基于样本的优化

张智杰[1,2]，孙晓明[1,2]，张家琳[1,2]，陈卫[3]
1. 中国科学院计算技术研究所，北京 100086；2. 中国科学院大学，北京 100049；
3. 微软亚洲研究院，北京 100080

摘要
基于样本的优化研究的是如何通过用于学习目标函数的样本数据直接优化目标函数。首先介绍这一问题的数学模型——样本优化模型，以及这个模型下的不可近似性结果；然后介绍若干方法和样本优化模型的变种，以绕过这个模型下的不可近似性结果，使得优化成为可能；接着着重介绍其中一个变种——结构化样本优化模型，并详细阐述该模型下的最大覆盖问题和影响力最大化问题的优化算法；最后总结全文，并展望这一问题的未来研究方向。

关键词
基于样本的优化；数据驱动的优化；结构化样本；最大覆盖问题；影响力最大化问题

中图分类号：TP30　　文献标识码：A　　doi: 10.11959/j.issn.2096-0271.2021051

Optimization from samples

ZHANG Zhijie[1,2], SUN Xiaoming[1,2], ZHANG Jialin[1,2], CHEN Wei[3]
1. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100086, China
2. University of Chinese Academy of Sciences, Beijing 100049, China
3. Microsoft Research Asia, Beijing 100080, China

Abstract
Optimization from samples studies how one can optimize objective functions from the sample data that one uses to learn them. Firstly, the mathematical model of this problem-optimization from samples model, as well as the inapproximability results under this model, was introduced. Secondly, some approaches and variants of OPS were introduced, in order to circumvent the impossibility results and make optimization possible. Thirdly, one of the variants-the optimization from structured samples model was focused on, and the algorithms for maximum coverage and influence maximization problem under it were introduced in details. Finally, the paper was concluded, and some future research directions for the problem were proposed.

Key words
optimization from samples, data-driven optimization, structured sample, maximum coverage problem, influence maximization problem

1 引言

为了解决实际生活中遇到的统筹优化问题，人们通常要建立一个问题模型，并确定模型的参数和优化目标函数，然后设计算法进行求解。然而，在大数据时代，许多应用场景无法提供足够的信息来确定模型参数和目标函数。人们只能通过观察到的历史样本数据来获取模型的信息，并进行优化。在这类场景下，人们通常使用机器学习的方法进行处理：首先近似地学习一个替代的目标函数，然后优化这个替代的函数。尽管这个方法在实际应用中获得了巨大的成功，但是在很多实际问题中，这个方法缺乏理论上的保证。事实上，它可能存在如下两个问题：① 即使针对原函数的优化问题是可求解或者可近似求解的，但是针对替代函数的优化问题也可能是不可近似的，这是因为替代函数可能丢失了一些原函数所具有的良好性质（如次模性）；② 即使替代函数是可近似的，而且从整体上看和原函数很接近，但是它的最优解相较于原函数的最优解也可能是一个很差的近似。这些担忧自然地引出了如下问题：人们是否真的能从一系列样本数据中求解目标函数的优化问题？

1.1 样本优化模型

组合优化问题通常具有如下形式：$\max\{f(S):S\in\mathcal{M}\}$，其中，目标函数$f:2^N\to\mathbb{R}$是一个定义在集合$N$的幂集合$2^N$上的集合函数，约束$\mathcal{M}\subseteq 2^N$。传统上，人们假定存在一个神谕（黑箱算法）$O_f$来访问目标函数$f$。将给定集合$S\subseteq N$作为输入，$O_f$会返回函数值$f(S)$。人们使用查询复杂度（即算法查询$O_f$的次数）来衡量算法的效率。这样的计算模型被称为查询模型。

为了回答基于样本的组合优化是否可能的问题，Balkanski E等人[1]定义了另一种计算模型——样本优化（optimization from samples，OPS）模型。

定义1（OPS模型） 给定参数$\alpha\in(0,1]$，如果存在算法A（不一定是多项式时间的），给定参数$\delta\in(0,1)$并将样本集$\{S_i,f(S_i)\}_{i=1}^{t}$作为输入，其中，S_i独立同分布于$\mathcal{D}$，$t\in\mathrm{poly}(|N|,1/\delta)$，$f\in\mathcal{F}$，算法$A$返回$S\in\mathcal{M}$，并满足

$$\Pr_{S_1,\cdots,S_t\in\mathcal{D}^t}\left[\mathbb{E}_A[f(S)]\geq\alpha\cdot\max_{T\in\mathcal{M}}f(T)\right]\geq 1-\delta \tag{1}$$

则称函数类$\mathcal{F}:2^N\to\mathbb{R}$在分布$\mathcal{D}$下对于约束$\mathcal{M}$是$\alpha$-可优化的。其中，$\alpha$被称为近似比，表示算法的解与最优解的比值。算法使用的样本数t被称为算法的采样复杂度。显然，样本分布$\mathcal{D}$会显著影响函数类$\mathcal{F}$在OPS模型下的可优化性。例如，当$\mathcal{D}$总是返回空集作为样本时，不可能对问题得到任何有意义的近似比。因此，人们转而希望在某些“合理的”样本分布下，优化是可能的。此外，对于在查询模型下具有常数近似比的问题，人们通常希望它在OPS模型下也具有常数近似比。对于这类问题，如果存在分布$\mathcal{D}$，当将给定多项式数量的独立同分布于$\mathcal{D}$的样本作为输入时，问题存在常数近似算法，则称它们（在OPS模型下）是可优化的；反之，则称它们是不可优化的。

样本优化模型在目标函数可优化且可学习的情况下最具研究价值。Balcan M F等人[2]首先定义了集合函数的PMAC（probably mostly approximately correct learnability）-可学习性。

定义2（PMAC-可学习性） 对于函数类$\mathcal{F}$和参数$\alpha\in(0,1)$，如果给定参数$\epsilon,\delta\in(0,1)$并将样本集$\{S_i,f(S_i)\}_{i=1}^{t}$作为

输入，其中，S_i独立同分布于$\mathcal{D}$，$t\in\mathrm{poly}\left(|N|,1/\delta,1/\epsilon\right)$，$f\in\mathcal{F}$，存在输出$\tilde{f}$，并满足

$$\Pr_{S_1,\cdots,S_t\sim\mathcal{D}^t}\left[\Pr_{S\sim\mathcal{D}}\left[\alpha\cdot\tilde{f}(S)\leqslant f(S)\leqslant\tilde{f}(S)\right]\geqslant 1-\epsilon\right]\geqslant 1-\delta \tag{2}$$

如果$\tilde{f}$在每个分布$\mathcal{D}$上都是α-PMAC-可学习的，则称$\mathcal{F}$在分布$\mathcal{D}$上是α-PMAC-可学习的。

由定义2可知，函数类$\mathcal{F}$是α-PMAC-可学习的意味着在大多数输入集合上（相对于分布$\mathcal{D}$而言），存在某种算法学习到的函数值与真实的函数值很接近。并且，人们通常要求这对于任意的分布$\mathcal{D}$均成立。而函数可优化性的定义只要求存在分布$\mathcal{D}$使之成立即可。

最后，覆盖函数和影响力函数是这一领域的重要研究对象，下面介绍它们的定义。给定二部图$G=(L,R,E)$，其中，L和R分别表示左右两边的点集，E表示点之间的边集。覆盖函数$f:2^L\to\mathbb{R}_+$定义为集合$S\subseteq L$的邻居的个数，即$f(S)=|N(S)|$。而最大覆盖问题要求选取最多k个左边的节点，并最大化它们覆盖的邻居数。换言之，它要求在基数约束下最大化一个覆盖函数，即$\max\{f(S):|S|\leqslant k\}$。

影响力函数是覆盖函数在一般有向图上的推广。它被定义在社交网络（有向图）$G=(V,E,\boldsymbol{p})$上，其中，V表示点集，E表示边集，$\boldsymbol{p}$表示概率向量，每条边$(u,v)\in E$具有概率$p_{uv}\in[0,1]$。每个节点存在激活和未激活两种状态。给定$t=0$的初始激活节点S_0（被称为种子集合），其他节点以如下方式被激活：在时刻t=1,2,3,…，首先令$S_t=S_{t-1}$；接着，对于每个节点$v\notin S_{t-1}$，令$N^{\mathrm{in}}(v)$表示v的入邻居，每个节点$u\in N^{\mathrm{in}}(v)\cap(S_{t-1}\setminus S_{t-2})$会以概率$p_{uv}$独立地激活节点$v$。$v$一旦被激活，就会被加入$S_t$中。节点被激活的过程是不可逆的，因此有$S_0\subseteq S_1\subseteq S_2\cdots$。一旦没有新的节点被激活，此过程终止。显然，这一过程最多进行$n-1$步。因此，可以使用$(S_0,S_1,\cdots,S_{n-1})$来表示激活节点的随机序列。上述传播过程被称为独立级联传播模型。给定S_0，定义$\Phi(S_0)=S_{n-1}$为最终的激活节点。影响力函数$\sigma:2^V\to\mathbb{R}_+$被定义为$\sigma(S)=\mathbb{E}\left[|\Phi(S)|\right]$，即种子集合$S$激活的节点数的期望。影响力最大化问题要求选取一个大小不超过k的种子集合，并最大化它激活的期望节点数，即$\max\{\sigma(S):|S|\leqslant k\}$[3]。

1.2 不可近似性结果

Balkanski E等人[1]在OPS模型下研究了最大覆盖问题的近似性，即在基数约束下最大化一个覆盖函数。此前，覆盖函数被证明是$(1-\epsilon)$-PMAC可学习的[4]。此外，在查询模型下，最大覆盖问题是$(1-\mathrm{e}^{-1})$-近似[5]的。因此，人们相信若采取“先学习后优化”的策略，最大覆盖问题在OPS模型下是可优化的。然而，令人惊讶的是，Balkanski E等人[1]证明了在OPS模型下，最大覆盖问题实际上是不存在常数近似的。换言之，尽管覆盖函数是可学习的，却不是可优化的。这使基于样本的组合优化问题得到一个否定性的回答。Balkanski E等人[1]的证明中构造了一类PMAC-可学习的覆盖函数，这类函数在绝大多数输入集合上能近似得很好，然而，这些近似良好的集合恰恰不是问题的最优解集，并且最优解与这些集合的函数值有较大差别。这解释了样本优化模型下不可近似性结果的由来。从概念上说，基于“先学习后优化”的思路，原问题通常可以被拆解为采样模型下的学习问题与查询模型下的优化问题。尽管这两个问题都是容易解决的，将它们结合起来却不能解决样本优化问题。这是因为这两个问题的子目标没有完全对应，

学习任务的子目标只要求在绝大多数集合上学得好，但这些学得好的集合恰恰是在优化意义上比较差的集合，因此对于原函数的优化没有帮助。

OPS模型十分容易被推广到其他优化问题上，而类似的不可近似性结果也出现在其他多个优化问题中。

众所周知，无约束次模函数①最小化问题可以在多项式时间内精确求解[6]。此外，当假定函数的取值在[0,1]之间时，可以证明以均等的概率返回空集或者全集，就能够得到一个1/2的加性近似[7]。针对这一问题，Balkanski E等人[7]定义了如下OPS模型。

定义3 给定参数$\epsilon\in[0,1/2]$，如果存在算法A，给定参数$\delta\in(0,1)$并将样本集$\{S_i,f(S_i)\}_{i=1}^{t}$作为输入，其中，S_i独立同分布于$\mathcal{D}$，$t\in\mathrm{poly}(|N|,1/\delta,1/\epsilon)$，$f\in\mathcal{F}$，算法$A$返回$S\subseteq N$，并满足

$$\Pr_{S_1,\cdots,S_t\in\mathcal{D}^t}\left[\mathbb{E}_A[f(S)]-\min_{T\subseteq N}f(T)\leqslant\epsilon\right]\geqslant 1-\delta \tag{3}$$

则称函数类$\mathcal{F}:2^N\to[0,1]$在分布$\mathcal{D}$下是ϵ-可优化的。

Balkanski E等人[7]证明了，在OPS模型下存在一类PAC-可学习的取值在[0,1]之间的次模函数，对于任意分布$\mathcal{D}$，将给定多项式数量的独立同分布于$\mathcal{D}$的样本作为输入，这类函数不存在$(1/2-O(1))$的加性近似。

上述不可近似性结果并不局限于组合优化中。众所周知，凸函数的最小化问题也是多项式时间可解的。针对这一问题，Balkanski E等人[8]定义了如下OPS模型。

定义4 给定参数$\epsilon\in[0,1/2]$，如果存在算法A，给定参数$\delta\in(0,1)$并将样本集$\{x_i,f(x_i)\}_{i=1}^{t}$作为输入，其中，x_i独立同分布于$\mathcal{D}$，$t\in\mathrm{poly}(n,1/\delta,1/\epsilon)$，$f\in\mathcal{F}$，算法$A$返回$\tilde{x}\in[0,1]^n$，并满足

$$\Pr_{x_1,\cdots,x_t\in\mathcal{D}^t}\left[\mathbb{E}_A[f(\tilde{x})]-\min_{x}f(x)\leqslant\epsilon\right]\geqslant 1-\delta \tag{4}$$

则称函数类$\mathcal{F}:[0,1]^n\to[0,1]$在分布$\mathcal{D}$下是$\epsilon$-可优化的。

Balkanski E等人[8]证明了，在OPS模型下，存在一类PAC-可学习的凸函数族$\mathcal{F}:[0,1]^n\to[0,1]$，对于任意分布$\mathcal{D}$，将给定多项式数量的独立同分布于$\mathcal{D}$的样本作为输入，这类函数不存在$(1/2-O(1))$的加性近似。这个界是紧的（相当于最优的），这是因为可以证明返回$x=(1/2,1/2,\cdots,1/2)$就能达到1/2的加性近似。

上述几个结果表明，许多在查询模型下可以优化的问题在采样模型下却是不可优化的，尽管从样本中可以学习到这些问题的目标函数。这说明了函数是可学习的并不意味着它是可优化的。

1.3 算法结果

后续有一系列工作尝试绕开OPS模型下的不可近似性结果[9-13]。这样的尝试大致可以分为3类。

第一类方法假设目标函数f拥有额外的性质。例如，Balkanski E等人[10]考虑了f是曲率为$c\in[0,1]$的单调②次模函数的情况。曲率[14]是衡量单调次模函数线性程度的一个度量。曲率越小，函数越接近线性。例如，线性函数和覆盖函数都满足单调性和次模性，但是线性函数的曲率为0，而覆盖函数的曲率为1。Balkanski E等人[10]证明了，在OPS模型下，当样本分布为约束$\mathcal{M}=\{S:|S|\leqslant k\}$上的均匀分布时，问题$\max\{f(S):|S|\leqslant k\}$存在$(1-c)/(1+c-c^2)$-近似，并且这个近似比是最优的。线性函数的曲率为0意味着线性函数即使在OPS模型下也是可以精确求解的。而覆盖函数的曲率为1，这个结果和OPS模型下最大覆盖问题的不可近似性并

① 对于$\forall S\subseteq T\subseteq N$，$u\notin T$，如果有$f(S\cup\{u\})-f(S)\geqslant f(T\cup\{u\})-f(T)$，则称函数$f:2^N\to\mathbb{R}$是次模的。

② 对于$\forall S\subseteq T$，如果$f(S)\leqslant f(T)$，则称函数$f:2^N\to\mathbb{R}$是单调的。

不矛盾。

影响力最大化问题是社交网络研究中的核心问题之一[3]。独立级联传播模型下的影响力函数是单调次模函数的一个重要实例，而覆盖函数又是此影响力函数的特例。因此，影响力函数在OPS模型下也是不可优化的。由于影响力函数被定义在社交网络$G=(V,E,\boldsymbol{p})$上，为了绕开OPS模型下的不可近似性结果，Balkanski E等人[9]考虑了带有社区结构的社交网络上的影响力函数。更具体地说，他们假设G是通过随机区块模型（stochastic block model）生成的，因此G可被高概率地划分为若干社区$C_1,C_2,\cdots,C_\ell$，且社区内部的边比较稠密，社区之间的边比较稀疏。他们证明了，对于这样生成的社交网络和约束$\mathcal{M}=\{S:|S|\leqslant k\}$上的均匀分布，影响力最大化问题存在常数近似算法。

可以发现，上述方法不改变OPS模型本身，但是通常要求目标函数具有良好的性质，因此其适用范围有所限制。

第二类方法弱化了优化目标。Rosenfeld N等人[13]提出了OPS模型的一个变种版本，称之为DOPS（distributional optimization from samples）模型。

定义5（DOPS模型） 给定参数$\alpha\in[0,1]$，如果存在算法A，对于任意分布$\mathcal{D}$，给定独立同分布于$\mathcal{D}$的样本集$\mathcal{T}=\{T_i\}_{i=1}^m$，参数$\epsilon,\delta\in(0,1)$并将另一批样本集$\{S_i,f(S_i)\}_{i=1}^t$作为输入，其中，$S_i$独立同分布于$\mathcal{D}$，$f\in\mathcal{F}$，$t\in\mathrm{poly}(|N|,m,1/\delta,1/\epsilon)$，算法$A$返回$A(\mathcal{T})\in\mathcal{T}$，并满足

$$\Pr_{S_1,\cdots,S_t\in\mathcal{D}^t}\left[\Pr_{\mathcal{T}\sim\mathcal{D}^m}\left[\mathbb{E}_A\left[f(A(\mathcal{T}))\right]\geqslant\alpha\cdot\max_{T\in\mathcal{T}}f(T)\right]\geqslant 1-\epsilon\right]\geqslant 1-\delta \tag{5}$$

则称函数类$\mathcal{F}:2^N\to\mathbb{R}$是$\alpha$-可优化的。

可以发现，在DOPS模型中，不存在约束$\mathcal{M}$，优化目标也不是寻找全局最优解$\max_{S\in\mathcal{M}}f(S)$。模型的优化目标是在函数值未知的大小为$m$的样本集$\mathcal{T}$中寻找函数值最大的样本。因此，优化目标取决于样本分布$\mathcal{D}$。算法可以使用另一批函数值已知的样本集$\{S_i,f(S_i)\}_{i=1}^t$来收集函数f的信息，并最终达成上述目标。需要注意的是，在OPS模型中，要求样本数t关于基集合的大小$|N|$是多项式的，$|N|$表示问题规模。因此，作为类比，在DOPS模型中，要求t关于m是多项式的。

Rosenfeld N等人[13]证明了一个集合函数类在DOPS模型下是α-可优化的，当且仅当它是α-PMAC-可学习的。这种解决方式恰恰利用了之前“可学习但不可优化”的矛盾之处。函数可学习说明替代函数在绝大多数地方和目标函数很接近，而这里的绝大多数是相对于分布$\mathcal{D}$而言的。如果分布$\mathcal{D}$较偏离函数最优解，则会导致即使替代函数整体上接近目标函数，在最优解附近可能也会偏离较远，进而使得全局优化目标很难达成。与之相对地，只针对函数值未知的样本定义的优化目标会更容易达成。但是这个解决方式最终达成的优化目标依赖于样本数据的分布，并不符合通常对集合函数的优化问题的要求。人们仍然希望相对合理的分布$\mathcal{D}$能为原目标函数的全局最优解提供一定的理论保证。

第三类方法既不假定目标函数满足额外的性质，也不弱化优化目标，而是假设样本携带额外的结构信息，这样的样本被称为结构化样本。Chen W等人[11]首先研究了这种方法，针对覆盖函数提出了OPS模型的一个变种版本——结构化样本优化（optimization from structured samples，OPSS）模型。

定义6（OPSS模型） 给定参数$\alpha\in[0,1]$，如果存在算法A，给定参数$\delta\in(0,1)$并将样本集$\{S_i,N_G(S_i)\}_{i=1}^t$作为输入，其中，S_i独立

同分布于$\mathcal{D}$，$t\in\mathrm{poly}\left(|G|,1/\delta\right)$，$N_G(S_i)$为$S_i$在二部图$G$上的邻居，算法$A$返回$S\in\mathcal{M}$，并满足

$$\Pr_{S_1,\cdots,S_m\in\mathcal{D}^m}\left[\mathbb{E}_A\left[f(S)\right]\geqslant\alpha\cdot\max_{T\in\mathcal{M}}f(T)\right]\geqslant1-\delta \tag{6}$$

则称覆盖函数类$\mathcal{F}=\left\{f_G:2^N\to\mathbb{R}\right\}$在分布$\mathcal{D}$下对于约束$\mathcal{M}$是$\alpha$-可优化的。

在OPSS模型中，算法不仅知道S_i覆盖的邻居数，还知道它具体覆盖了哪些节点，因此掌握了关于函数结构的部分信息。Chen W等人[11]证明了，当分布$\mathcal{D}$满足可行性、多项式大小的采样概率和负相关性这3个条件时，最大覆盖问题在OPSS模型下存在常数近似。因此，通过假设样本是结构化的，所得结果绕过了OPS模型下的不可近似性结果。

这一结果后来被推广到独立级联模型下的影响力函数最大化问题[12]。在OPSS模型下，算法的输入是结构化样本$\left\{S_{i,0},S_{i,1},\cdots,S_{i,n-1}\right\}_{i=1}^{t}$，其中$S_{i,0}\subseteq V$独立同分布于$\mathcal{D}$，给定$S_{i,0}$，$\left\{S_{i,1},\cdots,S_{i,n-1}\right\}$的产生遵循独立级联模型的传播过程。Chen W等人[12]证明了当分布是乘积分布时，影响力最大化问题存在常数近似。

可以发现，由于不同目标函数的结构各不相同，因此难以定义通用的OPSS模型，需要基于各个函数的结构特点给出具有针对性的定义。本文将着重介绍OPSS模型下的算法结果。

2 OPSS模型

2.1 最大覆盖问题

Chen W等人[11]为OPSS模型下的最大覆盖问题设计了如下算法。

算法1：最大覆盖问题的OPSS算法

输入：样本$\left\{S_i,N_G(S_i)\right\}_{i=1}^{t}$和约束$k\in\mathbb{N}_+$

令$T_1=S_1$

构造一个替代的二部图$\tilde{G}=\left(L,R,\tilde{E}\right)$，使得对于每个节点$u\in L$，$N_{\tilde{G}}(u)=\bigcup_{i:u\in S_i}N_G(S_i)$

令$T_2=A\left(\tilde{G},k\right)$，其中$A$表示标准最大覆盖问题的$k$-近似算法

以等概率返回T_1和T_2中的一个

算法1以相等的概率返回两个可行解T_1和T_2中的一个，其中$T_1=S_1$就是第一个样本，而T_2是通过在二部图$\tilde{G}$上运行标准最大覆盖问题的k-近似算法得到的。二部图$\tilde{G}$是原图G的一个近似，它是由样本$\left\{S_i,N_G(S_i)\right\}_{i=1}^{t}$构造出来的。对于节点$u\in L$，定义它在$\tilde{G}$上的邻居为$N_{\tilde{G}}(u)=\bigcup_{i:u\in S_i}N_G(S_i)$，用来近似它的真实邻居$N_G(S_i)$。

直观的算法设计如下：如果某个单元素集$\{u\}$从分布$\mathcal{D}$中被采样出来，那么算法能完全知晓$N_G(u)$的信息。然而，从$\mathcal{D}$中采样出来的可能是一个大集S，对于节点$u\in S$，$N_G(u)$的信息被隐藏在$N_G(S)$中。幸运的是，如果节点同时属于两个样本S_1、S_2，那么有$N_G(u)\subseteq N_G(S_1)\cap N_G(S_2)$。因此，算法1使用包含节点的样本的邻居的交集作为节点u的真实邻居的估计，以便尽可能地揭露u的真实邻居的信息。

上述构造表明，$N_{\tilde{G}}(u)$只可能高估$N_G(u)$，即$N_G(u)\subseteq N_{\tilde{G}}(u)$。然而，$N_{\tilde{G}}(u)$仍然可能不是$N_G(u)$的一个好的估计，即$N_{\tilde{G}}(u)\setminus N_G(u)$可能太大。一个极端的例子如下：假设对于某个节点$v\in L$，$\Pr[v\in S]=1$，那么，$N_{\tilde{G}}(u)$总是包含$N_G(u)\cup N_G(v)$中的所有元素，因此可能比$N_G(u)$大得多。由此可见，$T_2$在原图$G$上可能不是一个好的解。此外，观察上述例子可知，$N_{\tilde{G}}(u)\setminus N_G(u)$中的元素之所以会出现，是因为它们的邻居以高概率出现在样本S中。这意味着样本$T_1=S_1$能够以高概率覆盖学习误差

$N_{\bar{G}}(u)\setminus N_G(u)$。因此，$T_1\cup T_2$是原图$G$上一个好的近似解。为了保证可行性，算法随机选取其中的一个近似解，在期望的意义下，仍然能够得到一个好的近似解。

为了对算法1进行严格的理论分析，需要假定算法在如下假设下运行。

假设1 假设2^L上的分布$\mathcal{D}$满足如下3个条件。

- 可行性。样本$S\sim\mathcal{D}$总是可行的，即$|S|\leqslant k$。
- 多项式大小的采样概率。存在常数$c>0$，对于每个节点$u\in L$，$\Pr[u\in S]\geqslant 1/|L|^c$。
- 负相关性。对于$S\sim\mathcal{D}$，随机变量$X_u=1_{u\in S}$是负相关的，即$\forall T\subseteq L,u\notin T$，$\Pr\left[\vee_{u'\in T}(X_{u'}=1)|\ X_u=1\right]\leqslant\Pr\left[\vee_{u'\in T}(X_{u'}=1)\right]$。

上述3个条件都是非常自然的。特别地，第二个条件意味着L中的所有元素都有足够的概率被采样到。第三个条件直观上意味着u出现在样本中这一事件的发生会减少其他节点出现在样本中的概率。显然，这个条件降低了许多个节点同时出现在样本中的概率，有助于算法1揭示特定节点的邻居的信息。一些典型的分布均满足假设1，例如$\mathcal{F}_{\leqslant k}=\{S:|S|\leqslant k\}$上的均匀分布$\mathcal{D}_{\leqslant k}$以及$\mathcal{F}_{=k}=\{S:|S|=k\}$上的均匀分布$\mathcal{D}_k$。基于假设1，可以证明：

定理1 对于任意$\delta\in(0,1)$，给定任意标准最大覆盖问题的k-近似算法，令表示算法1返回的解，OPT表示原图G上的最优解。如果分布$\mathcal{D}$满足假设1且样本数$t\geqslant\frac{4|L|^c|R|}{\delta}\ln\frac{4|L||R|}{\delta}$，其中，$c$是假设1中的参数，那么

$$\Pr_{S_1,\cdots,S_t\sim\mathcal{D}}\left[\mathbb{E}\left[f_G(\mathrm{ALG})\right]\geqslant\frac{k}{2}\cdot f_G(\mathrm{OPT})\right]\geqslant 1-\delta \tag{7}$$

如果只要求常数近似比，则假设1中“可行性”的条件可以被放宽为$|S|=O(k)$。Chen W等人[11]还证明了如下结论。

- 当样本分布$\mathcal{D}$为均匀分布$\mathcal{D}_{\leqslant k}$或$\mathcal{D}_k$时，存在算法能够达到$(k-\epsilon)$-近似比。
- 存在满足假设1的某个分布$\mathcal{D}$，在这一分布下，不存在$(1/2+O(1))$-近似算法。这意味着当允许调用暴力搜索算法（$k=1$）时，算法1是OPSS模型下的最优算法。
- 移除假设1中的任意一个条件，存在满足剩下两个条件的某个分布$\mathcal{D}$，在这一分布下不存在常数近似算法。这意味着为了得到OPSS模型下最大覆盖问题的常数近似算法，假设1的3个条件都是必须满足的。

2.2 影响力最大化问题

Chen W等人[12]采取如下框架求解OPSS模型下的影响力最大化问题：首先学习边的概率，然后在学习到的社交网络上求解影响力最大化问题。其中，学习边概率的任务被称为网络推断（network inference）问题，它的严格定义如下：给定结构化样本$\{S_{i,0},S_{i,1},\cdots,S_{i,n-1}\}_{i=1}^t$，其中$S_{i,0}\subseteq V$独立同分布于$\mathcal{D}$，给定$S_{i,0}$，$\{S_{i,1},\cdots,S_{i,n-1}\}$的产生遵循独立级联模型的传播过程，$t\in\mathrm{poly}(|G|,1/\epsilon,1/\delta)$，要求计算一个概率向量$\hat{\boldsymbol{p}}$，使得

$$\Pr\left[\forall u,v\in V,|\hat{p}_{uv}-p_{uv}|\leqslant\epsilon\right]\geqslant 1-\delta \tag{8}$$

为了求解上述问题，Chen W等人[12]假设产生种子的分布是乘积分布，即对于样本$S\sim\mathcal{D}$，事件$u\in S$之间是相互独立的。

在是乘积分布的假设下，Chen W等人[12]提出了一个高效求解网络推断问题的方案。为了描述这一方案，需要定义一些符号。记$\{S_0,S_1,\cdots,S_{n-1}\}$是激活节点的随机序列。对于节点$u$，定义$q_u=\Pr[u\in S]$，表示$u$被选为种子的概率。对于节点$v\in V$，定义$\mathrm{ap}(v)=\Pr[v\in S_1]$，表示$v$在一个时间步之内

被激活的概率。节点v既有可能因为被选为种子而激活，也有可能被种子激活。因此，$\mathrm{ap}(v)$定义中的随机性既来自种子分布$\mathcal{D}$，也来自图G上第一个时间步之内的传播过程。此外，定义$\mathrm{ap}(v|\overline{u})=\Pr[v\in S_1|u\notin S_0]$，表示节点$u$不是种子时相应的条件概率。Chen W等人[12]的关键性观察如下。

引理1 给定任意$u,v\in V$且$u\neq v$，

$$p_{uv}=\frac{\mathrm{ap}(v)-\mathrm{ap}(v|\overline{u})}{q_u\left(1-\mathrm{ap}(v|\overline{u})\right)} \quad (9)$$

引理1的证明思路如下：在一个时间步之内，节点v或者被节点u之外的节点激活，或者被节点u激活。由于$\mathcal{D}$是乘积分布，可以得到

$$\mathrm{ap}(v)=\mathrm{ap}(v|\overline{u})+\left(1-\mathrm{ap}(v|\overline{u})\right)q_u p_{uv} \quad (10)$$

重新排列式(10)便可以得到引理1中的结果。

有了引理1后，就可以通过估计q_u、$\mathrm{ap}(v)$和$\mathrm{ap}(v|\overline{u})$来估计$p_{uv}$。

算法2：网络推断算法

输入：样本$\left\{S_{i,0},S_{i,1},\cdots,S_{i,n-1}\right\}_{i=1}^{t}$

对于所有$u,v\in V$，分别估计$\hat{q}_u$、$\widehat{\mathrm{ap}}(v)$和$\widehat{\mathrm{ap}}(v|\overline{u})$

令 $\hat{p}_{uv}=\dfrac{\widehat{\mathrm{ap}}(v)-\widehat{\mathrm{ap}}(v|\overline{u})}{\hat{q}_u\left(1-\widehat{\mathrm{ap}}(v|\overline{u})\right)}$

返回$\{\hat{p}_{uv}\}_{u,v\in V}$

在算法2中，$\hat{q}_u$、$\widehat{\mathrm{ap}}(v)$和$\widehat{\mathrm{ap}}(v|\overline{u})$分别表示$q_u$、$\mathrm{ap}(v)$和$\mathrm{ap}(v|\overline{u})$的估计，它们的计算方法如下：令$t_u=\left|\{i\in[t]:u\in S_{i,0}\}\right|$表示$u$是种子的样本的数量，$t^v=\left|\{i\in[t]:v\in S_{i,1}\}\right|$表示$v$在一步之内被激活的样本的数量，$t\dfrac{v}{u}=\left|\{i\in[t]:u\notin S_{i,0},v\in S_{i,1}\}\right|$表示$u$不是种子且$v$在一步之内被激活的样本的数量，那么$\hat{q}_u=t_u/t$、$\widehat{\mathrm{ap}}(v)=t^v/t$且$\widehat{\mathrm{ap}}(v|\overline{u})=t\dfrac{v}{u}/t_{\overline{u}}$。为了保证参数估计的准确性，算法2需要在如下假设下运行。

假设2 存在参数$\alpha\in(0,1]$和$\gamma\in(0,1/2]$，使得

- 对于所有$v\in V$，$\mathrm{ap}(v)\leqslant 1-\alpha$；
- 对于所有$u\in V$，$\gamma\leqslant q_u\leqslant 1-\gamma$。

在假设2下，可以证明如下定理。

定理2 在假设2下，令$\{\hat{p}_{uv}\}_{u,v\in V}$表示算法2返回的边的概率，$\{p_{uv}\}_{u,v\in V}$表示真实的边概率。给定$\epsilon,\delta\in(0,1)$，如果样本的数量$t\geqslant\dfrac{256}{\epsilon^2\alpha^2\gamma^3}\ln\dfrac{12n}{\delta}$，那么

$$\Pr\left[\forall u,v\in V,\left|\hat{p}_{uv}-p_{uv}\right|\leqslant\epsilon\right]\geqslant 1-\delta \quad (11)$$

接着介绍如何利用网络推断算法解决OPSS模型下的影响力最大化问题。Narasimhan H等人[15]证明了如下引理。

引理2 给定$S\subseteq V$和任意两个概率向量$\boldsymbol{p}$、$\hat{\boldsymbol{p}}$，并满足$\|\boldsymbol{p}-\hat{\boldsymbol{p}}\|_1\leqslant\epsilon/n$，用$\sigma^p$表示定义在图$G=(V,E,\boldsymbol{p})$上的影响力函数，那么

$$\left|\sigma^p(S)-\sigma^{\hat{p}}(S)\right|\leqslant\epsilon \quad (12)$$

可以发现，使用网络推断算法估计边概率，使得误差不大于$\epsilon k/(2n^3)$，结合引理2，就可以证明在学习到的社交网络上运行任何k-近似的影响力最大化算法，可以得到原图上的一个$(k-\epsilon)$-近似解。然而，这样的算法继承了网络推断算法的假设，因此对网络的结构和参数有所要求（注意到$\mathrm{ap}(v)\leqslant 1-\alpha$是一个关于分布$\mathcal{D}$和网络参数的条件），这意味着对影响力函数假设了额外的性质，有悖于定义OPSS模型的初衷。

为了得到一个在任何社交网络上均能运行的OPSS算法，Chen W等人[12]采取了处理最大覆盖问题时所用的技术。具体地说，算法首先对每个节点$v\in V$估计$\mathrm{ap}(v)$的值，并比较它们与给定阈值的大小。如果$\mathrm{ap}(v)$大于给定阈值，那就意味着节点v以高概率在一步之内被激活，此时，算法将它的所有入边的概率设置为1；如果$\mathrm{ap}(v)$小于此阈值，那么假设2的条件被满足，可以使用网络推断算法估计v的所有入边的概率。经过上述步骤可以得到一张新图。算法在

这张新图上运行k-近似的影响力最大化算法，并得到一个解。算法使用第一个样本作为另一个解，最终以等概率返回两个解中的一个。

算法3：影响力最大化问题的OPSS算法

输入：样本$\{S_{i,0},S_{i,1},\cdots,S_{i,n-1}\}_{i=1}^{t}$和约束$k\in\mathbb{N}_+$

for 每个$v\in V$ do

估计$\widehat{\mathrm{ap}}(v)$

if $\widehat{\mathrm{ap}}(v)\geqslant 1-\delta/4n$ then

对于所有$u\in V$，$\hat{p}_{uv}=1$

else

对于所有$u,v\in V$，分别估计$\hat{q}_u$、$\widehat{\mathrm{ap}}(v)$和$\widehat{\mathrm{ap}}(v|\bar{u})$

$$\hat{p}_{uv}=\frac{\widehat{\mathrm{ap}}(v)-\widehat{\mathrm{ap}}(v|\bar{u})}{\hat{q}_u\left(1-\widehat{\mathrm{ap}}(v|\bar{u})\right)}$$

end if

end for

令$T_1=S_{i,0}$，$T_2=A\left(G(\hat{p}),k\right)$

以等概率返回T_1和T_2中的一个

上述算法与最大覆盖问题的OPSS算法（算法1）的设计思路是一致的。$\mathrm{ap}(v)$接近1意味着节点v以高概率在一步之内被激活，因此网络推断算法的假设不被满足，算法无法学习到节点v的入边概率。幸运的是，这同时意味着任意一个样本都能够以高概率激活节点v。因此算法无须学习节点v的入边概率，而是直接把它们设置为1。这样的设置几乎不改变样本激活节点v的概率。

上述设计使得算法能够处理任意$\mathrm{ap}(v)$，从而移除了假设2中关于$\mathrm{ap}(v)$的条件。而作为代价，算法需要假设样本的期望大小不超过$k/2$，从而保证采样出来的样本高概率是可行的。因此，算法需要在下面的假设下运行。

假设3 存在参数$\gamma\in(0,1/2]$，使得

- 对于$S\sim\mathcal{D}$，$\mathbb{E}\left[|S|\right]=\sum_{u\in V}q_u\leqslant k/2$；
- 对于所有$u\in V$，$\gamma\leqslant q_u\leqslant 1-\gamma$。

显然，假设3的两个条件都是针对样本分布的，不针对社交网络。因此，算法3在任何社交网络都可以成功运行。可以证明，算法3是一个常数近似算法。

定理3 对于任意$\epsilon,\delta\in(0,1)$，给定任意标准影响力最大化问题的k-近似算法，令表示算法3返回的解表示原图G上的最优解。如果分布$\mathcal{D}$满足假设3且样本数$t=\tilde{\Omega}\left(\frac{n^8}{k^2}\right)$，那么

$$\Pr_{S_1,\cdots,S_t\sim\mathcal{D}}\left[\mathbb{E}\left[\sigma(\mathrm{ALG})\right]\geqslant\frac{k-\epsilon}{2}\cdot\sigma(\mathrm{OPT})\right]\geqslant 1-\delta \tag{13}$$

最后，若把假设3中的第一个条件改为“存在常数$c>0$，使得$\mathbb{E}\left[|S|\right]\leqslant ck$”，仍然能够通过修改算法得到一个常数近似比。如果$c=\epsilon\in(0,1/3)$，那么可以修改算法得到一个$(k-3\epsilon)$-近似。

3 未来研究方向

样本优化仍然有许多可以进一步研究的方向。

- 针对OPSS模型下的最大覆盖问题和影响力最大化问题，降低现有算法的查询复杂度。此外，目前影响力最大化的OPSS算法假设样本分布是乘积分布。如何突破这样的独立采样假设是一个十分重要的开放问题，一种可能的方法是将文中的方法与极大似然估计方法结合。
- 对更多的目标函数定义适当的结构化样本，并研究它们在OPSS模型下的近似性。一个直接的例子是线性阈值模型下的影响力最大化函数（笔者已经得到了这方面的初步结果）。可以发现，OPSS模型是一个表达能力丰富、能够挖掘函数内在结构性质的

模型。因此，在OPSS模型下研究各类优化问题是一个十分有潜力的研究方向。

● 研究更多的方法以绕开标准OPS模型下的不可近似性结果。更多这样的研究一方面有助于人们应对不同的应用场景，另一方面有助于人们理解样本数据与函数可优化性的内在联系。

● 研究从样本中优化凸函数的可能性。目前所有绕开OPS模型不可近似性结果的方法都是针对集合函数而言的。对于实函数，尤其是具有良好优化性质的凸函数，尚没有这方面的研究。考虑到凸函数在连续优化中的重要地位，对它的进一步研究是十分必要的。

4 结束语

本文总结了OPS模型及其变种模型下的不可近似性结果和算法成果，并展望了相关的未来研究方向。OPS模型是数据驱动的优化的重要研究方法之一，值得进行更加深入的研究。

参考文献:

[1] BALKANSKI E, RUBINSTEIN A, SINGER Y. The limitations of optimization from samples[C]//Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing. New York: ACM Press, 2017: 1016–1027.

[2] BALCAN M F, HARVEY N J A. Learning submodular functions[C]//Proceedings of the 43rd Annual ACM Symposium on Theory of Computing. New York: ACM Press, 2011: 793–802.

[3] KEMPE D, KLEINBERG J, TARDOS É. Maximizing the spread of influence through a social network[C]//Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2003: 137–146.

[4] BADANIDIYURU A, DOBZINSKI S, FU H, et al. Sketching valuation functions[C]//Proceedings of the 23rd Annual ACM-SIAM Symposium on Discrete Algorithms. Philadelphia: Society for Industrial and Applied Mathematics, 2012.

[5] NEMHAUSER G L, WOLSEY L A, FISHER M L. An analysis of approximations for maximizing submodular set functions–I[J]. Mathematical Programming, 1978, 14(1): 265–294.

[6] GRÖTSCHEL M, LOVÁSZ L, SCHRIJVER A. The ellipsoid method and its consequences in combinatorial optimization[J]. Combinatorica, 1981, 1(2): 169–197.

[7] BALKANSKI E, SINGER Y. Minimizing a submodular function from samples[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2017: 814–822.

[8] BALKANSKI E, SINGER Y. The sample complexity of optimizing a convex function[C]//Proceedings of the 2017 Conference on Learning Theory. [S.l.:s.n.], 2017: 275–301.

[9] BALKANSKI E, IMMORLICA N, SINGER Y. The importance of communities for learning to influence[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2017: 5864–5873.

[10] BALKANSKI E, RUBINSTEIN A, SINGER Y. The power of optimization from samples[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2016: 4024–4032.

[11] CHEN W, SUN X M, ZHANG J L, et al. Optimization from structured samples for coverage functions[C]//Proceedings of 2020 International Conference on Machine Learning. [S.l.:s.n.], 2020.

[12] CHEN W, SUN X M, ZHANG J L, et al. Network inference and influence maximization from samples[C]//Proceedings of 2021 International Conference on Machine Learning. [S.l.:s.n.], 2021.
[13] ROSENFELD N, BALKANSKI E, GLOBERSON A, et al. Learning to optimize combinatorial functions[C]// Proceedings of 2021 International Conference on Machine Learning. [S.l.:s.n.], 2018: 4374-4383.
[14] CONFORTI M, CORNUÉJOLS G. Submodular set functions, matroids and the greedy algorithm: tight worst-case bounds and some generalizations of the Rado-Edmonds theorem[J]. Discrete Applied Mathematics, 1984, 7(3): 251-274.
[15] NARASIMHAN H, PARKES D C, SINGER Y. Learnability of influence in networks[C]// Proceedings of the 29th Annual Conference on Neural Information Processing Systems. [S.l.:s.n.], 2015: 3168-3176.

作者简介

张智杰(1995-),男,中国科学院计算技术研究所博士生,主要研究方向为次模优化、公平分配等。

孙晓明(1978-),男,中国科学院计算技术研究所研究员,主要研究方向为量子计算、算法复杂性、社会网络近似算法、通信复杂性、判定树复杂性、组合数学等。

张家琳(1983-),女,中国科学院计算技术研究所研究员,主要研究方向为理论计算机科学、量子计算、近似算法、在线算法、算法博弈论。

陈卫(1968-),男,博士,微软亚洲研究院高级研究员,中国科学院计算技术研究所客座研究员,中国计算机学会大数据专家委员会和理论计算机科学专业委员会委员,IEEE Fellow,《大数据》期刊编委。主要研究方向为在线学习和优化、社交和信息网络、网络博奕论和经济学、分布式计算、容错等。

收稿日期: 2021-07-23
通信作者: 张家琳, zhangjialin@ict.ac.cn
基金项目: 国家自然科学基金资助项目(No. 61832003, No. 61872334),中国科学院先导研究专项资助项目(No. XDA27000000)
Foundation Items: The National Natural Science Foundation of China (No. 61832003, No. 61872334), The Strategic Priority Research Program of Chinese Academy of Sciences (No. XDA27000000)