(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 105631698 A
(43)申请公布日 2016.06.01

(21)申请号 201510484987.3

(22)申请日 2015.08.07

(30)优先权数据
14/552,047 2014.11.24 US

(71)申请人 奥多比公司
地址 美国加利福尼亚州

(72)发明人 P·S·托马斯 G·西奥查奥斯 M·加瓦姆扎德

(74)专利代理机构 北京市金杜律师事务所 11256
代理人 酆迅

(51)Int.Cl.
G06Q 30/02(2012.01)

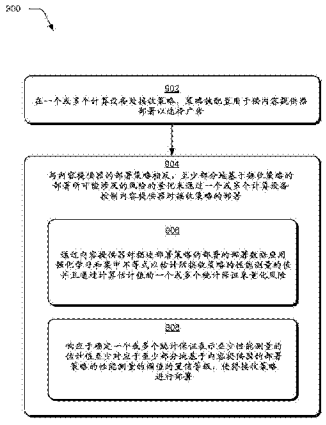
权利要求书2页 说明书18页 附图15页

(54)发明名称

用于策略部署的风险量化

(57)摘要

本发明的各实施方式总体上涉及用于策略部署的风险量化。具体地，描述了风险量化、策略搜索和自动安全策略部署。在一个或多个实施方式中，技术用于确定策略的安全性，诸如表示新策略将相对于当前部署的策略显示出增加的性能测量（例如，交互或转换）的置信等级。为了进行这种确定，利用强化学习和集中不等式，其生成和约束关于策略的性能测量的置信值并由此提供该性能的统计保证。这些技术可用于量化策略部署中的风险、基于估计性能和该估计中的置信等级选择策略用于部署（例如，其可以包括使用策略空间来减少数据处理量）、用于通过交互创建新策略（其中策略的参数被迭代调整并且评估这些调整的效果）等。



CN 105631698 A

1. 一种用于在用于识别和部署潜在的数字广告活动的数字介质环境中优化活动选择的方法，其中活动可以根据需求被改变、移除或替换，所述方法包括：

在一个或多个计算设备处接收策略，所述策略被配置为通过内容提供器进行部署来选择广告；以及

与所述内容提供器的部署策略相反，至少部分地基于在对接收策略的部署中可能涉及的风险的量化，通过所述一个或多个计算设备控制所述内容提供器对所述接收策略的部署，所述控制包括：

通过所述内容提供器对描述部署策略的部署的部署数据应用强化学习和集中不等式以估计所接收策略的性能测量的值并且通过计算估计值的一个或多个统计保证来量化风险；以及

响应于确定所述一个或多个统计保证表示至少所述性能测量的估计值至少对应于至少部分地基于所述内容提供器的部署策略的性能测量的阈值的置信等级，使得所接收的策略进行部署。

2. 根据权利要求 1 所述的方法，其中所述阈值至少部分地基于所述部署策略的测量性能和设置裕度。

3. 根据权利要求 2 所述的方法，其中设置所述阈值，使得所述接收策略的所述估计值显示出相对于所述部署策略的性能测量的改进。

4. 根据权利要求 1 所述的方法，其中所述置信等级和所述阈值是经由与所述一个或多个计算设备的用户接口的交互而用户可限定的。

5. 根据权利要求 1 所述的方法，其中所述集中不等式被配置为将限定阈值上方的估计值移动到所述限定阈值处。

6. 根据权利要求 1 所述的方法，其中所述集中不等式被配置为与所述估计值的随机变量的范围无关。

7. 根据权利要求 1 所述的方法，其中所述集中不等式被配置为塌陷所述估计值的随机变量分布的尾部，标准化所述随机变量分布，并且然后生成下限，从所述下限中提取所述估计值的原始随机变量的均匀平均值的下限。

8. 根据权利要求 1 所述的方法，其中所述策略被配置为被所述内容提供器用于至少部分地基于与访问内容的请求相关联的特性来选择用于与内容包括在一起的广告。

9. 根据权利要求 8 所述的方法，其中与所述请求相关联的所述特性包括发起所述请求的用户或设备的特性或者所述请求自身的特性。

10. 根据权利要求 8 所述的方法，其中使用特征矢量来表示所述特性。

11. 根据权利要求 1 所述的方法，其中所接收的部署数据不描述通过一个或多个实体对所述接收策略的部署。

12. 根据权利要求 1 所述的方法，其中所接收的部署数据还描述对所述接收策略的部署。

13. 一种系统，包括：

一个或多个计算设备，被配置为执行操作，所述操作包括与所述部署策略相反，至少部分地基于在对接收策略的部署中可能涉及的风险的量化来控制对所述接收策略的部署，所述控制包括：

对描述对所述部署策略的部署的部署数据使用强化学习和集中不等式，以估计所述接收策略的性能测量值并通过计算关于估计值的一个或多个统计保证来量化所述风险；以及

响应于确定所述一个或多个统计保证表示至少所述性能测量的估计值至少对应于至少部分地基于所述部署策略的性能测量的阈值的置信等级，使得用所接收的策略替换对所述部署策略的部署。

14. 根据权利要求 13 所述的系统，其中所述阈值至少部分地基于所述部署策略的测量性能和设置裕度。

15. 根据权利要求 14 所述的系统，其中设置所述阈值，使得所述接收策略的所述估计值显示出相对于所述部署策略的性能测量的改进。

16. 根据权利要求 13 所述的系统，其中所述置信等级和所述阈值是经由与所述一个或多个计算设备的用户接口的交互而用户可限定的。

17. 一种内容提供器，包括被配置为执行操作的一个或多个计算设备，所述操作包括：

基于与针对内容的请求相关联的一个或多个特性来部署策略以选择将与内容一起被包括的广告；以及

用另一策略替换所述部署策略，所述另一策略通过使用强化学习和集中不等式以处理部署数据并确定所述一个或多个统计保证表示至少所接收策略的性能测量的估计值至少对应于至少部分地基于所述部署策略的性能测量的阈值的置信等级来选择。

18. 根据权利要求 17 所述的内容提供器，其中所述阈值至少部分地基于所述部署策略的性能测量和设置裕度。

19. 根据权利要求 17 所述的内容提供器，其中设置所述阈值，使得所述接收策略的所述估计值显示出相对于所述部署策略的性能测量的改进。

20. 根据权利要求 17 所述的内容提供器，其中所述置信等级和所述阈值是经由与所述一个或多个计算设备的用户接口的交互而用户可限定的。

用于策略部署的风险量化

技术领域

[0001] 本发明的各实施方式总体上涉及计算机领域，具体地涉及用于策略部署的风险量化。

背景技术

[0002] 用户经由因特网接触越来越多的各种内容（诸如网页）。一种用于使内容提供器提供这些内容货币化的技术是通过加入广告。例如，用户可以访问包括各种广告的网页并且可以选择（例如，“点击”）感兴趣的广告来得到关于该广告中提到的商品或服务的附加信息。因此，商品或服务的提供器可以向内容提供器提供报酬用于包括广告以及用于潜在消费者选择广告。

[0003] 可以使用策略以选择哪些广告被呈现给特定用户或用户组。例如，可以收集描述用户、用户与内容的交互等的数据。然后，该数据可被策略用于确定哪些广告被呈献给用户，诸如增加用户将选择所包括广告中的一个或多个的可能性。然而，用于选择策略部署的传统技术不具有用于保证新选择的策略将比当前策略执行得更加好的机制。

[0004] 例如，存在被称为“策略脱离（off-policy）评价技术”的用于估计策略性能的传统解决方案。然而，这些传统的策略脱离评价技术不能以任何方式约束或描述这种评价的精度。例如，这些现有技术不提供新策略实际上要差于所部署策略的机会的知识。从而，这些传统技术可能潜在地损失收益以及源于较差表现策略的低效。

发明内容

[0005] 描述了风险量化、策略搜索和自动安全策略部署技术。在一个或多个实施方式中，这些技术用于确定策略的安全性，诸如表示新策略将相对于当前部署的策略显示出增加的性能（例如，交互或转换）测量的置信等级。为了进行这种确定，使用强化学习和集中不等式，其生成和约束关于策略的性能测量的置信值，因此提供该性能的统计保证。这些技术可用于量化策略部署中的风险，基于估计的性能和这种估计中的置信等级（例如，可以包括使用策略空间来减少被处理数据的量）选择用于部署的策略，用于通过交互（其中，策略的参数被迭代调整，并且这些调整的效果被评估等等）创建新策略。

[0006] 该发明内容部分以简化形式介绍了概念的选择，在以下具体实施方式部分进行进一步的描述。如此，该发明内容部分不用于表示所要求主题的主要特征，也不用于帮助确定所要求主题的范围。

附图说明

[0007] 参照附图描述具体实施方式。在附图中，参考标号最左边的数字表示参考标号首先出现的附图。说明书和附图中的不同实例中使用相同的参考标号可以表示类似或相同的项目。附图中表示的实体可以表示一个或多个实体，由此可以在讨论中以单个或多个实体形式来互换地进行参考。

[0008] 图 1 是可用于使用本文描述的技术的示例性实施方式的环境的示图。

[0009] 图 2 示出了详细示出强化学习模块的示例性实施方式的系统。

[0010] 图 3A 示出了策略的性能和置信的示图。

[0011] 图 3B 包括提供概率密度函数的经验估计的曲线。

[0012] 图 4 示出了不同的集中不等式函数的结果的图表。

[0013] 图 5 示出了确定策略参数的安全性的实例。

[0014] 图 6 示出了以下算法 1 的伪码的实例。

[0015] 图 7 示出了以下算法 2 的伪码的实例。

[0016] 图 8 示出了以下算法 3 的伪码的实例。

[0017] 图 9 是示出描述用于策略改进的风险量化的技术的示例性实施方式中的程序的流程图。

[0018] 图 10 是示出描述包括策略搜索的一个或多个部署策略的替换控制的示例性实施方式中的程序的流程图。

[0019] 图 11 是示出通过利用策略空间执行选择策略以替换部署策略来提高效率的示例性实施方式中的程序的流程图。

[0020] 图 12 是示出迭代生成新策略并用于替换部署策略的示例性实施方式中的程序的流程图。

[0021] 图 13 示出了执行策略改进技术和算法 3 的结果。

[0022] 图 14 表示 NAC 的性能与手动优化超参数进行比较的示例性结果。

[0023] 图 15 示出了算法 3 的应用的结果。

[0024] 图 16 示出了包括可以如所描述的和 / 或参照图 1 至图 15 使用的实施为任何类型的计算设备的示例性设备的各个部件以实施本文所描述技术的实施例的示例性系统。

具体实施方式

[0025] 概述

[0026] 策略被用于确定哪些广告被选择用于包括将被发送给特定用户的内容。例如，用户可以经由网络访问内容提供器以获取内容，诸如通过使用浏览器来获取特定网页。这种访问被内容提供器用于识别与这种访问相关的特性，诸如用户的特性（例如，人口统计资料）以及访问本身的特性（例如，日期、地理位置等）。这些特性被内容提供器使用策略进行处理以确定哪些广告将被选择用于包括在传输回用户的网页中。因此，策略可用于基于访问的不同特性选择不同的广告用于包括在内容中。

[0027] 然而，用户部署策略的传统技术不具有约束或量化新策略是否比当前部署的策略执行得更好的精度的机制。为此，这些传统技术通常迫使用户进行关于新策略是否具有更好性能的最佳猜测，例如使得增加广告的选择数量，使得增加用户购买商品或服务的转换的数量等等。

[0028] 因此，描述用于部署策略的风险可被量化的技术，其用于支持各种功能。例如，描述现有策略的部署的数据被访问和处理以确定新策略是否将相对于现有策略显示出提高的性能。这通过计算表示新策略的性能将至少满足限定值（例如，其可以基于部署策略的性能）的置信度的置信值来进行，因此用作该性能的统计保证。

[0029] 为了计算统计保证，集中不等式被用作以下强化学习的一部分。强化学习是机器学习的一种类型，其中软件代理被执行以在使累积奖的一些概念最大化的环境中采取动作。在该实例中，奖励是使测量的性能最大化以选择广告，诸如增加广告的选择数量（例如，“点击”）、广告的转换（例如，导致“购买”）等。

[0030] 集中不等式被用作强度学习的一部分以确保安全性，新策略显示出至少为部署策略的量的性能。例如，集中不等式被用于解决独立随机变量的函数与它们的期望值的偏离。因此，集中不等式提供了对这些分配的约束并且确保结果的精度。例如，如下面进一步描述的集中不等式可约束值使得阈值以上存在的值被移动到阈值处，可用于塌陷分布的尾部等等。

[0031] 以下，首先在算法 1 中表示集中不等式，其允许关于策略是否安全用于部署并由此选择广告而不降低性能的有效确定。第二，在算法 2 中表示安全批量强化学习算法，其被配置为利用强化学习和集中不等式来选择用于部署的策略。第三，在算法 3 中表示安全迭代算法，其被配置为使用强化学习和集中不等式通过参数和分析的迭代调整生成新策略以确定何时这些调整可能增加性能。即使算法 3 确保安全性，但其与通过使用策略空间如以下进一步描述的最先进的重度调整的非安全算法相比具有合理的采样效率。

[0032] 首先描述可采用本文描述的技术的示例性环境。然后，描述可以在示例性环境以及其他环境中执行的示例性程序和实施实例。从而，示例性程序的执行不限于示例性环境和实施实例，并且示例性环境不限于示例性程序的执行。

[0033] 示例性环境

[0034] 图 1 是可用于采用本文描述的强化学习和集中不等式的示例性实施方式中的环境 100 的示图。所示环境 100 包括内容提供器 102、策略服务 104 和客户设备 106，它们经由网络 108 相互通信耦合。实施这些实体的计算设备可以以各种方式进行配置。

[0035] 例如，计算设备可配置为桌上型计算机、膝上型计算机、移动该设备（例如，假设诸如平板或移动电话的手持结构）等。因此，计算设备包括从全资源设备（具有重要的存储器和处理器资源）（例如，个人计算机、游戏控制台）到低资源设备（具有有限的存储器和 / 或处理资源）（例如，移动设备）的范围。此外，尽管示出了单个计算设备，但计算设备还代表多个不同的设备，诸如被企业用于“在云上”执行操作的多个服务器，如内容提供器 102 和策略范围 104 所示并且参照图 16 所进一步描述的。

[0036] 客户设备 106 被示为包括通信模块 110，其表示经由网络 108 访问内容 112 的功能。通信模块 110 例如被配置为浏览器、能够联网的应用、第三方插件等。如此，通信模块 110 经由网络 108 访问内容提供器 102 的各种不同内容 112，其被示为存储在存储器 114 中。内容 112 可以以各种方式进行配置，诸如网页、图像、音乐、多媒体文件等。

[0037] 内容提供器 102 包括内容管理器模块 116，其表示管理内容 112 的提供的功能，从而包括哪些广告 118 与内容 112 一起被包括。为了确定哪些广告 118 包括内容 112，内容管理器模块 116 采用策略 120。

[0038] 当用户导航到诸如网页的内容 112 时，例如，包含用户的已知属性的列表被形成为特征矢量，其中特征矢量的值反映用户的当前状态或观察。例如，特征矢量的值可以描述开始访问内容 112 的用户的特性（例如，诸如年龄和性别的人口统计）和 / 或如何执行访问，诸如用于执行访问的客户设备 106 或网络 106 的特性、访问本身的特性（诸如时间、星

期几）、什么导致访问（例如，网页上链接的选择）等。

[0039] 因此，特征矢量被配置为表示用户的数量和被观察的访问的数字特征的 n 维矢量。以下，策略 120 基于关于用户的被观察当前状态（例如，通过上述特征矢量表示）的判定来执行动作。例如，内容管理器模块 116 首先观察用户的状态，然后使用策略 120 判定将采取何种动作。在所示情况下，可能的动作是哪些广告 118 被选择用于被客户设备 106 显示。因此，如果存在十个可能的广告，则在该实例中存在十种可能的动作。

[0040] 策略 120 的性能可以通过各种方式进行测量。例如，性能被定义为与广告 118 的用户交互的测量（例如，用户“点击”的频繁程度），因此在以下讨论中越高越好。在另一实例中，性能被定义为广告 118 的转换率，例如在选择广告 118 之后购买商品或服务，因此在该实例中也是越高越好。应该注意，不同的策略可具有不同的性能。例如，一些策略可导致对广告的高点击率，而其他策略不会。随后，该实例的目标是部署具有最好可能性能的策略 120，即支持最多的交互、转换等等。

[0041] 为了确保安全策略被部署至少显示性能的限定等级（例如，至少等于部署策略的性能以及限定裕度），策略服务 104 利用策略管理模块 122。策略管理模块 122 代表生成策略 120 和 / 或计算统计保证以确保策略 120 对于部署来说是安全的（例如，至少显示出先前部署的策略的性能等级）的功能。

[0042] 该功能的实例被示为强化学习模块 124，其被用于部署强化学习技术来保证新策略的部署将相对于当前使用的策略（即，部署策略）具有改进。强化学习是机器学习的类型，其中软件代理被执行以在使累计奖励的一些概念最大化的环境中采取动作，在这种情况下使策略 120 的性能最大化以选择导致相关商品或服务的用户交互（例如，点击）或转换的广告 118。

[0043] 例如，强化学习模块 124 使用强化学习来生成新策略将相对于部署策略显示出增加的性能的置信值并由此提供这种增加性能的统计保证。以各种方式生成置信值，诸如通过内容提供器 102 使用描述先前策略（即，现有或当前策略）的部署的部署数据。强化学习模块 124 然后使用新策略来处理该部署数据以计算统计保证，如此可以在不具有新策略的实际部署的情况下进行。以这种方式，内容提供器 102 被保护不受潜在坏策略的部署的影响，而这种坏策略会通过较低的交互和 / 或转换而导致降低的收益。

[0044] 作为统计保证的计算的一部分，强化学习模块 124 使用置信不等式 126，诸如确保新策略至少显示出部署策略的量的“安全性”。集中不等式被用于解决统计保证的置信度的函数与其预期（即，期望值）的偏离。这用于约束置信值的分布，并由此提高统计保证的精度。例如，集中不等式可以约束置信值，使得阈值之上的置信值被移动到阈值处，可用于塌陷分布的尾部等等。以下描述集中不等式和强化学习的进一步讨论。

[0045] 如此，以下使用强化学习来支持与用于选择广告的策略 120 的选择和生成相关联的各种不同功能或其他功能。例如，强化学习和集中不等式被用于通过使用统计保证基于先前策略的部署数据量化新策略的部署中涉及的风险的量。在另一实例中，强化学习和集中不等式用于选择多个策略（如果具有的话）中的哪些被部署以替代当前策略。在又一实例中，强化学习和集中不等式被用于通过迭代技术（包括策略的参数调整以及使用部署数据计算统计保证）生成新策略。以下描述并在对应附图中示出这些和其他实例的进一步讨论。

[0046] 尽管以下描述了广告的选择，但本文所描述的技术可用于各种不同类型的策略。其他策略使用的实例包括市场效应系统、新闻推荐系统、患者诊断系统、神经义肢控制、自动药品管理等中的寿命值优化。

[0047] 图2示出了详细示出强化学习模块124的示例性实施方式中的系统200。系统200被示为包括第一实例202、第二实例204和第三实例206。在第一实例中，部署策略208被用于选择广告118包括内容112(例如，网页)，其如先前所述被传输至客户设备106的用户。因此，部署数据210被策略管理模块122收集，其描述内容提供器102对部署策略208的部署。

[0048] 在这种情况下，策略管理模块112还提出了新策略212用于替换部署策略208。然后，策略管理模块122利用强化学习模块124来确定是否部署新策略212，其包括使用参照图1所描述的集中不等式126的使用以增加新策略的可能性能的统计保证的精度。如果新策略212是“坏的”(例如，具有低于部署策略208的性能分数)，则新策略212的部署例如由于失去用户交互、转换和上述其他性能测量而昂贵。

[0049] 为了执行这种确定，策略管理器模块122访问部署数据210，其描述图1的内容提供器102使用部署测量208。这种访问用于基于新策略212具有比部署策略208更好的性能的置信度来预测是否部署新策略212。以这种方式，这种预测在不具有新策略212的实际部署的情况下进行。

[0050] 在所示实例中，强化学习模块124包括置信评估模块214，其表示生成统计保证216的功能，其实例在以下被描述为算法1和“安全”。通过使用集中不等式，统计保证216被用于基于被图1的集中不等式126约束的部署数据210使用针对新策略212计算的置信值量化新策略212的部署的风险。这提高了相对于传统技术的精度。因此，不同于传统技术，统计保证216指示由强化学习模块124学习的置信值表示的估计是正确的置信量。例如，给出部署策略208、来自部署策略208的部署的部署数据210以及性能等级“f_{min}”，通过限定估计精度的统计保证216来表示新策略212性能处于至少“f_{min}”的等级的置信度。

[0051] 如图3A所示，考虑示图300。水平轴是“f_{min}”，其是策略的性能。垂直轴是置信度，并且部署策略208在示图300中具有性能302。使用从部署策略208的部署收集的部署数据210来评估新策略212，其导致示图300中绘制的置信值304。置信值304表示性能至少为水平轴上指定的值的置信度，并由此为该性能的统计保证。在所示实例中，性能为至少0.08的置信度几乎为1。性能为至少0.086的置信度接近0。应该注意，这不意味着新策略212的实际性能不是这么好，而是意味着还不能利用任何实际置信度来保证性能。

[0052] 该实例中的统计保证的置信值304支持强论证来部署新策略212，因为该值表示新策略212将比部署策略208执行得更好的高置信度。在该实例中表示实际部署的新策略212的性能306也在示图300中示出。可以在以下算法1的讨论中找到并且在对应附图中示出该实例的进一步讨论。

[0053] 在第二实例204中，还示出了描述部署策略208的部署的部署数据210。在该实例中，策略改进模块218用于处理多个策略220以进行策略选择222，其具有性能大于部署策略208的相关统计保证。如前所述，传统方法不包括生成统计保证的技术，其中一个策略将相对于另一个显示出改进。如此，难以使用这些传统方法来证明新策略的部署，尤其是由于坏策略的部署会是昂贵的(例如，具有低点击率)。

[0054] 由策略改进模块 218 实施以进行这种选择的功能被称为“策略改进算法”并且在以下还称为“算法 2”。在该实例中，策略改进模块 218 搜索一组策略 220 并且如果选择被确定为“安全”则进行策略选择 222。如果策略 220 的性能好于性能等级（例如，“f_{min}”）并且在置信等级内（例如，“1-δ”），则选择是安全的。

[0055] 可通过用户来限定性能等级（例如，“f_{min}”）和置信等级（例如，“1-δ”）。例如，用户选择“δ = 0.5”且“f_{min}= 1.1 乘以（部署策略的性能）”意味着以 95％的置信度保证性能的 10％改进。因此，如果可以根据安全的定义保证是安全的，则策略改进模块 218 将在该实例中仅建议新策略。策略改进模块 218 可以以各种方式来进行这种确定，诸如采用在第一实例 202（例如，以下为算法 1）中描述的置信评估模块 214。

[0056] 在第三实例 206 中，示出了用于安全策略部署的自动系统。在先前实例中，描述了数据用于选择策略的分布，例如作为其采用现有数据并提出单个新策略的“批量”。然而，在该实例中，描述了上述分布的迭代版本，其功能被示为可用于生成新策略 226 的策略生成模块 224。例如，迭代可用于调整策略的参数，利用置信度的限定等级确定具有调整的策略是否将比部署策略 208 显示出更好的性能，如果是，则部署新策略 226 作为替换。因此，策略生成模块 224 被配置为进行一系列改变以生成新策略 226，诸如连续多次应用由策略改进模块 218 所表示的功能，添加记录本来跟踪对策略参数进行的改变。

[0057] 在第二实例 204 中，针对部署策略 208 在一时间段（例如，一月）内收集部署数据 210 以进行新策略 220 的策略选择 222。在第三实例 206 中，收集部署数据 210 直到找到新策略 226 为止，然后策略管理模块 122 使得立即切换到执行新策略 226，例如来替代部署策略 208。可以针对多个“新”策略重复该处理以替换部署策略。以这种方式，可以通过容易地实施新策略 26 来实现改进的性能，可以在以下实例中的“算法 3”和“代达罗斯 (Daedalus)”的描述中找到进一步的描述。

[0058] 实施示例

[0059] 用“S”和“A”表示可能状态和动作的集合，其中状态描述对内容（例如，用户或用户访问的特性）的访问，以及动作源于使用策略 120 进行的判定。尽管以下使用马尔克夫判定处理 (MDP)，但通过用观察结果代替状态，结果可以直接利用反应策略对 POMDP 执行。假设奖励被约束“$r_t \in [r_{min}, r_{max}]$”，并且“$t \in \mathbb{N}$”被用于索引时间，从“t = 1”开始，其中相对于状态具有一些固定分布。表达“π(s, a, θ)”被用于表示当使用策略参数“$\theta \in \mathbb{R}^{n_\theta}$”时状态“s”下的动作“a”的可能性（密度或质量），其中“n_θ”是整数，策略参数空间的维度。

[0060] 假设“$f:\mathbb{R}^{n_\theta} \rightarrow \mathbb{R}$”是将策略 120 的策略参数看作“π(.,., θ)”的期望返回值，即，对于任何“θ”来说，

[0061] $$f(\theta) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^{t-1} r_t \middle| \theta\right],$$

[0062] 其中，“γ”是指定随时间的奖励的折扣的 [0,1] 间隔中的参数。问题可以包括有限范围，其中每个轨迹在“T”时间步内到达终端状态。因此，每个轨迹“τ”是状态（或观察结果）、动作和奖励的排序集合：“$\tau = \{s_1, a_1, r_1, s_2, a_2, r_2, \cdots, s_T, a_T, r_T\}$”。为了简化

分析，不损失普遍性，可以进行返回值“$\sum_{t=1}^{T}\gamma^{t-1}r_t$”总是在间隔 [0,1] 中的要求。这可以通过缩放和转换奖励来实现。

[0063] 获取数据集合“D”，其包括“n”个轨迹，用策略参数来标记，如下生成它们：

[0064] $D=\{(\tau_i\ \theta_i):i\in\{1,\ldots,n\},\ \tau_i \text{ generated using } \theta_i\}$，

[0065] 其中，“θ_i”表示第 i 个参数矢量，“θ”不是“θ”的第 i 个元素。最后，获取“$f_{min}\in\mathbb{R}$”和置信等级“$\delta\in[0,1]$”。

[0066] 当利用置信度“1-δ”确定“$f(\theta)>f_{min}$”时，如果仅提出了新策略参数“θ”，则认为算法是安全的。如果利用置信度“1-δ”确定“$f(\theta)>f_{min}$”测量参数“θ”（与算法相对）被认为是安全的。注意，说明策略是安全的是关于给出一些数据的策略的信任的申明而不是关于策略本身的申明。此外，注意，确保“θ”是安全的等效于确保利用显著等级“δ”拒绝“$f(\theta)\leqslant f_{min}$”的假设。这种置信度和假设测试框架被采用是因为其没有意义来讨论“$Pr(F(\theta)>f_{min})$”或“$Pr(f(\theta)>f_{min}|D)$”，因为“$f(\theta)$”和“f_{min}”都不是随机的。

[0067] 假设“Θ^{D}_{safe}”表示给出数据“D”的安全策略参数的集合。首先，确定什么分析将可能被用于考虑可用数据“D”（即，部署数据 210）生成最大“Θ^{D}_{safe}”。如果“$\Theta^{D}_{safe}=\emptyset$”，则算法返回“没有找到解”。如果“$\Theta^{D}_{safe}\neq\emptyset$”，则以下是被配置为返回新策略参数的算法“$\theta'\in\Theta^{D}_{safe}$”，其被评估为“最好的”：

[0068]
$$\theta'\in\arg\max_{\theta\in\Theta^{D}_{safe}} g(\theta,D) \qquad (1)$$

[0069] 其中，“$g(\theta,D)\in\mathbb{R}$”基于提供的数据“D”指定“θ”如何“好”（即，新策略参数）。典型地，“g”将是“$f(\theta)$”的评估值，但是允许针对任何“g”进行。“g”的另一实例是类似于“f”的函数，但是其考虑返回值的变化。注意，即使等式 (1) 使用“g”，但安全保证是坚定的，因为其使用真实（未知，并且总是未知）期望返回值“$f(\theta)$”。

[0070] 最初，描述了考虑一些数据“D”，并且产生策略参数的单个新集合“θ’”，因此从多个策略中选择新策略的批量技术。这种批量方法可以扩展到迭代方法，如以下进一步描述的，其进行多个策略改进，然后自动和立即进行部署。

[0071] 生成 $f(\theta)$ 的无偏估计值

[0072] 以下技术利用从使用行为策略“θ_i”生成的每个轨迹“$\tau\in D$”生成无偏估计值“$f(\theta)$ 的 $f(\theta,\tau,\theta_i)$”的能力。重要的采样被用于如下生成这些无偏估计值：

[0073]
$$\hat{f}(\theta,\tau,\theta_i):=\underbrace{\prod_{t=1}^{T}\frac{\pi(s_t,a_t,\theta)}{\pi(s_t,a_t,\theta_i)}}_{\text{importance weight}}\underbrace{\sum_{t=1}^{T}\gamma^{t-1}r_t}_{\text{return}} \qquad (2)$$

[0074] 注意，在 (2) 中没有出现除以 0，因为如果“ π (s$_t$, a$_t$, θ$_i$) = 0”则在轨迹中不选择“a$_t$”。然而，为了实施将被应用的重要采样，要求对于所有“s”和“a”来说“ π (s, a, θ)”为 0，其中“ π (s,a, θ$_i$) = 0”。如果不是这种情况，则来自“ θ$_i$”的数据可以不被用于评估“ θ ”。直观地，当评估策略在“s”中执行“a”时，如果行为策略在状态“s”中从不执行动作“a”，则不存在关于输出的信息。

[0075] 对于每个 θ$_i$，$\hat{f}(\theta,\tau,\theta_i)$是通过使用“ θ$_i$”采样“ τ ”然后使用等式 (2) 计算的随机变量。由于重要采样是无偏的，因此对于所有“i”，

[0076] $$\mathrm{E}[\hat{f}(\theta,\tau,\theta_i)] = f(\theta)$$

[0077] 因为最小的可能返回值为 0 且重要权重是非负的，所以重要权重返回值约束到 0 以下。然而，当“ θ ”导致在动作不可能在“ θ i”以下的状态中可能的动作时，重要权重返回值可以较大。因此，“$\hat{f}(\theta,\tau,\theta_i)$”是约束到 0 以下的随机变量，具有 [0:1] 间隔中的期望值，并且就有较大的上限。这意味着$\hat{f}(\theta,\tau,\theta_i)$可以具有相对较长的尾部，如图 3B 的示例性示图 350 所示。

[0078] 曲线 352 是关于简化且“T = 20”的登山 - 汽车领域的“$\hat{f}(\theta,\tau,\theta_i)$”的概率密度函数 (PDF) 的经验估计。垂直轴对应于概率密度。稍后在以下讨论中描述曲线 304。行为策略参数“ θ i”产生次优策略并且沿着从“ θ i”开始的自然策略梯度选择评估策略参数“ θ ”。在该实例中通过生成 100,000 个轨迹、计算对应的重要权重返回值、然后将它们传输至密度函数来评估概率密度函数 (PDF)。关于重要权重返回值的最紧上限近似为 $10^{9.4}$，尽管最大观察重要权重返回值近似为 316。采样平均接近 $0.2 \approx 10^{0.7}$。注意，水平轴被算法地缩放，例如十进制。

[0079] 集中不等式

[0080] 为了确保安全性，如上所述采用集中不等式 126。集中不等式 126 被用作置信值的约束，并由此用于提供性能的统计保证，例如至少对应于限定值的策略的性能测量的估计值。集中不等式 126 可以采用各种不同的形式，诸如 Chernoff-Hoeffding 不等式。该不等式用于计算每个策略被约束的每条轨迹上的采样平均 (平均$\hat{f}(\theta,\tau,\theta_i)$)，例如与真实平均“f(θ)”偏离的不太远。

[0081] 每个集中不等式都在以下表示为应用于“n”和独立和相同分布的随机变量“X$_1$, …, X$_n$”，其中对于所有“i”来说“X$_i$∈ [0,b]”且“E[X$_i$] = μ ”。在这些技术的情况下，这些“X$_i$”对应于使用相同行为策略和“ μ = f(θ)”的“n”个不同轨迹的“$\hat{f}(\theta,\tau,\theta_i)$”。集中不等式的第一实例是 Chernoff-Hoeffding(CH) 不等式：

[0082] $$\Pr\left(\mu \geq \frac{1}{n}\sum_{i=1}^{n} X_i - \frac{b}{\sqrt{n}}\sqrt{\frac{\ln(1/\delta)}{2}}\right) \geq 1-\delta \qquad (3)$$

[0083] 在第二实例中，表示 Maurer 和 Pontil 的经验伯恩斯坦 (MPeB) 不等式，其用如下

采样变量替换伯恩斯坦不等式中的真实（该设置为未知）变量：

[0084]

$$\Pr\left(\mu \geq \frac{1}{n}\sum_{i=1}^{n} X_i - \frac{b}{n-1}\left(\frac{7\ln(2/\delta)}{3}\right) - \frac{1}{n}\sqrt{\frac{\ln(2/\delta)}{n-1}\sum_{i,j=1}^{n}(X_i - X_j)^2}\right) \geq 1-\delta \tag{4}$$

[0085] 在第三实例中，安德森（AM）不等式在以下被示为使用 Dvoretzky-Kiefer-Wolfowitz 不等式，其如下通过 Massart 找到最优常数：

[0086]

$$\Pr\left(\mu \geq z_n - \sum_{i=0}^{n-1}(z_{i+1} - z_i)\min\left\{1, \frac{i}{n} + \sqrt{\ln\left(\frac{2}{1-\delta}\right)\frac{1}{2}n}\right\}\right) \geq 1-\delta, \tag{5}$$

[0087] 其中，“z_1、z_2，…，z_n”是“X_1，X_2，…，X_n”的顺序统计且“$z_0 = 0$”。即，“Z_i”是随机变量“X_1，X_2，…，X_n”的采样，它们进行排序使得“$z_1 \leq z_2 \leq \cdots z_n$”且“$z_0 = 0$”。

[0088] 注意，等式（3）仅考虑随机变量的采样平均，而等式（4）考虑采样平均和采样变量。这使得等式（4）减少了范围“b”的英系那个，即，在等式（4）中，范围除以“n-1”，而在等式（3）中，其除以“$\sqrt{n}$”。等式（4）仅考虑采样平均和采样变量，等式（5）考虑整个采样累计分布函数。这使得等式（5）仅依赖于最大观察采样而不依赖“b”。这在一些情况下可以是显著的改进，诸如图 3 所示的示例性情况，其中最大观察采样近似为 316 同时“b”近似为 $10^{9.4}$。

[0089] 在另一实例中，上面将 MPeB 不等式示为扩展为与随机变量的范围无关。这导致新不等式，其将 MPeB 不等式的期望特性（例如，没有相同分布的随机变量的一般紧密型和适应性）与 AM 不等式的期望特性（例如不直接依赖于随机变量的范围）进行组合。还移除了确定关于最大可能重要权重返回值的紧密上限的需求，这可以包括域专用特性的专业考虑。

[0090] MPeB 不等式的扩展利用两种方式。第一种方式是移除分布的上尾部降低其期望值。第二种方式是如果同时专用于具有相同平均值的随机变量则 MPeB 不等式可以被概括为处理具有不同范围的随机变量。因此，随机变量分布的尾部塌陷，并且在该实例中标准化随机变量，使得可以应用 MPeB 不等式。然后，MPeB 不等式用于生成下限，从中提取原始随机变量的均匀平均值的下限。在以下定理 1 中提供所得到的集中不等式。

[0091] 用于塌陷分布的尾部然后约束新分布的平均值的方法类似于约束截顶或缩尾均值估计量。然而，在截顶均值丢弃一些阈值以上的每个采样的情况下，本技术中的采样从阈值上方移动到精确位于阈值，这类似于计算缩尾均值，除了阈值不依赖于数据。

[0092] 在定理 1 中，假设“$X = (X_1, \cdots X_n)$”是独立随机变量的矢量，其中“$X_i \geq 0$”且所有“X_i”都具有相同的期望值“μ”。假设对于所有“i”来说，“$\delta > 0$”并选择任何“$c_i > 0$”。然后，具有至少为“$1-\delta$”的概率：

[0093]

$$\mu > \underbrace{\left(\sum_{i=1}^{n}\frac{1}{c_i}\right)^{-1}}_{\text{scaling term}}\left(\underbrace{\sum_{i=1}^{n}\frac{Y_i}{c_i}}_{\text{sample mean}}\underbrace{-\sqrt{\frac{\ln(2/\delta)}{(n-1)}\sum_{i,j=1}^{n}\left(\frac{Y_i}{c_i}-\frac{Y_j}{c_j}\right)^2}-\frac{7n\ln(2/\delta)}{3(n-1)}}_{\text{terms that, after being scaled, go to zero as } n\to\infty}\right) \quad (6)$$

[0094] 其中，“$Y_i = \min\{X_i, c_i\}$”。

[0095] 为了应用定理 1，对于每个“ci”（阈值超过其）选择值，塌陷“Xi”的分布。为了简化该任务，选择单个“$c \in \mathbb{R}$”并且对于所有“i”来说设置“$c_i = c$”。当“c”太大时，其放松约束，就像大范围“b”一样。当“c”太小时，其降低“Y_i”的真实期望值，这也放松了约束。因此，最佳“c”平衡了“Y_i”的范围与“Y_i”的真实平均之间的折中。所提供的随机变量被划分为两组“D_{pre}”和“D_{post}”。“D_{pre}”用于估计最佳标量阈值，作为（该等式中的最大函数是具有标量“c”的等式 (6) 的右侧）：

[0096]

$$c \in \arg\max_{c}\frac{1}{n}\sum_{i=1}^{n}Y_i - \sqrt{\frac{2\ln(2/\delta)}{n^2(n-1)}\left(n\sum_{i=1}^{n}Y_i^2-\left(\sum_{i=1}^{n}Y_i\right)^2\right)}-\frac{7c\ln(2/\delta)}{3(n-1)} \quad (7)$$

[0097] 回忆“$Y_i = \min\{X_i, c_i\}$”，使得等式 (7) 中三个项目的中每一项都依赖于“c”。一旦从“D_{pre}”中形成最佳“c”的估计值，则使用“D_{post}”中的采样和优化“c”值应用定理 1。在一个或多个实施方式中，发现使用“D_{pre}”中采样的 1/3 和“D_{post}”中的剩余 2/3 在已知真实平均值在 [1,0] 中、“$c \geq 1$”的情况下执行得很好。当一些随机变量被相同分布时，可以确保变量以 1/3 在“D_{pre}”且 2/3 在“D_{post}”中进行划分。在一个或多个实施方式中，这种用于确定多少点包括在 D_{pre} 中的自组方案被改善以针对每个随机变量选择不同的“c_i”。

[0098] 图 3B 中的曲线 354 示出了当选择“c”时的折中。其对于平均值“f(θ)”给出 95% 的置信下限，对于值“c”的（垂直轴）通过水平轴来指定。一个或多个实施方式中的最佳“c”值在 10^2 左右。曲线 304 继续在水平轴下方。在这种情况下，当“$c = 10^{9.4}$”时，不等式退化成 MPeB 不等式，其对 -129703 的平均值产生 95% 的置信下限。

[0099] 使用用于创建图 3B 的 100000 个采样，利用 1/3、2/3 数据划分使用定理 1 以及 CH、MpeB 和 AM 不等式计算平均值的 95% 置信下限。还得到和测试塌陷 -AM 不等式，其是 AM 不等式的扩展以使用本文描述的方案，其中塌陷“X_i”成为“Y_i”且从数据的 1/3 中优化“c”值。在图 4 所示图表 400 中提供的结果。类似于通过重要采用所生成的，比较示出了用于长尾分布的集中不等式的功率。还示出了 AM 不等式不从应用于 MpeB 不等式的塌陷方案中获益。

[0100] 确保策略搜索中的安全性

[0101] 为了确定策略参数“θ”对于给定的提供数据“D”是否安全，来自部分 4 的集中不等式被应用于重要的权重返回值。为了简化，如图 5 的实例 500 所示，当使用“D”中的轨迹和提供的阈值“c”来估计“θ”时，假设“$f_1(D, \theta, c, \delta)$”为通过定理 1 生成的“f(θ)”的置信下限“1-δ”，其中，“n”是“D”中的轨迹的数量。如图 6 的实例 600 所示，在算法 1 中提供确定“θ”对于“D”是否安全的伪码。

[0102] Oracle 约束策略搜索

[0103] 上面描述了确定策略参数是否安全的技术，然后选择适当的对象函数“g”并且使用该函数找到最大化“g”的安全参数。任何策略脱离评估技术可用于“g”，诸如对风险敏感的“g”，其喜欢具有较大期望返回值的“θ”但也具有返回值的较小变化。为了简化，以下用于“g”的权重重要采样：

[0104] $$g(\theta, D) := \sum_{i=1}^{n} \frac{\hat{f}(\theta, \tau_i, \theta_i)}{\prod_{t=1}^{T} \frac{\pi(s_t, a_t, \theta)}{\pi(s_t, a_t, \theta_i)}}.$$

[0105] 根据等式(1)选择“θ’”是约束优化问题的形式，因为用于“Θ^{D}_{safe}”的采样分析表示不可用。此外，会员oracle可用，利用其使用算法1来确定“θ”是否为“Θ^{D}_{safe}”。当“n_θ”较小时，使用栅格搜索或对于每个可能“θ”的随机搜索，该约束优化问题被暴力破解。然而，随着“n_θ”的增长，该技术变得棘手。

[0106] 为了克服该问题，自然策略梯度算法用于将搜索减少到多个约束线搜索。直观地，代替搜索每一个“$\mathbb{R}^{n_\theta}$”，从期望与策略空间的安全区域相交的每个行为策略“θ”中选择单个方向$\tilde{\nabla} f(\theta)$，并且执行这些方向上的搜索。从每个行为策略中选择的方向是广义的自然策略梯度。尽管不保证广义自然策略梯度指向安全区域，但其是合理的方向选择，因为该方向上的点使得期望返回值更快速地增加。尽管可以使用用于计算广义自然策略梯度的任何算法，但在该实例中使用具有LSTD的偏置自然评估决策。通过强力解决约束线搜索问题。

[0107] 在算法2中提供了用于该算法的伪码，在图7中示出了其实例700，其中如果“A”为真则指示函数“1_A”为1，否则为0。

[0108] 多策略改进

[0109] 策略改进技术使用上面讨论中的批量方法，其被应用与现有数据集合“D”。然而，可以通过提取新安全策略参数来以递增方式使用技术。用户可以在每次迭代时选择改变“f_{min}”，例如反映至今找到的最好策略或最近提出的策略的性能的估计。然而，在本文描述的伪码中，假设用户不改变“f_{min}”。

[0110] 假设“θ_0”表示用户的初始策略参数。如果“$f_{min} = f(\theta_0)$”，则可以说明具有提出的每个策略将至少与用户持续使用初始策略一样好的高置信度。如果“f_{min}”是“$f(\theta_0)$”的评估值，则可以说明具有提出的每个策略将至少与用户策略的观察性能一样好的高置信度。用户还可以选择“f_{min}”低于“$f(\theta_0)$”，这对算法给出更大的自由度来探索同时保证性能不劣化到低于指定等级。

[0111] 算法保持策略参数的列表“C”，其被确认为安全。如参照图2所描述的，当生成新轨迹时，算法使用“C”中的策略参数，其被期望执行得最好以生成新策略226。在算法3中表示用于该在线安全学习算法的伪码，在图8中示出其实例800，其也在图中表示为Daedalus。关于以下程序描述这些和其他实例的进一步讨论。

[0112] 示例性程序

[0113] 以下讨论描述了使用先前描述的系统和设备实施的技术。每个程序的方面都可以以硬件、固件或软件或它们的组合来实施。程序被示为框的集合，它们执行由一个或多个设备执行的操作并且不是必须限于用于由各个框执行操作所示的顺序。在以下讨论的部分

中，将参照图 1 至图 8。

[0114] 图 9 示出了描述用于策略改进的风险量化的技术的示例性实施方式。接收策略，其被配置用于被内容提供器部署以选择广告（框 902）。在一种情况下，技术员通过与内容管理器模块 116 的交互（诸如通过针对策略的特性参数的用户接口）创建策略。在另一种情况下，自动地创建策略而不使用用户干涉，诸如通过内容管理器模块 116 自动调整参数来创建新策略，其具有显示出性能测量的改进的潜力，诸如交互（例如，“点击”）的数量、转换率等等。

[0115] 与内容提供器的部署策略相反，至少部分地基于接收策略的部署所可能涉及的风险的量化来控制内容提供器接收部署（框 904）。如前所述，内容提供器 102 使用策略不是静止的，其中策略被频繁改变，新策略更好利用关于接收通过使用策略选择的广告的用户的已知信息。在该实例中，通过使用统计保证来控制部署，其中新策略将增加性能的测量（例如，交互或转换的寿命值）并且降低新策略将引起性能和对应收益的降低的风险。

[0116] 控制基于通过内容提供器对描述部署策略的部署的部署数据应用强化学习和集中不等式以估计所接收策略的性能测量的值并且通过计算估计值的一个或多个统计保证来量化风险（框 906）。控制还包括响应于确定一个或多个统计保证表示至少性能测量的估计值至少对应于至少部分地基于内容提供器的部署策略的性能测量的阈值的置信等级，使得接收策略进行部署（框 908）。换句话说，当基于统计保证将策略确定为安全时，以上述方式部署策略。

[0117] 例如，内容管理器模块 116 管理用于部署策略的部署数据，然后使用该数据作为用于评估接收策略的部署的风险的基础，因此在没有实际部署新策略的情况下进行。在另一实例中，如果接收策略已经被部署，则策略管理模块利用来自先前策略的数据和从部署新策略累计的数据。

[0118] 不同于仅估计策略的性能而不具有关于估计精度的任何保证的现有技术，策略管理模块 122 通过使用强化学习和集中不等式提供了性能的估计以及估计不是过估计的统计保证。即，策略管理模块 122 通过统计保证提供策略将执行得与估计一样好的概率并由此用于量化策略部署中的风险。

[0119] 如关于定理 1 和算法 1 所描述的，策略管理模块 122 应用的定理 1 使用描述任何数量的先前或当前部署的策略的部署的数据和阈值等级 f_{min}，并产生所接收的策略的真实性能至少为 f_{min}，即性能测量的阈值等级的概率。

[0120] 对于算法 1，用户可以指定置信等级（例如，如上所述的 $1-\delta$）和性能测量的阈值 f_{min}。如果可以至少利用设置的置信等级（例如，$1-\delta$）进行其真实性能至少为 f_{min} 的保证，策略被确认为安全的。因此，算法 1 可以使用定理 1 来确定策略是否是安全的，作为策略管理模块 122 的处理的部分，通过使用强化学习和集中不等式，其中将接收策略（例如，写为上述 θ）、部署数据 D 以及性能测量的阈值 f_{min} 和置信等级（例如，$1-\delta$）作为输入并返回真或假来表示策略是否安全。

[0121] 因此，在该实例中，首先使用强化学习模块 124 和集成不等式 126 由策略管理模块 122 处理接收策略以量化与其部署相关联的风险。风险的量化及其用于控制策略的部署提供了显著的优点，其中危险或风险策略可以在部署之前被标记。注意，这不仅帮助避免坏（即，表现不佳）策略的部署，这提供了生成新策略和选择技术的自由度，而不害怕坏策略

的部署，以下描述并在对应附图中示出进一步讨论。

[0122] 图 10 示出了描述涉及策略搜索的一个或多个部署策略的替换控制的示例性实施方式中的程序 1000。控制利用多个策略中的至少一个策略替换用于选择广告的内容提供器的一个或多个部署策略（框 1002）。如上所述，强化学习和集中不等式可用于确定部署新策略是否是安全的。在该实例中，这些技术被应用于从策略中进行选择以确定哪些策略（如果有的话）将被部署。

[0123] 控制包括搜索多个策略以定位被确认安全替换一个或多个部署策略的至少一个策略，如果至少一个策略的性能测量大于性能的阈值测量并且在如通过使用强化学习和集中不等式对一个或多个部署策略生成的部署数据计算的一个或多个统计保证所表示的置信度的限定等级内，则至少一个策略被确认为安全（框 1004）。例如，策略管理模块 122 使用描述任何数量的先前或当前部署的策略的部署的数据以及阈值性能等级 f_{min}，并产生所接收策略的真实性能至少为 f_{min}，即性能测量的阈值等级的概率。在该实例中，该技术被应用于多个策略以确定哪些策略满足该要求，如果是这样的话，确定哪些策略可能显示出最好的性能，例如由交互或转换的数量所限定的寿命值。

[0124] 响应于被确认安全替换一个或多个其他策略的至少一个所述策略的定位，使得用至少一个所述策略替换一个或多个其他策略（框 1006）。例如，策略服务 104 可以向内容提供器 102 传输指示来从部署策略切换至所选策略。在另一实例中，作为内容提供器 102 本身的一部分来实施该功能。还可以采用技术来改进这种选择的计算的效率，在以下描述并在对应附图中示出其实例。

[0125] 图 11 示出了通过利用策略空间执行策略的选择来替换部署策略以提高效率的示例性实施方式的程序 1100。选择多个策略中的至少一个策略来替换用于选择与内容一起包括的广告的内容提供器的一个或多个部署策略（框 1102）。在该实例中，通过利用描述策略的策略空间来执行选择。

[0126] 例如，选择包括访问表示多个策略中的对应策略的多个高维矢量（框 1104）。例如，多个高维矢量描述被策略基于请求的特性进行广告选择以访问包括广告的内容中所使用的参数。

[0127] 在多个策略的策略空间中计算期望指向期望安全的区域的方向，其中所述区域包括具有大于性能的阈值测量且在置信度的限定等级内的性能测量的策略（框 1106）。选择多个策略中的至少一个策略，其具有对应于该方向的高维矢量并显示出性能测量的最高等级（框 1108）。被期望为指向安全区域的方向是广义的自然策略梯度 (GeNGA)，其是使得性能以相对于策略空间中的其他区域以最快方式增加的策略空间中的方向的估计值。执行被该方向约束的搜索，使得对于与方向相对应的高维矢量来执行线搜索。这些线搜索是低维度的，并且可以被强力破解，由此提高这些策略的定位中的效率。

[0128] 根据对应于方向的策略，如图 9 所述，基于性能测量和置信等级从这些策略中定位策略。策略管理模块 122 使用强化学习和集中不等式来基于性能的阈值测量和由统计保证表示的置信度的限定等级确定哪些策略对于部署来说是最安全的。以这种方式，策略管理模块 122 自动搜索新策略来通过使用安全区域进行部署，因此降低了数据处理量，并且安全区域中的策略可显示出比当前部署的策略显著更好的性能等级。这些技术还可以用于自动地生成新策略而不需要用户交互，在以下描述并在对应附图中示出其实例。

[0129] 图 12 示出了迭代地生成新策略并用于替换部署策略的示例性实施方式的程序 1200。控制利用多个策略中的至少一个策略替换用于选择广告的内容提供器的一个或多个部署策略（框 1202）。在该实例中，替换包括使用迭代技术生成用于替换部署策略的新策略。作为该处理的一部分包括统计保证技术来确保这种部署的安全性。

[0130] 迭代地收集描述一个或多个部署策略的部署的部署数据（框 1204）。如前所述，部署数据 210 描述部署策略 208 的部署，其可以包括或不包括描述新策略的部署的数据。

[0131] 迭代地调整一个或多个参数来生成可用于选择广告的新策略（框 1206）。例如，参数作为策略的一部分而包括并且表示策略如何基于与请求相关联的特性选择广告。特性可用于描述请求的起源（例如，用户和 / 或客户设备 106）、请求本身的特性（例如，时间）等等。因此，在该实例中，策略管理模块 122 的策略生成模块 224 迭代地调整这些参数，并且以各种组合来形成新策略。继续图 11 的实例，这些调整可用于进一步细化策略空间的安全区域，使得调整参数进一步朝向该安全区域偏置新策略，即，使得表示策略的高维矢量更接近地与安全区域对齐。

[0132] 使用具有调整后的一个或多个参数的新策略对描述一个或多个部署策略的部署的部署数据应用强化学习和集中不等式，来估计所述新策略的性能的测量的值并计算所估计值的一个或多个统计保证（框 1208）。这种应用被用于确定新策略将增加新策略相对于部署策略的性能测量的置信等级。

[0133] 响应于确定所述一个或多个统计保证表示至少性能测量的估计值对应于至少部分地基于一个或多个部署策略的性能测量的阈值的置信等级，使得所述新策略中的一个或多个新策略进行部署（框 1210）。例如，策略生成模块 224 被配置为迭代地调用策略改进模块 218，并且在置信度的限定等级内识别改进的阈值等级的情况下引起新策略的部署。

[0134] 在一个或多个实施方式中，如果发现新策略的部署具有较低性能，则策略管理模块 122 终止新策略的部署并部署不同的新策略、返回到先前部署的策略等。因此，在该实例中，策略生成模块 224 自动搜索新的安全策略来部署。此外，不同于参照图 11 所描述的实例，通过自动地调整参数来递增地执行该实例并且不需要用户交互。

[0135] 示例性情况研究

[0136] 以下描述三种情况研究。第一情况研究表示针对第一情况研究选择简单栅格世界的结果。第二情况研究表明第三算法对于部分可观察性来说是稳健的。第三情况研究使用系统识别技术以近似真实的世界数字市场应用。

[0137] 4×4 栅格世界

[0138] 该实例开始于具有确定转换的 4×4 栅格世界。每个状态都导致 -0.1 的奖励，除了最右下的状态（其导致 0 的奖励并且为末端）。如果终端状态还没有准备达到并且“$\gamma = 1$”，则在“T”个步骤之后终止插曲 (Episode)。最佳策略的期望返回值为 -0.5。当“$T = 10$”时，最差策略具有“-1”的期望返回值，当“$T = 20$”时，最差策略具有“-2”的期望返回值，以及当“$T = 30$”时，最差策略具有“-3”的期望返回值。选择手工制造的初始策略，其执行得很好但留有改进的余地，并且“f_{min}”被设置为该策略的期望返回值的估计（注意，“f_{min}”随着“T”变化）。最后，“$k = 50$”且“$\delta = 0.05^5$”。

[0139] 图 13 示出了关于该问题的执行该策略改进技术和算法 3 的结果 1300。两种情况下的所有报告期望返回值都通过使用每个策略生成 10^5 个轨迹并计算蒙特卡洛返回值来

计算。示出了当“T = 20”时由批量策略改进技术生成的策略的期望返回值。初始策略具有 -1.06 的期望返回值，并且最佳策略具有 -0.5 的期望返回值。在顶部实例中还示出了来自三个试验的标准错误条。在底部实例中，利用各种“T”示出了由算法 3 和 NAC 以及相对于 1000 个插曲所生成的策略的期望返回值 (NAC 曲线用于“T = 20”)。每条曲线相对于十个试验求平均，并且最大的标准错误为 0.067。曲线将 1000/k-20 个调用扩展到策略改进技术。

[0140] 算法 3 与使用 LSTD 的偏置自然评估决策 (NAC) 相比，在每个插曲之后被修改为清楚的合格轨迹。尽管 NAC 不是安全的，但其提供了基线来示出算法 3 可以添加其安全保证而不牺牲显著量的学习速度。结果是尤其印象深刻的，因为为 NAC 示出的性能使用手动调整的步长和策略更新频率，而对于算法 3 没有调整超参数。注意，由于集中不等式的选择，性能不会随着最大轨迹长度的增加而劣化。

[0141] 注意，与利用几千个轨迹中的几百个实现的策略改进技术的批量应用相比，算法 3 使用几百个轨迹实现较大的期望返回值。这突出了算法 3 的显著特性，其中轨迹趋于从策略空间的越来越好的趋于中进行采样。与使用初始策略生成所有轨迹相比，这种探索提供了关于更好策略的值的更多信息。

[0142] 数字市场 POMDP

[0143] 第二情况研究包括产品的个别化广告的公司优化。在每个周期（时间步），公司具有三种选择：推销、售卖和 NULL。推销动作表示产品的推销而不具有生成中间销售（例如，提供关于产品的信息）的直接意图，这导致市场损失。售卖动作表示具有生成中间销售的直接意图的产品推销（例如，提供关于产品的销售）。NULL 动作表示不推销产品。

[0144] 用户行为的底层模型基于新近和频率方案。新近“r”是指用户进行购买需要多长时间，而频率“f”是指用户进行了多少次购买。为了更好地建模用户行为，向模型添加真实值项，用户统计 (cs)。该项依赖于用户与公司的整体交互并且不可观察，即，公司没有方式来对其进行测量。这种隐藏状态变量允许更多感兴趣的动力研究。例如，如果公司试图在用户购买产品之后的一周期中向用户售卖产品，则“cs”可以降低（购买产品的用户可能不喜欢在几个月之后看到更低价格的广告，但是可能喜欢不基于打折的促销）。

[0145] 所得到的 POMDP 具有 36 种状态和一个真实值隐藏状态、3 个动作，“T = 36”且“γ = 0.95”。选择“k = 50”、“δ = 0.05”的值，并且初始策略执行得很好但具有改进的余地。其期望返回值近似为 0.2，而最佳策略的期望返回值近似为 1.9 且最差策略的期望返回值近似为 -0.4。选择“f_{min}= 0.18”的值，这表示不多于 10％的收益劣化是可接受的。

[0146] 图 14 表示示例性结果 1400，其再次与具有手动优化的超参数的 NAC 的性能进行比较。为了强度 NAC 不是安全算法，当步长是手动优化值的两倍时也示出 NAC 的性能。该实例示出了算法 3 相对于传统 RL 算法的优势，尤其对于高风险应用来说。再次，对于算法 3 来说不调整超参数。尽管 NAC 以优化的超参数执行得很好，但这些参数通常未知，并且在针对良好超参数的搜索期间可以执行不安全的超参数。此外，即使利用优化的超参数，NAC 也不提供安全性保证（尽管经验上说是安全的）。

[0147] 使用真实世界数据的数字市场

[0148] Adobe® 市场云是强有力的工具集合，其允许公司完全使用自动和手动解决方

案来利用数字市场。Adobe® 目标工具的一个部件允许广告和活动的用户专用目标。当用户请求包含广告的网页时，基于包含用户的所有已知特性的矢量来计算示出哪个广告的判定。

[0149] 该问题趋向于视为土匪问题，其中代理人处理每个广告作为可能动作并且其试图最大化用户点击广告的概率。尽管该方法是成功的，但其不是必须也使每个用户在他或她的寿命期间点击的总数最大化。已经表明，该问题的更有远见的强化学习方法可以显著改进目光短浅的土匪解决方案。

[0150] 产生真实值特征的矢量 31，其提供关于用户的所有可用信息的压缩表示。广告被分为两个高级组，代理人从中进行选择。在代理人选择广告之后，用户点击 (+1 的奖励）或者不点击 (0 的奖励），并且描述描述的特征矢量被更新，选择“T = 10”。

[0151] 在该实例中，奖励信号是稀疏的，使得如果总是以 0.5 的概率选择每个动作，则奖励大约 0.48％的转换，因为用户总是不点击广告。这意味着大多数轨迹不提供反馈。此外，用户是否点击接近随机，使得返回值具有相对较高的变化。这导致梯度和自然梯度估计的大变化。

[0152] 使用具有三阶解耦傅里叶基础的 Softmax 动作选择，算法 3 被应用于该领域。进行“ δ = 0.05”的选择，其中“f_{min}= 0.48”并且初始策略被使用得稍好于随机。仅基于其中没有优化超参数的先验运行时间考虑来选择“k = 100000”的值。在图 15 中提供了结果 1500。在五个试验上平均点，并且提供标准错误条。在 500000 先验（即，用户交互）上，算法 3 能够安全地增加点击概率，从 0.49％到 0.61％ -a 24％改进。该使得研究表明如何将算法 3 用于实际世界应用的详细模拟。不仅可以由于其安全保证而负责地部署，而且其实现对实践时间标度可进行安全学习的显著的数据效率。

[0153] 示例性系统和设备

[0154] 图 16 示出了以 1600 表示的示例性系统，其包括代表可实施本文描述的各种技术的一个或多个计算系统和 / 设备的示例性计算设备 1602。这通过包括策略管理模块 122 来示出。例如，计算设备 1602 可以是服务提供器的服务器、与客户（例如，客户设备）相关联的设备、芯片上系统和 / 或任何其他适当的计算设备或计算系统。

[0155] 如图所示，示例性计算设备 1602 包括处理系统 1604、一个或多个计算机可读介质 1606 以及一个或多个 I/O 接口 1608，它们相互通信耦合。尽管未示出，但计算设备 1602 可以进一步包括系统总线或其他数据和命令传送系统，它们将各个部件相互耦合。系统总线可以包括不同总线结构的任何一种或组合，诸如存储总线或存储控制器、外围总线、通用串行总线和 / 或利用各种总线架构中的任何一种的处理器或本地总线。还预期各种其他实例，诸如控制和数据线。

[0156] 处理系统 1604 表示使用硬件执行一个或多个操作的功能。因此，处理系统 1604 被示为包括硬件元件 1610，其可以被配置为处理器、功能块等。这可以包括硬件的实施方式作为使用一个或多个半导体形成的专用集成电路或其他逻辑设备。硬件元件 1610 不被形成它们的材料或其中使用的处理机制所限制。例如，处理器可以由半导体和 / 或晶体管组成（例如，电子集成电路 (IC)）。在这种情况下，处理器可执行指令可以是电可执行指令。

[0157] 计算机可读存储介质 1606 被示为包括存储器 / 存储 1612。存储器 / 存储 1612 表示与一个或多个计算机可读介质相关联的存储器 / 存储能力。存储器 / 存储 1612 可以包括

易失性介质（诸如随机存取存储器（RAM））和/或非易失性介质（诸如只读存储器（ROM）、闪存、光盘、磁盘等）。存储器/存储1612可以包括固定介质（例如，RAM、ROM、固定硬盘驱动等）以及可移除介质（例如，闪存、可移除硬盘驱动、光盘等）。计算机可读介质1606可以下面进一步描述的各种其他方式来配置。

[0158] 输入/输出接口1608表示允许用户向计算设备1602输入命令和信息的功能，并且还允许使用各种输入/输出设备将信息呈现给用户和/或其他部件或设备。输入设备的实例包括键盘、光标控制设备（例如，鼠标）、麦克风、扫描仪、触摸功能（例如，被配置为检测物理触摸的电容或其他传感器）、相机（例如，可使用可见或不可见波长，诸如红外频率，以根据不涉及触摸的姿势来识别移动）等。输出设备的实例包括显示设备（例如，监控器或投影仪）、扬声器、打印机、网络卡、触觉响应设备等。因此，可以以下面进一步描述的各种方式来配置计算设备1602以支持用户交互。

[0159] 可以在软件、硬件元件或程序模块的一般条件下描述各种技术。通常，这种模块包括例行程序、程序、对象、元件、部件、数据结构等，它们执行特定的任务或实施特定的抽象数据类型。本文使用的术语“模块”、“功能”和“部件”通常表示软件、固件、硬件或它们的组合。本文所描述的技术的特征是不依赖于平台的，意味着可以在具有各种处理器的商业计算平台上实施技术。

[0160] 所描述模块和技术的实施方式可以存储在一些形式的计算机可读介质上或者横跨一些形式的计算机可读介质进行传输。计算机可读介质可以包括各种可被计算设备1602访问的介质。通过实例但不限制，计算机可读介质可以包括“计算机可读存储介质”和“计算机可读信号介质”。

[0161] “计算机可读存储介质”可表示与仅进行信号传输、载波或信号本身相比能够进行信号的永久和/或非暂态存储的介质和/或设备。因此，计算机可读存储介质表示非信号承载介质。计算机可读存储介质包括硬件，诸如易失性和非易失性、可移除和不可移除介质和/或以适合于存储信息（诸如计算机可读指令、数据结构、程序模块、逻辑元件/电路或其他数据）的方法或技术实施的存储设备。计算机可读存储介质的实例可以包括但不限于RAM、ROM、EEPROM、闪存或其他存储技术、CD-ROM、数字通用盘（DVD）或其他光学存储器、硬盘、磁带盒、磁带、磁盘存储或其他磁性存储设备或其他存储设备、可触介质或者适合于存储期望信息并可以被计算机访问的制造品。

[0162] “计算机可读信号介质”可以表示信号承载介质，其被配置为例如经由网络向计算设备1602的硬件传输指令。信号介质通常可以具体化计算机可读指令、数据结构、程序模块或调制数据信号的其他数据（诸如载波、数据信号或其他传输机制）。术语“调制数据信号”表示具有以编码信号中的信息的这种方式设置或改变其特性中的一个或多个的信号。通过实例但不限制，通信介质包括有线介质（诸如有线网络或直接有线连接）和无线介质（诸如声学、RF、红外和其他无线介质）。

[0163] 如前所述，硬件元件1610和计算机可读介质1606表示以硬件形式实施的模块、可编程设备逻辑和/或固定设备逻辑，其在一些实施例中可以用于实施本文所描述技术的至少一些方面，诸如执行一个或多个指令。硬件可以包括集成电路或芯片上系统、专用集成电路（ASIC）、现场可编程门阵列（FPGA）、复杂可编程逻辑设备（CPLD）的部件以及硅或其他硬件的其他实施方式。在这种情况下，硬件可以操作为处理设备，其处理由硬件具体化的指令

和 / 或逻辑所限定的程序以及用于存储指令用于执行的硬件（例如，先前描述的计算机可读存储介质）。

[0164] 前述的组合还可以用于实施本文描述的各种技术。因此，软件、硬件或可执行模块可以实施为在一些形式的计算机可读存储介质上和 / 或通过一个或多个硬件元件 1610 具体化的一个或多个指令和 / 或逻辑。计算设备 1602 可以被配置为实施与软件和 / 或硬件模块相对应的特定指令和 / 或功能。因此，被计算设备 1602 可执行为软件的模块的实施方式可以至少部分地以硬件来实现，例如通过使用计算机可读存储介质和 / 或处理系统 1604 的硬件元件 1610。指令和 / 或功能可被一个或多个制造品（例如，一个或多个计算设备 1602 和 / 或处理系统 1604）执行 / 操作以实施本文描述的技术、模块和实例。

[0165] 本文所述的技术可以被计算设备 1602 的各种结构来支持并且不限于本文所述技术的具体实例。该功能还可以全部或部分通过使用分布式系统来实施，诸如下面描述的经由平台 1618 在“云”1614 上。

[0166] 云 1614 包括和 / 或表示用于砖 1616 的平台 1618。平台 1618 抽象云 1614 的硬件（例如，服务器）和软件的潜在功能。资源 1616 可以包括应用和 / 或数据，其可以在远离计算设备 1602 的服务器上执行计算机处理的同时被利用。资源 1616 还可以包括在因特网上提供和 / 或通过用户网络（诸如蜂窝或 Wi-Fi 网络）提供的服务。

[0167] 平台 1618 抽象资源和功能以连接计算设备 1602 与其他计算设备。平台 1618 还可以用于抽象资源的缩放以提供针对经由平台 1618 实施的资源 1616 的遭遇请求的对应等级。因此，在互连设备实施例中，本文所描述的功能的实施方式可以在系统 1600 中分布。例如，可以部分地在计算设备 1602 上以及经由抽象云 1614 的功能的平台 1618 来实施功能。

[0168] 结论

[0169] 尽管以特定的结构特征和 / 或方法逻辑动作描述了本发明，但应该理解，所附权利要求中限定的发明不是必须限于所描述的特定特征或动作。此外，具体特征和动作被公开为实施所要求发明的示例性形式。

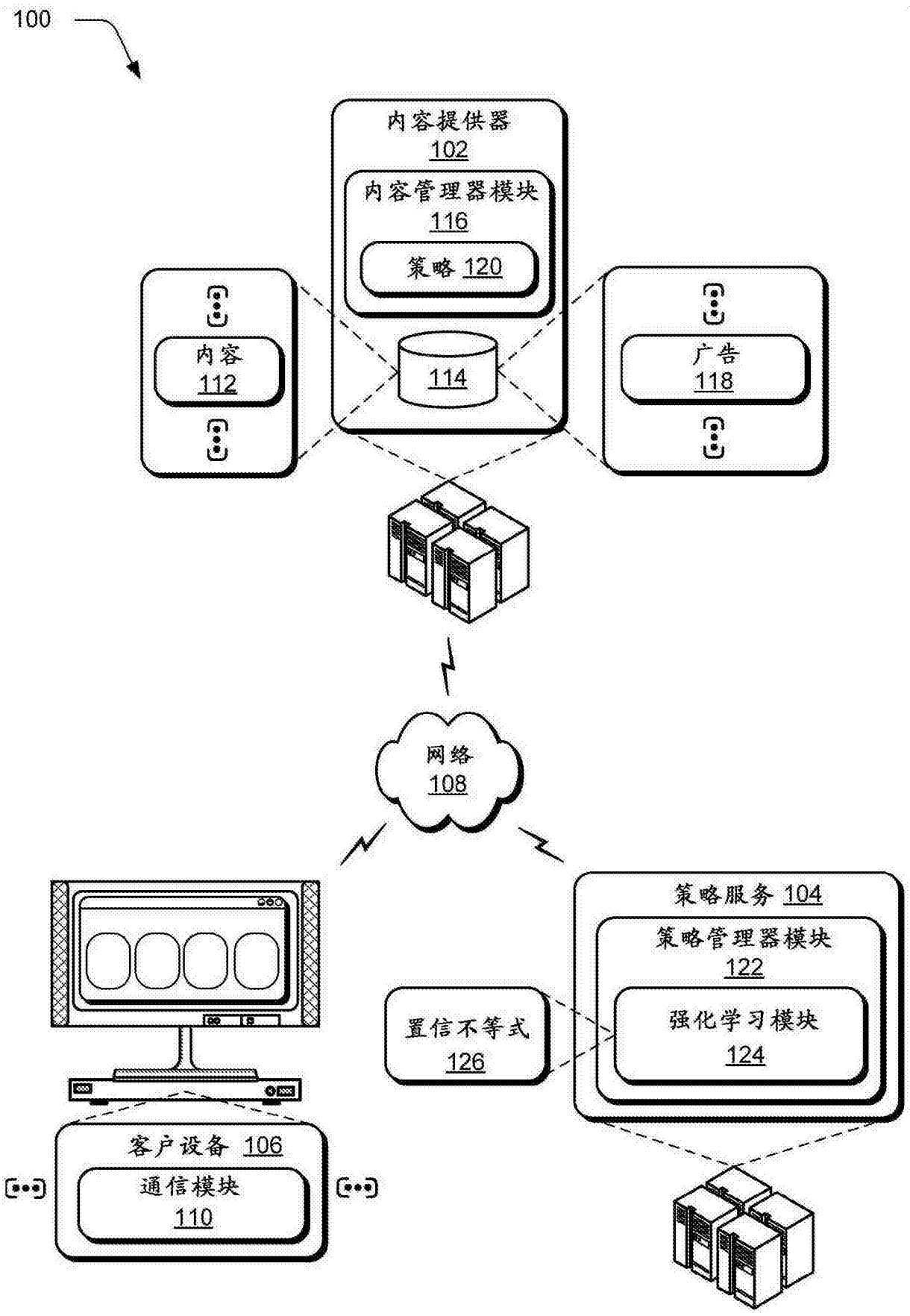

100
内容提供器
102
内容管理器模块
116
策略 120
114
内容
112
广告
118
网络
108
策略服务 104
策略管理器模块
122
强化学习模块
124
置信不等式
126
客户设备 106
通信模块
110

图 1

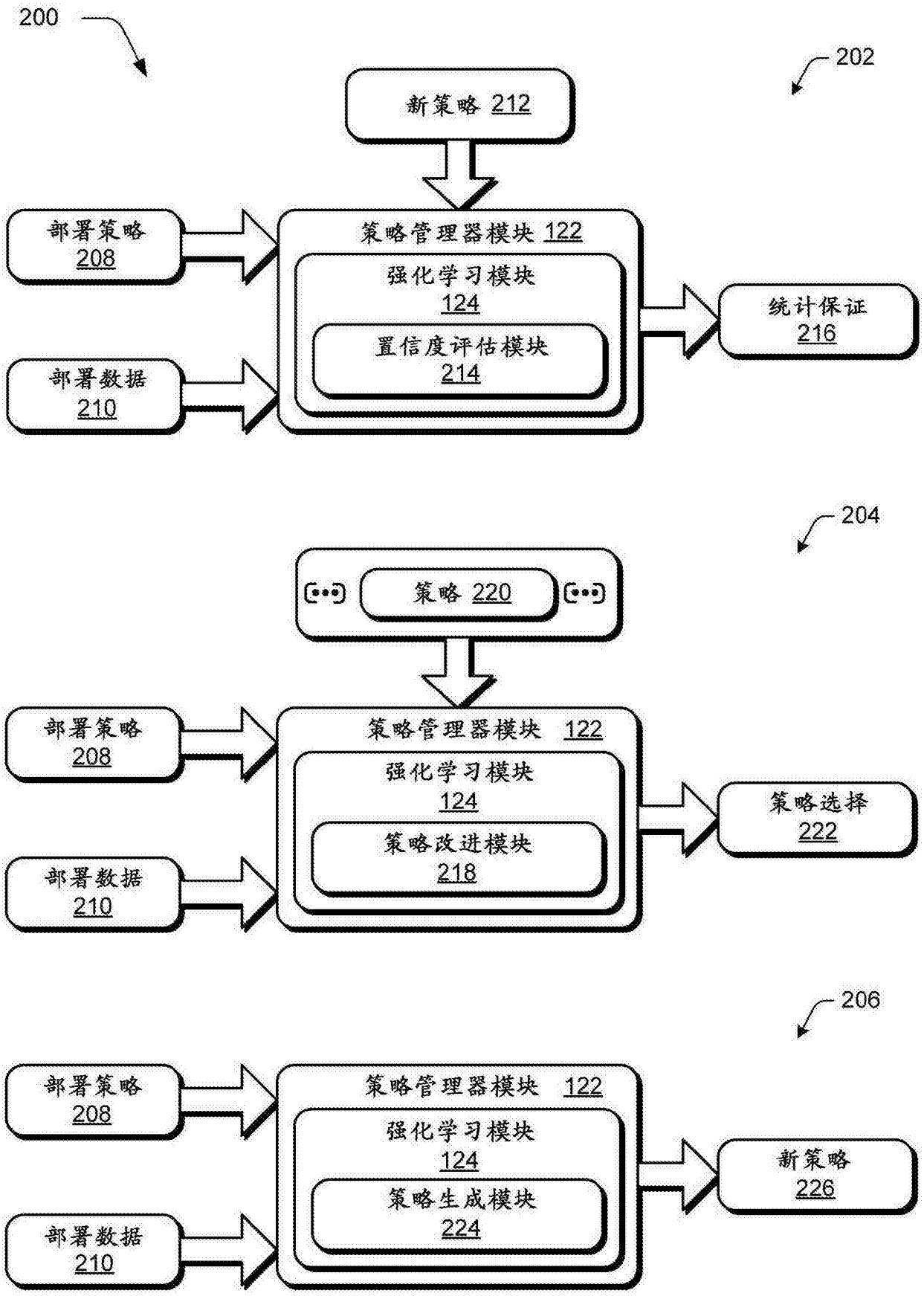

200
202
新策略 212
部署策略 208
策略管理器模块 122
强化学习模块 124
置信度评估模块 214
统计保证 216
部署数据 210
204
策略 220
部署策略 208
策略管理器模块 122
强化学习模块 124
策略改进模块 218
策略选择 222
部署数据 210
206
部署策略 208
策略管理器模块 122
强化学习模块 124
策略生成模块 224
新策略 226
部署数据 210

图 2

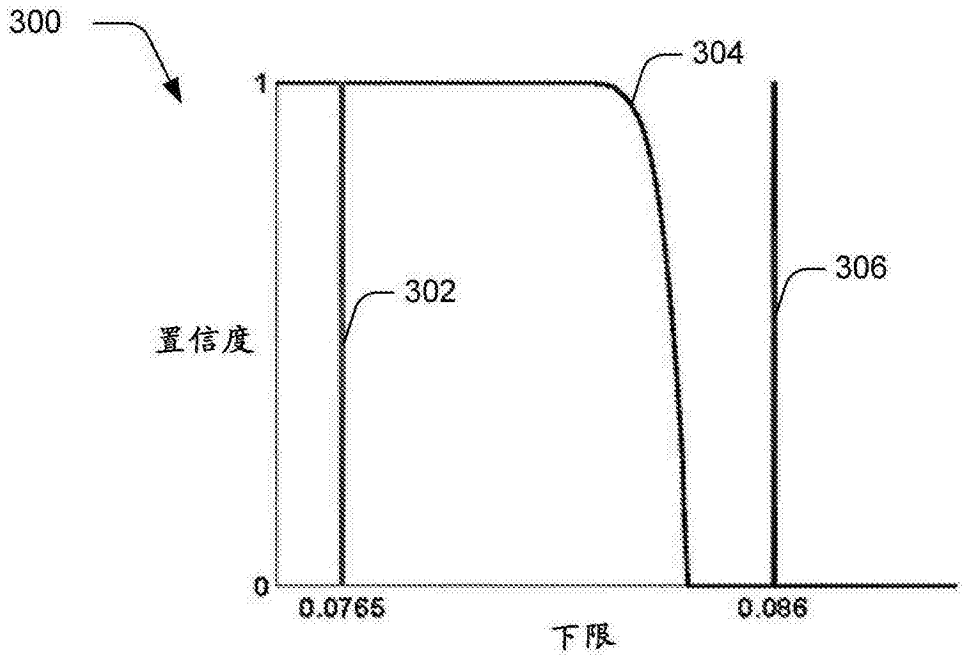


图 3A

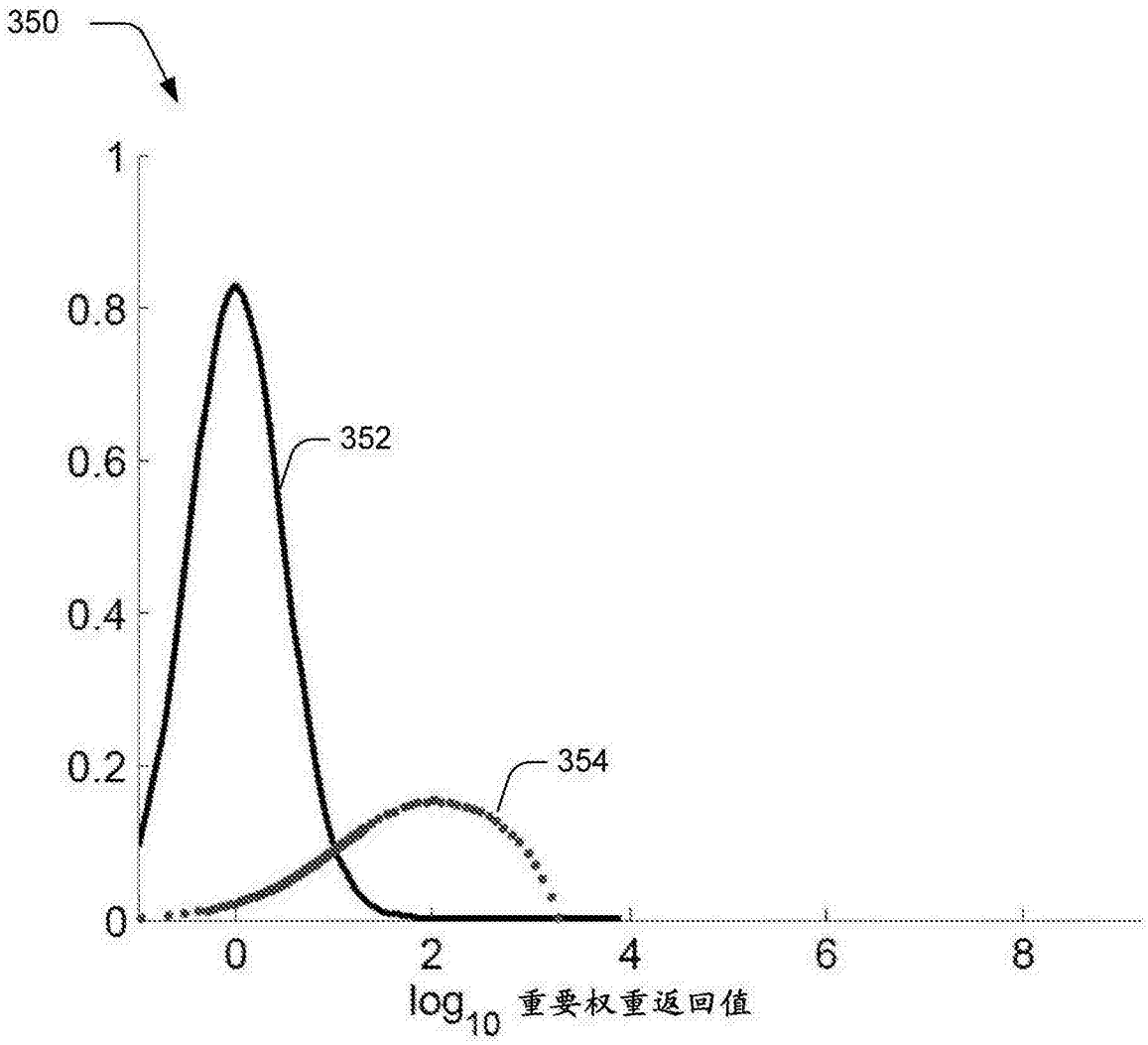


图 3B

400

	定理1	CH	MPeB	AM	塌陷-AM
针对平均值的95%置信度下限	0.145	-5.831×10^{6}	−129703	0.092	0.087

图 4

500

$$\hat{f}(D,\theta,c,\delta) := \frac{1}{n}\sum_{\tau_i \in D}^{n} \min\{\hat{f}(\theta,\tau_i,\theta_i),c\} - \sqrt{\frac{2\ln(2/\delta)}{n^2(n-1)}\left(n\sum_{\tau_i \in D}^{n} \min\{\hat{f}(\theta,\tau_i,\theta_i),c\}^2 - \left(\sum_{\tau_i \in D}^{n} \min\{\hat{f}(\theta,\tau_i,\theta_i),c\}\right)^2\right)} - \frac{7c\ln(2/\delta)}{3(n-1)}$$

图 5

600

算法1: $\mathrm{IsSAFE}(\theta, D, \hat{f}_{min}, \delta)$ — 返回在给定D的情况下指定θ是否安全的布尔值

$D_{pre} = D_{test} = \emptyset$;
for each *behavior policy* **do**
  Add [1/3] of the trajectories from the behavior policy to D_{pre} and the other [2/3] to D_{test};
$c \in \max\{1, \arg\max_c \hat{f}(D_{pre}, \theta, c, \delta)\}$;
return $\hat{f}_{min} < \hat{f}(D_{test}, \theta, c, \delta)$;

图 6

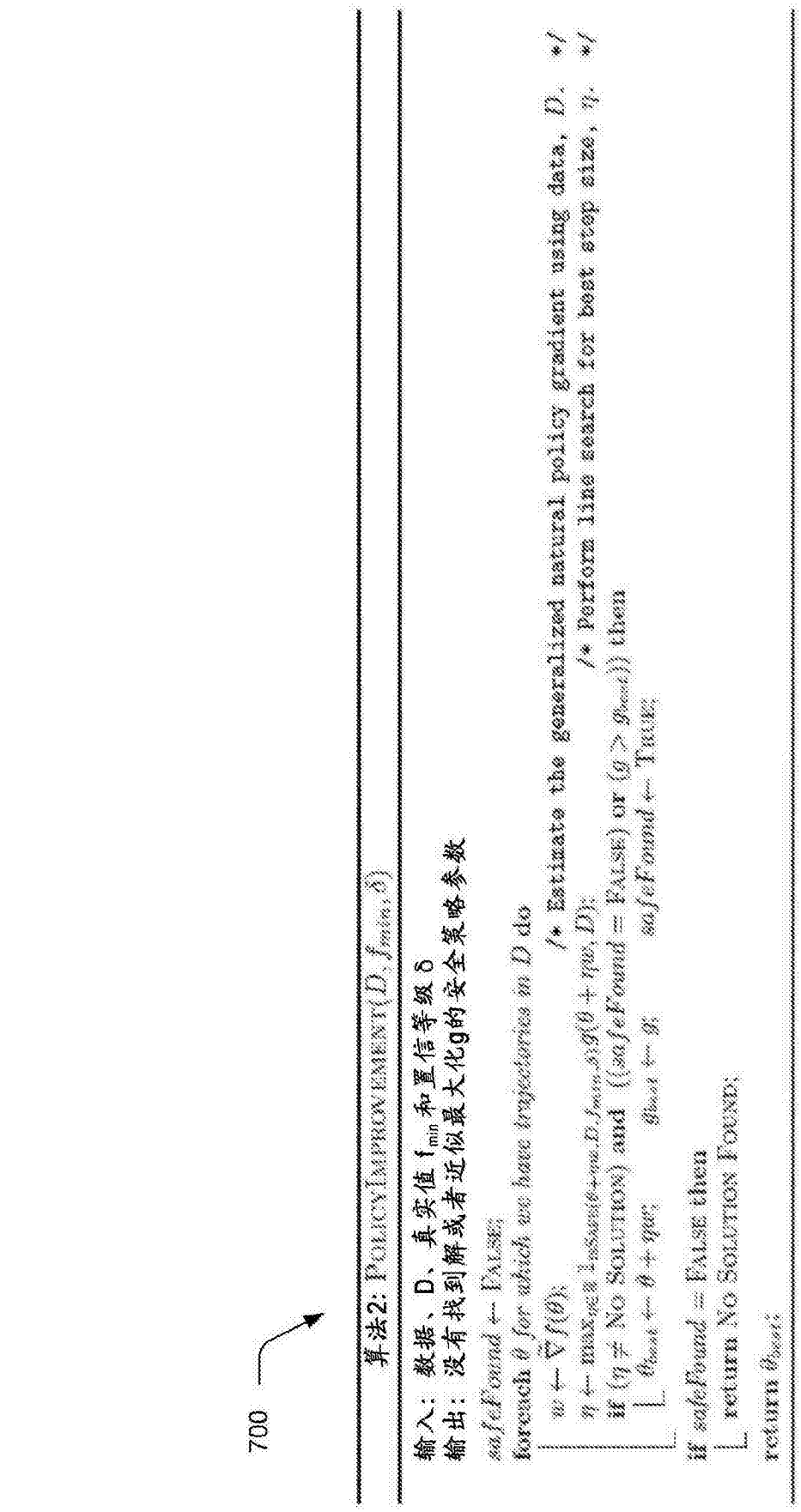

700
算法2: POLICYIMPROVEMENT(D, f_{min}, δ)
输入：数据、D、真实值 f_{min} 和置信等级 δ
输出：没有找到解或者近似最大化g的安全策略参数
safeFound ← FALSE;
foreach θ for which we have trajectories in D do
w ← ∇f(θ); /* Estimate the generalized natural policy gradient using data, D. */
η ← maxη LineSearch(η, D, f_min, δ)g(θ + ηw, D); /* Perform line search for best step size, η. */
if (η ≠ NO SOLUTION) and ((safeFound = FALSE) or (g > g_best)) then
θ_best ← θ + ηw; g_best ← g; safeFound ← TRUE;
if safeFound = FALSE then
return NO SOLUTION FOUND;
return θ_best;

图 7

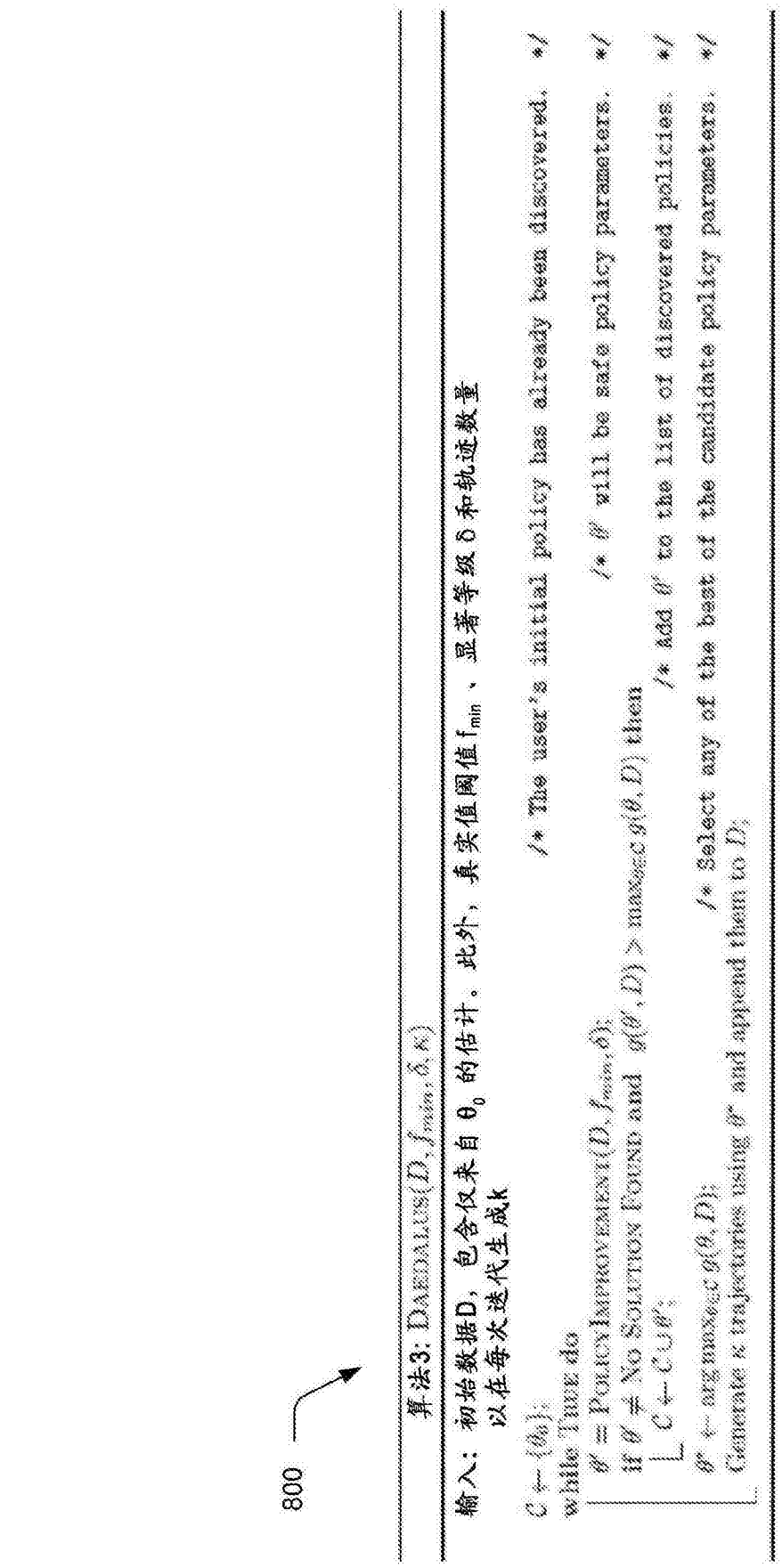

800
算法3: DAEDALUS(D, f_{min}, δ, k)
输入：初始数据D，包含仅来自 θ_0 的估计。此外，真实值阈值 f_{min} 、显著等级 δ 和轨迹数量
以在每次迭代生成k
$C \leftarrow \{\theta_0\}$; /* The user's initial policy has already been discovered. */
while TRUE do
$\theta' = \text{POLICYIMPROVEMENT}(D, f_{min}, \delta)$; /* θ' will be safe policy parameters. */
if $\theta' \neq$ NO SOLUTION FOUND and $g(\theta', D) > \max_{\theta \in C} g(\theta, D)$ then
$C \leftarrow C \cup \theta'$; /* Add θ' to the list of discovered policies. */
$\theta' \leftarrow \text{argmax}_{\theta \in C} g(\theta, D)$; /* Select any of the best of the candidate policy parameters. */
Generate k trajectories using θ' and append them to D;

图 8

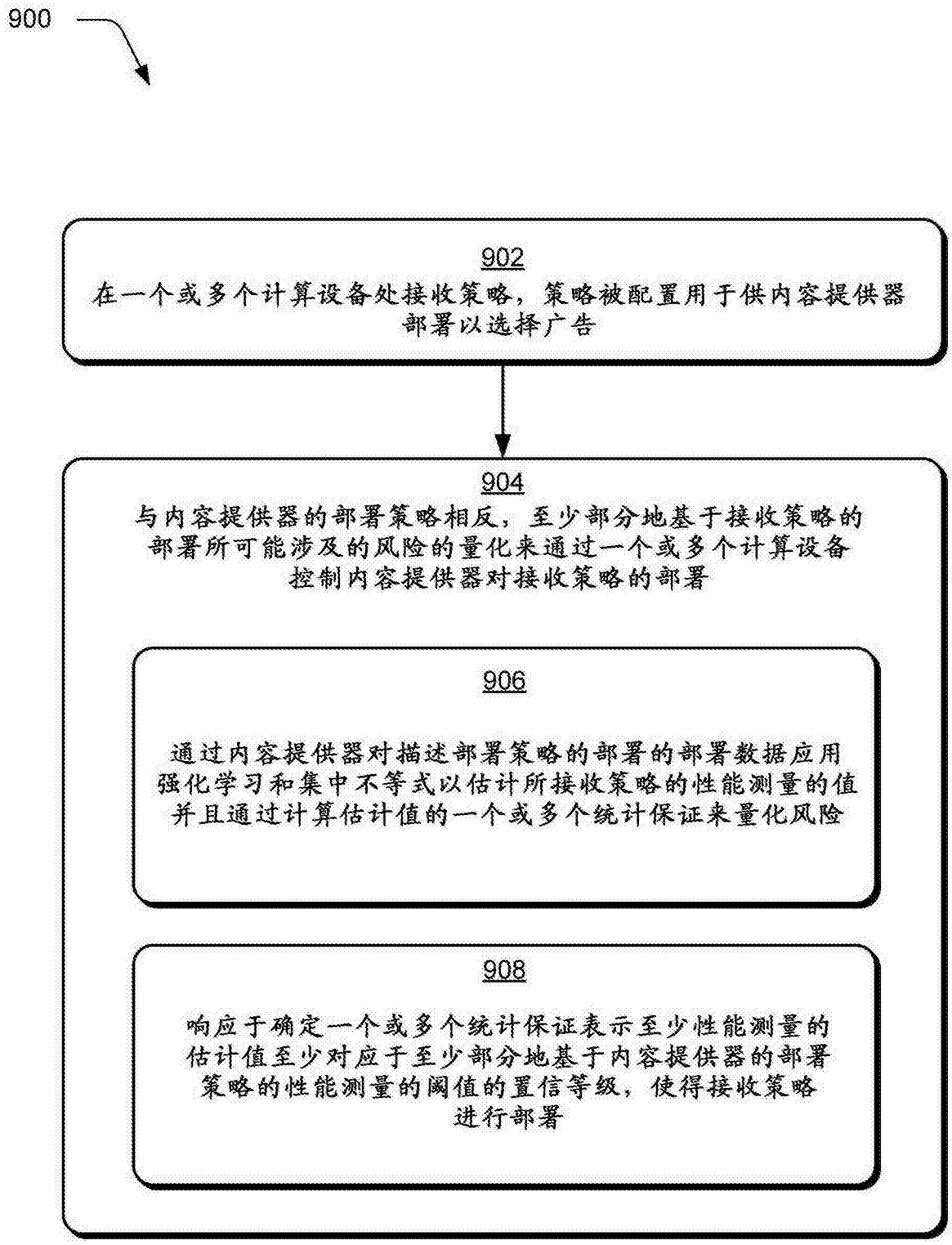

900
902
在一个或多个计算设备处接收策略，策略被配置用于供内容提供器部署以选择广告
904
与内容提供器的部署策略相反，至少部分地基于接收策略的部署所可能涉及的风险的量化来通过一个或多个计算设备控制内容提供器对接收策略的部署
906
通过内容提供器对描述部署策略的部署的部署数据应用强化学习和集中不等式以估计所接收策略的性能测量的值并且通过计算估计值的一个或多个统计保证来量化风险
908
响应于确定一个或多个统计保证表示至少性能测量的估计值至少对应于至少部分地基于内容提供器的部署策略的性能测量的阈值的置信等级，使得接收策略进行部署

图 9

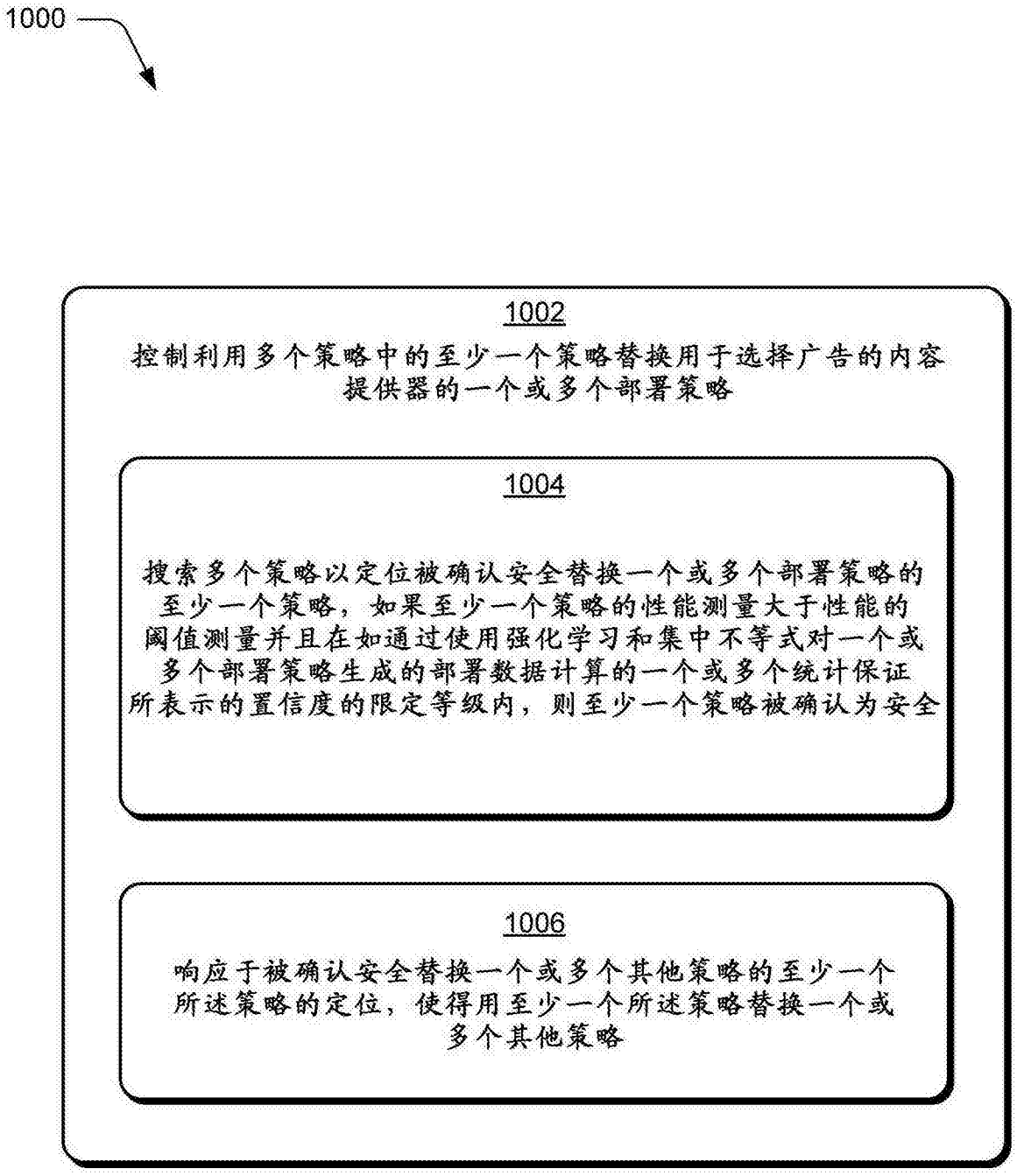

1000
1002
控制利用多个策略中的至少一个策略替换用于选择广告的内容提供器的一个或多个部署策略
1004
搜索多个策略以定位被确认安全替换一个或多个部署策略的至少一个策略，如果至少一个策略的性能测量大于性能的阈值测量并且在如通过使用强化学习和集中不等式对一个或多个部署策略生成的部署数据计算的一个或多个统计保证所表示的置信度的限定等级内，则至少一个策略被确认为安全
1006
响应于被确认安全替换一个或多个其他策略的至少一个所述策略的定位，使得用至少一个所述策略替换一个或多个其他策略

图 10

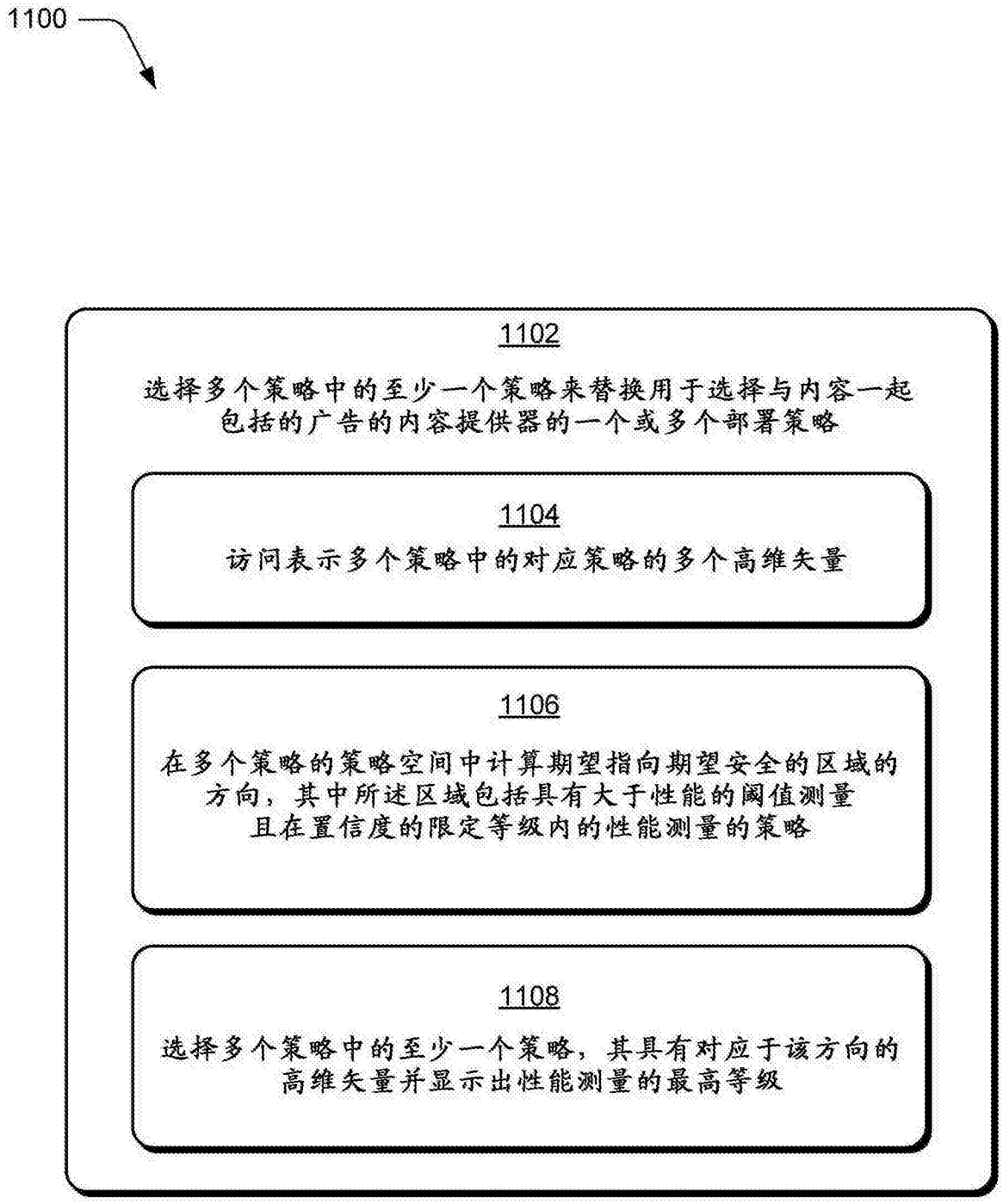

1100
1102
选择多个策略中的至少一个策略来替换用于选择与内容一起包括的广告的内容提供器的一个或多个部署策略
1104
访问表示多个策略中的对应策略的多个高维矢量
1106
在多个策略的策略空间中计算期望指向期望安全的区域的方向，其中所述区域包括具有大于性能的阈值测量且在置信度的限定等级内的性能测量的策略
1108
选择多个策略中的至少一个策略，其具有对应于该方向的高维矢量并显示出性能测量的最高等级

图 11

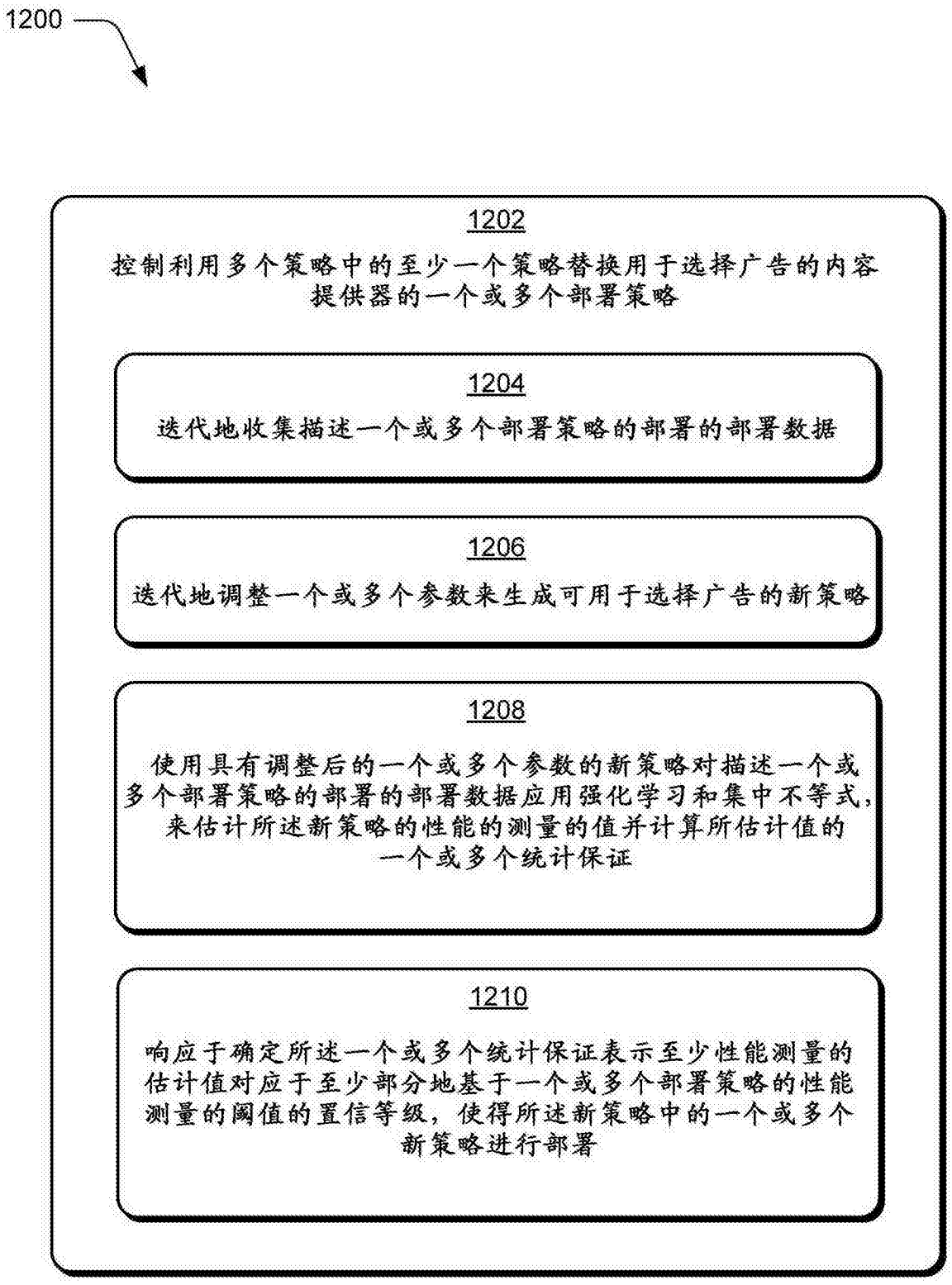

1200
1202
控制利用多个策略中的至少一个策略替换用于选择广告的内容提供器的一个或多个部署策略
1204
迭代地收集描述一个或多个部署策略的部署的部署数据
1206
迭代地调整一个或多个参数来生成可用于选择广告的新策略
1208
使用具有调整后的一个或多个参数的新策略对描述一个或多个部署策略的部署的部署数据应用强化学习和集中不等式，来估计所述新策略的性能的测量的值并计算所估计值的一个或多个统计保证
1210
响应于确定所述一个或多个统计保证表示至少性能测量的估计值对应于至少部分地基于一个或多个部署策略的性能测量的阈值的置信等级，使得所述新策略中的一个或多个新策略进行部署

图 12

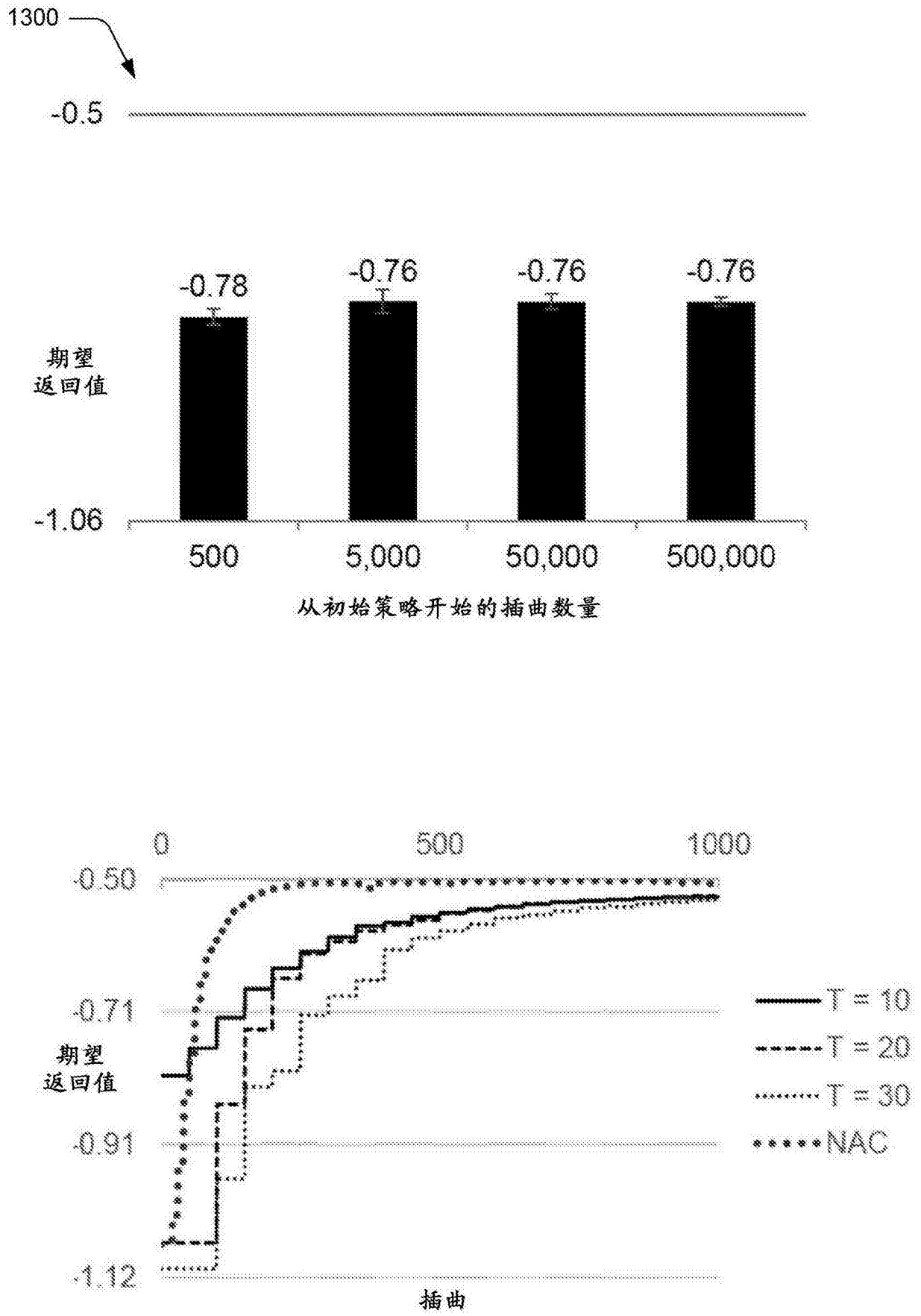

1300
-0.5
-0.78
-0.76
-0.76
-0.76
期望
返回值
-1.06
500
5,000
50,000
500,000
从初始策略开始的插曲数量
0
500
1000
-0.50
-0.71
期望
返回值
-0.91
-1.12
插曲
T = 10
T = 20
T = 30
NAC

图 13

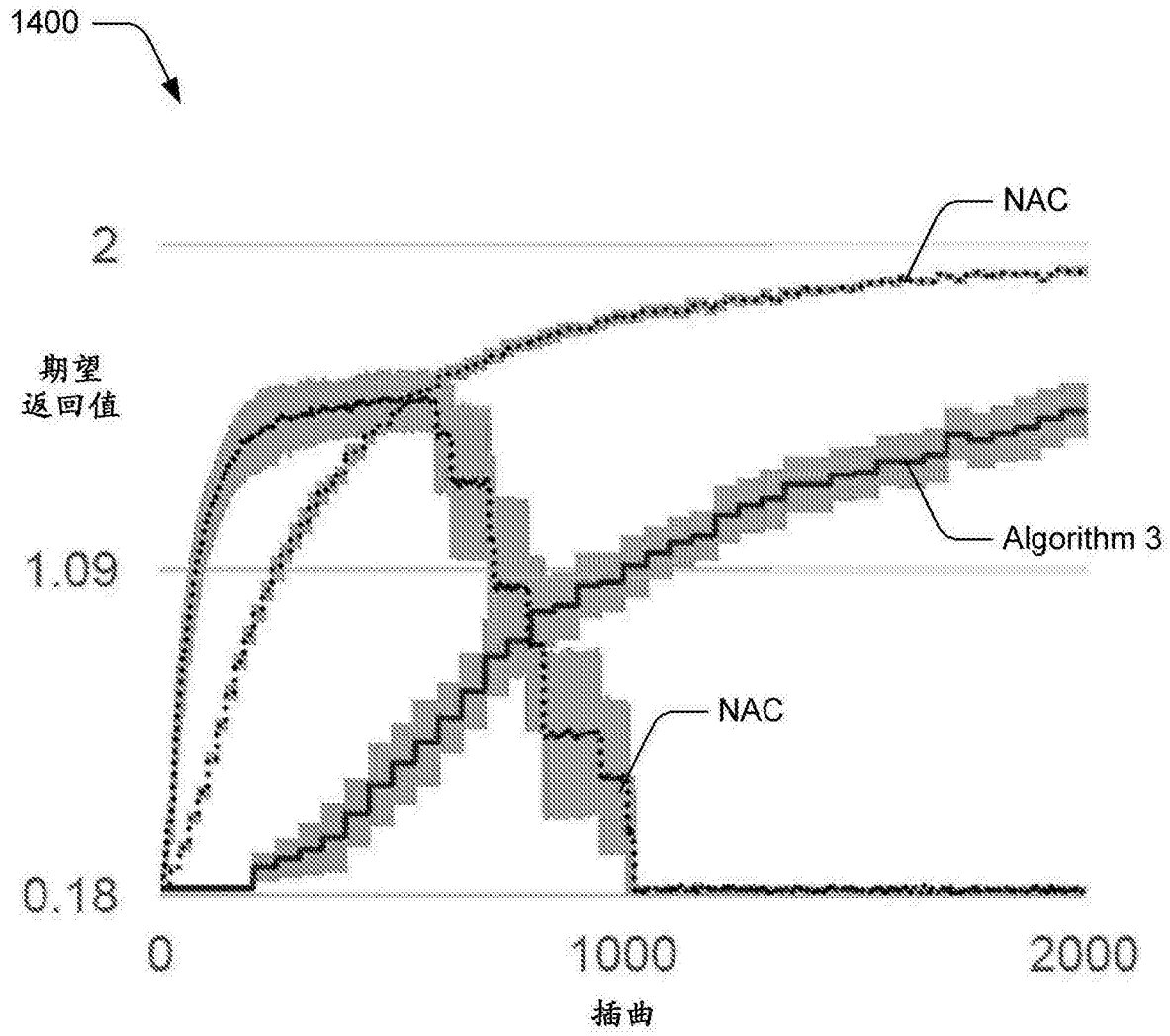


图 14

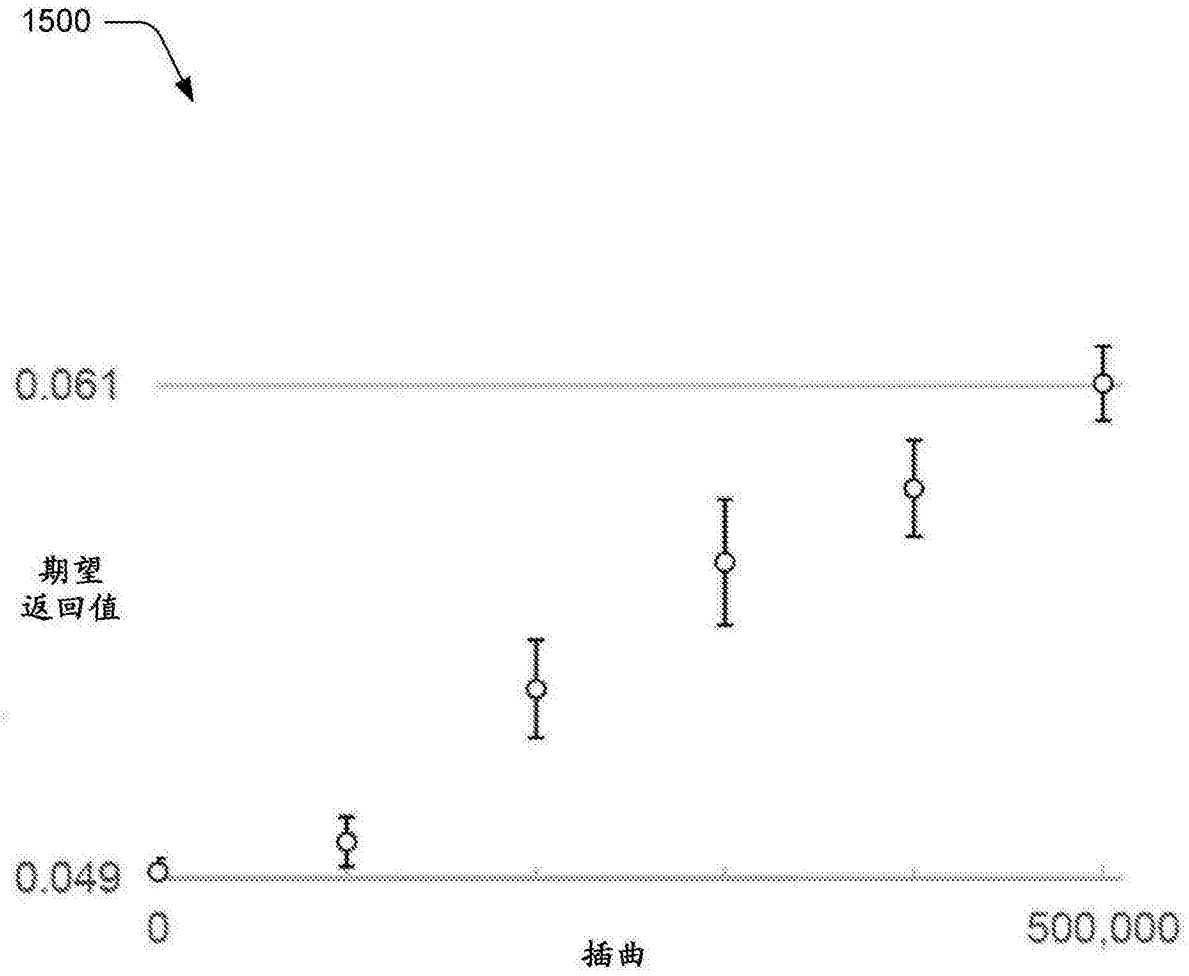


图 15

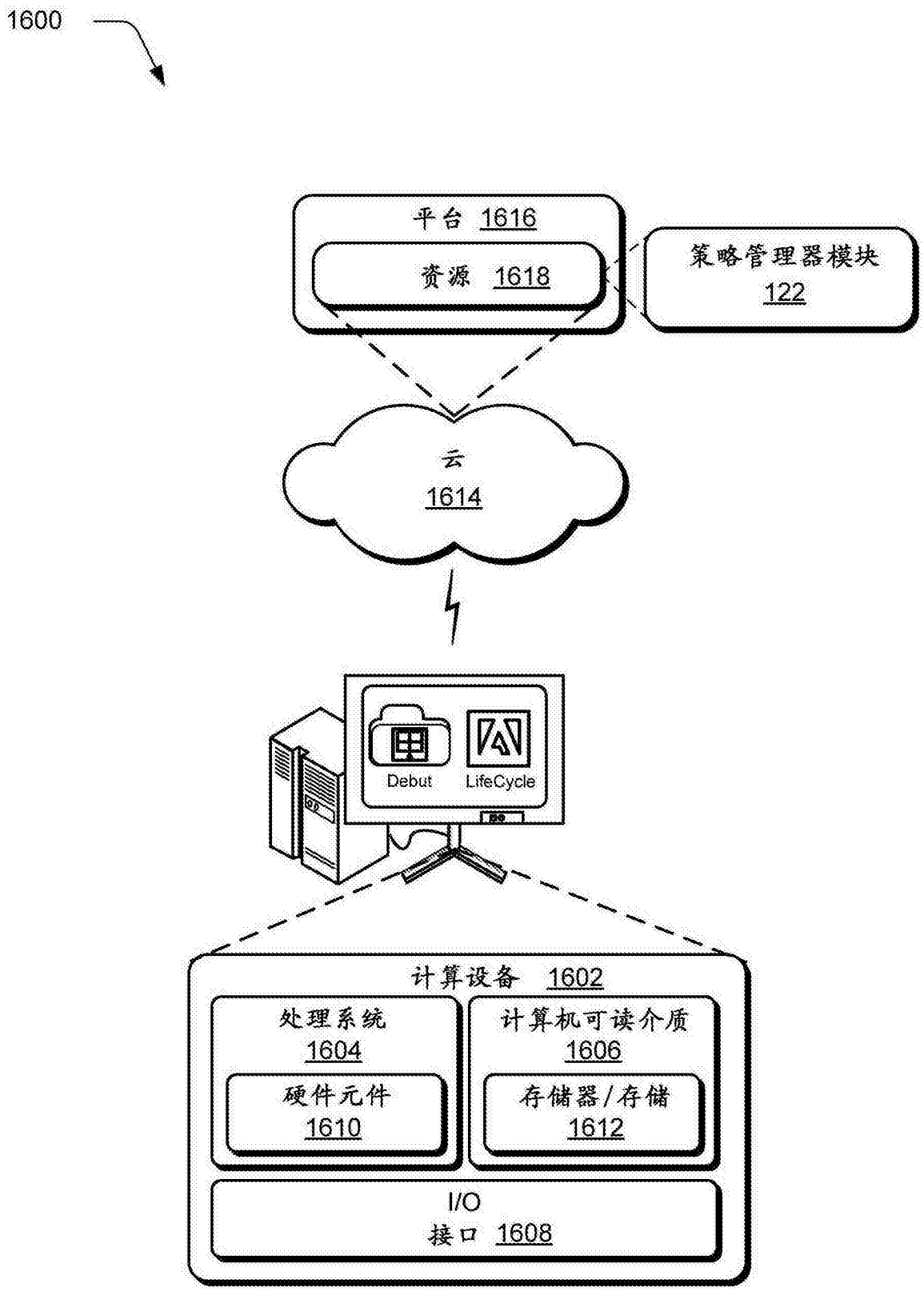

1600
平台 1616
资源 1618
策略管理器模块
122
云
1614
Debut
LifeCycle
计算设备 1602
处理系统
1604
硬件元件
1610
计算机可读介质
1606
存储器/存储
1612
I/O
接口 1608

图 16