

# Cakewalk Sampling

Uri Patish, Shimon Ullman

Weizmann Institute of Science, Rehovot, Israel

## Abstract

We study the task of finding good local optima in combinatorial optimization problems. Although combinatorial optimization is NP-hard in general, locally optimal solutions are frequently used in practice. Local search methods however typically converge to a limited set of optima that depend on their initialization. Sampling methods on the other hand can access any valid solution, and thus can be used either directly or alongside methods of the former type as a way for finding good local optima. Since the effectiveness of this strategy depends on the sampling distribution, we derive a robust learning algorithm that adapts sampling distributions towards good local optima of arbitrary objective functions. As a first use case, we empirically study the efficiency in which sampling methods can recover locally maximal cliques in undirected graphs. Not only do we show how our adaptive sampler outperforms related methods, we also show how it can even approach the performance of established clique algorithms. As a second use case, we consider how greedy algorithms can be combined with our adaptive sampler, and we demonstrate how this leads to superior performance in k-medoid clustering. Together, these findings suggest that our adaptive sampler can provide an effective strategy to combinatorial optimization problems that arise in practice.

## 1 Introduction

Combinatorial optimization is one of the foundational problems of computer science. Though in general such problems are NP-hard (Papadimitriou 2003), it is often the case that locally optimal solutions can be useful in practice. In clustering for example, a common objective is to divide a given set of examples into a fixed number of groups so as to minimize the distances between group members. Since enumerating all the possible groupings is usually intractable, local search methods such as k-means (MacQueen and others 1967) are frequently used to approach such problems. In many problems however, methods that transform one solution to another can be highly sensitive to their initialization. In some cases this is a result of applying a local search to a problem which has multiple local optima. In others, the search space is simply disconnected, and transforming one valid solution to another is only possible within a small sub-space. In such

cases, the quality of the final solution is determined by how the search is initialized. One common heuristic around this is to sample few initial solutions, and to apply the search multiple times. However, the success of this heuristic mostly depends on the sampling distribution that produces these initial solutions. Thus, if we can have an algorithm that adapts a sampling distribution towards solutions that are associated with good objective values, then we might be able to use it to find good local optima. Such a method could potentially sample locally optimal solutions on its own, or be used as an algorithm that learns how to initialize a particular local search.

One type of algorithms which seem suitable for the task, and which have drawn considerable interest in the last few years are policy gradient methods (Sutton and Barto 2017). Such methods construct a parametric sampling distribution over the search space, and optimize the expected value of some objective function by applying gradient updates in the parameters' space. On the surface, when provided with the right sampling distribution, such methods can access any valid solution, and can therefore provide a strategy that is suitable to our setting. Nonetheless, a closer inspection reveals these methods are highly sensitive to perturbations of the objective function. In particular, the objective values directly affect the sign and the magnitude of the gradient, making these methods notoriously hard to tune. Since the objective in this construction is essentially a random variable whose distribution changes from problem to problem, finding a general rule for tuning them seems impractical. Following this understanding, we propose to circumvent this sensitivity by utilizing a generic surrogate objective function that has the following two properties. First, the surrogate should preserve the set of locally optimal solutions. Second, the surrogate should have a predetermined distribution for every possible objective. Once in this form, such constructions can provide us with a generic adaptive sampler. With this idea in mind, we show how the empirical cumulative distribution function (CDF henceforth) of the original objective can be used to construct such surrogate objectives, and we present a version which makes the basis of our method. Since the crux of our method is based on capitalizing on the CDF of the original objective, we refer to our method as CAKEWALK which stands for CumulativELY Weighted Likelihood.

We start by considering adaptive sampling methods for combinatorial optimization problems in section 2, and proceed to present Cakewalk in section 3. In section 4 we discuss how Cakewalk is related to policy-gradient methods in reinforcement learning, to multi-arm bandit algorithms, and to the cross-entropy (CE henceforth) method. Since ideally we would like to have an adaptive sampler which can recover locally optimal solutions on its own, we use the problem of finding inclusion maximal cliques in undirected graphs as a controlled experiment for testing this property in a non-trivial setting. For that matter, in section 5 we investigate how to apply such methods to the clique problem, and we report experimental results on a dataset of 80 graphs that is regularly used as a benchmark for clique algorithms. In section 6 we consider how Cakewalk can be combined with greedy algorithms, and we demonstrate such a use case on k-medoid clustering, the combinatorial counterpart of k-means. We test how Cakewalk compares to two greedy algorithms commonly used to approach the problem on 38 small datasets, and we show how using Cakewalk for learning how to initialize these methods produced the best performance. We then conclude with a few final remarks in section 7.

## 2 Background

We construct an adaptive sampler for combinatorial optimization problems, and start by stating the problem. Let  $f$  be an objective function which we need to maximize, and let  $\mathbf{x} \in [M]^N$  be a string that describes  $N$  items such that each  $x_j$  is one of a discrete set of  $M$  items. In this text we denote discrete sets  $\{1, \dots, K\}$  using  $[K]$ . Our goal is to search a possibly constrained space  $\mathcal{X} \subseteq [M]^N$  for some  $\mathbf{x}^*$  that achieves an optimal  $f(\mathbf{x}^*) = y^*$  (for constrained problems  $\mathcal{X} \subsetneq [M]^N$ ). Since  $\mathcal{X}$  is discrete and high-dimensional, in general this problem is NP-hard (maximum clique can be reduced to this description), hence we focus only on finding locally optimal solutions. For the purpose of defining locally optimal solutions, we rely on a neighborhood function  $\mathcal{N}$  that maps each  $\mathbf{x}$  to its neighboring set. For example, if  $\mathcal{X} = \{0, 1\}^N$ , then a neighborhood function could be  $\mathcal{N}(\mathbf{x}) = \left\{ \mathbf{x}' \in \mathcal{X} \mid \sum_{i=1}^N |x_i - x'_i| = 1 \right\}$ . Note however that the methods we describe treat  $f$  as a black-box, and do not require  $\mathcal{N}$  for their operation. Our goal is to find some locally optimal solution  $\mathbf{x}^* \in \mathcal{X}_f^*$  where the set of locally optimal solutions is defined as  $\mathcal{X}_f^* = \{ \mathbf{x} \in \mathcal{X} \mid \forall \mathbf{x}' \in \mathcal{N}(\mathbf{x}). f(\mathbf{x}) \geq f(\mathbf{x}') \}$ . Preferably, we would like to find some  $\mathbf{x}^*$  whose objective value  $y^* = f(\mathbf{x}^*)$  is as large as possible, though in general, this cannot be guaranteed.

We describe a learning algorithm for problems of this type. Let  $\mathbf{X}$  be a random variable that is defined over  $\mathcal{X}$ , and which is distributed according to a parametric distribution  $\mathbb{P}_\theta$  that the algorithm maintains. In addition, let  $\mathbf{Y}$  be a random variable that is defined over the values of the objective function  $f$ , i.e.  $\mathbf{Y} = f(\mathbf{X})$ . We emphasize that in this text we refer to random variables using capital English letters in bold such as  $\mathbf{X}$  or  $\mathbf{Y}$ , and we use  $\mathbf{x}$  and  $y$  to refer to elements in their appropriate sample spaces (deterministic quantities).

The algorithm we describe iteratively samples solutions according to  $\mathbb{P}_\theta$ , and it updates the parameters  $\theta \in \mathbb{R}^d$  which govern  $\mathbb{P}_\theta$  in a manner that reflects the quality of those solutions. Initially  $\mathbb{P}_\theta$  is set to have high entropy, but as the algorithm progresses, the entropy in the distribution is decreased until eventually only few solutions become likely (for a discussion of entropy as measure of uncertainty see (Cover and Thomas 2012)). At this point, sampling some  $\mathbf{x}$  from  $\mathbb{P}_\theta$  should return some locally optimal solution with high probability. Since we discuss an iterative algorithm that at each iteration  $t$  updates the parameters  $\theta_t$ , we refer to the random variables that are associated with  $\mathbb{P}_{\theta_t}$  by  $\mathbf{X}_t$  and  $\mathbf{Y}_t$ . Lastly, as a short hand notation, we refer to  $\mathbb{P}_\theta(\mathbf{X} = \mathbf{x})$  simply by  $\mathbb{P}_\theta(\mathbf{x})$ .

Since we learn a distribution function, we say that our learning objective  $J(\theta)$  is to maximize the expectation over  $\mathbf{x} \sim \mathbb{P}_\theta$  of the original objective which we denote as  $\mathbb{E}_\theta[\mathbf{Y}]$ . To find the parameters  $\theta$  which maximize  $J(\theta) = \mathbb{E}_\theta[\mathbf{Y}]$ , we derive a gradient ascent algorithm which relies on estimates of  $\nabla_\theta \mathbb{E}_\theta[\mathbf{Y}]$ . To calculate the gradient, we use the log-derivative trick,  $\nabla_\theta \mathbb{E}_\theta[\mathbf{Y}] = \mathbb{E}_\theta[\mathbf{Y} \nabla_\theta \log \mathbb{P}_\theta(\mathbf{X})]$ , which allows us to estimate  $\mathbb{E}_\theta[\mathbf{Y} \nabla_\theta \log \mathbb{P}_\theta(\mathbf{X})]$  through Monte Carlo sampling (Wasserman 2013). Traditionally, at each iteration  $t$ , a large sample  $S_t = \{ \mathbf{x}_t^k, y_t^k \}_{k=1}^K$  of some fixed size  $K$  is sampled using  $\mathbb{P}_{\theta_t}$ . Denoting this estimate by  $\Delta_t$ , then the update at iteration  $t$  takes the following form,

$$\theta_t = \theta_{t-1} + \eta_t \Delta_t \quad (1)$$

$$\Delta_t = \frac{1}{K} \sum_{k=1}^K [y_t^k \nabla_\theta \log \mathbb{P}_\theta(\mathbf{x}_t^k)] \quad (2)$$

where  $\eta_t$  is a positive learning rate parameter that is predetermined. We describe the update step using a vanilla gradient update mostly for illustrative purposes, though in practice any gradient based update such as Adam (Kingma and Ba 2014) or AdaGrad (Duchi, Hazan, and Singer 2011) can be used instead.

While this stochastic optimization scheme can theoretically converge to a local maximum of  $J$  (Williams 1992), in practice it is highly sensitive to choices of  $K$  and  $\{\eta_t\}_{t=1}^T$ , and to the distributions of  $\{\mathbf{Y}_t\}_{t=1}^T$  (exemplified in the next section). One way to handle this sensitivity is to draw large samples in each iteration, which can reduce the variance of the gradient estimator (in the combinatorial setting, this might require exponentially sized samples). However, even if we increase the sample size, we still need to find a rule that adjusts  $\eta_t$  to the distribution of  $\mathbf{Y}_t$  if we are to produce a generic sampler. Thus, we approach this problem differently, and consider instead how can we adjust the distribution of the objective regardless of the sample size. We focus on online updates (setting  $K = 1$ ), and accordingly drop the superscript  $k$  when referring to  $\mathbf{x}_t^k$  and  $y_t^k$  for the remainder of the text.

## 3 Cakewalk

We start by examining equations 1 and 2, and observing that we update  $\theta_t$  by making a step  $\eta_t y_t$  in the gradient's direction  $\nabla_\theta \log \mathbb{P}_\theta(\mathbf{x}_t)$ . Thus, the sign and magnitude of  $\eta_t y_t$

essentially determine whether we increase or decrease the (log) likelihood of  $\mathbf{x}_t$ , and to what extent we do so. Such direct dependence on the objective values could make our sampler susceptible to perturbations of the objective function. For example, suppose that we have two functions such that  $f_2(\mathbf{x}) = cf_1(\mathbf{x})$  for every  $\mathbf{x}$ , with  $c$  being some fixed positive constant. Clearly,  $\mathcal{X}_{f_1}^* = \mathcal{X}_{f_2}^*$ , nonetheless, sampling and updating the parameters using equations 1 and 2 would change the magnitude of the gradient updates by a factor  $c$ . Though one can adjust the learning rates to the particularities of some given objective, such an approach would require that we tune our method on a case by case basis.

More generally, it appears that the distributions of  $\{\eta_t \mathbf{Y}_t\}_{t=1}^T$  play a critical role in our gradient process. If for example  $|\eta_t \mathbf{Y}_t|$  is unbounded from above for all  $t$ , we might take steps that are too large which may cause the gradient process to diverge. Steps that are too small are unfavorable as well, as these will maintain too much entropy in  $\mathbb{P}_\theta$ , and due to the discrete nature of  $\mathcal{X}$ , finding good  $\mathbf{x}$ s can take exponentially many examples. Since in general we do not know ahead of time the distribution of each  $\mathbf{Y}_t$ , if we follow the construction presented in section 2, we will not be able to determine the series  $\{\eta_t\}_{t=1}^T$  in a manner that would fit all scenarios. This reasoning leads us to conclude that if we wish to obtain generic updates, we must come up with some surrogate objective function which preserves  $\mathcal{X}_f^*$ , and for which we can determine the distributions of  $\{\mathbf{Y}_t\}_{t=1}^T$  ahead of time. To that end, we introduce a weight function  $w$  that when composed over  $f$  (i.e.  $w \circ f$ ) produces a surrogate objective that meets these criteria.

Preserving the original set of optimal solutions is the easy part, as we just need to require that  $w$  will be monotonically increasing, and that would imply that  $\mathcal{X}_f^* \subseteq \mathcal{X}_{w \circ f}^*$  (and strict monotonicity would ensure that  $\mathcal{X}_f^* = \mathcal{X}_w^*$ ). The harder part is to construct  $w$  in a manner that would fix the distribution of  $w(\mathbf{Y}_t)$  for all  $t$ . Nonetheless, basic probability tells us that if  $F_t$  is the CDF of  $Y_t$ , then  $F_t(\mathbf{Y}_t)$  is uniformly distributed on  $[0, 1]$  (Wasserman 2013). Since every CDF is monotonic increasing, if we construct  $w$  using  $F_t$ , we can preserve the original set of optimal solutions. More importantly, if we can estimate  $F_t$ , we could use it to produce our surrogate objective as it **would fix the surrogate’s distribution once and for all**, thus making significant progress towards our goal. Since  $w(Y_t) \sim U(0, 1)$  might not be ideal, we can utilize another monotonic increasing function  $g$  for which  $g(F_t(\mathbf{Y}_t))$  can be distributed differently. For purposes that we specify next, we also require that  $g$  will be bounded.

Since we do not have access to  $F_t$  in general, as was the case with the gradient, we need to estimate it from data. Fortunately enough, since the image of  $f$  is one dimensional (an optimization objective), order statistics can supply us with highly reliable non-parametric estimates for each  $F_t$ . At this point however, it is worth considering how can we estimate  $F_t$  without drawing a large sample at each iteration. Due to equation 2, if we use a sampling distribution for which  $\|\nabla_\theta \log \mathbb{P}_\theta(\mathbf{x}_t)\|$  is bounded, then since  $w(y_t)$  is bounded as well,  $\|\Delta_t\|$  will be bounded for every  $\mathbf{x}_t$  and  $y_t$ . This im-

plies that we can control how different the parameters will be between any two iterations: for any two iterations  $t$  and  $t - k$  where  $k \in [t - 1]$ , we can make  $\|\theta_t - \theta_{t-k}\|$  as small as we want simply by changing  $\eta_t$ . Thus, instead of drawing a large sample in each iteration, we can say the last objective values  $y_{t-1}, \dots, y_{t-k}$  are approximately i.i.d from  $\mathbb{P}_{\theta_{t-1}}$ . Therefore, if we use small enough learning rates, we can use  $\hat{F}_{t-1}(y) = \frac{1}{k} \sum_{i=1}^k \mathbb{I}[y_{t-i} < y]$  as an estimator for  $F_{t-1}$ , where  $\mathbb{I}[\cdot]$  is the indicator function. In our experiments, using some fixed learning rate  $\eta \in (0, 1)$  along with  $k = \frac{1}{\eta}$  seem to work quite well. Overall, the updates we suggest have the following form,

$$\Delta_t = g\left(\hat{F}_{t-1}(y_t)\right) \nabla_\theta \log \mathbb{P}_\theta(\mathbf{x}_t) \quad (3)$$

## Surrogate Objectives

In this section we focus on a single iteration  $t$ , and thus, drop the subscript  $t$  when discussing two possible weight functions. One simple option is to use the empirical CDF  $\hat{F}$  directly, which would make  $\hat{F}(\mathbf{Y})$  uniform discrete on  $[0, 1]$ . However, this surrogate has a major drawback: it leads to an increase in the likelihood of every example it sees. This creates a bias towards  $\mathbf{x}$ s that have already been sampled, compared with  $\mathbf{x}$ s that were not, even though their associated objective value might be better. Since  $\mathcal{X}$  grows exponentially with  $N$ , examples that are drawn early in the process can influence the course of the optimization dramatically. Following this reasoning, we adjust  $\hat{F}$  so that it would increase the likelihood of only half of the examples, and decrease the likelihood of the other half. To do so, we make  $\hat{w}(y) = 2\hat{F}(y) - 1$ . By construction, it follows that  $\hat{w}(\mathbf{Y})$  is uniform discrete on  $[-1, 1]$ . In this fashion, when applied with some fixed learning rate,  $\hat{w}$  determines whether the likelihood of some example will be increased or decreased, and to what extent. Notably, this is achieved along with full specification of the distribution of  $\hat{w}(\mathbf{Y})$ . This is a major advantage compared with, for example, transforming  $\mathbf{Y}$  with its estimated z-score, as in this case we cannot determine how  $w(\mathbf{Y})$  is distributed, nor can we guarantee that  $|w(\mathbf{Y})|$  is bounded (leading to a risk of divergence, and disrupting the online estimation of  $\hat{w}$ ). We summarize Cakewalk with  $\hat{w}$ , and any gradient addition rule *Add* (this includes hyperparameters) in algorithm 1.

## 4 Related Work

Cakewalk is closely related to policy gradient methods. The research on these methods was initiated by Williams with REINFORCE (Williams 1988), an algorithm which we consider as the prototype to Cakewalk, and which provides Cakewalk with convergence guarantees. Most of the work on policy gradient methods derives from REINFORCE, essentially discussing how to transform the objective in various scenarios. Most commonly these involve a baseline estimate  $\hat{\mu}$  of  $\mathbb{E}(\mathbf{Y})$  that can be used to make  $\mathbb{E}(\mathbf{Y} - \hat{\mu}) = 0$ , or a problem specific model for  $\mathbf{Y}$  as is done in actor-critic methods (Sutton and Barto 2017). While sometimes useful, in these constructions the distribution of the objective

---

**Algorithm 1** Cakewalk

---

**input**  $f, \mathbb{P}_\theta, k, Add$   $\triangleright$  objective function  $f$ , sampling distribution  $\mathbb{P}_\theta$ , integer  $k$ , gradient addition rule  $Add$   
initialize  $\theta_0$   
**while** not converged,  $t = 1, 2, \dots$  **do**  
     $\mathbf{x}_t \sim \mathbb{P}_{\theta_{t-1}}$   $\triangleright$  sampling an example  
     $y_t = f(\mathbf{x}_t)$   $\triangleright$  objective value  
    **if**  $t > k$  **then**  
         $w_t = 2 \left( \frac{1}{k} \sum_{i=1}^k \mathbb{I}[y_{t-i} < y_t] \right) - 1$   
         $\Delta_t = w_t \nabla_{\theta} \log \mathbb{P}_{\theta}(\mathbf{x}_t)$   
         $\theta_t = Add(\theta_{t-1}, \Delta_t)$   
    **end if**  
**end while**  
**return**  $\mathbf{x}^*$  which had the highest  $y^*$

---

usually remains unknown, and as a result these methods require careful tuning that is mostly done through trial and error. Cakewalk on the other hand uses a surrogate objective whose distribution is predetermined, and therefore can be applied to a variety of problems in the same way.

Cakewalk is also reminiscent of multi-arm bandit algorithms. In the bandit setting, a learner is faced with a sequential decision problem, where in each round an arm is chosen, and each arm is associated with some non-deterministic loss. Initially suggested by Thompson (Thompson 1933), this setting has been explored extensively with the notable successes of the UCB algorithm (Auer 2002; Auer, Cesa-Bianchi, and Fischer 2002) for cases where the losses are stochastic, and Exp3 (Auer et al. 1995; 2002) for when they can even be determined by an adversary. Over the years these have become a basis for a wide variety of algorithms (Bubeck, Cesa-Bianchi, and others 2012) for various settings which even extend to cases that involve high dimensional structured arms (Awerbuch and Kleinberg 2004; McMahan and Blum 2004; Cesa-Bianchi and Lugosi 2012). The key difference between the bandit and the optimization setting is that the losses associated with each of the arms are non-deterministic, and thus in the bandit setting the main challenge is to balance estimating the statistics associated with each of the arms, with exploiting the information that was already gathered. In the optimization setting however, the goal is to find the best deterministic solution using the least number of steps. Thus, in spite of the apparent similarity, it is this fundamental difference that separates the optimization from bandit settings, and which leads to fundamentally different algorithms.

Cakewalk is also related to the CE method, an iterative algorithm for adapting an importance sampler towards some event of interest. CE was initially introduced by Rubinstein for estimating low probability events (Rubinstein 1997), and later adapted to combinatorial optimization problems (Rubinstein 2001). In CE, at each iteration a large set of examples are sampled. Then, the examples are sorted according to some performance measure (objective function in optimization), and the examples that belong to some highest percentile are selected. For the update step, the distribution's parameters are set to the maximum likelihood estimate of the

selected examples. In this sense, CE is similar to Cakewalk as both methods are only sensitive to the objective values order. However, CE requires large samples at each iteration, and it requires distributions for which maximum likelihood estimates can be produced efficiently (usually this means in closed form). Cakewalk on the other hand only requires a single example at each iteration, and can be applied with any differentiable sampling distribution. Thus, not only can Cakewalk be considerably less costly than CE, its potential applications are much broader.

## 5 Maximum Clique

In this section, we study whether adaptive samplers can recover locally optimal solutions. We emphasize that our goal in this section is mostly to investigate this question, rather than compete with iterative algorithms that transform solutions, and search the input space directly. We study this question on a NP-hard problem instead of problem in which the global optimum can be found in polynomial time, as it is important to verify that such methods can recover non-trivial optima in challenging scenarios. We focus on the problem of finding inclusion maximal cliques, as the notion of inclusion maximal cliques naturally entails what neighborhood function should be used to judge local optimality. Formally, a graph  $G$  is a pair  $(V, E)$  where  $V = [N]$  is a set of vertices, and  $E \subseteq V \times V$  is a set of edges.  $G$  is undirected if for every  $(i, j) \in E$  it follows that  $(j, i) \in E$ . A clique in an undirected graph is a subset of vertices  $U \subseteq V$  such that each pair of which is connected by an edge. An inclusion maximal clique  $U$  is such that there is no other  $v \in V \setminus U$  for which  $U \cup \{v\}$  is also a clique.

We design an objective that could inform algorithms that only rely on function evaluations how densely connected is some subgraph, and which favors larger subgraphs. We refer to this objective as the soft-clique-size function, and denote it by  $f_{SCS}$ . For our purposes, we say the space  $\mathcal{X} = \{0, 1\}^N$  correspond to strings which determine membership in some subgraph  $U$ . Let  $\mathbf{x} \in \mathcal{X}$ , then for each vertex  $j \in V$ , we say that  $j \in U$  if and only if  $x_j = 1$ , and accordingly we denote such subgraphs by  $U_{\mathbf{x}}$ . If some  $U_{\mathbf{x}}$  is a clique, for every  $i, j \in U_{\mathbf{x}}, i \neq j$  it follows that  $(i, j) \in E$ , and therefore  $\sum_{i, j \in U_{\mathbf{x}}, i \neq j} \mathbb{I}[(i, j) \in E] = |U_{\mathbf{x}}|(|U_{\mathbf{x}}| - 1)$ . As a

consequence, for a general subgraph  $U_{\mathbf{x}}$  dividing the LHS by the RHS produces a subgraph density term. However, simply returning a density term would not indicate to an algorithm it should prefer larger subgraphs over smaller ones. Accordingly, we add a parameter  $\kappa \in [0, 1]$  that rewards larger subgraphs, and we change the denominator to  $|U_{\mathbf{x}}|(|U_{\mathbf{x}}| - 1 + \kappa)$ . To see why higher  $\kappa$  can reward larger cliques we focus on the case that  $|U_{\mathbf{x}}| \geq 2$ , and observe that for  $U_{\mathbf{x}}$  which is clique our subgraph density term will be 1 when  $\kappa = 0$ . However, when  $\kappa = 1$ , the subgraph density will be  $\frac{|U_{\mathbf{x}}|(|U_{\mathbf{x}}|-1)}{|U_{\mathbf{x}}|^2} = \frac{|U_{\mathbf{x}}|-1}{|U_{\mathbf{x}}|}$ , and thus, the larger  $U_{\mathbf{x}}$  is, the closer this ratio is to 1. In this manner, increasing  $\kappa$  gives larger subgraphs a 'boost' compared to smaller ones. This of course comes at a price, as it could be that some subgraph which is not a clique will have a higher score than

Table 1: Rate of locally optimal solutions, higher is better

	Exp3	REINF	REINF <sub>B</sub>	REINF <sub>Z</sub>	OCE <sub>0.01</sub>	OCE <sub>0.1</sub>	CW $\hat{F}$	CW $\hat{w}$
SGA	0.000*	0.001*	0.001*	0.097	0.001*	0.000*	0.002*	0.042
AdaGrad	0.000*	0.000*	0.352*	0.427*	0.077*	0.691*	0.164*	<b>0.835</b>
Adam	0.000*	0.000*	0.525*	0.616*	0.106*	0.353*	0.184*	0.753

Table 2: Rate of inclusion maximal cliques, higher is better

	Exp3	REINF	REINF <sub>B</sub>	REINF <sub>Z</sub>	OCE <sub>0.01</sub>	OCE <sub>0.1</sub>	CW $\hat{F}$	CW $\hat{w}$
SGA	0.000	0.000	0.000	0.138	0.000	0.000	0.000	0.037
AdaGrad	0.000*	0.000*	0.637*	0.688*	0.063*	0.875	0.175*	<b>0.912</b>
Adam	0.000*	0.000*	0.662*	0.787	0.100*	0.412*	0.212*	0.887

some smaller subgraph which is a clique (only for  $\kappa = 0$  a score of 1 necessarily means that  $U_{\mathbf{x}}$  is clique). Empirically, we see that the algorithms we have tested are not very sensitive to the value of  $\kappa$ . Lastly, to avoid division by zero for cases  $|U_{\mathbf{x}}| < 2$ , we wrap the denominator with  $\max(\cdot, 1)$ . Altogether, our soft-clique-size function is as follows,

$$f_{SCS}(\mathbf{x}, G, \kappa) = \frac{\sum_{i,j \in U_{\mathbf{x}}, i \neq j} \mathbb{I}[(i, j) \in E]}{\max(|U_{\mathbf{x}}|(|U_{\mathbf{x}}| - 1 + \kappa), 1)}$$

## Experimental Results

As a benchmark for the clique problem, we used 80 undirected graphs that were published as part of the second DIMACS challenge (Johnson and Trick 1996). Each graph was generated by a random generator that specializes in a particular graph type that conceals cliques in a different manner. The graphs contain up to 4000 nodes, and are varied both in their number of nodes and in their edge density. We tested each method on all 80 graphs, letting it maximize the soft-clique-size function using various values of  $\kappa$ . Since a-priori we do not know which  $\kappa$  will lead some method towards an inclusion maximal clique, we have executed each method with each of the values 0.0, 0.1, . . . , 1.0 as  $\kappa$ . We have executed a method for 100  $|V|$  samples (hence runtime is fixed per graph), and recorded the following items at the execution’s end. We recorded the best solution that was found during an execution, along with its objective value, and the sample number in which that solution was found.

In terms of the methods tested, following the discussion on related work, we experimented with a version of CE, three versions of REINFORCE, and of the bandit algorithms we used Exp3. As a sampling distribution, we followed Rubinstein’s construction that assumes independence between the different dimensions, and we used  $N$  softmax distributions defined over the  $N$  dimensions of a problem (the distribution is fully specified in the supplementary material). Since we focus on online algorithms, for CE, we derived a surrogate objective that causes the parameters to update only when we encounter an example whose objective value belongs to some  $\rho$ -highest percentile (also specified in the supplementary material). For this surrogate, we get an online algorithm that operates like CE with parameter smooth-

ing (De Boer et al. 2005), and thus we refer to it as OCE with O standing for online. We applied OCE with two values that were suggested by Rubinstein,  $\rho = 0.1$  and  $\rho = 0.01$ , and refer to these as OCE<sub>0.1</sub> and OCE<sub>0.01</sub>. Next, we experimented with three versions of REINFORCE. First is the vanilla version, second is a version where the mean  $\hat{\mu}$  is subtracted from  $y$  as a baseline, and a third uses the objective’s estimated z-score  $\frac{y - \hat{\mu}}{\hat{\sigma}}$ . We refer to these by REINF, REINF<sub>B</sub>, and REINF<sub>Z</sub>. For Cakewalk, we used both the unscaled empirical CDF  $\hat{F}$ , and its scaled counterpart  $\hat{w}$ , denoting these as CW  $\hat{F}$  and CW  $\hat{w}$  respectively. Note however that the former is only used for comparison, and that we identify Cakewalk with the latter. For estimating  $\hat{\mu}$ ,  $\hat{\sigma}$  and  $\hat{F}$ , we have used the last 100 objective values. We emphasize that both REINF<sub>B</sub>, and REINF<sub>Z</sub> are important comparisons as these methods only transform the objective values, but they do not fix the distribution of the objective as CE and Cakewalk do. For the gradient update steps, we have used vanilla stochastic gradient ascent (SGA henceforth), AdaGrad, and the Adam gradient updates. The latter two updates are considered scale invariant, and could therefore help Exp3, REINF, and REINF<sub>B</sub> handle changes in the objective’s scale. Altogether, we have tested 8 adaptive samplers, 3 gradient updates, on 80 graphs, and 11 values of  $\kappa$ , leading to a total of 21120 separate executions. We specify the complete experimental details in the supplementary material.

We analyzed 4 performance measures for each of the 8 samplers, and the 3 gradient update types, and accordingly we report results in four  $3 \times 8$  tables. In the following, we refer to each combination of a sampler and gradient update as a method. First, we examined whether a locally optimal solution was found using a simple neighborhood function. To that end, given a result  $\mathbf{x}$  in some graph, we compared  $\mathbf{x}$  to every other  $\mathbf{x}'$  such that  $\sum_i |x_i - x'_i| = 1$ , and checked that no  $\mathbf{x}'$  in that graph achieved higher soft-clique-size than  $\mathbf{x}$ . We report the rates at which locally optimal solutions were found in table 1. Then, we proceeded to test if the returned solutions were inclusion maximal. Since the soft-clique-size does not guarantee convergence to cliques, for every graph, we tested whether a method returned at least one inclusion

Table 3: Best-sample to total-samples ratio, lower is better

	Exp3	REINF	REINF <sub>B</sub>	REINF <sub>Z</sub>	OCE <sub>0.01</sub>	OCE <sub>0.1</sub>	CW $\hat{F}$	CW $\hat{w}$
SGA	-	-	0.654	0.907	0.874	0.945	0.939	0.927
AdaGrad	-	-	0.821*	0.821*	0.966*	0.820*	0.926*	0.657
Adam	-	-	0.743*	0.731*	0.835*	0.697*	0.741*	<b>0.619</b>

Table 4: Largest-returned-clique to largest-known-clique ratio, higher is better

	Exp3	REINF	REINF <sub>B</sub>	REINF <sub>Z</sub>	OCE <sub>0.01</sub>	OCE <sub>0.1</sub>	CW $\hat{F}$	CW $\hat{w}$
SGA	0.000	0.000	0.000	0.135	0.000	0.000	0.000	0.038
AdaGrad	0.000*	0.000*	0.538*	0.567*	0.062*	0.738	0.161*	<b>0.756</b>
Adam	0.000*	0.000*	0.577	0.657	0.091*	0.364*	0.190*	0.737

maximal clique when applied with some  $\kappa$ . We report the rates at which inclusion maximal cliques were found in table 2. Since some methods find their best solution earlier than others, to analyze the sampling efficiency of each method we calculated the ratio of the best-sample and the total-samples used in that execution. Since this comparison only makes sense when controlling for the quality of the solution, we excluded REINF and Exp3 from it as they did not return locally optimal solutions. We report average best-sample to total-samples ratios in table 3. To ensure returned solutions are not trivial (say cliques of size 2), for each graph, we compared the largest inclusion maximal clique found by that method, and compared it to the best known solution for that graph, using results from (Nguyen 2017). We report average largest-found-clique to largest-known-clique ratios in table 4. Lastly, we performed multiple hypothesis tests to compare every sampler to CW  $\hat{w}$  in all the experimental conditions using one sided sign test (Gibbons and Chakraborti 2011). To control the false discovery rate (Wasserman 2013), we determined the significance threshold at a level of  $10^{-2}$  using the Benjamini-Hochberg method (Wasserman 2013). In all the tables in this section, when a method is outperformed in a statistically significant manner by CW  $\hat{w}$  this is denoted by \*. The best sampler in each table is emphasized using bold fonts.

The results in tables 1 and 2 clearly support our main proposition: in our setting, a surrogate objective function whose distribution is predetermined significantly improves the robustness of a sampler, and accordingly improves the rate at which locally optimal solutions are found. Both CW  $\hat{w}$  and OCE<sub>0.1</sub> rely on such surrogates, and both outperform Exp3 and all versions of REINFORCE which do not. Importantly, since the previous comparison also includes REINF<sub>Z</sub>, it follows that it is not enough to simply normalize the objective values, and it is better to actually fix the distribution of the objective. Nonetheless, not all distributions are as effective (OCE<sub>0.01</sub> and CW  $\hat{F}$  did not perform as well), and of the ones that we have tested, uniform on  $[-1, 1]$  as used by CW  $\hat{w}$  achieved the best results. CW  $\hat{w}$  clearly outperforms OCE<sub>0.1</sub> in table 1, and the latter only comes close in the more permissive comparison which selects the best result out of 11 different executions (different

values of  $\kappa$ ) as reported in table 2. In terms of sampling efficiency, the results in table 3 show that even though OCE<sub>0.1</sub> can recover locally optimal solutions, it is not as efficient as CW  $\hat{w}$  which finds the best solution considerably faster. When considering the various gradient updates, CW  $\hat{w}$  with AdaGrad produces the best results in almost all measures (CW  $\hat{w}$  with Adam converges slightly faster, though at the cost of worse optimality rates). This is unsurprising as AdaGrad’s classical use case is sparse data (indicator vectors in this case). Lastly, the comparisons to the best known results in table 4 show that the recovered solutions are far from trivial, and that Cakewalk might even approach the performance of problem specific algorithms which have access to a complete specification of the problem.

## 6 K-Medoid Clustering

In this section we demonstrate how Cakewalk can be used as a method that learns how to initialize greedy algorithms. The key idea we utilize in this section is the following. Since Cakewalk only relies on function evaluations, it does not matter if we let it optimize some function  $f : \mathcal{X} \rightarrow \mathbb{R}$ , or a composition  $f \circ T$  where  $T : \mathcal{X}' \rightarrow \mathcal{X}$  is some deterministic transformation ( $\mathcal{X}'$  can be identical to  $\mathcal{X}$ , or some other space that specifies possible initializations of a procedure  $T$ ). As long as some input  $x$  is associated with some fixed objective value  $y = f(T(x))$ , Cakewalk will be able to optimize it. Thus, we can treat a deterministic greedy algorithm as such  $T$ , and use Cakewalk to optimize its initialization.

To demonstrate such a usage we study the k-medoids (Hastie et al. 2009) problem, the combinatorial counterpart of k-means. As in the k-means, we are given a set of  $m$  data points, and our goal is to divide these into  $k$  clusters which minimize the points’ distances to a set of cluster representatives. In contrast to k-means, in k-medoids the representatives must be a subset of the original points that we are given. Thus, one can think of the problem as selecting  $k$  representatives from the  $m$  data points, and in the general case where we allow points to represent more than one cluster, the solution space becomes  $[m]^k$ . Since in k-medoids the representatives are a subset of the data points, it is enough to consider as input a distance matrix  $D \in \mathbb{R}_+^{m \times m}$  where  $D_{i,j}$  is

Table 5: K-medoid clustering results

	Rank1	P-Value	Rank2	P-Value	Rank3	P-Value	Rank4
Objective	CWV 1.0002	0.003816*	PAM 1.0025	7.276e-12*	CW 1.0426	1.419e-10*	VOR 1.6373
Evaluations	VOR 1.0000	3.638e-12*	CWV 1764.1380	3.638e-12*	CW 4352.3860	0.1279	PAM 4922.3370

the distance between point  $i$  and  $j$ , and  $\mathbb{R}_+$  is the set of non-negative reals. Given a set of representatives  $x \in [m]^k$ , each point  $i$  is assigned to the representative  $x_j$  which minimizes the distance  $D_{i,x_j}$  to it. In this formulation, the k-medoids optimization problem can be stated as follows,

$$\text{minimize}_{x \in [m]^k} \sum_{i=1}^m \left[ \min_{j \in [k]} \{D_{i,x_j}\} \right]$$

Since the problem is combinatorial, going over all the possible solutions quickly becomes intractable, and greedy algorithms are usually used to approach the problem. Of these, probably the two most commonly used algorithms are the Voronoi iteration (Hastie et al. 2009), and the more computationally expensive, Partitioning Around Medoids (Kaufman and Rousseeuw 2009) (PAM henceforth).

## Experimental Results

Using a publicly available collection (White 2017), we gathered 38 datasets that had between 500 and 1000 data points. In each dataset we extracted all the numerical attributes, and used these to represent each data point. Then, for each dataset we calculated pairwise Mahalanobis distances, using diagonal covariance matrices (Bishop 2006). At this point, we were able to test the aforementioned algorithms on these datasets. Specifically, we used the Voronoi iteration and PAM, as well as vanilla Cakewalk. As an example for a use case where Cakewalk is combined with a greedy method, we also used Cakewalk with the Voronoi iteration. We did not combine Cakewalk with PAM as it is considerably more computationally expensive than the former. In the result table which we specify next, we refer to these by VOR, PAM, CW, and CWV. We applied Cakewalk using the best configuration found in section 5, and the same factorized distribution. We applied each of the 4 algorithms on all datasets with  $k = 10$ , and recorded the best objective value that was returned, as well as the number of objective evaluations that were performed. We specify the complete experimental details in the supplementary material.

In the analysis our goal was to produce two statistically significant rankings of the methods tested. First, we ranked the methods by the objective values they returned. Second, as the former criterion is influenced by the number of objective function evaluations, we also recorded the number of evaluations each method performed. To produce rankings for both criteria, we compared the 4 methods in a manner that is invariant to the specifics of a given dataset. For that matter, we first calculated the ratio between a method’s score

(objective value, or number of evaluations) in some dataset, and the minimal score achieved by any method on that dataset. Then, we averaged these ratios on all 38 datasets, and used that averaged ratio as a method’s score. This provided us with two scores for each of the 4 methods. Sorting these averaged ratios provided us with a ranking which we could then test for statistical significance. For each two consecutive methods in a ranking, we performed one sided sign test using the original values measured on the 38 datasets. This procedure produced 3 p-values for the ranking of the objective values, and another 3 for the number of evaluations. Lastly, to control the false discovery rate, we determined the significance threshold at a level of  $10^{-2}$  using the Benjamini-Hochberg method. We report the results of this analysis in table 5, where rank1 one is the best method, and rank4 is the worst. In table 5, the ranking in terms of objective values is displayed in the first row, and the ranking in terms of number of evaluations is displayed in the second row. For each ranking, presented are the methods names, along with the p-values for the difference between each pair. When the difference between a pair of methods is statistically significant, this is denoted by \*. In addition, under each method are the averaged ratios used to produce the ranking.

These results demonstrate that combining Cakewalk with a greedy algorithm can produce a method that outperforms the components that make it up. Notably, here we combined Cakewalk with the Voronoi iteration, the weaker of the two greedy methods we tested, and that already produced the best results. Furthermore, vanilla Cakewalk outperformed the Voronoi iteration, showing that Cakewalk can outperform some greedy algorithms as these might be limited by the transformations they apply, and the initializations they rely on. In terms of function evaluations, it appears that providing Cakewalk with the Voronoi iteration leads to faster convergence compared to vanilla Cakewalk, another positive outcome for this combined optimization strategy.

## 7 Conclusion

In this paper we presented Cakewalk, a generic adaptive sampler for combinatorial optimization problems. We demonstrated how Cakewalk outperforms similar adaptive samplers, and how Cakewalk can be combined with greedy algorithms to produce highly effective optimizers. We believe that future research will prove Cakewalk’s effectiveness in combinatorial problems that arise in practice, as well as in other domains such as continuous non-convex optimization, and reinforcement learning.

## 8 Acknowledgment

Supported by Israeli Science Foundation grant 320/16.

### References

- Auer, P.; Cesa-Bianchi, N.; Freund, Y.; and Schapire, R. E. 1995. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Foundations of Computer Science, 1995. Proceedings., 36th Annual Symposium on*, 322–331. IEEE.
- Auer, P.; Cesa-Bianchi, N.; Freund, Y.; and Schapire, R. E. 2002. The nonstochastic multiarmed bandit problem. *SIAM journal on computing* 32(1):48–77.
- Auer, P.; Cesa-Bianchi, N.; and Fischer, P. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine learning* 47(2-3):235–256.
- Auer, P. 2002. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research* 3(Nov):397–422.
- Awerbuch, B., and Kleinberg, R. D. 2004. Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, 45–53. ACM.
- Bishop, C. M. 2006. *Pattern recognition and machine learning*. Springer.
- Bubeck, S.; Cesa-Bianchi, N.; et al. 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning* 5(1):1–122.
- Cesa-Bianchi, N., and Lugosi, G. 2012. Combinatorial bandits. *Journal of Computer and System Sciences* 78(5):1404–1422.
- Cover, T. M., and Thomas, J. A. 2012. *Elements of information theory*. John Wiley & Sons.
- De Boer, P.-T.; Kroese, D. P.; Mannor, S.; and Rubinstein, R. Y. 2005. A tutorial on the cross-entropy method. *Annals of operations research* 134(1):19–67.
- Duchi, J. C.; Hazan, E.; and Singer, Y. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research* 12:2121–2159.
- Gibbons, J. D., and Chakraborti, S. 2011. Nonparametric statistical inference. In *International encyclopedia of statistical science*. Springer. 977–979.
- Hastie, T.; Tibshirani, R.; Friedman, J.; Hastie, T.; Friedman, J.; and Tibshirani, R. 2009. *The elements of statistical learning*, volume 2. Springer.
- Johnson, D. S., and Trick, M. A. 1996. *Cliques, coloring, and satisfiability: second DIMACS implementation challenge, October 11-13, 1993*, volume 26. American Mathematical Soc.
- Kaufman, L., and Rousseeuw, P. J. 2009. *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons.
- Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *CoRR* abs/1412.6980.
- MacQueen, J., et al. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, 281–297. Oakland, CA, USA.
- McMahan, H. B., and Blum, A. 2004. Online geometric optimization in the bandit setting against an adaptive adversary. In *International Conference on Computational Learning Theory*, 109–123. Springer.
- Nguyen, T. H. 2017. Clique benchmark instances. <http://cse.unl.edu/nguyen/npbenchmarks/cliq.html>. Last update: 2017-02-01, Accessed: 2018-05-16.
- Papadimitriou, C. H. 2003. *Computational complexity*. John Wiley and Sons Ltd.
- Rubinstein, R. Y. 1997. Optimization of computer simulation models with rare events. *European Journal of Operational Research* 99(1):89–112.
- Rubinstein, R. Y. 2001. Combinatorial optimization, cross-entropy, ants and rare events. *Stochastic optimization: algorithms and applications* 54:303–363.
- Sutton, R. S., and Barto, A. G. 2017. Reinforcement learning an introduction—second edition, in progress (draft).
- Thompson, W. R. 1933. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* 25(3/4):285–294.
- Wasserman, L. 2013. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media.
- White, J. M. 2017. Julia package for loading many of the data sets available in r. <https://github.com/johnmyleswhite/RDatasets.jl>. Last update: 2018-04-15, Accessed: 2018-05-15.
- Williams, R. J. 1988. On the use of backpropagation in associative reinforcement learning. In *Proceedings of the IEEE International Conference on Neural Networks*, volume 1, 263–270. San Diego, CA.
- Williams, R. J. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* 8(3-4):229–256.