Gumbel-softmax Optimization (GSO)

A Simple General Framework for Combinatorial Optimization Problems on Graphs

Jing Liu

School of Systems Science, Beijing Normal University

October 11, 2019

Introduction

Background

- Many problems in real life can be converted to combinatorial optimization problems (COPs) on graphs, that is to find a best node state configuration or a network structure such that the designed objective function is optimized under some constraints.

Introduction

Background

- Many problems in real life can be converted to combinatorial optimization problems (COPs) on graphs, that is to find a best node state configuration or a network structure such that the designed objective function is optimized under some constraints.
- Usually these problems are notorious for their hardness to solve because most of them are NP-hard or NP-complete.

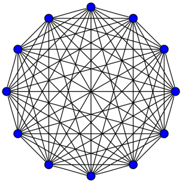
Introduction

Examples of COPs

- Sherrington-Kirkpatrick (SK) model: a celebrated spin glasses model defined on a complete graph. The objective function (ground state energy) is defined as:

$$E(s_1, s_2, \cdots, s_N) = -\sum_{1 \leq i < j \leq N} J_{ij} s_i s_j \tag{1}$$

where $s_i \in \{-1, +1\}$ and $J_{ij} \sim \mathcal{N}(0, 1/N)$ is the coupling strength between two vertices. The number of all possible configurations is 2^N and minimizing the object function is an NP-hard problem.



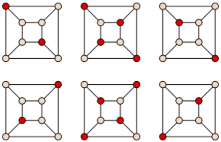
Introduction

Examples of COPs

- MIS problem: Finding the largest subset $\mathcal{V}' \subseteq \mathcal{V}$ such that no two vertices in $\mathcal{V}'$ are connected by an edge in $\mathcal{E}$.The Ising-like objective function consists of two parts:

$$E(s_1, s_2, \cdots, s_N) = -\sum_i s_i + \alpha \sum_{ij \in \mathcal{E}} s_i s_j, \tag{2}$$

MIS problem is also an NP-hard problem.



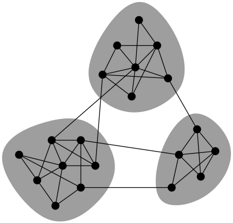
Introduction

Examples of COPs

- Modularity: Modularity is a graph clustering index for detecting community structure in complex networks. In general cases where a graph is partitioned into K communities, the objective is to maximize the following modularity:

$$E(s_1, s_2, \cdots, s_N) = \frac{1}{2M} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2M} \right] \delta(s_i, s_j), \tag{3}$$

It is suggested that maximizing modularity is strongly NP-complete.



Introduction

Traditional methods

- Simulated annealing (SA)
- Genetic algorithm (GA)
- Extremal optimization (EO)
- ...

Introduction

Traditional methods

- Simulated annealing (SA)
- Genetic algorithm (GA)
- Extremal optimization (EO)
- ...

Remark

However, these methods suffer from:

- slow convergence
- limited to system size up to thousand

Introduction

Recent efforts

- Recent efforts focus on machine/deep learning methods, which is based on automatic differentiation techniques.
- Usually these methods belong to supervised learning, containing two stages of problem solving: first training the solver and then testing.
- For example, Li et al.[1] used graph convolution networks (GCNs) to train a heuristic solver for some NP-hard problems on graphs.

[1] Li, Z., Chen, Q., Koltun, V., 2018. Combinatorial optimization with graph convolutional networks and guided tree search, in: Advances in Neural Information Processing Systems. pp. 539548.

Introduction

Recent efforts

- Recent efforts focus on machine/deep learning methods, which is based on automatic differentiation techniques.
- Usually these methods belong to supervised learning, containing two stages of problem solving: first training the solver and then testing.
- For example, Li et al.[1] used graph convolution networks (GCNs) to train a heuristic solver for some NP-hard problems on graphs.

Remark

Although relatively good solutions can be obtained efficiently, it takes a long time for training the solver and the quality of solutions depends heavily on the quality and the amount of the data for training, which is hardly for large graphs.

[1] Li, Z., Chen, Q., Koltun, V., 2018. Combinatorial optimization with graph convolutional networks and guided tree search, in: Advances in Neural Information Processing Systems. pp. 539548.

Introduction

Our approach

The difficulty of solving COPs using deep/machine learning without traning

Sampling operation introduces stochasticity and is non-differentiable. Thus we cannot use automatic differentiation techniques.

Introduction

Our approach

The difficulty of solving COPs using deep/machine learning without traning

Sampling operation introduces stochasticity and is non-differentiable. Thus we cannot use automatic differentiation techniques.

We adopt a reparameterization trick developed in machine learning community called Gumbel-softmax, which provides another approach for differentiable sampling.

Methodology

Mean field approximation

We assume that vertices in the network are independent and the joint probability of a configuration $(s_1, s_2, \cdots, s_N)$ can be written as a product distribution:

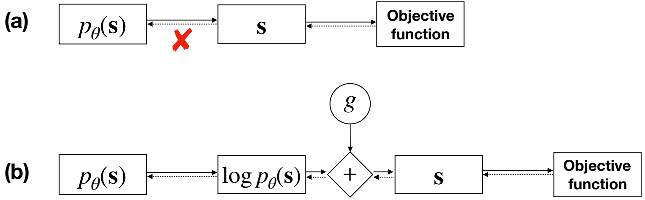
$$p_\theta(s_1, s_2, \cdots, s_N) = \prod_{i=1}^{N} p_{\theta_i}(s_i). \tag{4}$$

The probability $p_{\theta_i}(s_i)$ is a Bernoulli or multinoulli distribution parameterized by parameters θ_i.

Methodology

Gumbel-softmax

Gumbel-softmax, a.k.a. concrete distribution, provides an alternative approach to tackle the difficulty of non-differentiability. For a Bernoulli distribution, instead of sampling a hard one-hot vector$[0, 1]$ or $[1, 0]$, Gumbel-softmax gives a continuous proxy like $[0.01, 0.99]$.



Methodology

Gumbel-softmax Optimization (GSO)

Gumbel-softmax Optimization (GSO) Algorithms

1. Initialization for N vertices: $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_N)$
2. Sample from $p(s_i)$ simultaneously via Gumbel-softmax technique and then calculate the objective function $E(s_1, s_2, \cdots, s_N)$
3. Backpropagation to compute gradients $\partial E(\mathbf{s}; \boldsymbol{\theta})/\partial \boldsymbol{\theta}$ and update parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_N)$ by gradient descent

Methodology

Gumbel-softmax Optimization (GSO)

Gumbel-softmax Optimization (GSO) Algorithms

1. Initialization for N vertices: $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_N)$
2. Sample from $p(s_i)$ simultaneously via Gumbel-softmax technique and then calculate the objective function $E(s_1, s_2, \cdots, s_N)$
3. Backpropagation to compute gradients $\partial E(\mathbf{s}; \boldsymbol{\theta}) / \partial \boldsymbol{\theta}$ and update parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_N)$ by gradient descent

Batch version

We can simultaneously initialize N_{bs} different initial values and calculate N_{bs} objective functions. When the training procedure is finished, we select result with best performance from N_{bs} candidates.

Experiment

Experimental settings

- For SK model, we optimize the ground state energy with various sizes ranging from 256 to 8192.
- For MIS problem, we use citation network datasets: *Cora, Citeseer and PubMed* and treat them as undirected networks.
- For modularity optimization, we use four real-world datasets: *Zachary, Jazz, C.elegans and E-mail.*

Experiment

Experimental settings

We compare our proposed method to other classical optimization methods and state-of-the-art deep learning approaches:

- Simulated annealing (SA): a general optimization method inspired by Metropolis-Hastings algorithm;
- Extremal optimization (EO): a heuristic designed to address combinatorial optimization problems;
- Structure2Vec Deep Q-learning (S2V-DQN)[1]: a reinforcement learning method to address optimization problems over graphs;
- GCNs[2]: a supervised learning method based on graph convolutional networks (GCNs).

[1] Khalil, E., Dai, H., Zhang, Y., Dilkina, B., Song, L., 2017. Learning combinatorial optimization algorithms over graphs, in: Advances in Neural Information Processing Systems. pp. 63486358.

[2] Li, Z., Chen, Q., Koltun, V., 2018. Combinatorial optimization with graph convolutional networks and guided tree search, in: Advances in Neural Information Processing Systems. pp. 539548.

Experiment

Results: SK model

Table: Results: SK model

N	I	EO[1]		SA		GSO ($N_{bs} = 1$)	
		E_0	time	E_0	time (s)	E_0	time (s)
256	5000	-**0.74585(2)**	$\sim$ 268s	-0.7278(2)	1.28	-0.7270(2)	**0.75**
512	2500	-**0.75235(3)**	$\sim$ 1.2h	-0.7327(2)	3.20	-0.7403(2)	**1.62**
1024	1250	**0.7563(2)**	$\sim$ 20h	-0.7352(2)	15.27	-0.7480(2)	**3.54**
2048	400	-	-	-0.7367(2)	63.27	-**0.7524(1)**	**5.63**
4096	200	-	-	-0.73713(6)	1591.93	-**0.7548(2)**	**8.38**
8192	100	-	-	-	-	-**0.7566(4)**	**26.54**

[1] Boettcher, S., 2005. Extremal optimization for Sherrington-Kirkpatrick spin glasses. The European Physical Journal B-Condensed Matter and Complex Systems 46, 501505.

Experiment

Results: SK model

Table: Results: SK model

N	I	GD (Adam)		GD (L-BFGS)		GSO ($N_{bs}=1$)		GSO ($N_{bs}=128$)	
		E_0	time (s)	E_0	time (s)	E_0	time (s)	E_0	time (s)
256	5000	-0.6433(3)	2.84	-0.535(2)	2.29	-0.7270(2)	0.75	**-0.7369(1)**	**0.69**
512	2500	-0.6456(3)	2.87	-0.520(3)	2.56	-0.7403(2)	1.62	**-0.7461(2)**	**1.61**
1024	1250	-0.6466(4)	3.22	-0.501(5)	**2.73**	-0.7480(2)	3.54	**-0.7522(1)**	4.09
2048	400	-0.6493(2)	3.53	-0.495(8)	**3.06**	-0.7524(1)	5.63	**-0.75563(5)**	12.19
4096	200	-0.6496(5)	4.62	-0.49(1)	**3.55**	-0.7548(2)	8.38	**-0.75692(2)**	39.64
8192	100	-0.6508(4)	16.26	-0.46(2)	**4.82**	-0.7566(4)	26.54	**-0.75769(2)**	204.26

Experiment

Results: SK model

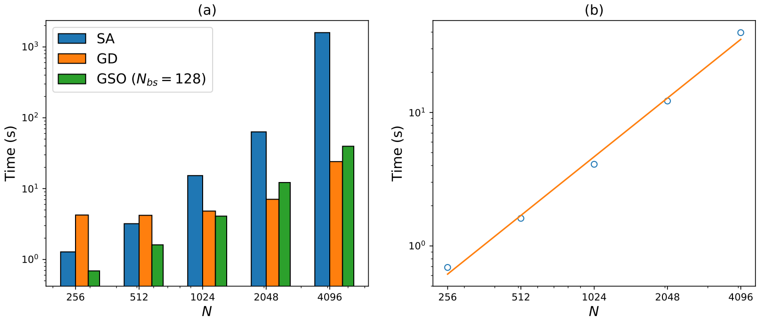


Figure: (a): The time for simulated annealing (SA), gradient descent (GD) with Adam optimizer and our proposed method ($N_{bs} = 128$) on optimization of ground state energy of SK model. (b) A log-log plot of time versus system size N and the slope is 1.46, which indicates that the algorithmic cost is less than $\mathcal{O}(N^2)$.

Experiment

Results: MIS problem

Table: Results on MIS problems.

Graph	size	S2V-DQN[1]	GCNs[2]	GD (L-BFGS)	Greedy	GSO
Cora	2708	1381	**1451**	1446	**1451**	**1451**
Citeseer	3327	1705	**1867**	1529	1818	1802
PubMed	19717	15709	**15912**	15902	**15912**	15861

[1] Khalil, E., Dai, H., Zhang, Y., Dilkina, B., Song, L., 2017. Learning combinatorial optimization algorithms over graphs, in: Advances in Neural Information Processing Systems. pp. 63486358.

[2] Li, Z., Chen, Q., Koltun, V., 2018. Combinatorial optimization with graph convolutional networks and guided tree search, in: Advances in Neural Information Processing Systems. pp. 539548.

Experiment

Results: MIS problem

Table: Results on MIS problems.

Graph	size	S2V-DQN[1]	GCNs[2]	GD (L-BFGS)	Greedy	GSO
Cora	2708	1381	**1451**	1446	**1451**	**1451**
Citeseer	3327	1705	**1867**	1529	1818	1802
PubMed	19717	15709	**15912**	15902	**15912**	15861

Remark

- Our method obtained much better results compared to the sophisticated S2V-DQN.
- Although our results are not competitive with GCNs, we must stressed that it is a supervised learning algorithm. Besides, it also adopts graph reduction techniques and a parallelized local search algorithm. Our method, however, requires none of these tricks.

[1] Khalil, E., Dai, H., Zhang, Y., Dilkina, B., Song, L., 2017. Learning combinatorial optimization algorithms over graphs, in: Advances in Neural Information Processing Systems. pp. 63486358.

[2] Li, Z., Chen, Q., Koltun, V., 2018. Combinatorial optimization with graph convolutional networks and guided tree search, in: Advances in Neural Information Processing Systems. pp. 539548.

Experiment

Results: Modularity

Table: Results on modularity optimization.

Graph	size	Newman [1]		EO[2]		GSO	
		Q	No. comms	Q	No. comms	Q	No. comms
Zachary	34	0.3810	2	0.4188	4	**0.4198**	4
Jazz	198	0.4379	4	**0.4452**	5	0.4451	4
C. elegans	453	0.4001	10	**0.4342**	12	0.4304	8
E-mail	1133	0.4796	13	**0.5738**	15	0.5275	8

1 Newman, M.E., 2006. Modularity and community structure in networks. Proceedings of the national academy of sciences 103, 85778582.

2 Duch, J., Arenas, A., 2005. Community detection in complex networks using extremal optimization. Physical review E 72, 027104.

Experiment

Results: Modularity

Table: Results on modularity optimization.

Graph	size	Newman [1]		EO[2]		GSO	
		Q	No. comms	Q	No. comms	Q	No. comms
Zachary	34	0.3810	2	0.4188	4	**0.4198**	4
Jazz	198	0.4379	4	**0.4452**	5	0.4451	4
C. elegans	453	0.4001	10	**0.4342**	12	0.4304	8
E-mail	1133	0.4796	13	**0.5738**	15	0.5275	8

Remark

The difficulty of optimizing modularity is that sampling from categorical distributions becomes harder with the increase of number of communities.

[1] Newman, M.E., 2006. Modularity and community structure in networks. Proceedings of the national academy of sciences 103, 85778582.
[2] Duch, J., Arenas, A., 2005. Community detection in complex networks using extremal optimization. Physical review E 72, 027104.

Conclusion

- In this work, we have presented a novel optimization method, Gumbel-softmax optimization (CSO), for solving combinatorial optimization problems on graph.
- Our experiment results show that our method has good performance on all four tasks and also take advantages in time complexity.
- However, there is much space to improve our algorithm on accuracy. We also note that our methods can find other applications, e.g., structure optimization.