Benchmark Test of Black-box Optimization Using D-Wave Quantum Annealer *

Ami S. KOSHIKAWA　Masayuki OHZEKI　Tadashi KADOWAKI
Kazuyuki TANAKA

In solving optimization problems, objective functions generally need to be minimized or maximized. However, objective functions cannot always be formulated explicitly in a mathematical form for complicated problem settings. Although several regression techniques infer the approximate forms of objective functions, they are at times expensive to evaluate. Optimal points of “black-box” objective functions are computed in such scenarios, while effectively using a small number of clues. Recently, an efficient method by using inference with a sparse prior for a black-box objective function with binary variables has been proposed. In this method, a surrogate model was proposed in the form of a quadratic unconstrained binary optimization (QUBO) problem, and was iteratively solved to obtain the optimal solution of the black-box objective function. In this study, we employ the D-Wave 2000Q quantum annealer, which can solve QUBO by driving the binary variables by quantum fluctuations. The D-Wave 2000Q quantum annealer does not necessarily output the ground state at the end of the protocol owing to the freezing effect during the process. We investigate the effects from the output of the D-Wave quantum annealer in performing black-box optimization. We demonstrate a benchmark test by employing the sparse Sherrington–Kirkpatrick (SK) model as the black-box objective function, by introducing a parameter controlling the sparseness of the interaction coefficients. By comparing the results of the D-Wave quantum annealer with those of the simulated annealing (SA) and semidefinite programming (SDP), we found that the D-Wave quantum annealer and SA exhibit superiority in black-box optimization with SDP. On the other hand, we did not find any advantage of the D-Wave quantum annealer over the SA. As far as in our case, no significant effects by quantum fluctuation were found.

Key words :
Black-box optimization, Bayesian optimization of combinatorial structures (BOCS), Quantum annealing, D-Wave 2000Q

1. Introduction

Black-box optimization is a method to optimize complex and expensive intractable functions, as well as functions without derivatives or explicit forms. Such functions appear in many problems in various fields such as material informatics[1)], machine learning[2)], and robotics[3)]. A systematic way to perform black-

＊著者の了解を得て J. Phys. Soc. Jpn. 90, 064001 (2021) より転載

box optimization is Bayesian optimization[4]. In this method, data points are randomly chosen to generate a training dataset for inferring the black-box objective function. A regression model is then constructed to predict a relation between the input variables and the black-box objective function in the training dataset. Once the regression model is trained, an acquisition function is set up on its basis, which selects the next data point in a solution space from the trained model. The optimal solution of the acquisition function is used to evaluate the black-box objective function and to obtain a new data point of it. When this value is evaluated, the regression model is retrained with new data. These steps are performed iteratively to pursue desired solutions, namely, the optimal point of the black-box objective function.

Bayesian optimization is applied mostly to black-box objective functions with continuous variables, because the optimization of the acquisition function is relatively straightforward. It may be applied to black-box objective functions with discrete variables as well. A significant bottleneck appears in problems with discrete variables, where the resultant acquisition functions also contain discrete variables. It is generally problematic to solve acquisition functions with discrete variables. Optimization problems with discrete variables often belong to the NP-hard class. It takes an extremely long time to solve them using any algorithms. In a previous study, Bayesian optimization of combinatorial structures (BOCS)[5] was proposed as a promising algorithm to evaluate the global minimum of black-box functions. In particular, a sparse prior was employed to efficiently perform regression in the Bayesian inference. The acquisition function was assumed as a quadratic unconstrained binary optimization (QUBO). Notably, relaxation to semidefinite programming (SDP) was used in the optimization phase, which can attain approximate solutions in a reasonable amount of time.

Recently, D-Wave Systems has developed a device[6] that physically implements quantum annealing (QA)[7]. It is a metaheuristic to obtain the ground state of Ising spin glasses belonging to QUBO problems, and this device is now available commercially. Because various combinatorial optimization problems can be formulated as Ising models[8], this D-Wave device has been used in the real world to solve a multitude of practical problems[9)–11)]. The device uses niobium rings as quantum bits (qubits) with programmable local fields and mutual inductance of two qubits, so that the device can solve QUBO problems. Solving a QUBO problem is equivalent to finding a ground state of an Ising spin glass, because binary variables can be rewritten as spin variables. The first stage of QA is initialized in the trivial ground state of the driver Hamiltonian. The quantum effect involved in the driver Hamiltonian is gradually turned off and then ends so that only the classical Hamiltonian with a nontrivial ground state remains. One of the standard choices of the driver Hamiltonian consists only of the x element of the Pauli matrices called the transverse field. When the transverse field changes sufficiently slowly, the quantum adiabatic theorem ensures that we can find the nontrivial ground state at the end of QA[12)–14)]. Numerous studies[15)16)] have shown that QA outperforms simulated annealing (SA)[17)], which utilizes the thermal fluctuation and solves the combinatorial optimization problems. In the context of machine learning, in which various optimization problems are solved during training, QA leads to a different type of value in the output solution known as the generalization performance, as shown in the literatures[18)–20)].

A previous study by Kitai et al. on black-box optimization using the D-Wave device has used the factorization machine[21)], which is used for recommendation systems and can be formulated in QUBO. They focused on the metamaterial design,

and evaluated the figure-of-merit in their metamaterial simulation. In this study, we test the D-Wave quantum annealer in the black-box optimization using BOCS. In particular, the D-Wave quantum annealer does not necessarily output the ground state at the end of the procedure, partly because the connectivity realized in D-Wave 2000Q is a sparse structure called a Chimera graph. To embed a desired graph expressing the structure of the problem on the Chimera graph, redundant qubits with chain structures are used to enhance the connectivity. This is because a single qubit possesses only six connections on average. We use a heuristic tool called minorminer[22] to embed the complete graph into the Chimera graph. Since qubits in the same chain must have the same up or down direction of their magnetic moments, interactions between the qubits are inferred as ferromagnetic interactions. However, qubits in the same chain often do not have aligned magnetic moments, and this makes the solution undetermined. We resolve these broken chains by a majority vote of the directions. This is one of the reasons why the performance of QA in D-Wave 2000Q is unreliable. To achieve better performance of QA in D-Wave 2000Q, various techniques were proposed previously[23)–25)]. In addition, several techniques that avoid many interactions between variables were proposed[26)27)] .The performance of the D-wave quantum annealer is affected by the freezing effect, which appears because of a lack of sufficient quantum fluctuations for driving binary variables at the last stage of QA[28)]. In addition, the thermal effect affects the dynamics of the spin variables as well as quantum fluctuation nontrivially[29)]. Thus, the output is generally deviated from the ground state, especially for the hard optimization problems. Therefore, several protocols employ a non-adiabatic counterpart beyond the standard protocol of QA[30)–33)], with the thermal effect[34)].

In BOCS, the fully connected Ising model is set as the acquisition function. In general, the model includes a hard optimization problem. Thus, the resulting solution from the D-Wave quantum annealer is not necessarily the ground state. The deviation from the ground state is expected to affect the performance of BOCS. We investigate the effect from the quantum device, while comparing the performance of BOCS when optimizing the acquisition function by SA. It is worth noting that SA does not always yield the ground state of the acquisition function depending on the schedule of decreasing a control parameter, i.e., temperature, although, in the protocol of QA, we also tune the transverse field to control the quantum fluctuation. Therefore, the comparison roughly demonstrates the difference between thermal and quantum fluctuations. In this study, we mainly focus on the performance of BOCS depending on the solver in the optimization phase of the acquisition function. We compare the results of BOCS by SDP, which leads to an approximate minimizer of the acquisition function and previously proposed in the original paper on BOCS[5)] as a solver in the optimization phase, with those by SA and QA. Also in the literature on BOCS, the performance employing SA in the optimization phase was investigated. It was not necessarily hard to solve the black-box objective functions, which were random spin systems including the Sherrington–Kirkpatrick (SK) model with decaying interactions in distance measured in indices of spin variables, which is essentially onedimensional Ising spin glass with long-range interactions. In the original paper, the superiority of BOCS by SDP compared with that by SA was reported. However, it is insufficient to discuss the performance of BOCS by solving the problems that appeared in the previous study. In this paper, we change the problem setting into a harder one, a sparse SK model as the black-box objective function. In this sense, the setting is completely different from that in the original study. In addition, we utilize the D-Wave

quantum annealer to perform the optimization in BOCS, not only by SA in a classical computer because this is also a candidate of the solvers employed in BOCS.

In the procedure of BOCS, we have no information on the interaction strengths of the black-box objective function. In BOCS, we iteratively find low-energy state of the acquisition function as a candidate of the optimal solution of the black-box objective function, while the coefficients in the acquisition function are changed.

To investigate the performance of BOCS from a different perspective, we consider the case when the form of the black-box objective function is known. Then we may perform regression to infer only the coefficients to reveal the objective function. The inferred coefficients lead to a good approximation of the black-box objective function. Then, we optimize the resulting approximate function and attain the good estimator of the minimizer of the objective function. In previous studies, the regression method was proposed only from pairs of spin configuration and the corresponding energy value[35)36)]. The method works particularly well for Ising spin glass with sparse interactions. An analytical study using a sophisticated replica method and a numerical verification revealed a relationship between the number of the zero-value coefficients and the number of data needed to reconstruct all the coefficients. If the interaction matrix is sparse, fewer data are needed than the number of coefficients. In BOCS, similarly to this regression method, we utilize the sparse prior distribution to infer the coefficients of the surrogate model, but the form of the objective function is unknown. In this study, we set the SK model as the black-box objective function. We assume that the surrogate model takes the same form as that of the black-box objective function. In other words, in this case, we know the form of the black-box objective function. The comparison of BOCS with the regression and optimization clarifies the performance to infer the sparse interactions of the black-box objective function. We measure the efficiency of BOCS for the sparse SK model in terms of the necessary number of data, which corresponds to the number of iterations in BOCS, to find the optimal solution of the black-box objective function. In the previous study, although BOCS was proposed as the black-box optimization technique for sparse interactions in the objective function, the performance of BOCS depending on the sparseness remained unclear. To solve this problem, we change the sparseness of the SK models and focus on the necessary number of data before finding the ground state by investigating the success rates in finding a minimum.

The remainder of this paper is organized as follows. We explain the BOCS method in Sect. 2, we discuss the numerical experiments in Sect. 3, we compare the performances between BOCS and regression with a sparse prior in Sect. 4, and summarize in Sect. 5.

2. Method

We assume that the input $\vec{x}$ is a vector of binary variables, its ith element is denoted by x_i, and N is the dimension of $\vec{x}$. Each $\vec{x}$ provides an observation y containing a finite error σ. Our goal is to find $\vec{x}$ that minimizes a black-box function. Since we cannot determine an explicit form of the black-box function, we employ a surrogate model and train it using the values of the black-box objective function. Any objective function on the N-dimensional binary variables can be expressed by using up to the N-th polynomial of $\vec{x}$, although this polynomial requires $O(2^N)$ data to fix all the parameters. This huge amount of data cannot be collected in practical situations. We may thus cut the polynomials in finite orders. When the surrogate model contains higher-order terms than

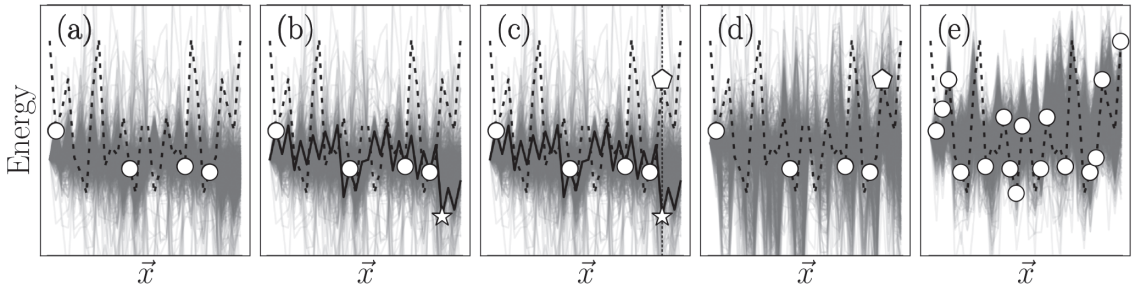


Fig. 1 Schematic of BOCS algorithm. (a) Evaluation of a black-box function (dashed line) at four data points (open circles), and training a surrogate model with the four data points. The grey lines show surrogate models with regression parameters sampled 1000 times. (b) Construction of an acquisition function (solid line) by sampling a regression parameter from the posterior distribution. An open star represents the optimal solution of the acquisition function. (c) Evaluation of the black-box function at the new data point (open pentagon). (d) Retraining the surrogate model by using the five data points (open circles and open pentagon). (e) Trained surrogate model with 16 data points (open circles).

quadratic ones, we do not optimize it efficiently in general. In addition, the approximate surrogate model up to the secondorder terms can be solved by SDP or by using the D-Wave quantum annealer efficiently in the optimization phase. We thus set a quadratic model as the surrogate model:

$$\tilde{f}(\vec{x}) = \alpha_0 + \vec{x}^{\top} Q_\alpha \vec{x}, \tag{1}$$

where α_0 is a real value, and Q_α is an upper triangular matrix.

In this algorithm, we compute a posterior distribution for the model parameters by following the framework of Bayesian inference. By sampling from the posterior distribution, we construct an acquisition function that indicates the next data point to choose from a solution space. We find the minimum of this acquisition function by using some optimization solvers. Then, we obtain the new data from the black-box function with the $\vec{x}$ value. This minimizes the acquisition function. We retrain the model with data, including the new value from the black-box objective function. This algorithm iteratively searches for the global minimum by updating the data points. Fig. 1 shows a schematic of the BOCS algorithm.

2.1 Construction of the acquisition function

We assume that a few data points are attained, because the evaluation of objective functions is expensive. We then consider the sparse Bayesian linear regression to take uncertainties of the regression parameters $\vec{\alpha} = [\alpha_0, Q_{\alpha 11}, Q_{\alpha 22}, \ldots, Q_{\alpha 12}, Q_{\alpha 13}, \ldots]$ and observation noise σ. From observations of several data points $\{\vec{x}^{(i)}, y^i\}_{i=1,2\ldots}$, we compute a posterior distribution over $\vec{\alpha}$ as

$$\mathcal{P}(\vec{\alpha}|\vec{y}, X) \propto \mathcal{P}(\vec{\alpha})\mathcal{P}(\vec{y}|X, \vec{\alpha}) \quad X \in \{0,1\}^{p \times D}, \tag{2}$$

where we construct a matrix X from the vector $\vec{x}^{(i)}$ as $\vec{X}^{(i)} = [1, x_1^{(i)}, x_2^{(i)}, \ldots, x_1^{(i)} x_2^{(i)}, x_2^{(i)} x_3^{(i)}, \ldots]$. In addition, $p = 1 + N + N(N\text{-}1)/2$ and D represents the number of data points. Then, we set a likelihood function and a prior distribution over the parameter $\vec{\alpha}$. The likelihood function is given by a Gaussian distribution with a variance σ^2 as

$$\mathcal{P}(\vec{y}|X, \vec{\alpha}) = \mathcal{N}(\vec{\alpha}X, \sigma^2 I). \tag{3}$$

Since the number of elements of $\vec{\alpha}$ is $O(N^2)$, the number of data also needs $O(N^2)$ to estimate the regression parameters. Otherwise, we will obtain high-variance estimators. To avoid obtaining uncertain parameters even when the data are scarce or the input dimension is large, we set a prior distribution. We here use a horseshoe distribution as the prior distribution to omit the hyperparameters to perform BOCS. This prior distribution is capable of efficiently inferring sparse parameters in the model even if the number of

data is small:[37]

$$\begin{aligned} \alpha_k|\beta_k^2,\tau^2,\sigma^2 &\sim \mathcal{N}(0,\beta_k^2,\tau^2,\sigma^2) \quad k=1,\ldots,p \\ \tau,\beta_k &\sim \mathcal{C}^+(0,1) \quad k=1,\ldots,p \\ \mathcal{P}(\sigma^2) &= \sigma^{-2}, \end{aligned} \tag{4}$$

where $C^+(0,1)$ is the standard half-Cauchy distribution. This formulation, however, cannot realize efficient sampling. Following the method developed by Makalic and Schmidt[38], the half-Cauchy distribution can be expressed with the inverse-Gamma distribution to introduce auxiliary parameters. The inverse-Gamma distribution is a conjugate prior for normal distribution. The modified formulation is a closed form from which we can efficiently sample parameters from this distribution with complexity $O(p^3)$. We use a faster algorithm[39] $[O(D^2p)]$ that is exactly the same as that in the formulation with auxiliary parameters.

As tested in the previous study, one may compute $\vec{\alpha}_{MLE}$ by using a maximum likelihood estimation[5]. Then BOCS with $\vec{\alpha}_{MLE}$ showed purely exploitative behavior and failed to evaluate the optimal solution. In this study, we set coefficients Q_α by sampling from the posterior distribution $P(\vec{\alpha}|X,\vec{y})$, so that the BOCS algorithm shows exploring behavior in the solution space. This is inspired by Thompson sampling in the context of bandit problems, which often shows better performance of the surrogate model attained by the maximum likelihood estimation.

Notice that the horseshoe prior efficiently estimates the sparse parameters in the surrogate model. Thus, the above formulation of BOCS exhibits better performance in the case that the black-box objective function is inherently the sparse interactions when we write its explicit form in a quadratic form. Similarly, as a prior distribution, we may use the Laplace distribution, which is typically chosen for estimating the sparse parameters. Owing to the existence of the hyperparameter in the Laplace distribution, a fair performance comparison between the Laplace distribution and the horseshoe distribution is generally difficult. The result from the Laplace distribution can be changed by tuning the hyperparameter and may become closer to the result from the horseshoe distribution. The superiority of the sparse prior can be expected from the sharpness of the shape of the prior distribution. From this aspect, the horseshoe prior has a good property to infer the sparse parameters compared with the Laplace distribution.

2.2 Optimization solver

Once Q_α is fixed, we optimize the acquisition function (1) to select the new evaluation point. The main contributor of this study is the D-Wave 2000Q quantum annealer that solves discrete quadratic problems as well as SA and SDP.

Note that it can solve only 64-variable problems on a complete graph owing to the sparsity of the hardware graph on the quantum processing unit in the D-Wave quantum annealer. The current quantum annealer (D-Wave advantage) has 5000+ qubits and it implements 180 variables on a complete graph.

Although the solvable size of the D-Wave quantum annealer is limited, the natural computation performed in the D-Wave quantum annealer following the protocol of QA outputs a near-optimal solution in a relatively short time of about 20μs. In this sense, we may expect that the D-Wave quantum annealer can be a fast solver of QUBO. In general, it takes a relatively long time to solve the exact solution of QUBO as in the explanation of SDP.

We explain the three solvers used in this study: SA, SDP, and QA by D-Wave 2000Q.

Simulated annealing SA is a metaheuristic utilizing thermal fluctuations in computation. A spin configuration corresponding to binary variables starts from a random state in a solution space, and one spin in the configuration is flipped following the Metropolis–Hastings algorithm[40]. The energy difference between the initial state and the one-spin

特集

flipped state is denoted by ΔE. If $\Delta E < 0$, the state is updated to the flipped one; otherwise, it is updated with a probability $e^{-\Delta E/T}$. The parameter T is diminished in every iteration. When T is large, SA updates the spin configuration regardless of ΔE, and the system moves in a wide range in the solution space. When T becomes smaller, it magnifies the energy landscape and falls into the local minimum, because the state will not update without small $\Delta E > 0$ values. SA can thus be trapped into a local minimum in general. When the speed to control the temperature is very slow, SA can attain the ground state of the system, namely, the optimal solution. Practically, a relatively quick sweep of the temperature can lead to the optimal solution in most cases.

Semidefinite programming We briefly describe the application of SDP in solving discrete optimization problems. We consider the following quadratic constrained problem in general as follows.

$$\begin{aligned} &\text{minimize}_{\vec{y}} && \sum_{i,j} C_{ij} y_i y_j + d \\ &\text{subject to} && \sum_{i,j} D_{ijk} y_i y_j = b_k, \quad k = 1, 2, \ldots, K \\ & && \vec{y} \in \mathbf{R}^n \end{aligned} \tag{5}$$

Here, $y_i y_j$ can be regarded as the (i,j)-element of the Gram matrix, whose eigenvalues are non-negative. This problem can thus be transformed to a SDP as

$$\begin{aligned} &\text{minimize}_{X} && \mathrm{Tr}(CX) + d \\ &\text{subject to} && \mathrm{Tr}(D_k X) = b_k, \quad k = 1, 2, \ldots, K \\ & && X \succeq 0. \end{aligned} \tag{6}$$

Here, $X \succeq 0$ denotes that X is a semidefinite matrix, which has non-negative eigenvalues. The resultant minimization problem is a convex optimization problem. Thus, we readily attain the optimal solution. The minimum value of this convex optimization problem is the same as that of the original one. We use the property of SDP to attain an approximate solution of our acquisition function. The optimization of the acquisition function is as follows:

$$\begin{aligned} \arg\min_{\vec{x}} \tilde{f}(\vec{x}) &= \arg\min_{\vec{x}} \left(\sum_i Q_{\alpha ii} x_i + \sum_{i<j} Q_{\alpha ij} x_i x_j \right) \\ &= \arg\min_{\vec{x}} (\vec{a}^{\top} + \vec{x}^{\top} A \vec{x}). \end{aligned} \tag{7}$$

We then replace each binary variable x_i with $\sigma_i = 2x_i - 1$, and the minimization problem can then be written as

$$\begin{aligned} &\arg\min_{\vec{\sigma}'} \vec{\sigma}'^{\top} C \vec{\sigma}', \quad \vec{\sigma}' = [\vec{\sigma}^{\top}, \sigma_0]^{\top} \in \{-1, 1\}^{N+1} \\ &C = \begin{bmatrix} \tilde{A} & \vec{c} \\ \vec{c}^{\top} & 0 \end{bmatrix} \\ &\tilde{A} = A/4, \quad \vec{c} = \vec{a}/4 + A^{\top}\vec{1}/4. \end{aligned} \tag{8}$$

We relax the binary variable σ_i to a vector y_i on the N+1-dimensional unit sphere. We can then rewrite Eq. (8) as Eq. (5) with $d = 0$, $D_{ijk} = \delta_{ij}\delta_{ik}$, $\vec{b} = \vec{1}$, and $k = N+1$. We instead solve the resultant SDP as a relaxation problem to generate the approximate solution. We binarize the attained approximate solution using an adequate binarization technique as described in a previous study[5]. In general, it takes an exhaustive time to solve the original optimization problem with binary variables. However, in BOCS, we must iteratively optimize the acquisition function. We should thus employ approximate techniques for performing BOCS in a reasonable amount of time.

D-Wave 2000Q quantum annealer D-Wave 2000Q is a commercial quantum annealer from D-Wave Systems, which physically implements the Ising model with the transverse field. By mapping combinatorial optimization problems into the two-body Ising model, we can find near-optimal solutions based on QA in a few microseconds. In particular, at the end of QA in the D-Wave quantuannealer, the weak quantum fluctuations by the transverse field cannot drive the spins.

This is known as the freezing phenomenon[28]. Thus, the spin configuration often deviates from the ground state.

In addition, the connectivity realized in D-Wave 2000Q is a sparse structure called a Chimera graph. We use a heuristic tool called minorminer[22] to

embed the complete graph into the Chimera graph. Redundant qubits are formed into chain structures to realize a larger graph connectivity. Although the magnetic moments of redundant qubits in a chain structure should take the same direction, they often take different directions. To obtain a solution with these redundant qubits, the direction is adopted by a majority vote. This is also the reason why the performance of QA in the D-Wave 2000Q worsens on dense graph problems. Another postprocess to find a better solution by fixing the broken chain is minimizing energy. This technique often finds a lower energy state than the results after a majority vote. In this study, we choose the majority vote to perform the benchmark without any further improvements from the default setting of using the D-Wave 2000Q. Although there are a few reasons that diminish the output from the D-Wave 2000Q differing from the ground state, it can often yield the ground state in various types of optimization problems described in QUBO with a small number of variables.

3. Numerical Experiment

We perform numerical experiments employing the SK model as a black-box objective function. The SK model belongs to the NP-hard problem depending on the parameters described by spins $\vec{\sigma} \in \{-1,1\}^N$ and interactions J_{ij}:

$$\mathcal{H} = -\frac{1}{N}\sum_{i<j} J_{ij}\sigma_i\sigma_j. \tag{9}$$

The interaction coefficients are randomly selected as

$$J_{ij} \sim (1-\rho)\mathcal{N}(0,+0) + \rho\mathcal{N}(0,1). \tag{10}$$

The standard definition of the SK model is that the coefficient $1/N$ should be $1/\sqrt{N}$. However, we introduce a parameter to control sparseness $\rho \in (0,1]$ and set the coefficient as $1/N$. If the parameter ρ is close to zero, the number of zero elements in the J_{ij} matrix increases. BOCS is expected to work well because it implements the sparse prior. The performance of BOCS should be discussed in two ways. The first is inference. The value of ρ affects the performance of inference in BOCS. This is independent of the optimization solvers. The other way is in optimization. The sparse connectivity of the Ising spin glass set in the black-box objective function affects the performance of optimization solvers. In this study, we use SDP to quickly generate the approximate solution of the acquisition function. SA and QA on the D-Wave quantum annealer do not always reach the ground state but directly solve the acquisition function. Notice that we perform iterations in SA in fixed steps during the optimization phase. In addition, QA on the D-Wave quantum annealer performs only in a fixed amount of time.

In this study, we treat SK models with $N = 20$ spins and 10 initial datasets $\{\vec{x}^{(i)}, y^{(i)}\}_{i=1,\ldots,10}$. The parameter ρ varies from 0.1 to 1.0 in increments of 0.1, and we investigate its dependence on performance. For every ρ value, we generate 50 instances and compute each problem for 10 runs.

Fig. 2 shows the residual energy after t BOCS iteration steps at $\rho = 0.1$ (a), 0.5 (b), and 1.0 (c). The residual energy is calculated by subtracting the global minimum through brute-force computation (E_{glob}) from the minimum value in the obtained data (E_{min}). The curves indicate the mean values of all trials, and the shaded areas indicate the standard deviation. Each curve and hatch type refer to solvers used in the optimization phase: the solid, dashed, and dashed-dotted curves show results from D-Wave 2000Q, SA, and SDP, respectively. In addition, the dotted curve shows the result of random search (RS), which randomly chooses a data point at each iteration step. The performance of BOCS with any optimization solver is superior to that of RS. In our SK models, there is no significant difference between using SA and D-Wave 2000Q, whereas BOCS with SDP does

特集

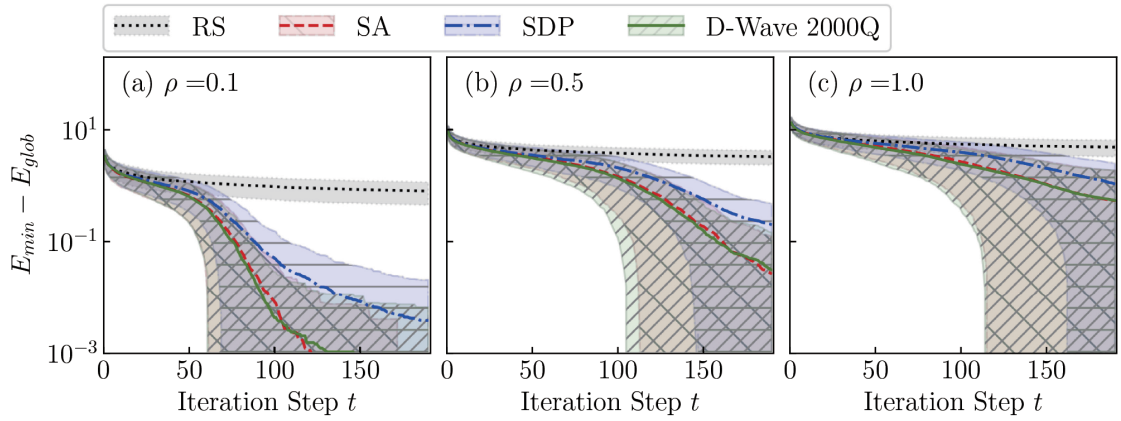


Fig. 2 (Color online) Subtraction of the global minimum (E_{glob}) from the minimum value in the dataset (E_{min}) at ρ = 0.1 (a), ρ = 0.5 (b), and ρ = 1.0 (c). Each curve represents the average of all trials, and each hatch stands for corresponding standard deviations. The solid, dashed, and dash-dotted curves respectively indicate the results of BOCS with D-Wave 2000Q, SA, and SDP.

not efficiently decrease the resulting energy. This is in contrast to the previous result in the original paper on BOCS[5].

The D-Wave 2000Q generates a low-energy state at the end of the protocol. The low-energy state is governed by the Gibbs–Boltzmann distribution with a finite value of the transverse field. In a sense, the resultant spin configuration is affected by the quantum fluctuation. Furthermore, SA in a finite number of steps remains in thermal fluctuation in its resultant solutions. Therefore, the comparison between the results obtained by SA and D-Wave focuses on the difference between the thermal and quantum fluctuations. However, we did not find a significant difference between the thermal and quantum fluctuations in our experimental setup.

We hereafter focus on the comparison between the results by SDP and SA. The difference between the two cases is mainly the deviation from the tentative ground state of the acquisition function. We simply assume that BOCS by SDP falls into a local minimum in the acquisition function at each step, which explains why the performance of BOCS by SDP becomes worse. In BOCS, the balance between exploration and exploitation of the Bayesian inference is important. As we introduce the sampling technique in BOCS for increasing the exploratory property inspired by the Thompson sampling, BOCS by SDP sometimes approaches the optimal solution. However, SA (and D-Wave) outperforms SDP in our problem setting, owing to the better performance to attain the low-energy state by SA (and D-Wave) of the acquisition function. In our problem setting, we employ the sparse SK model. Thus the acquisition function also takes the similar spin glass problem during the process of BOCS. This is one of the reasons why SDP shows worse performance than SA (and D-Wave).

The difference between SDP and SA (and D-Wave) becomes small as the ρ value increases. This implies that the exploratory space becomes narrower as ρ takes higher values. In other words, the solution space is divided into the states around many deep local minima; thus, the exploratory space cannot be sufficiently broadened even by using thermal or quantum fluctuations. The difficulty in solving the hard problems appears in the performance in BOCS.

To investigate the dependence of the performance of BOCS on ρ, we plot a success probability of finding the global minimum in Fig. 3. Regardless of the optimization solvers used, the number of iteration steps before finding the minimum value increases with increasing ρ. The number of iteration steps consists

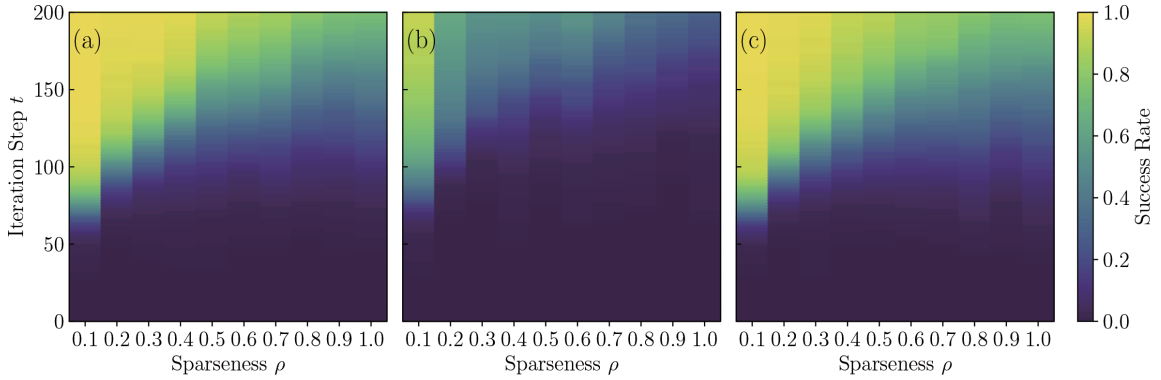


Fig. 3 (Color online) Success rate of finding the global minimum as a function of ρ and the iteration step when SA (a), SDP (b), or D-Wave 2000Q (c) is used.

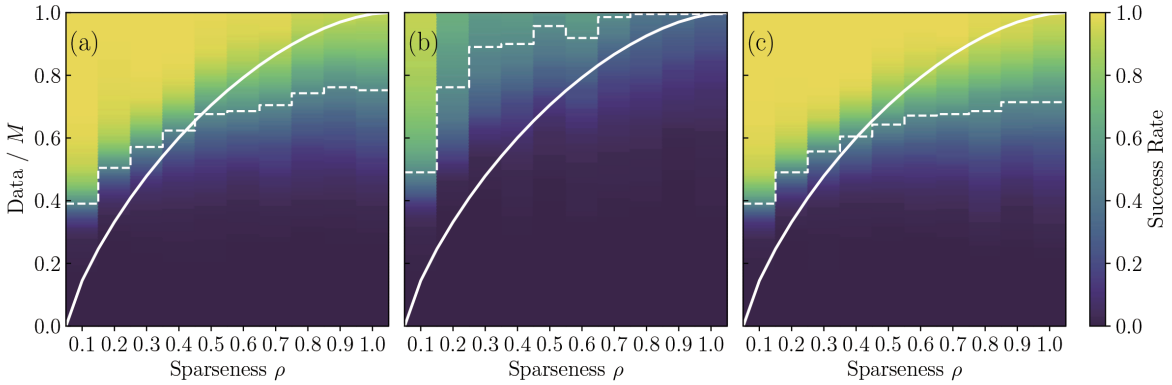


Fig. 4 (Color online) Success rate of finding the global minimum as a function of ρ and the number of data points divided by the number of coefficients when SA (a), SDP (b), or D-Wave 2000Q (c) is used. The solid curve shows the results of the replica analysis. The dashed line shows where the success rate is 50% when BOCS is used. The lower curve shows a better performance for a given ρ . The performances of SA and D-Wave 2000Q are insensitive to the value of ρ compared with the results of the previous study[35].

of the number of data points used in BOCS and the number of duplications. Moreover, we replace the number of iteration steps with the number of data points as shown in Fig. 4. Then, we find the same dependence on ρ as that in Fig. 3.

4. Comparison with Regression

If the form of the black-box objective function is known a priori whereas its coefficients in the quadratic form and some parameters are unknown, one may infer only the coefficients from the regression data attained. This means that once the required number of data points is collected at random, one can reconstruct the coefficients, and the minimum solution can be obtained with an appropriate optimization solver. To reconstruct the coefficients, we solve the following equation:

$$\min_{\tilde{J}} \|\tilde{\boldsymbol{J}}\|_1 \quad \text{subject to } \boldsymbol{E} - S\tilde{\boldsymbol{J}} = \boldsymbol{0}, \tag{11}$$

where $\tilde{J}$ is the coefficient vector, its element is J_{ij}, S is the spin-data matrix, the ith spin data vector is denoted by $S^{(i)} = [\sigma_1^{(i)}\sigma_2^{(i)},\ldots,\sigma_{N-1}^{(i)}\sigma_N^{(i)}]$, $S \in \{0,1\}^{D\times N(N-1)/2}$, and E is the energy-data vector. Here, $E^{(i)}$ is equal to the energy of a spin configuration $S^{(i)}$. In the literature[35], the replica method revealed the relationship between ρ and the required number of data points that can reconstruct J_{ij}. The analysis was validated by numerical experiments. To compare the performance of BOCS and the previous results[35], we convert the iteration steps in Fig. 3 into the number of obtained data points. BOCS performs a random

search at the beginning, because its surrogate model has coefficients with large uncertainties, and the acquisition function varies markedly at each step. When the data are collected partially, the acquisition function moderately changes. Therefore, new data may not always be obtained at each step, and previously obtained data may be selected instead. The number of iteration steps is not equivalent to the number of data points. Fig. 4 shows the success rate as a function of ρ where the number of data points is divided by the number of J_{ij} parameters. The heat map indicates the success rate for obtaining the minimum value. The dashed line indicates a border where the success rate is 50% when using BOCS. The solid white curve is the previous results in the literature[35]. The area above this line indicates that the number of obtained data points is sufficiently large to reconstruct J_{ij} by regression. The other area indicates the number of data points that cannot reconstruct J_{ij}; thus, this solid curve can be regarded as a phase transition line. Since the results given by the previous study are typical reconstruction limits[35], in the case where $N \to \infty$, the success and the failure areas are clearly separated. However, our numerical experiments of BOCS are for finite-number tests. Therefore, the borders of the success area are rather ambiguous. If SA (and D-Wave) is used for the optimization solver in a complicated problem, BOCS is more likely to find the minimum value with a relatively small number of data points. Under certain situations, BOCS occasionally yields better performance than the typical reconstruction limit that was revealed in the previous study.[35] We propose a few reasons that explain this observation. The first is that BOCS only focuses on finding the ground state. The regression does not directly derive the ground state. A typical reconstruction limit is the performance of inferring the correct coefficients of the black-box objective function. Once we find the correct coefficients, we can obtain the exact ground state of the black-box objective function. One might find the ground state from approximate values of the coefficients when the number of data points is insufficient. However, below the typical reconstruction limit, a marked change in the coefficients appears compared with the correct coefficients, because the system undergoes phase transition. Therefore, we cannot optimistically expect that the ground state of the black-box objective function is attained. Another reason is the difference between the sparse priors. In the previous study, the L_1 norm was used for inference of the sparse coefficients[35]. In addition, the hyperparameter for the Laplace distribution was, in some sense, optimized in their analysis. However, BOCS utilizes the horseshoe distribution, which may infer the sparse coefficients more efficiently.

Notice that the results seem to be beyond the theoretical reconstruction limit as $\rho > D/M$. This would be a finite-size effect. We are not arguing that our results suggest any advantage of BOCS beyond the theoretical reconstruction limit in the region where ρ takes a relatively large value. That being said, the possibility remains that BOCS, by making use of the horseshoe prior, reaches the theoretical reconstruction limit. We will be investigating this further in a future study.

5. Summary and Future Directions

Black-box optimization aims at reducing the value of objective functions that are expensive to evaluate, and it has broad applications in various fields such as machine learning and robotics. In this study, we tested BOCS by setting the SK model as a black-box objective function and evaluated its minimum value. In the optimization phase of BOCS, we proposed using the D-Wave 2000Q quantum annealer, which is expected to return near-optimal solutions in constant time, regardless of its problem size up to

the limit of the capacity to embed the problem. In particular, the D-Wave quantum annealer outputs the low-energy state affected by a finite strength of the quantum fluctuation. Similarly to the D-Wave quantum annealer, we use SA in a finite number of steps. The comparison between SA and the D-Wave quantum annealer clarifies the effects of the thermal and quantum fluctuations in BOCS. BOCS iteratively evaluates the tentative acquisition function. Thus, we expected that both thermal and quantum effects were present in the results of BOCS. Despite our hypothesis, we did not find a significant difference between the thermal and quantum fluctuations when using BOCS. In addition, we also employed SDP in the optimization phase of BOCS. The results obtained by SA (and D-Wave) showed better performance than those obtained using SDP in the SK model. This is possibly due to the degree of deviation from the ground state of the tentative acquisition function.

We also compare the number of required data points to find the minimum value of the black-box objective function by BOCS and regression with the L_1 norm. Although BOCS needs more data points than regression to obtain the minimum value in most of the cases, there is a possibility of finding the minimum value with a smaller number of data points than in the case of regression with the L_1 norm, when the black-box has a dense structure. Although the L_1 norm is used to perform regression to infer the sparse parameters, we employed the horseshoe prior in BOCS. This is possibly explained by the difference between the priors used and will be investigated in a future study. We emphasize that our problem setting is slightly different from that in regression under the assumption of the form of the black-box objective function. We search for the minimum without knowledge of the details of the structure of the cost function. In addition, our results suggest that our approach can find only the minimum more efficiently than in the case of regression with the L_1 norm finding the parameters to express the objective function. In this study, we set the SK model as the black-box objective function, which is of the same form as the acquisition function. In other words, the black-box objective function can be expressed by the acquisition function in principle. The performance of BOCS has not been sufficiently investigated when the black-box objective function is not of the same form as the acquisition function. We will also investigate this further. In addition, note that, although we here choose the SK model as the benchmark test, the application of BOCS is not restricted to the objective function with two-body interactions. Beyond two-body interactions, BOCS is applicable in principle. The performance of BOCS diminished by mismatch between the objective function and the surrogate model will be the next scope along the same line as this study.

One of the reasons to employ the D-Wave quantum annealer as the optimization solver is the use of quantum fluctuations. To investigate the quantum fluctuations, we may use the quantum Monte Carlo simulations. In addition, we may investigate the nontrivial effect of quantum fluctuations, known as the non-stochastic Hamiltonian[41)-45)]. However, longer times are required for each optimization. Thus, we consider using the approximate message-passing algorithm depending on the form of the acquisition function[46)]. However, in this study, there is no significant difference between thermal and quantum fluctuations in finding the approximate solution of the acquisition function. Thus, we may employ other Ising solvers such as the CMOS annealer[47)], the Fujitsu Digital Annealer[48)], TOSHIBA simulated bifurcation algorithm[49)], and the FPGA for performing quantum Monte Carlo simulation efficiently[50)]. In addition, the current D-Wave 2000Q performs hybrid computation up to 20,000 variables on a complete graph. By using

these solvers and employing BOCS, we compute more complicated optimization problems with a large number of variables beyond our investigations in study.

It appears that utilizing a distribution generated from D-Wave devices can improve acquisition functions in the BOCS algorithm, because D-Wave devices return nearoptimal solutions within a few microseconds. One of the topics for future work will be finding a surrogate model suited for a distribution generated from D-Wave devices. We conclude that no significant difference between thermal and quantum fluctuations is observed in our results. However, the sampling from the D-Wave quantum annealer actually generates the output affected by the quantum fluctuation. Therefore, more suitable applications of the quantum device in the framework of Bayesian inference should be considered. This is another future research problem.

Acknowledgement

The authors would like to thank Masamichi J. Miyama and Shuntaro Okada for fruitful discussions. The present work was financially supported by JSPS KAKENHI Grant Nos. 18H03303, 18J20396, 19H01095, and 20H02168 and the JST-CREST (No. JPMJCR1402) for the Japan Science and Technology Agency, the Next Generation High-Performance Computing Infrastructures and Applications R&D Program of MEXT and by the MEXT Quantum Leap Flagship Program Grant Number JPMXS0120352009.

References

1) A. R. Oganov and C. W. Glass, J. Chem. Phys. **124**, 244704 (2006).
2) J. Snoek, H. Larochelle, and R. P. Adams, Proc. 25th Int. Conf. Neural Information Processing Systems - Volume 2, NIPS'12, 2012, p. 2951.
3) D. Floreano and F. Mondada, Neural Networks **11**, 1461 (1998).
4) D. R. Jones, M. Schonlau, and W. J. Welch, J. Global Optim. **13**, 455 (1998).
5) R. Baptista and M. Poloczek, in *Proceedings of the 35th International Conference on Machine Learning*, ed. J. Dy and A. Krause (2018) Proceedings of Machine Learning Research, Vol. 80, p. 462.
6) 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).
7) T. Kadowaki and H. Nishimori, Phys. Rev. E **58**, 5355 (1998).
8) A. Lucas, Front. Phys. **2**, 5 (2014).
9) M. Ohzeki, A. Miki, M. J. Miyama, and M. Terabe, Front. Comput. Sci. **1**, 9 (2019).
10) F. Neukart, G. Compostella, C. Seidel, D. von Dollen, S. Yarkoni, and B. Parney, Frontiers in ICT **4**, 29 (2017).
11) M. H. Amin, E. Andriyash, J. Rolfe, B. Kulchytskyy, and R. Melko, Phys. Rev. X **8**, 021050 (2018).
12) S. Suzuki and M. Okada, J. Phys. Soc. Jpn. **74**, 1649 (2005).
13) S. Morita and H. Nishimori, J. Math. Phys. **49**, 125210 (2008).
14) M. Ohzeki and H. Nishimori, J. Comput. Theor. Nanosci. **8**, 963 (2011).
15) G. E. Santoro, R. Martoňák, E. Tosatti, and R. Car, Science **295** , 2427 (2002).
16) R. Martoňák, G. E. Santoro, and E. Tosatti, Phys. Rev. E **70**, 057701 (2004).
17) S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, Science **220**, 671 (1983).
18) M. Ohzeki, S. Okada, M. Terabe, and S. Taguchi, Sci. Rep. **8**, 9950 (2018).
19) C. Baldassi and R. Zecchina, Proc. Natl. Acad. Sci. U.S.A. **115**, 1457 (2018).
20) S. Arai, M. Ohzeki, and K. Tanaka, arXiv:2102.08609 [cond-mat,disnn].
21) K. Kitai, J. Guo, S. Ju, S. Tanaka, K. Tsuda, J. Shiomi, and R. Tamura, Phys. Rev. Res. **2**, 013319 (2020).
22) J. Cai, W. G. Macready, and A. Roy, arXiv:1406.2741 [quant-ph].
23) S. Okada, M. Ohzeki, M. Terabe, and S. Taguchi, Sci. Rep. **9**, 2098 (2019).
24) S. Okada, M. Ohzeki, and K. Tanaka, J. Phys. Soc. Jpn. **89**, 094801 (2020).
25) S. Okada, M. Ohzeki, and S. Taguchi, Sci. Rep. **9**, 13036 (2019).
26) M. Ohzeki, Sci. Rep. **10**, 3126 (2020).
27) M. Kuramata, R. Katsuki, and K. Nakata, arXiv:2012.10135 [quant-ph].
28) M. H. Amin, Phys. Rev. A **92**, 052323 (2015).
29) Y. Bando, Y. Susa, H. Oshiyama, N. Shibata, M. Ohzeki, F. J. GómezRuiz, D. A. Lidar, S. Suzuki, A. del Campo, and H. Nishimori, Phys. Rev. Res. **2**, 033369 (2020).
30) M. Ohzeki, Phys. Rev. Lett. **105**, 050401 (2010).
31) M. Ohzeki, H. Nishimori, and H. Katsuda, J. Phys. Soc. Jpn. **80**, 084002 (2011).
32) M. Ohzeki and H. Nishimori, J. Phys.: Conf. Ser. **302**, 012047 (2011).
33) R. D. Somma, D. Nagaj, and M. Kieferová, Phys. Rev. Lett.

109, 050501 (2012).
34) T. Kadowaki and M. Ohzeki, J. Phys. Soc. Jpn. **88**, 061008 (2019).
35) C. Takahashi, M. Ohzeki, S. Okada, M. Terabe, S. Taguchi, and K. Tanaka, J. Phys. Soc. Jpn. **87**, 074001 (2018).
36) M. Ohzeki, C. Takahashi, S. Okada, M. Terabe, S. Taguchi, and K. Tanaka, Nonlinear Theory Its Appl., IEICE **9**, 392 (2018).
37) C. M. Carvalho, N. G. Polson, and J. G. Scott, Biometrika **97**, 465 (2010).
38) E. Makalic and D. F. Schmidt, IEEE Signal Process. Lett. **23**, 179 (2016).
39) A. Bhattacharya, A. Chakraborty, and B. K. Mallick, Biometrika **103**, 985 (2016).
40) W. K. Hastings, Biometrika **57**, 97 (1970).
41) Y. Seki and H. Nishimori, Phys. Rev. E **85**, 051112 (2012).
42) Y. Seki and H. Nishimori, J. Phys. A **48**, 335301 (2015).
43) M. Ohzeki, Sci. Rep. **7**, 41186 (2017).
44) S. Arai, M. Ohzeki, and K. Tanaka, Phys. Rev. E **99**, 032120 (2019).
45) S. Okada, M. Ohzeki, and K. Tanaka, J. Phys. Soc. Jpn. **88**, 024802 (2019).
46) M. Ohzeki, J. Phys. Soc. Jpn. **88**, 061005 (2019).
47) M. Yamaoka, C. Yoshimura, M. Hayashi, T. Okuyama, H. Aoki, and H. Mizuno, IEEE J. Solid-State Circuits **51**, 303 (2016).
48) S. Tsukamoto, M. Takatsu, S. Matsubara, and H. Tamura, Fujitsu Sci. Tech. J. **53**, 8 (2017).
49) H. Goto, K. Tatsumura, and A. R. Dixon, Sci. Adv. **5**, eaav2372 (2019).
50) H. M. Waidyasooriya, M. Hariyama, M. J. Miyama, and M. Ohzeki, J. Supercomput. **75**, 5019 (2019).

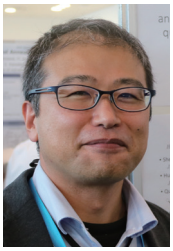
著者

越川 亜美
こしかわ あみ

東北大学　大学院情報科学研究科
応用情報科学専攻　博士課程
機械学習，量子アニーリングの応用研究に従事

大関 真之
おおぜき まさゆき

東北大学　大学院情報科学研究科
情報基礎科学専攻　教授
量子機械学習・量子アニーリングの産業応用の研究に従事



門脇 正史
かどわき ただし

株式会社デンソー
AI 研究部 基盤技術研究室
量子コンピューティング・量子アニーリングの応用研究に従事

田中 和之
たなか かずゆき

東北大学　大学院情報科学研究科
応用情報科学専攻　教授
確率的情報処理の研究に従事