7-1-2023

Why Bump Reward Function Works Well In Training Insulin Delivery Systems

Lehel Dénes-Fazakas
Óbuda University, denes-fazakas.lehel@uni-obud.hu

Lásló Szilágyi
Sapientia University, lalo@ms.sapientia.ro

Gyorgy Eigner
eigner.gyorgy@uni-obuda.hu

Olga Kosheleva
The University of Texas at El Paso, olgak@utep.edu

Vladik Kreinovich
The University of Texas at El Paso, vladik@utep.edu

See next page for additional authors

Comments:

Technical Report: UTEP-CS-23-40

Why Bump Reward Function Works Well In Training Insulin Delivery Systems

Lehel Dénes-Fazakas, László Szilágyi, Gyorgy Eigner, Olga Kosheleva, Vladik Kreinovich, and Nguyen Hoang Phuong

Abstract Diabetes is a disease when the body can no longer properly regulate blood glucose level, which can lead to life-threatening situations. To avoid such situations and regulate blood glucose level, patients with severe form of diabetes need insulin injections. Ideally, the system should automatically decide when best to inject insulin and how much to inject. To find the optimal control, researchers applied

Lehel Dénes-Fazakas
Physiological Controls Research Center and
Applied Informatics and Applied Mathematics Doctoral School
Óbuda University, Bécsi Street 96/B, Budapest H-1034, Hungary
e-mail: denes-fazakas.lehel@uni-obuda.hu

László Szilágyi
Physiological Controls Research Center
Óbuda University, Bécsi Street 96/B, Budapest H-1034, Hungary
and
Computational Intelligence Research Group
Sapientia University, Tg. Mureş, Romania, e-mail: lalo@ms.sapientia.ro

Gyorgy Eigner
Physiological Controls Research Center and
Biomatics and Applied Artificial Intelligence Institute
John von Neumann Faculty of Informatics
Óbuda University, Bécsi Street 96/B, Budapest H-1034, Hungary
e-mail: eigner.gyorgy@uni-obuda.hu

Olga Kosheleva
Department of Teacher Education, University of Texas at El Paso, 500 W. University
El Paso, Texas 79968, USA, e-mail: olgak@utep.edu

Vladik Kreinovich
Department of Computer Science, University of Texas at El Paso, 500 W. University
El Paso, Texas 79968, USA, e-mail: vladik@utep.edu

Nguyen Hoang Phuong
Artificial Intelligence Division, Information Technology Faculty, Thang Long University
Nghiem Xuan Yem Road, Hoang Mai District, Hanoi, Vietnam
e-mail: nhphuong2008@gmail.com

machine learning with different reward functions. It turns out that the most effective learning occurred when they used the so-called bump function. In this paper, we provide a possible explanation for this empirical result.

1 Formulation of the problem

What is diabetes. All living creatures need energy to function. To many cells in a human body, energy comes from glucose that is delivered to these cells by the blood flow. The absorption of glucose into the cells is regulated by a special hormone called *insulin*.

When the body does not produce enough insulin – the illness know as *diabetes* – it hinders the ability of cells to get energy and can lead to life-threatening situations.

Need for insulin injections. To avoid dangerous situations, a natural idea is to inject insulin into the body when the insulin level becomes dangerously low – and we can detect that, since in this case the cells do not absorb the glucose and thus, the blood glucose level becomes too high.

It is desirable to have automatic insulin injections. In healthy patients, the body itself decides how much insulin is needed. In the absence of such an automatic biologic regulation, at present, the patients themselves decide when to inject insulin and how much to inject, based on some general recommendations.

The effectiveness of these general recommendations is different for different patients. It is therefore desirable to have automatic systems individually trained to becomes maximally effective for each patient. Such systems are indeed being actively developed, trained, and tested all over the world.

Empirical fact: bump reward function works the best. The purpose of the system is to keep the patient's blood sugar level x within the desired interval $[\underline{x}, \overline{x}]$. For training the automatic insulin delivery system, we can, in principle, use different reward functions. In [2], researchers compared the results of using different reward functions, and found out that the most effective is the so-called *bump* reward function that is equal to 0 outside the desired interval and to

$$b(x) = \exp\left(-\frac{c}{(x-\underline{x})\cdot(\overline{x}-x)}\right)$$

for values x within this interval.

Natural question. A natural question is: why the bump functions works best?

What we do in this paper. In this paper, we provide a possible explanation for the effectiveness of the bump function.

2 Analysis of the Problem and the Resulting Explanation

What is a natural reward function? We want the patient to feel healthy. At each moment of time, the only information that we have about the patient is the patient's blood glucose level x. Based on this level, we can only determine the probability $p(x)$ that the patient feels healthy. This probability is what we want to maximize, i.e., that should be our reward function.

Clearly, when the value x is outside the given interval, something is wrong, so the corresponding probability is 0 (or close to 0). So, to find an appropriate reward function, we need to find the probabilities $p(x)$ corresponding to values x from the given interval.

We need to select probabilities based on partial information. In many practical situations, probabilities are determined experimentally, as corresponding frequencies; see, e.g., [5]. However, in our case, we do not have enough statistics, so we need to select the probability distribution based on whatever information we have. For this purpose, let us recall how, in general, a probability distribution is determined based on partial information.

How probability distribution is determined based on partial information: reminder. In many practical situations, we only have partial information about probabilities.

For example, we may know that there are two possible situations, but we have no information which of the two situations is more probable. In such situations, a reasonable idea is to assign equal probability to both situations. Similarly, if we have n possible situations, and we have no reason to believe that one of them is more probable, a reasonable idea is to assign, to all of them, equal probability $1/n$. This natural idea is known as *Laplace Indeterminacy Principle*.

This principle can be described in a slightly different way. If we have two alternatives, we have an uncertainty, in which to determine which is a correct one, we need to ask one binary (= "yes"-"no") question. If we have 2^n alternatives, then we need n binary questions to uniquely determine the alternative.

- When we select equal probabilities, the average number of questions needed to determine the situation remains the same.
- However, if we selected unequal probabilities, then, on average, the number of questions becomes smaller.

For example, if we assign probability 1 to one of the alternatives and 0 to all others, we need 0 questions to find the alternative – it is the one whose probability is 1. So, in this case we kind of cheat, we insert artificial certainty where there was none.

Similarly, if we have partial information about probabilities, i.e., if there is a whole set of probability distributions that is consistent with available information, then a reasonable idea is not to cheat, not to add artificial certainty, but to preserve the original uncertainty.

- In the discrete case, a natural measure of uncertainty is the average number of binary questions that is needed to uniquely determine the alternative.

- In the continuous case, a similar natural measure is the average number of binary question that is needed to determine the unknown value x with a given accuracy $\varepsilon > 0$.

In both cases, there are distributions for which this average number of questions is smaller – but selecting them would be artificially adding certainty. What we need is the distribution that best reflects the original uncertainty, i.e., for which the average number of questions is as large as possible. It turns out that the average of question is described by Shannon's entropy

$$S = -\int f(x) \cdot \ln(f(x))\, dx, \tag{1}$$

where $f(x)$ is the corresponding probability density function; see, e.g., [1, 4]. Thus, a reasonable idea is to select, from each class of probability distributions, the distribution with the largest possible entropy. This *Maximum Entropy* approach has indeed led to many successful applications; see, e.g., [3].

Let us apply this general idea to our case: first idea. We have a class of probability distributions located on the interval $[\underline{x}, \overline{x}]$. If we do not make any assumptions about the distribution, then the only constraint on the probability density function is that the overall probability is 1:

$$\int f(x)\, dx = 1. \tag{2}$$

To maximize entropy under this constraint, it is natural to use Lagrange multiplier method, i.e., to reduce the corresponding constraint optimization problem to an equivalent unconstrained optimization problem of maximizing the function

$$-\int f(x) \cdot \ln(f(x))\, dx + \lambda \cdot \left(\int f(x)\, dx - 1 \right), \tag{3}$$

for an appropriate value λ. This value – known as *Lagrange multiplier* – is determine by the condition that the optimal function $f(x)$ satisfies the constraint (1).

The solution to this unconstrained optimization problem can be obtained by using the known fact from calculus – that the maximum of an expression is attained when all its derivatives are equal to 0. Differentiating the expression (3) with respect to each unknown $f(x)$ and equating the derivative to 0, we conclude that

$$-\ln(f(x)) - 1 + \lambda = 0,$$

i.e., that $f(x) = \exp(\lambda - 1) = \text{const}$. So, we get a uniform distribution on the desired interval – in perfect accordance with the above argument (that used Laplace Indeterminacy Principle).

Limitations of the first idea. From the mathematical viewpoint, this is reasonable, but from the viewpoint of our problem, it is not reasonable at all:

- For the uniform distribution, the probability of being healthy is exactly the same whether we are in the middle of the desired interval or close to one of its endpoints.
- However, in practice, if the value of the blood glucose level start getting closer to the threshold, this should be a sign to be alarmed – so the probability of being healthy should be smaller close to the endpoints.

This means that we cannot get a reasonable distribution if we do not impose any constraints. We need to impose some constraints if we want a reasonable result.

Second idea. We need to add constraints, and constraints reflect partial information that we have. What do we know about the probability distribution? We rarely know individual characteristic, but often, from observations, we know averages.

So, a seemingly natural idea is to add a constraint that we know the average value $\widetilde{x}$ of the quantity x:

$$\int x \cdot f(x)\,dx = \widetilde{x}. \tag{4}$$

By applying the same Lagrange multiplier method to the problem of maximizing entropy (1) under constraints (2) and (3), we arrive at the problem of optimizing the following expression:

$$-\int f(x) \cdot \ln(f(x))\,dx + \lambda_1 \cdot \left(\int f(x)\,dx - 1\right) + \lambda_2 \cdot \left(\int x \cdot f(x)\,dx - \widetilde{x}\right),$$

for which equating derivatives to 0 leads to

$$-\ln(f(x)) = 1 + \lambda_1 + \lambda_2 \cdot x = 0,$$

i.e., to $f(x) = \exp((\lambda_1 - 1) + \lambda_2 \cdot x)$.

Limitations of the second idea. The resulting formula has the same limitations as the first idea:

- we want the probability of healthiness to tend to 0 as approach the endpoints,
- but this is not happening here.

How can we modify the second idea? Let us take into account that the same physical quantity can be described by different numerical values. First, we can select a different measuring unit: e.g., the height of 2 m becomes 200 if we use centimeters. Second, we can select a different starting point: the current year 2023 can become year 2014 in Ethiopian calendar that uses a different starting date.

Finally, many quantities are ratios – e.g., blood glucose level is the ratio of the amount of glucose in blood to the corresponding amount of blood. In such cases, we can reverse the ratio and also get a meaningful description of the same quantity; for example:

- we can have velocity $v = d/t$ – which is the ratio of distance to time – and we can have *slowness* $1/v = t/d$;

- we can have resistance $R = V/I$ – which is the ratio of voltage to current – and we can have *conductivity* $1/R = I.V$, etc.

First try. If we simply change the measuring unit or the starting point, the situation does not change: fixing the mean value of the re-scaled quantity $k \cdot x$ or $x + x_0$ is equivalent to fixing the mean value of the quantity x itself.

What is we reverse the formula for the blood sugar level and consider the mean value of this reverse

$$\int \frac{1}{x} \cdot f(x)\,dx = \widetilde{r}.$$

then the corresponding constraint optimization leads to

$$-\ln(f(x)) - 1 + \lambda_1 + \lambda_2 \cdot \frac{1}{x} = 0,$$

i.e., to

$$f(x) = \exp\left((\lambda_1 - 1) + \frac{\lambda_2}{x}\right).$$

This is still not exactly we want.

Second try leads to the desired explanation. What if we take into account both the possibility of taking a reverse *and* the probability of changing the starting points. It is reasonable to use both endpoints $\underline{x}$ and $\overline{x}$ as starting points. Thus:

- we get the re-scaled values $x - \underline{x}$ and $x - \overline{x}$ (or, better, $\overline{x} - x$, to keep the values non-negative), and
- reversing these re-scaled values leads to the following two constraints:

$$\int \frac{1}{x - \underline{x}} \cdot f(x)\,dx = \widetilde{r}_- \tag{5}$$

and

$$\int \frac{1}{\overline{x} - x} \cdot f(x)\,dx = \widetilde{r}_+. \tag{6}$$

Maximizing the expression (1) under constraints (5) and (6) leads to the following unconstrained optimization problem:

$$-\int f(x) \cdot \ln(f(x))\,dx + \lambda_1 \cdot \left(\int f(x) - 1\right) +$$

$$\lambda_- \cdot \left(\int \frac{1}{x - \underline{x}} \cdot f(x)\,dx - \widetilde{r}_-\right) + \lambda_+ \cdot \left(\int \frac{1}{\overline{x} - x} \cdot f(x)\,dx - \widetilde{r}_+\right).$$

For this expression, equating its derivatives to 0 leads to

$$\ln(f(x)) = -1 + \lambda_1 + \frac{\lambda_-}{x - \underline{x}} + \frac{\lambda_+}{\overline{x} - x}. \tag{7}$$

Here:

- the value λ_- reflects the importance of the left endpoint of the desired interval,
- while the value λ_+ reflects the uncertainty of the right endpoint.

Both endpoints are important. Going beyond each of these two endpoints can be life-threatening, and we have no reason to assume that one of the endpoints is more important. Thus, in line with the Laplace Indeterminacy Principle, it makes sense to assume that these two values are equal: $\lambda_- = \lambda_+$. In this case, the formula (7) takes the form

$$\ln(f(x)) = \text{const} + \frac{\text{const}}{(x-\underline{x})\cdot(\overline{x}-x)}.$$

So, we get exactly the bump function expression for the probability values $f(x)$!

Conclusion. Thus, we have indeed explained the effectiveness of the bump reward function: the explanation is that this function naturally follows from first principles.

Acknowledgments

This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), HRD-1834620 and HRD-2034030 (CAHSI Includes), EAR-2225395, and by the AT&T Fellowship in Information Technology.

It was also supported by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478, and by a grant from the Hungarian National Research, Development and Innovation Office (NRDI).

References

1. B. Chokr and V. Kreinovich. "How far are we from the complete knowledge: complexity of knowledge acquisition in Dempster-Shafer approach." In R. R. Yager, J. Kacprzyk, and M. Pedrizzi (Eds.), *Advances in the Dempster-Shafer Theory of Evidence*, Wiley, N.Y., 1994, pp. 555–576.
2. L. Dénes-Fazakas, M. Siket, L. Szilágyi, G. Eigner, and L. Kovacs, ´ "Investigation of reward functions for controlling blood glucose level using reinforcement learning", *Proceedings of the IEEE 17th International Symposium on Applied Computational Intelligence and Informatics SACI 2023*, Timişoara, Romania, May 23–26, 2023, pp. 387–392.
3. E. T. Jaynes and G. L. Bretthorst, *Probability Theory: The Logic of Science*, Cambridge University Press, Cambridge, UK, 2003.
4. H. T. Nguyen, V. Kreinovich, B. Wu, and G. Xiang, *Computing Statistics under Interval and Fuzzy Uncertainty*, Springer Verlag, Berlin, Heidelberg, 2012.
5. D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman and Hall/CRC, Boca Raton, Florida, 2011.