Derivaton of the QUBO formulation for sparse estimation

Tomohiro Yokota[1], Jun Ohkubo[1], Makiko Konoshima[2], Hirotaka Tamura[2]

[1]*Graduate School of Science and Enginnering, Saitama University, 255 Shimo-Okubo, Sakura-ku, Saitama-shi, 338-8570, Japan*
[2]*Fujitsu Laboratories Ltd., 4-1-1 Kawasaki, Kanagawa 211-8558, Japan*

In recent years, annealing machine have been developed, and their use methods are considered in fields such as optimization problems and machine learning. In the annealing machine, it is necessary to express the problem to be dealt with in QUBO(Quadratic Unconstrained Binary Optimization) formulation and implement it as hardware. However, since the general method of rewriting to the QUBO format is not known yet, it needs to be derived individually. In this paper, we derive the QUBO formulation for the l1-norm (absolute value function) used in sparse estimation. As a result of experiment, it was possible to predict that one variable could be reduced from the result of numerical experiment by applying the Legendre transformation and Wolf-duality theorem to l1norm. By reviewing the formulation we were actually able to reduce one variable.

1. Introduction

In recent yeaars, sepecial machines for annealing have benn developed such as D-Wave Systems Inc.'s D-Wave[1,2)] and Fujitsu's Digital Annealer[3)] and methods for use in various fields are being considered. The annealing machine acceptes Ising model parameters as input, but there are many functions thathave been represented yet because systematic derivation of the Ising model has been shown. In the previous researches, the robust q-loss function is derived in the QUBO form using the Legendre transform,[4)] and its performance is evaluated using the classification problem as an example. Also, in the recently published papaer on the derivation of the ReLU function in the QUBO form,[5)] the derivation has made using the Wolfe-duality theorem,[6)] so that the Legendre transformation alone is insufficient. On the other hand, sparse estimation is used in the fields of image processing and machine learning, and it is an important research in reducing the number of data, selecting related data from high-dimensional and complex data and making it simple. The purpose of this research is to derive an l1-norm, which is a regularization term of Lasso[7)] used in sparse estimation, in the QUBO form using the derivation method in.[5)] Furthemore, we evaluate whether the derived result is correct using simulated anneaing.

2. Background

In this section, we describes the knowledge and algorithms used in the exeriment.

2.1 QUBO and Ising model

Since the QUBO formulation and the Ising model are equivalent, we can be converted to other form if we can be represented one side. The Ising model is represented as follows:

$$H = -\sum_{i,j} J_{i,j}\sigma_i\sigma_j - \sum_i h_i\sigma_i \tag{1}$$

where $\sigma_i \in \{-1,+1\}$is a spin variable for i-th spin, $J_{ij} \in \mathbb{R}$ a quadratic term of i and j, and $h_i \in \mathbb{R}$ a liner term of i. We can easily converted the Ising model to QUBO formulation, which uses binary variable $q_i \in \{0,1\}$, by applying $q_i = \frac{\sigma_i+1}{2}$ and QUBO formulation is represented as follows:

$$H = -\sum_{i,j} \tilde{J}_{i,j}q_iq_j - \sum_i \tilde{h}_i\sigma_i \tag{2}$$

2.2 Simulated Annealing

Simulated annealing(SA) is optimization algorithm, and it find the minimum of the objective function efficiently. SA repeatedly minimize the objective function by moving the variables randmoly (in this papaer, it moves at a fixed size) at each iteration. Also, it does not necessarily reject non-minimizing movement, but accepts it with a certain acceptance probability. It is possible to avoid convergence to the local minimum. Here, the acceptance probability is inversely proportional to the inverse temperature (execution time) and the size of the energy difference before and after movement. Algorithm.1 show the SA algorithm. We uses the temperature function $T(n+1) = \alpha T(n) (0 < \alpha < 1)$ and run the SA algorithm until the temperature becomes small.

Algorithm 1 Simulated Annealing

```
Require: variables $m, t, z_1, z_2$, initila temperature $T_0$, lower
    limit temperature $T_{limit}$, initial random state with $-10 \leq$
    $m \leq 10, -1 \leq t \leq 1, 0 \leq z_1 \leq 10$ and $0 \leq z_2 \leq 10$
Ensure: $t, z_1, z_2$ of local minimum when $m$ is given.
for $n \leftarrow 1$ to $T(n) < T_{limite}$ do
    Evaluate the cost $L(m, t, z_1, z_2)$ at the current state
    Let $t', z'_1$ and $z'_2$ be random move of $t, z_1$ and $z_2$ and
    evaluate the cost $L(m, t', z'_1, z'_2)$
    $\Delta \leftarrow L(m, t', z'_1, z'_2) - L(m, t, z_1, z_2)$
    $p \leftarrow \exp(\Delta / T(count))$
    Update $t, z_1, z_2$ to $t', z'_1, z'_2$with probability $\min(1, p)$,
    otherwise do not update
end for
```

2.3 Lasso

Lasso(Least Absolute Shrinkage and Selection Operator) is a method to estimate regression coefficients sparsely by using l1-norm regularization terms. The cost function of Lasso is represented as follows:

$$\frac{1}{2n} * ||||y - Xw||||_2^2 + alpha||||w||||_1 \tag{3}$$

where . Here, since the Lasso regularization term is not continuous, updating by partial differentation with respect to parameter α can not be obtain. So, in the next section, we will explain how to obtain the update formula of *alpha* using the Coordinate Descent.

2.3.1 Coordinate descent

Coordinate Descent is an algorithm, which updates by repeating the procedure of fixing the values of $p - 1$ elements among p parameters and updating only one remaining parameter for all elements.

3. Derivation of l1-norm in QUBO formulation

We define the function $f(m)$ as follows:

$$f(m) = -\min(-m, m) \tag{4}$$

A function form of $f(m)$ is shown in Fig.1. We convert $f(m)$ to quadratic form by applying the Legendre transform to it. Using the method used in the papaer, the one converted to quadratic form is represented as follows:

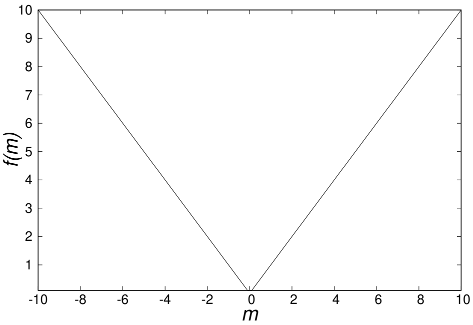

Fig. 1. Outline of l1-norm

$$F(m) = -\min_t \{-mt\} \quad \text{subject to} \quad 1 \le t \le 1 \tag{5}$$

We could express the quadratic form of $f(m)$ as (5), but there is the min function is proceded by a minus sign, which makes it difficult to solve an optimization problem that combines multiple cost functions. Let the other cost function be $C(m)$, and the combination with $F(m)$ is as follows:

$$\begin{aligned} \min_m \{C(m) + F(m)\} &= \min_m \left\{ C(m) - \min_t \{-mt\} \right\} \\ &\neq \min_{m,t} \{C(m) - (-mt)\} \end{aligned}$$

Hence, it is not in the form of minimization problem for both m and t. In precious researche,[5] this problem was solved by applying Wolfe dual theorem to (5). By applying this theorem, the dual problem of the optimization problem (5) is represented as follows:

$$F'(m) = \max_{t,z_1,z_2} \{-mt - z_1(t+1) + z_2(t-1) \tag{6}$$

$$\text{subject to} \quad \begin{cases} -m - z_1 + z_2 = 0, \\ -1 \le t \le 1, z_1 \ge 0, z_2 \ge 0 \end{cases}$$

In order to remove the equality constraint $(-m - z_1 + z_2 = 0)$, it is enough to add the following penalty term of the square of it. Therefore, the optimization problem (6) can be represented as follows:

$$\begin{aligned} F'(m) = \min_{t,z_1,z_2} \{&mt + z_1(t+1) - z_2(t-1) \\ &+ M(-m - z_1 + z_2)^2\} \end{aligned} \tag{7}$$

$$\text{subject to} \quad -1 \le t \le 1, z_1 \ge 0, z_2 \ge 0$$

where M is a constant and take a large value to ensure the equality constraint $(-m-z_1+z_2 = 0)$ to be satisfied, and the remaining inequality consraints conditions $(-1 \le t \le 1, z_1 \ge 0$ and $z_2 \ge 0)$ can be easily realized by expanding these variables t, z_1 and z_2, in the binary expressions which satisfy the corresponding domain constraints. We will varify in the next section that the (7) is correctly fomulated.

3.1 Results

In this section, we verify that the formulation is correct by optimizing problem (7) with SA Algorithm. 1. Each variable moves by +0.001 or -0.001 with the same probability for each iteration. The 10,000 times execution results when $T(0) = 1000$, $\alpha = 0.9999$ and running algorithm until $T(n) < 0.001$ are as shown in the Fig.2.

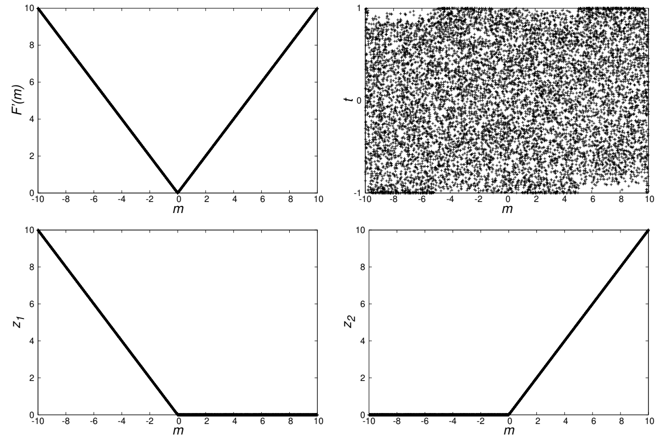

Fig. 2. The figures for upper left, upper right, lower left and lower right show the results when $F'(m), t, z_1, z_2$ are optimized for each m respectively.

From this result, we can confirm that l1-norm can be obtained by optimizing (7). Also, if we look at each variable t, z_1 and z_2 at optimization, we can see the following: It may be possible to reduce one variable by reviewing the (7) because t is not converged alothough z_1 and z_2 converge to a specific value.

4. Reviewing formulation

We can think of the following from the results of the numerical experiments in the previous section: The variable t seems to be taking a random value rather than settling to the optimal value, so we will consider whether we can eliminate t

by reviewing the formulation. The cost function can be transformed as follows using equality constraint.

$$\begin{aligned} F'(m) &= \min_{t,z_1,z_2} \{mt + z_1(t+1) - z_2(t-1) \\ &\quad + M(-m - z_1 + z_2)^2\} \\ &= \min_{t,z_1,z_2} \{mt + z_1(t+1) - (m+z_1)(t-1) \\ &\quad + M(-m - z_1 + z_2)^2\} \\ &= \min_{z_1,z_2} \{z_1 + (m + z_1) + M(-m - z_1 + z_2)^2\} \\ &= \min_{z_1,z_2} \{z_1 + z_2 + M(-m - z_1 + z_2)^2\} \qquad (8) \end{aligned}$$

This conversion from (7) to (8) is possible because the penalty term, $M(-m - z_1 + z_2)^2$, forces the equality constraint to be satisfied.

4.1 Results

The result of experimenting the optimization problem under the same experimental conditions as Section3.1 with (8) as the objective function is as shown in Fig.3. From this result, the outline of $F'(m), z_1$ and z_2 when optimizing (7) did not change from Fig.3, so we seem that removing the variable t does not affect the result.

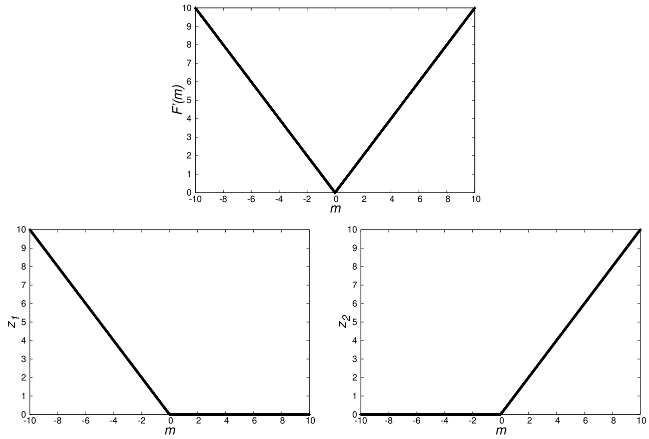

Fig. 3. The figures for upper, lower left and lower right show the results when $F'(m), z_1, z_2$ are optimized for each m respectively.

5. Comparetive experiment

5.1 Results

elements	CD	Annealing	
		α =	α=0.9999995
”crim”	-0.00000000		-0.001
”zn”	0.00000000		0.000
”indus”	-0.00000000		-0.002
”chas”	0.00000000		0.014
”nox”	-0.00000000		-0.008
”rm”	2.71542789		2.703
”age”	-0.00000000		0.000
”dis”	-0.00000000		-0.000
”rad”	-0.00000000		-0.002
”tax”	-0.00000000		-0.001
”ptratio”	-1.34428304		-1.281
”black”	0.18036988		0.200
”lstat”	-3.54677609		-3.560

1) M. W. Johnson, M. H. S. Amin, S. Gildert, T. Lanting, F. Hamze, N. Dickson, R. Harris, A. J. Berkley, J. Johansson, P. Bunyk, E. M. Chapple, C. Enderud, J. P. Hilton, K. Karimi, E. Ladizinsky, N. Ladizinsky, T. Oh, I. Perminov, C. Rich, M. C. Thom, E. Tolkacheva, C. J. S. Truncik, S. Uchaikin, J. Wang, B. Wilson and G. Rose, Nature. **473**, 194 (2011).

2) P. I. Bunyk, E. Hoskinson, M. W. Johnson, E. Tolkacheva, F. Altomare, A. J. Berkley, R. Harris, J. P. Hilton, T. Lanting, J. Whittaker, IEEE Trans. Appl. Supercond. **24**, 1700110 (2014).

3) M. Aramon, G. Rosenberg, E. Valiante, T. Miyazawa, H. Tamura, and H. G. Katzgraber, arXiv:1806.08815.

4) V. Denchev, N.Ding, S. V. N Vishwanathan, and H. Neven, in Proceedings of the 29th International Conference on Machine Learning, p.863 (2012).

5) Go Sato, Makiko Konoshima, Takuya Ohwa, Hirotaka Tamura and Jun Ohkubo, arXiv:1911.03829.

6) P. Wolfe, Quart. Appl. Math. **19**, 239 (1961).

7) Robert Tibshirani, Regression Shrinkage and Selection via the Lasso, J. R. Statist. Soc, B, 58(1):267-288 (1996).