{d^2}{d\epsilon^2} F\left(f + \epsilon\eta\right)\Big|_{\epsilon=0} \\ &= \int \left\{ \eta^2 \frac{\partial^2 \phi}{\partial f^2} + \left(\eta'\right)^2 \frac{\partial^2 \phi}{\partial f'^2} + 2\eta\eta' \frac{\partial^2 \phi}{\partial f \partial f'} \right\} dx \end{aligned} \tag{10}$$

Similar to normal functions, the sign of the second variation can be used to ascertain if the stationary point of the functional corresponds to either maxima or minima. If $Q(f, \eta)$ is positive definite at the stationary point then it is minima.
One way $Q(f, \eta)$ can be positive definite is when the integrand itself is positive definite over the entire integration interval. While that is a valid possibility, it is too prescriptive.

Necessary and sufficient conditions for second functional positivity

Expressing $Q(f,\eta)$ as follows

$$\begin{aligned} Q(f,\eta) &= \int \left\{ A\eta^2 + 2B\eta\eta' + C\left(\eta'\right)^2 \right\} dx \\ &\quad where\ A = \frac{\partial^2\phi}{\partial f^2}\ ;\ B = \frac{\partial^2\phi}{\partial f \partial f'}\ ;\ C = \frac{\partial^2\phi}{\partial f'^2} \\ &= \int \left\{ A\eta^2 + B\frac{d}{dx}\left(\eta^2\right) + C\left(\eta'\right)^2 \right\} dx \\ &= \int \left\{ A\eta^2 - B'\eta^2 + C\left(\eta'\right)^2 \right\} dx \end{aligned}$$

Nb: The last simplification is achieved with integration by parts and noting η vanishes on boundaries The necessary and sufficient conditions for $Q(f,\eta)$ to be positive requires fairly involved mathematics and is nicely covered in text book (page 100-110) on Variational Calculus The two necessary conditions for second variant to be positive definite are as follows.

Condition 1:

$$C > 0$$

Condition 2:

The following linear differential has
no other solution but a trivial solution $v = 0$

$$-\frac{d}{dx}\left(C\eta'\right) + \left(A - B'\right)\eta = 0 \tag{11}$$

The role of sign of scaling parameter T in equation (1) explained

Referring back to our original Lagrangian in equation 1 the second variation can be calculated as follows.

$$\begin{aligned} Q(p,\eta) &= \lim_{\epsilon\to 0}\frac{d^2}{d\epsilon^2}\left\{ \begin{array}{c} \int f(x)\left\{p(x)+\epsilon\eta(x)\right\}dx + T\int \left\{p(x)+\epsilon\eta(x)\right\} ln\left\{p(x)+\epsilon\eta(x)\right\}dx \\ + \int \left\{p(x)+\epsilon\eta(x)\right\}dx \end{array} \right\} \\ &= \lim_{\epsilon\to 0}\frac{d}{d\epsilon}\left\{ \int f(x)\eta(x)dx + T\int \eta(x)\left[ln\left\{p(x)+\epsilon\eta(x)\right\}+1\right]dx + \int \eta(x)dx \right\} \\ &= T * \int \frac{\eta^2(x)}{p(x)}dx \end{aligned}$$

From the above derivation it is clear that sign of the second variation of the Lagrangian of our interest (Eq. 1) is same as that of T because $p(x)$ and η^2 are always positive. This explains our observed trends (with few sample functions) where stationary points turned out to be minima when T is positive, and maxima when T is negative. Since the above derivation is valid for any generic function $f(x)$ we can use this result to seek either maxima or minima for the MaxEnt problem by appropriate choice of sign for T.

Distributed MonteCarlo Tree Search Algorithm

$$\begin{aligned}\mathcal{L}\left(q_i, \lambda_i \quad i = 1..N\right) = & \int G(x_1, x_2, ...x_N) \prod_{i=1}^{N} q_i(x_i) \quad dx_1 dx_2 ..dx_N \\ & + T \sum_{i=1}^{N} \int q_i(x_i) ln(q_i(x_i)) dx_i \\ & + \sum_{i=1}^{N} \lambda_i \left\{ \left(\int q(x_i) dx_i \right) - 1 \right\}\end{aligned} \tag{12}$$

The above Lagrangian is formed to enable minimise the global objective function G which is a function of all the variables $x_1, x_2 ...x_N$. The probability distribution function (PDF) of each agent is represneted by $q_i(x_i)$. It is assumed that the approximate collective PDF can be represented as product of distribution functions of each agent. This assumption enables decentralization of each agents actions while allowing optimization of the global objective function G. The stationary conditions for Lagrangian in eqn. 12 with respect of $q_k(x_k)$ and λ_k can be written as follows.

$$\left.\begin{aligned}\nabla_{q_k} \mathcal{L} &= \int G(x_1, x_2, ...x_N) \prod_{i \neq k} q_i(x_i) \quad \prod_{i \neq k} dx_i + T \left\{1 + ln(q_k(x_k))\right\} + \lambda_k \\ &= E(G|x = x_k) + T \left\{1 + ln(q_k(x_k))\right\} + \lambda_k = 0 \\ \nabla_{\lambda_k} \mathcal{L} &= \left\{ \left(\int q(x_k) dx_k \right) - 1 \right\} = 0\end{aligned}\right\}_{k=1..N} \tag{13}$$

The Lagrangian parameter λ_k can be elemenated from the above two conditions and it is trivial show that

$$q_k(x_k) = \frac{e^{-E(G|x=x_k)/T}}{\int e^{-E(G|x=x_k)/T} dx_k} \tag{14}$$

Newton method to minimize Lagrangian based on collective probability $p(x)$

In this seciton the iteration scheme to find the stationary point with respect to $p(x)$ to the Lagrangian in equation 1 is derived based on Newton method. The Newton iteration template for any arbitray function $f(x)$ is $x_{new} = x_0 - \frac{\nabla f(x_0)}{\nabla^2 f(x_0)}$. Although this example is based on normal functions, it appears that one can use this strategy in functional space ???.
The first and second variations of the Lagrangian of interest have already been derived. Althogh these are integral expressions, choosing the arbitray function η to be a delta function $\delta(x = x')$ the point wise first and second derivatives of the functional can be realized. (This statement needs some review ... by mathematician for rigour)

$$
\begin{aligned}
\nabla_p \mathcal{L}|_{p=p^0} &= f(x) + T\left\{1 + ln(p^0(x))\right\} + \lambda \\
\nabla_p^2 \mathcal{L}|_{p=p^0} &= \frac{T}{p^0(x)} \\
p^{new}(x) &= p^0(x) - \frac{f(x) + T\left\{1 + ln(p^0(x))\right\} + \lambda}{\frac{T}{p^0(x)}} \\
p^{new}(x) &= -p^0(x)\left\{\frac{f(x)}{T} + ln(p^0(x)) + \frac{\lambda}{T}\right\}
\end{aligned}
$$

Integrating over the domain and noting that $\int p^{new}(x)dx = 1; \int p^0(x)dx = 1$

$$
1 = -\frac{1}{T}\int p^0(x)f(x)dx - \int p^0(x)ln(p^0(x))dx - \frac{\lambda}{T} \tag{15}
$$

The λ value can now be expressed as follows

$$
\lambda = -T - \int p^0(x)f(x)dx - T\int p^0(x)ln(p^0(x))dx = -T - E(f, p^0) + TS(p^0)
$$

Finally the p_{new} can now be written as

$$
\begin{aligned}
p^{new}(x) &= -p^0(x)\left\{\frac{f(x)}{T} + ln(p^0(x)) - 1 - \frac{1}{T}E(f, p^0) + S(p^0)\right\} \\
\frac{p^{new}(x)}{p^0(x)} &= 1 - S(p^0) - ln(p^0(x)) - \frac{1}{T}\left\{f(x) - E(f, p^0)\right\}
\end{aligned}
$$

Armed with the above template to seek extrema for a Lagrangian based on collective probability distribution, it is relatively straightforward to apply this template to the Lagrangian with collective probability replaced by product distributions as shown in eqn. 12.

Entropy Term in DMCTS

It is worth noting that the entropy term in the Lagrangian (eqn. 12) is strictly not based the collective probability. The exact entropy of the collective probability distribution should have been as follows.

$$\begin{aligned} & T \int \left\{ \prod_{i=1}^{N} q_i(x_i) \right\} ln \left\{ \prod_{i=1}^{N} q_i(x_i) \right\} \quad dx_1 dx_2 .. dx_N \\ & = T \sum_{k=1}^{N} \int \left\{ \prod_{i=1}^{N} q_i(x_i) \right\} ln(q_k(x_k)) \quad dx_1 dx_2 .. dx_N \end{aligned} \tag{16}$$

Replacing the collective probability ($\prod_{i=1}^{N} q_i(x_i)$) in the integral with individual probability associated with each term in the summation, the collective entropy term becomes

$$= T \sum_{k=1}^{N} \int q_k(x_k) ln(q_k(x_k)) dx_k \tag{17}$$

This virtually amounts to minimising cumulative cross entropy (often referred to as Kulbec entropy differential) of all distributions relative to the collective. Another way to interpret this is to say that instead of minimizing true entropy of the collective distribution, we minimize the sum of individual entropies. Since sum of any arbitray isolated systems entropies is always less than entropy of the combined system without isolation, the maxEnt search is biased to some extent. It appears that rationale for introduction of this restriction is to eliminate cross term derivatives in Jacobian during Newton steps.