ICML 2019 会议笔记
美国加利福尼亚州长滩

大卫・阿贝尔*
david_abel@brown.edu

2019年6月

目录

*http://david-abel.github.io

这份文件包含了我在ICML会议上参加的活动期间所做的笔记，会议地点在美国加利福尼亚州长滩。请随意传阅，并在发现任何拼写错误或其他需要更正的地方时给我发送电子邮件至david_abel@brown.edu。

1 会议亮点

这一轮我大部分时间都在强化学习会议上度过（可惜错过了所有的主题演讲），所以我的反思（和笔记）主要集中在强化学习方面：

1. 关于离线策略评估和离线策略学习有很多出色的工作（例如，Hanna等人的工作[35]，Le等人的工作[49]，Fujimoto等人的工作[26]，Gottesman等人的工作[32]，以及第6.3节的演讲）。这些问题设置非常重要，因为我（和许多其他人）预计强化学习应用将会产生大量来自次优策略的数据。

2. 探索再次成为热门话题，这是理所当然的（参见Mavrin等人的工作[57]，Fatemi等人的工作[25]，Hazan等人的工作[37]，Shani等人的工作[76]）。 除了离线评估（和其他一些问题），这是强化学习中的基础问题之一，我们目前正处于一个有望取得重大进展的良好位置。

3. 一些非常好的工作继续澄清分布式强化学习[10]（参见[74, 57, 67]的工作）。

4. 周五举行的面向气候变化的人工智能研讨会非常棒，参与人数众多（我参加的演讲只有站立空间）。 我在之前的会议上也说过这个，但是：众所周知，存在着非常重要的问题，而机器学习的工具在其当前形式下可以非常有效地解决这些问题。

5. 我真的认为我们需要在强化学习中标准化评估。并不是说我们只需要一种方法来做评估，或者只需要一个领域，但目前评估协议的差异太大。

6. 喜欢RL for real life研讨会上的小组讨论（见第6.2.1节）

2 6月10日星期一：教程

开始了！我来参加PAC-Bayes教程的下半场。

2.1 教程：PAC-Bayes理论（第二部分）

演讲者是Benjamin Guedi和John Shawe-Taylor。

第一部分回顾：Shawe-Taylor和Williamson [77]对贝叶斯估计进行了PAC [86]分析（也见图1）。之后不久，McAllester [58]提出了第一个PAC-Bayesian界限：

定理1.（*McAllester [58]*）对于任何先验P，$\delta \in (0,1]$，我们有：

$$\Pr\left(\forall_{Q\in\mathcal{H}} : R_{out}(Q) \leq R_{in}(Q) + \sqrt{\frac{D_{KL}\left(Q \mid\mid P\right)) + \ln\frac{2\sqrt{m}}{\delta}}{2米}}\right) \geq 1-\delta, \tag{1}$$

其中 $\mathcal{H}$是假设空间， m是样本数量， R_{out}是假设在测试数据上的风险， $R_{in}(h)$是假设在训练数据上的风险， P是先验概率， Q是后验概率。

PAC-Bayes：一个灵活的学习理论框架！与回归、线性分类和支持向量机有紧密联系，用于转导学习，还可以在强化学习中使用[24]等等。

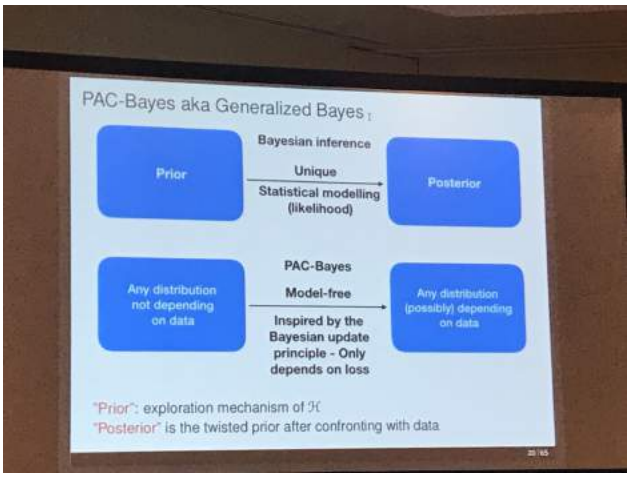


图1：贝叶斯和PAC-Bayes的区别

2.1.1 PAC-Bayes理论

问：PAC-Bayes如何推动学习？

答：首先，回顾一下：

$$R_{out}(Q) \leq R_{in}(Q) + F(Q), \tag{2}$$

或者：

$$\text{未见数据上的错误} \leq \text{样本上的错误} + \text{复杂度项}. \tag{3}$$

这定义了一种获得新算法的基本策略：

$$h \sim Q^* \tag{4}$$
$$Q^* \in \operatorname{arginf}_{Q \ll P} \left\{ R_{in}(Q) + F(Q) \right\}. \tag{5}$$

提出了一个优化问题：可以通过求解或近似求解来找到好的解决方案！

PAC-Bayes解释了著名算法：

- 使用sigmoid损失和KL正则化的Adaboost可以重新解释为PAC-Bayesian界的最小化器[5]。

- 此外，最小化器为：

$$\left\{ R_{in}(Q) + \frac{KL}{\lambda} \right\},$$

  是著名的Gibbs后验分布：

$$Q_\lambda(h) \propto \exp(-\lambda R_{in}(h)) P 9h), \quad \forall_{h \in \mathcal{H}}.$$

  当 $\lambda \to 0$时，我们得到一个平坦的后验分布，当 $\lambda \to \infty$时，我们得到期望风险最小化（ERMs）的Dirac质量。

定理2.

$$\log \int \exp \phi dP = \sup_{Q \ll P} \left\{ \int \phi dQ - D_{KL}(Q \mid\mid P) \right\}.$$

证明.首先需要KL散度的变分定义[18]

$$-D_{\mathrm{KL}}(Q \mid\mid G) = -\int \log \left(\frac{dQ}{dP} \frac{dP}{dG} \right) dQ \tag{6}$$
$$= -D_{\mathrm{KL}}(Q \mid\mid P) + \int \phi dp - \log \int \exp \phi dP. \tag{7}$$

注意KL> 0, $Q \to -D_{\mathrm{KL}}(Q \mid\mid P)$在 $Q = G$时达到最大值。 因此，取 $\phi = -\lambda R_{in}$:

$$Q_\lambda(h) \propto \exp(-\lambda R_{in}(h)) P(h), \quad \forall_{h \in \mathcal{H}}. \qquad \square$$

问：非独立同分布数据怎么办?

A: 当然！我们放弃独立同分布和有界损失的假设。首先，需要矩：

> **定义1**（矩）：分布的p阶矩为：
>
> $$M_p := \int \mathbb{E}\left[|R_{in}(h) - R_{out}(h)|^p \right] dP(h).$$

还可以利用 f-散度，它是KL散度的一般化。

定理3.对于任意分布Q，固定 $\phi_p : x \to x^p$，当$p > 1$，q 以至少$_{1-}$ δ的概率，对于任意分布 Q:

$$|R_{out}(Q) - R_{in}(Q)| \leq (\frac{M_q}{\delta}^{1/q} (D_{\phi_{p-1}}(Q, P) + 1)^{1/p}.$$

要点：我们可以使用 f-散度 (D_ϕ 来限制泛化误差 $_{p-1}$ 和矩 (M_q)。
证明策略需要：1) Jensen不等式，2) 测度变换，3) Holder不等式，和4) Markov不等式。

Oracle Bounds; Catoni [13] 推导了Gibbs后验的PAC-Bayesian界限。

2.1.2 PAC-Bayes和任务意识

注意：PAC-Bayesian界限表达了经验准确性和复杂度度量之间的权衡。

问：那么，我们如何改进我们得到的界限呢？ 我们如何选择合适的先验分布，以便我们可以1) 控制复杂度，和2) 确保良好的性能?

→所以：我们能选择一个“更好”的先验分布吗？ (不查看测试数据本身？)

主要思想：使用部分数据来学习如何选择先验分布。

可以在支持向量机中使用PAC-Bayes：

- 假设先验和后验都是球形高斯分布（先验以原点为中心，后验以单位SVM权重向量的缩放μ为中心）。
- 这意味着泛化误差界中的KL项为$\mu^2/2$（参见定理1）。
- 可以计算后验分布的随机误差，它的行为类似于软间隔，缩放μ在边缘损失和KL之间进行权衡。
- 该界对所有μ都成立，因此选择μ来优化界限。

问：但是我们如何学习SVM的先验知识?

- 界限取决于先验和后验之间的距离。
- 更好的先验意味着更紧的界限。
- 思路：用部分数据学习先验P。
- 将学习到的先验引入界限。
- 用剩余数据计算随机误差：PrPAC。
- 可以进一步进行：1）在选择的方向上缩放先验τ-PrPAC，或者2）调整SVM以优化新的界限：$\eta-$Prior SVM。

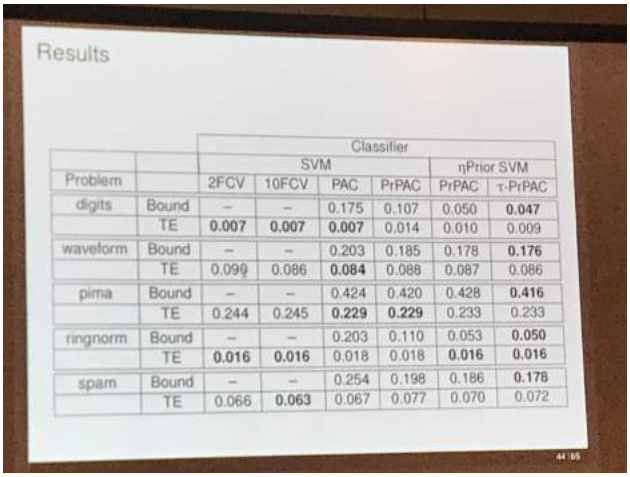


图2：应用不同的PAC-Bayes先验选择方法进行实验的结果。

上述方法收紧边界的结果：见图2。

结果要点：

1. 边界非常紧！
2. 通过这些新边界进行模型选择与10折交叉验证一样好。
3. 最佳边界不一定意味着最佳模型选择。

   →我们没有完全捕捉到正确的东西（但肯定捕捉到了一些正确的东西）。

接下来是：分布定义的先验：

- 考虑P和Q是Gibbs-Boltzmann分布：

$$P_\gamma(h) = \frac{1}{Z}\exp(-\gamma R_{out}(h) \qquad Q_\gamma(h) = \frac{1}{Z}\exp(-\gamma R_{in}(h)).$$

- 由于我们无法将其应用于单个权重向量，因此这些分布很难处理。

  根据Catoni [13]的研究，我们可以得出：

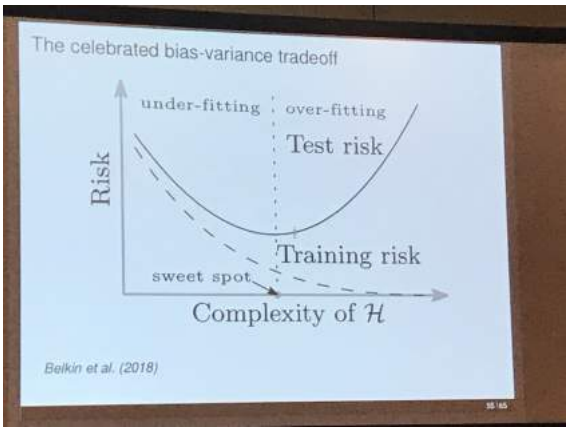
$$D_{\mathrm{KL}}\left(R_{in}(Q_\lambda) \mid\mid R_{out}(Q_\lambda)\right) \lesssim \frac{1}{m}\left(\gamma/\sqrt{m} + \gamma^2/4m\right),$$

  其中$\lesssim$忽略了右侧的对数项（Dave: 这是我（糟糕的）缩写方式，不是他们的）。关于稳定性：

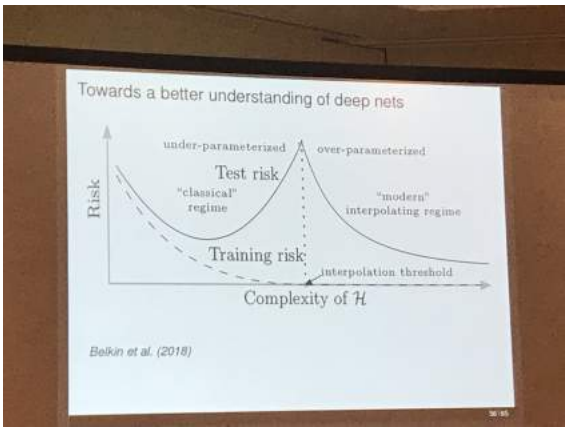
- 如果样本大小为 m时，算法 A的敏感性为 β，则可以限制泛化误差[11]。
- 问：算法输出高度集中，是否意味着结果更强大?

  →是的！我们可以推导出（更紧密的）界限，这取决于分布敏感先验和精心选择的后验之间的KL散度。

  开放领域！有很大的空间来探索一种新的泛化误差分析方法。



(a) 过拟合的经典观点



(b) 新范式?

图3：过拟合的经典观点（左），以及深度神经网络可能避免过拟合的新提议（右），来自Belkin等人[9]。

最后的案例研究：我们能否利用这些来分析深度神经网络?

问：深度学习是否打破了我们所知的统计范式?

→在大规模数据集上训练的神经网络实现了零训练误差，这对它们的性能来说并不乐观。 然而! 它们在测试集上的误差也往往非常低。

思路：也许我们可以使用PAC-Bayes来解释这种现象。

Dziugaite和Roy [22]以这种方式推导出了极其紧密的深度学习泛化误差界限；

- 基于扩大“吸引盆地”的训练
- 因此，不能衡量正常训练的良好泛化

问：训练集中包含多少信息?

A1：Achille和Soatto [1]研究了深度神经网络权重中存储的信息量。 过拟合可能与存储在权重中的编码训练集的信息有关，而不是数据生成分布。

A2: 信息瓶颈准则 [] 可能控制这些信息，并可能导致更紧密的 PAC-Bayes 界限。

结论：

- PAC-Bayes 起源于两个领域：1) 统计学习理论，和 2) 贝叶斯学习。
- 广泛推广了这两个领域，并指出了有希望的方向。

- PAC-Bayes 理论可以激发新的理论分析灵感，也可以推动算法设计（特别是当理论证明困难时）。

.....

2.2 教程：元学习

演讲者是 Chelsea Finn 和 Sergey Levine。

动机:从少量数据中学习。

→最近的进展在大规模多样化的数据集中蓬勃发展，因为它允许广泛的泛化）（参见：BERT，AlexNet）

→现有方法需要大量数据集。但是，有一些问题：

1. 如果我们没有大型数据集怎么办?
2. 如果我们想要一个通用的人工智能系统在世界上怎么办?
3. 如果我们的数据呈现长尾分布怎么办?

要点：这些设置开始违反了标准的监督学习设置。

例子：用绘画进行少样本学习，观众能够“推测”出新绘画的画家。

问题：我们如何实现这一点?

回答：嗯，之前的经验！我们并不是没有先前的经验就做到这一点的。 我们之前遇到过类似的问题/任务/图像。

问题：我们如何让机器完成这个任务?

回答：嗯，我们可以通过以下方式对结构进行编码： {建模图像形成、几何、任务特定特征、超参数选择}等等

> 要点：我们能否从先前的经验中明确学习到导致有效的下游学习的先验知识?

大纲：

- 问题陈述
- 元学习算法
- 元学习应用
- 元强化学习

2.2.1 两种视角看待元学习

问题：我们如何形式化元学习问题?

A1：机械观点！模型读取整个数据集并对新数据点进行预测。 训练这个网络使用一个“元”数据集，它本身由许多数据集组成。

A2：概率观点：从一组（元训练）任务中提取先验信息，以便有效学习任务。 学习一个新任务使用这个先验训练集来推断最可能的后验参数。

→A1对于实现更方便，A2对于理解更方便。

定义**2**（监督学习）：给定数据 D，找到参数 ϕ：

$$\arg\max_{\phi} \Pr(\phi \mid D),$$

应用贝叶斯规则，得到：

$$\arg\max_{\phi} \log \Pr(D \mid \phi) + \Pr(\phi).$$

定义**3**（元学习）：找到参数 θ，使我们能够快速解决新的任务，给定一堆数据集 D元训练：

$$\arg\max_{\phi} \log \Pr(\phi \mid D, D_{\text{元训练}}) = \arg\max_{\theta} \log \int_{\Theta} \Pr(\phi \mid D, \theta) \Pr(\theta \mid D_{\text{元训练}}).$$

也就是说，我们假设 θ是 ϕ的充分统计量。

通常很难在一般情况下进行这种分解：通常采用最大后验概率（MAP）方法。

元学习的目标：找到一个适当的参数集 θ，给定元数据集，使其最大可能性最大化。所以：

$$\theta^* = \arg\max_{\theta} \log \Pr(\theta \mid D_{\text{元训练}}).$$

值得注意的是， D元训练可能包含不同的任务（具有相关结构）。

示例：想要对新的数据集进行分类。首先，进行元学习；

$$\theta^* = \arg\max_{\theta} \log \Pr(\theta \mid D_{\text{元训练}}).$$

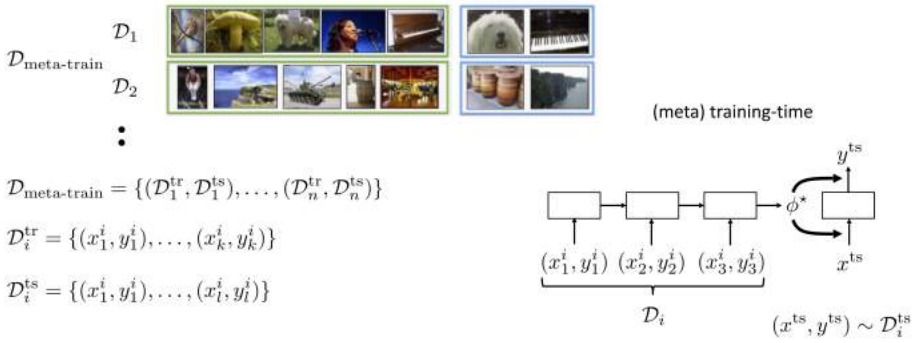


图4：元学习范式

然后，调整并找到一些附近的参数， ϕ^*：

$$\phi^* = \arg\max_{\phi} \log \Pr(\phi \mid D, \theta^*).$$

关键思想：我们的训练过程基于一个简单的机器学习原则：测试和训练条件必须匹配”（参见Vinyals等人的工作[88]）。

因此，为了使我们上述的元学习概念有意义，我们还需要保留每个训练任务的测试条件。

因此，元学习可以写成：

$$\theta^* = \max_{\theta} \sum_{i=1}^{n} \log \Pr(\theta_i \mid D_i^{\text{test}}.$$

密切相关的问题设置：

1. 多任务学习：元学习的特殊情况，其中 $\theta = \phi$.
2. 超参数优化：也可以看作是具有 θ=超参数和 ϕ=网络权重的元学习。

2.2.2 元学习算法

在讨论算法之前，让我们讨论一下评估。

主要思想：参考Lake等人的研究[48]，介绍了Omniglot数据集。 思路：大量的类别，每个类别只有少量的示例。 包含来自50个不同字母表的1623个字符，每个字符有20个实例。

→提出了few-shot判别和few-shot生成问题。 最初的few-shot学习方法采用贝叶斯模型和非参数方法。 其他类似的数据集：CIFAR，CUB，Mini-ImageNEt。

问题：如何评估元学习算法?

答案：进行常规的元训练，并确定生成的模型是否能够在不同（保留）任务上进行快速泛化。

**元学习算法的一般步骤：

1. 选择$\Pr(\phi_i \mid D_i^{\text{train}}, \theta)$的形式。
2. 选择如何使用 $D_{\text{meta-train}}$来优化 θ的最大似然目标。方法1：黑盒适应。 关键思想是

$$\text{训练一个神经网络来表示 } \Pr(\phi_i \mid D_i^{\text{train}}), \theta)。$$

$\rightarrow$例如，可以使用RNN来表示 f_θ，给定一堆这些元训练数据集。
然后，我们可以使用标准的监督学习进行训练：

$$\max_\theta \sum_{T_i} L(f_\theta(D_i^{\text{train}}, D_i^{\text{test}}))。$$

挑战：输出所有神经网络参数似乎不可扩展 $\rightarrow$但是，我们不需要输出神经网络的所有参数，只需要输出足够的统计信息。

问题：我们如何将其构建为优化过程?

方法2：通过优化获取 ϕ_i：

$$\max_{\phi_i} \log \Pr(D_i^{\text{train}} \mid \phi_i) + \log \Pr(\phi_i \mid \theta)。$$

元参数 θ作为先验。什么形式的先验? 一个成功的形式：从其他任务中学习
θ。

目标：学习一个参数向量 θ，能够有效地传输，即使在新任务上进行微调也容易/有用。 为此，解决以下问题；

$$\min_\theta \sum_i L(\theta - \alpha \nabla_\theta L(\theta, D_i^{\text{训练}}), D_i^{\text{测试}}.$$

通用算法：

1. 采样任务 T_i.
2. 从 D_i中采样不相交的数据集
3. 优化 $\phi_i \leftarrow \theta - \alpha \nabla_\theta L(\theta, D_i$训练)
4. 使用 $\nabla\ \theta L(\theta, D_i$训练)更新$_\theta$).

因此，我们剩下两种方法：优化与黑盒适应

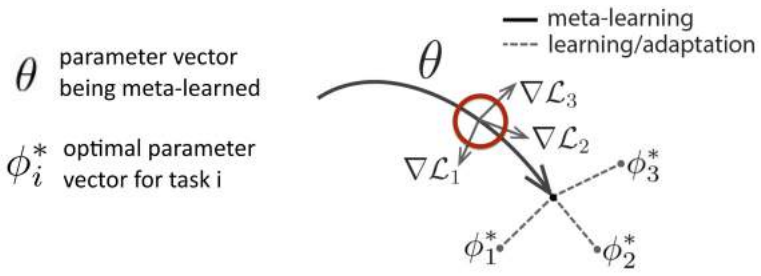


图5：模型无关元学习

→模型无关元学习（MAML），在优化中，可以看作是一个带有嵌入梯度运算符的计算图（见图5：

$$y^{ts} = f_{\text{MAML}}(D_i^{\text{训练}}, x^{\text{测试}} \tag{8}$$
$$= f_{\phi_i}(x^{ts}), \tag{9}$$

其中 $\phi_i = \theta - \alpha\nabla_\theta L(\theta, D_i^{\text{train}})$.

实验：比较MAML与黑盒方法（SMAIL，MetaNetworks）。倾向于在领域适应/外推任务中获得更高的性能。
→任务是对Omnioglot数据集中的数字进行一定程度的剪切，并检查性能的衰减。 MAML的性能很少衰减。

问：这种学习结构是否有代价?

答：实际上没有！MAML函数可以近似任何函数$D_{_i_}\text{tra}_i\text{n}$，$x_\text{te}^\text{s}\text{t}$的函数。 假设：非零$\alpha$，损失函数不会丢失关于标签的信息，数据点在$D_{_i_}\text{tra}_i\text{n}$中。 是独特的。

∴MAML具有归纳偏差的好处，而不失表达能力。

另一种方法：基于概率的优化推理解释。 发现：MAML是一种隐式先验，大致是梯度下降+提前停止，得到一个隐式高斯先验。

Q: 可以使用其他形式的先验知识来表达元学习吗?

A: 当然可以！可以在学习到的特征上进行贝叶斯线性回归，或者进行闭式/凸优化（如岭回归或逻辑回归）。

挑战:

- 我们如何选择对内部梯度步骤有效的架构？ →但是，可以通过渐进神经搜索（使用MAML）来克服这个问题。
- 二阶元优化可能会出现不稳定性。 →有很多解决方案可用：1）仅优化内部子集，2）解耦学习率，3）引入上下文变量。

方法3：非参数方法。 在低数据情况下，非参数方法简单且往往效果良好。

→在元测试时：少样本学习 ≡低数据情况。 在元训练期间，仍然希望是参数化的。

Q: 我们可以使用产生有效的非参数学习器的参数化元学习器吗?

A: 是的！通过将测试数据与训练图像进行比较，使用非参数学习器。

关键思想：学习一个度量空间，以便在测试时进行更有效的比较和预测。

要点：每种方法都有一些优点/缺点，详见图??。

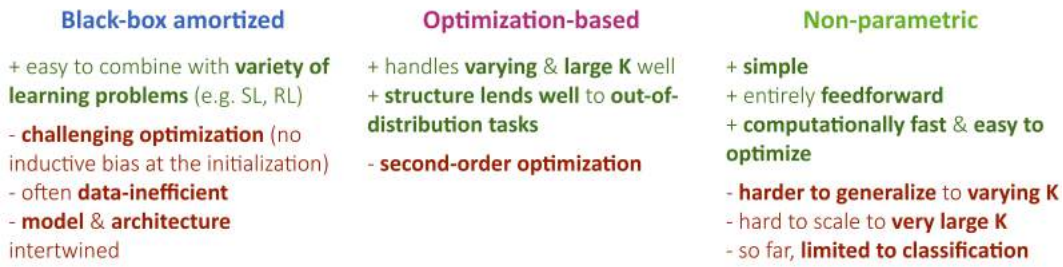


图6：不同元学习方法的优点和缺点。

方法4：贝叶斯元学习。

假设我们有参数先验分布$\Pr(\theta), \Pr(\phi_i)$，我们能够采样$\phi_i \sim \Pr(\phi_i | x_i_\text{tra}_i\text{n}, y_i_\text{tra}_i\text{n})$吗?

简单的想法：使用神经网络对h产生一个高斯分布，其中h是网络的一些相关权重（如最后一层）。

Q: 好的，但贝叶斯优化的元学习呢?

A: 当然！有很多方法可以实现这一点。 一种想法是将$\Pr(\phi_i|\theta)$建模为高斯分布，在训练时进行变分推断（参见Rav_i和Beatson [70]）。 另一种方法：仅对最后一层进行基于梯度的推断，使用SVGD来避免高斯建模假设（参见L_iu和Wang的工作[54]）。

→ 关键思想; 近似计算 $\Pr(\quad \phi_i \mid \theta, x_i^{\text{train}}, y_i^{\text{train}} \quad)$ 通过MAP推理。非常粗糙，但非常方便!

进一步阅读; Garnelo等人[28]，Kim等人[46]，Ravi和Beatson[70]。
应用：

- 视觉: 少样本图像生成，图像到图像的转换，生成新视角。

•模仿学习/强化学习: 单次逆向强化学习，基于演示的优化推理。

•语言: 适应新程序，适应新语言，适应对话代理的新角色。

2.2.3 元强化学习

问: 为什么期望元强化学习有用?

答: 嗯，强化学习的主要挑战！几乎都与现有方法的样本低效有关。
对于一个真实机器人来说，应用TRPO算法可能需要几天或几周的时间才能开始取得任何进展（学会行走）。首先，一些背景知识；

> 定义**4**（马尔可夫决策过程（M**D**P））：M**D**P是一个四元组：<$\mathcal{S}$, $\mathcal{A}$, R, P>，其中$\mathcal{S}$是状态集合，$\mathcal{A}$是动作集合，P：$\mathcal{S} \times \mathcal{A} \to Pr(\mathcal{S})$表示转移函数，$R$：$\mathcal{S} \times \mathcal{A} \times \mathcal{S} \to R$表示奖励函数。

目标是让代理学习一个策略π：$\mathcal{S} \to \mathcal{A}$，以最大化长期预期奖励：

$$\theta^* = \arg\max_\theta \mathbb{E}_{\pi_\theta}\left[R(\tau \mid \pi_\theta)\right], \tag{10}$$

其中τ是π_θ采取的轨迹，$R(\tau|\pi_\theta)$选择π_θ获得的奖励。

“每个强化学习算法简介”：通过以下方式之一找到π_θ：1）直接学习一个好的策略，2）学习一个值函数，或者3）学习一个模型并使用它找到一个好的策略。

迄今为止的元学习：学习 θuch以使 $\phi_i = f_\theta(D_i^{\text{train}})$ 对于测试 D_i^{test}良好。

因此，元强化学习问题如下；

> 定义**5** (元强化学习)：具有强化学习目标的元学习问题。也就是说，学习 θ^*:
>
> $$\theta^* = \arg\max_\theta \sum_{i=1}^{n} \mathbb{E}_{\pi_{\phi_i}}\left[R(\tau \mid \pi_{\phi_i})\right],$$
>
> 其中 $\phi_i = f_\theta(M_i)$

问：那么我们如何生成元训练MDP M_i?
答：一个想法是只选择相关任务（想象我们想要一个家务机器人，定义 M_i为小而简单的相关任务

现在，一些元强化学习算法：类似于监督元学习，我们还可以基于黑盒方法获得基于元强化学习的算法（这里称为“循环策略”）.

主要问题：我们如何实现 $f_\theta(M_i)$，其中 M_i是一个MDP？ 好的， $f_\theta(M_i)$应该做什么？

A: 有几件事情：1）通过 M_i的经验改进策略，2）选择如何交互（也就是说，元强化学习必须选择如何探索）。

$\rightarrow$这基本上只是运行带有RNN的元强化学习。

Q: 但是元强化学习算法如何有效地进行探索？ 在这种情况下，这基本上是免费的：在训练中优化所有使用的剧集的奖励会导致良好的探索。 Dave: 我个人不认同这一点。需要多读一些！

下一个观点：将元强化学习视为一个优化问题。 标准强化学习将其表述为一个策略梯度问题；

$$\theta^* = \arg\max_\theta \mathbb{E}[R(\tau)]$$

然后：

$$\theta^{k+1} \leftarrow \theta_k + \alpha \nabla_{\theta_k} J(\theta)$$

下一个方法：部分可观察强化学习的元学习！需要一个更丰富的环境模型：

> 定义**6**（部分可观察马尔可夫决策过程（POM**D**P））：POM**D**P是一个六元组：$\langle S, A, O, \mathrm{P}, E, \mathrm{r}\rangle$，实际上是一个M**D**P，但函数$E$根据当前状态生成来自$O$的观测。 代理只能感知到观测$o \in O$，而不是状态。

也就是说，代理不能看到关于世界的所有相关信息，只能看到基于状态生成的一些观测o。

关键思想：解决POMDP问题非常困难！但是，它类似于元学习。 寻找一个策略$\pi_\theta(a|s,z)$，其中我们不知道s，只知道z。 工作原理如下：1）从$\hat{p}(z_t|s1{:}_t, a1{:}_t, \mathrm{r}1{:}_t)$中采样$z$，2）根据$\pi_\theta(a|s,z)$行动，假设$z$是正确的。

关于元强化学习的三个观点：

1. 只需使用RNN：概念简单易懂，但容易过拟合且难以优化。
2. MAML方法：良好的外推能力，但复杂（需要大量样本）。
3. POMDP方法：简单有效，但也容易过拟合。

但是它们并没有那么不同！POMDP方法实际上只是在RNN方法中增加了一个额外的隐藏变量，而MAML方法只是对 f_θ的设计的特定选择。

2.2.4 元学习中的挑战和前沿

有许多令人兴奋的前进方向，但也面临许多挑战。

挑战1：元过拟合-元学习需要一个任务分布。 一些元学习方法可能会对这些任务分布过拟合。

挑战2：任务设计-通常这些任务分布必须手动选择，或者不足以鼓励正确的行为。 很难以正确的方式选择任务分布！

挑战3：了解哪些算法过拟合——有很多不同的方法（黑盒、基于优化、非参数化），但我们不知道哪种算法最容易受到元过拟合的影响。

问：我们还能用元学习做什么？

答：嗯，我们可以自动提出生成适当任务分布的任务。
$\Longrightarrow$ “无监督元学习”，指的是在元训练期间不使用手动指定的任务分布或标签来高效解决任务的算法。

挑战4：记忆！一些算法可能只会记住相关测试任务的解决方案。

挑战5：我们如何确定哪些任务信息应该在输入中，哪些应该在数据中？ 广泛的任务分布可能很有用，但会使探索变得困难。 如果它们太狭窄，就不够代表性。

戴夫：今天必须离开一会儿！

.....

3 6月11日星期二：主会议

我今天到达了，准备参加最佳论文奖演讲。

3.1 最佳论文演讲：挑战学习中的假设性解缠表示

弗朗西斯科·奥利维尔·巴赫姆的演讲。

核心问题：我们能否以纯无监督的方式学习解缠表示？

背景：表示学习。 考虑一幅风景图片。 表示学习的目标是学习一个函数 $f: \mathcal{X} \to \Phi$，将每个感兴趣的项目（例如图像）转化为一组捕捉图像重要特征的特征。

问题：在表示中我们可能想要什么？

答案：很多东西！但有一个想法：解缠表示。

定义7（解缠表示）：因素的单一变化应该导致表示的单一变化。

例子：考虑放置在场景中的几何形状，每个形状具有不同的颜色/形状/大小。 如果我们改变原始图像的一个关键方面（例如正方形的颜色或大小），表示也应该以完全一种方式改变。

重点：无监督学习来解开表示（因此，关键是不能观察到真实的变化因素）。

贡献：

1. 理论结果：对于任意数据，无监督学习解开表示是不可能的。

2.大规模实验研究：我们能否在不查看标签的情况下学习解开表示？

   →可以调和并且仍然有效！定理是关于最坏情况的数据。 我们的数据可能具有正确的结构，我们仍然可以学习这些解开表示。

实验：六种方法，七个数据集，六个度量标准，产生了$10{,}000$个训练模型和$150{,}000$个分数。

→需要关注的问题：

Q1 应该使用哪种方法?

Q2 如何选择超参数?

Q3 如何从一组训练模型中选择最佳模型?

实验结果：

1. (回答问题1的方向): 随机种子和超参数似乎比目标函数的选择更重要。

2. (回答问题2的方向): 无法确定可以用作经验法则的明显趋势（但是，在任务之间进行转移似乎有所帮助）。 →超参数在不同数据集和度量之间的转移似乎并没有显著影响。

3. (回答问题3的方向): 无监督模型选择仍然是一个关键挑战!

→含义: 从糟糕的超参数设置中获得的良好运行结果很容易超过从良好超参数设置中获得的糟糕运行结果。

主要要点：

- 归纳偏见和监督的作用应该明确说明
- 应该展示出解缠结的具体实际好处
- 具有多个数据集的可靠、可重复的实验设置至关重要。
- 代码:https://github.com/google-research/disentanglement_lib/。 还发布了10,000个预训练模型。

.....

3.2 投稿演讲：深度强化学习

接下来我们有5分钟的贡献演讲。 它们非常短，所以我猜我会更难转录。

3.2.1 DQN和时间离散化 [82]

演讲者是Corentin Tallec。

要点：在深度强化学习中，时间离散化确实很重要!

→常规的强化学习算法在高帧率下会导致失败。 可扩展性通常受算法限制；更好的硬件、传感器、执行器可以导致新的超参数。

问：为什么近似连续的Q-learning会失败?

答：随着 δ_t（时间离散化），$Q^\pi(s,a) \to V^\pi(s)$。因此，在极限情况下，Q不再依赖于动作！所以我们无法进行策略改进。

问：我们能解决这个问题吗？

答：是的！可以查看所有这些数量，形成新的算法（更多信息，请参阅论文！）。

.....

3.2.2 非线性分布梯度 TD 学习 [67]

演讲者是Chao Qu。

背景：分布式强化学习考虑了长期回报的随机性质，$Z(s,a)$。

这项工作：考虑梯度**T**D学习的分布对应。性质：

- 在非策略下收敛
- 在非线性函数逼近下收敛。

→新算法：分布式GTD2。使用时间分布差异代替GTD2的时间差异。

定理4.在温和的假设下，分布式*GTD2*（*D-GTD2*）在极限中收敛。

实验：证明D-GTD2确实收敛。

.....

3.2.3 使用发散校正组合熵策略 [38]

Jonathan Hunt的演讲。

问：人们如何同时解决像杂耍和单轮车这样的复杂运动控制任务？

答：嗯，也许我们倾向于独立学习每个任务，然后将它们合并在一起。

注意：不是像选项一样！我们不是先做A再做B，而是一个新的 $A \circ B$任务。

问题：给定训练任务 T_1和 T_2，我们想解决一些合并的任务 $T_b = T_1 + T_2$。

先前的工作：广义策略改进[8]和组合乐观主义[？]。
将现有的想法应用于这种转移设置：

1. 后继特征
2. 广义策略改进

3. 发散校正

→在连续动作空间中组合任务的新算法。

.....

3.2.4 TibGM: 一种用于强化学习的图模型方法 [**2**]

演讲者是Tameem Adel。

起源：用于增加可转移性泛化和采用精确目标和清晰解释的图形建模方法。

→通过提出一个信息论目标，旨在最大化"局部"奖励和更全局的度量（与转移相关）来实现这一目标。

贡献：

- 基于图形模型
- 证明了新目标与典型奖励最大化目标之间的对应关系。
- 基于信息论的预训练过程，重点是探索。
- 在16个基准任务上取得了最先进的结果。

.....

3.2.5 多智能体对抗逆强化学习 [**93**]

演讲由Lantao Yu进行。

动机：强化学习代理的性能依赖于奖励函数的质量。

→但是，设计正确的奖励函数可能很困难！

一个解决方案：从专家示范中学习奖励函数（如模仿学习）。

问：但是为什么我们要关心奖励学习？

答：嗯，一些优势：1）科学探究（理解动物行为），2）奖励函数是任务的最简洁、最稳健和最可转移的描述，3）如果我们想在新环境中重新优化策略，它可以帮助我们。

→在多智能体环境中，这些特性更加可取！

定义**8**（单智能体逆强化学习）：找到一个解释专家行为的奖励函数

但是：难以处理！太多的奖励函数可以解释相同的行为。

将这个设置推广到多智能体使用马尔可夫博弈：

定义**9**（马尔可夫博弈）：M**D***P*的多智能体推广*[52]*

马尔可夫博弈的解概念是纳什均衡：

定义**10**（纳什均衡）：当没有代理能够通过改变自己的策略来获得更高的预期回报时。

方法：引入了逻辑随机最佳响应均衡（LSBRE）。 优化伪-
似然目标。

实验：策略模仿性能。 新方法在合作/沟通和竞争任务中达到了最先进水平。

总结：马尔可夫博弈的新解决方案概念，引发了新的兴趣度量。 提出了第一个多智能体最大熵逆强化学习框架。

.....

3.2.6 连续强化学习的策略整合 [**44**]

动机：神经网络中的灾难性遗忘！网络倾向于忘记先前任务的信息。

→即使在连续学习过程中，RL中也会发生，因为探索/策略的改变导致已观察状态的分布随时间变化。

代理应该应对：数据分布的离散和连续变化！

贡献：策略整合代理。 一群代理通过PPO进行训练，并通过KL教师强化损失进行连接。

→最终存储的策略是对所有其他训练的策略进行提炼的结果。 确保它不会偏离先前的性能太多。

实验：交替任务以探索遗忘的影响。 发现他们的算法优于所有其他PPO的变体。

未来工作：1）研究如何在训练过程中优先考虑重要的记忆，2）适应离线学习。

3.2.7 无探索的深度强化学习离线策略评估 [**26**]

考虑：使用相同的离线策略算法（DDPG）在相同的数据集上采用两种不同的方法（橙色和蓝色）。

代理：1）橙色-与环境交互（标准强化学习循环），2）蓝色-仅使用环境数据但不实际交互！然而，它们的性能是不同的。

问：为什么它们会有所不同?

答：外推误差：$Q(s,a,) \leftarrow r = \gamma Q(s',a')$，其中（$s,a,r,s'$）来自数据集。 基本上，你可能会有一个糟糕的目标$Q(s',a')$。

> 定义**11**（外推误差）：试图在没有足够访问（s，a）对的情况下评估 π。

解决方案：批量约束强化学习：只选择 π，以便选择数据集中的（s，a）对并最大化性能。

→新算法：批量约束深度Q学习（BCQ）。

1. 通过生成模型模仿数据集
2. $\pi(s) = \arg\max_a Q(s,\, a)$
3. 加入一些额外的魔法

发现BCQ在性能上大大优于一些现有方法（DDPG），而且非常稳定。

.....

3.2.8 随机专家蒸馏 [**90**]

王若涵的演讲

> 定义**12**（模仿学习）：从有限的专家演示中进行策略学习

有用的原因：直观且高效的技能转移，可以捕捉个体演示者的风格/偏好。

→最近的IRL框架：生成对抗模仿学习（GAIL）。

但是，优化挑战有：1）训练不稳定性，和2）样本效率低。

主要贡献：随机专家蒸馏（RED）。 使用专家策略的估计支持作为奖励的模仿学习框架。

基于新的奖励函数优化新的轨迹损失，并证明在极限情况下它近似正确。

实验**1**：在MuJoCo中，相比于GAIL等其他方法，找到高训练稳定性和良好的样本效率。

实验**2**：使用人类行为作为训练数据的自动驾驶任务。 代理学习跟随训练者的偏好（速度、车道偏好等）。

.....

3.2.9 重新审视 Softmax Bellman 操作符 [79]

演讲者是赵松。

回想一下，贝尔曼算子是一个收缩映射。

→Mellowmax算子[7]也是一个收缩映射：

$$max_{a'}Q(s',\, a') \rightarrow \sum_{a'} \frac{\exp \tau\, Q}{\sum_b \exp \tau\, Q(s',\, b)}$$

问题：softmax真的像酸牛奶一样糟糕吗？（感谢Ron Parr）。

思路：将DDQN的目标网络中的max函数与softmax相结合。

→导致SQDN的出现，在大多数Atari游戏中获得更高的分数比DQN。

因此：有必要对softmax进行分析！

定理**5.**分析表明：良好的性能，收敛保证，较小的误差，降低边界，以及过高估计误差与τ（逆温度）单调增加。

.....

3.3 贡献演讲：强化学习理论

更多强化学习！

3.3.1 用于高效探索的分布式强化学习 [57]

由Hengshuai Yao演讲。

要点：强化学习中存在许多不确定性来源-估计（在有限数据情况下），但也有其他因素，如环境随机性，游戏中对手的玩法等。

→基于不确定性的探索策略：不确定性下的乐观主义。 在多臂赌博机中，根据以下选择臂：

$$a = \arg\max_k \hat{\mu}_k + c_k,$$

其中$\hat{\mu}_k$是估计的平均回报，而 c_k是对该估计的不确定性的某种度量（类似于UCB）.

但是！使用天真的探索奖励并不总是有效：永远偏爱具有高内在不确定性的动作。

新的想法：在面对不确定性时使用乐观主义，在不确定性的权重上执行衰减计划。 这确保随着收集到更多证据，代理开始更多地进行利用。

.....

3.3.2 通过重要性采样的乐观策略优化 [62]

Matteo Papini的演讲。

问题：策略优化。

- 参数空间Θ $\subseteq \mathbb{R}^d$
- 每个 $\theta \in \Theta$对应一个参数化策略
- 每个策略引发一种分布。 关于轨迹的 $R(\tau)$回报。
- 目标是找到最大化预期回报的参数：

$$\text{最大化}_{\theta \in \Theta} \mathbb{E}[R(p_\tau \mid \theta)].$$

然而，挑战在于探索是困难的！

问：如果我们将其视为多臂赌博机怎么办?
那么：

- 臂：参数 θ
- 回报：预期回报 $J(\theta)$
- 实际上是连续多臂赌博机。
- 通过轨迹分布利用臂之间的相关性。

•使用重要抽样更新候选策略的回报。

主要结果：新算法“Optimist”具有次线性遗憾：

$$\text{遗憾} \quad (T) - \tilde{O}(\sqrt{dT})。$$

代码：https://github.com/wolfLo/optimist

.....

3.3.3 神经逻辑强化学习 [41]

演讲由Shan Luo主持。

深度强化学习的两个主要挑战：1）我们如何将从一个任务中学到的策略推广到另一个任务？2）我们如何解释这些学到的策略？

方法：使用背景知识来学习概念和关系，例如grandfather(x,y)。
为此，使用可微分的归纳逻辑编程。
→思路：使用策略梯度学习逻辑规则（一种新的架构，DILP，使用REINFORCE进行训练）。

实验：对于许多设置，与MLP代理（基准）相比，获得高奖励。

.....

3.3.4 在 MDP 中学习协作 [68]

演讲由Goran Radanovic进行。

动机：人工智能与人类的合作-考虑一个辅助AI帮助人们解决某个任务。

→这些代理可能有共同的目标，但在某种程度上以不同的方式看待任务。

形式化模型：双代理MDP，代理有承诺。

→目标是设计一个学习算法（对于第一个非人类AI），以实现次线性遗憾。

挑战：从A1的角度来看，由于A2（人）的存在，世界看起来像是一个非平稳的MDP。

主要贡献：基于最新性偏见的“双专家”算法，针对这种情况进行了改进。

定理6.（主要结果）该算法的遗憾值以 O的速度衰减 $\left(T^{\frac{1}{4}}\right)$.

.....

3.3.5 预测-校正策略优化 [15]

由Ching-An Cheng发表的演讲。

问题：情节式策略优化。 因此，代理人试图优化某个策略 $\pi(a \mid s)$，以实现高回报。

→一个目标：样本效率。 在任何真实交互之前，我们应该花时间进行规划/思考。

问：为什么我们应该使用模型?

答：可以总结过去的经验，可以更加样本高效，并且可以在没有真实世界交互的情况下高效优化策略。

问：为什么不使用模型?

答：嗯，模型总是不精确的！模型的弱点可以被利用来优化策略。

问：我们能否调和这两个阵营?

A: 当然！这篇论文：PicColo（见演讲标题）。 基于"不应完全信任模型，而只利用正确的部分"的思想，提出了一种元算法

→如何实现这一目标?

- •将策略优化框架化为可预测的在线学习（POLL）
- •设计了一种基于规约的算法来重用已知算法。
- •将其翻译回来后，得到了一种策略优化的元算法。

在线学习：考虑一个学习者和一个对手，学习者每隔一段时间选择一个决策 $\pi_n \in \Pi$ every。 对手选择一个损失函数来最小化性能，反复进行。

→常见的性能度量是遗憾。思路：将策略优化定义为在线学习过程。

→算法上：可以尝试典型的无遗憾算法（镜像下降），但不是最优的！希望学得更快。 因此，将可预测性视为预测未来梯度的能力。 引入以下模型：

定义**13**(预测模型):一个估计未来损失梯度的函数

$$\Phi_n(\pi) \approx \nabla \ell_n(\pi)$$

现在，基于这个可预测性模型开发一个算法。

→希望从可预测性转向在线学习。这就是PicColo！

假设我们有一个可预测的学习问题，思路是将其转化为对抗性问题。所以：

$$\ell_n(\cdot) = \hat{\ell}_n(\cdot) + \Delta_n(\cdot),$$

这是一个预测损失（第一项）和误差（第二项）的组合。 戴夫：我没听清楚第二个误差来源是什么
PicColO：两个步骤—1) 预测步骤 ($\pi_n = \hat{\pi}_n - \eta_n \hat{g}_n$，和2) 修正步骤—ˆ $\pi_{n+1} = \eta_n(g_n - \hat{g}_n,$
其中 g是回报。

实验：在MuJoCo任务上将Pi**C**olo与其他各种算法进行比较。

总结：

- PicColO可以使无模型算法更快，但没有偏差。
- 预测模型可以被视为注入先验的统一接口
- 由于PicCoLO是为一般可预测的在线学习而设计的，我们期望将其应用于其他问题和领域。

.....

3.3.6 通过元逆强化学习学习意图先验 [**91**]

Kelvin Xu的演讲。

动机：我们经常假设我们有一个明确定义的奖励函数！

问：一个代理如何从一个或少数几个演示中推断奖励?

→这项工作！使用以前任务的演示来引导可以在新任务中使用的任务先验。

主要思想：利用先前任务的信息加速逆向强化学习。

→建立在MAML的基础上（参见元学习教程！）。

实验：S**prite**世界环境和第一人称导航任务。在这两种情况下，从元训练任务中学习任务先验，并在测试任务中快速学习。

→结果：即使只给出少量示范，表现也非常好。

.....

3.3.7 DeepMDP: 学习RL的后期空间模型 [30]

Carles Gelada的演讲。

目标：为强化学习找到简单的表示。

方法：学习一个潜在空间模型 $M = \langle \mathcal{S}, \mathcal{A}, R, T \rangle$。基于两个损失函数；

$$\bar{R}(s,a) \approx R(s,a), \forall s,a \tag{11}$$

$$\bar{T}(s' \mid s,a) \approx T(s' \mid s,a), \forall s,a,s' \tag{12}$$

利用这些损失函数，确保 ϕ 是一个好的表示，即只丢弃坏/无用的策略。

实验1：甜甜圈世界。一个圆的图像。 嵌入函数最终找到了一个类似的表示。

实验2：Atari与C51，找到超越基准的改进。

.....

3.3.8 重要性采样策略评估 [35]

Josiah Hanna的演讲。

注意：最近有很多实证强化学习的成功案例!

但是：要使强化学习成功，我们必须问："强化学习代理如何从少量经验中获得最大收益？"

核心贡献：研究强化学习策略评估子问题中的重要采样。
更具体地说：用估计值替换重要采样的分母，并通过实证和理论证明这是合理的。

强化学习中的典型重要采样：

$$OIS(\pi, D) = \frac{1}{m} \sum_{i=1}^{n} \Pi_{t=0}^{L} \frac{\pi(a_t \mid s_t)}{\pi_D(a_t \mid s_t} \sum_{t=0}^{L} \gamma^t R_t$$

他们用行为策略性能的MLE估计值替换了分母：

$$新的 - IS(\pi, D) = \frac{1}{m} \sum_{i=1}^{n} \Pi_{t=0}^{L} \frac{\pi(a_t \mid s_t)}{\hat{\pi}_D(a_t \mid s_t} \sum_{t=0}^{L} \gamma^t R_t$$

论文提供了理论和实验证明这是一件好事。

.....

3.3.9 从学习者中学习 [40]

Alexis Jacq的演讲。

目标：通过观察他人学习来学习最佳行为。

假设：学习是优化一个正则化目标函数：

$$J(\pi) = \mathbb{E}_{\pi}\left[\sum_{t} \gamma^{t}(r(s_t,\, a_t) + \alpha H(\pi(\cdot \mid s_t))\right].$$

状态动作对的值由正则化贝尔曼方程的不动点给出。
此外，softmax是策略的改进（参见Haarnoja等人[33]）。

实验：在MujoCo上找到改进。

.....

3.3.10 在时间尺度上分离价值函数 [72]

Joshua Romoff的演讲。

多步回报：

$$G_t^k = \sum_{i=0}^{k-1} \gamma^i r_{t+1} + \gamma^k V(s_{t+k}).$$

可以选择 k在偏差-方差之间进行权衡。或者， λ返回值：

$$G_t^{\lambda} := (1-\lambda)\sum_{k=1}^{\infty} \lambda^{k-1} G_t^k.$$

许多任务都是折扣的：当 $\gamma \to 1$时，训练 V_{γ}很困难。

问：为什么使用折扣因子?

答：嗯，这使问题变得更简单。 只考虑未来几个奖励，所以代理可以更容易地学会做出决策。

$\to$但是，会增加很多偏差！有时我们并不真正关心早期奖励。

退一步：理想的强化学习代理应该怎么做？ 它会快速学会做得好！理想情况下：希望1）使

用一个大的折扣因子进行学习，但是2）建立在计算和样本效率高的方法之上。

解决方案：定义一系列 λ值：

$$\Delta; = \{\gamma_0, \gamma_1, \ldots, \gamma_Z\},$$

其中 $\gamma_i \leq \gamma_{i+1}, \forall_i$.

然后，使用这个 γ序列形成一系列贝尔曼方程：

$$W_0 : r_t + \gamma_0 W_0(s_{t+1}) \tag{13}$$
$$W_{i>0} : (\gamma_i - \gamma_{i-1}) V_{\gamma_{i-1}}(s_{t+1}) + \gamma_i W_i(s_{t+1}). \tag{14}$$

算法：TD(Δ)，等同于标准的TD(λ).

分析：在一些条件下，等同于标准的TD(λ).
→那么，为什么要做这个新的、有序的事情呢?
→嗯，使用线性函数逼近时是等价的。 并且：1）每个 W使用相同的学习率，2）每个 W使用相同的 k步长、 γ和 λ。
→（当然，不必以相同的方式设置这些参数）。

可以将TD(Δ)扩展到深度强化学习：1）共享网络多个输出，2）按照描述训练 W，3）使用 W的相同值而不是 V的值。

实验：在Atari上进行测试，将TD(Δ)与PPO和PPO+（PPO加上额外参数）进行对比。

→通过绘制值函数进一步探索$TD(\Delta)$的学习内容。

他们没有尝试的事情（但我们应该尝试！）：

- 添加更多 W。
- Q-learning扩展（他们尝试的都是on-policy方法）。
- 值函数的模型架构：可以尝试使用更少的参数等。
- 分布式变体。

总结：

1. 基于 γ分解价值函数
2. 使用贝尔曼方程进行训练
3. 提高样本效率

.....

3.3.11 在RL中学习动作表示 [14]

Yash Chandak的演讲。

问题：考虑具有数千个可能动作的强化学习问题！（应用可能是医疗治疗、广告或投资组合管理）。

→ 可能通过学习选项[81]来捕捉这一点，但通常会导致学习太多的选项。

问：我们如何学习良好的动作表示，而不是学习太多？

关键见解：

1. 动作不是独立的离散量
2. 它们的行为具有低维结构。
3. 可以独立于奖励学习这种结构。
4. 因此，代理可以在这种新的行为空间中行动，行为可以推广到类似的动作。

新算法：1）监督学习动作表示，2）使用策略梯度学习这些表示的内部策略。

实验：将算法应用于各种具有丰富行动空间的领域（例如Photoshop中的推荐），获得一致良好的性能。

.....

3.3.12 贝叶斯对抗风险最小化 [**55**]

本次演讲由本·伦敦（Ben London）进行。

问题：从已记录的数据中学习以推荐音乐。

挑战：

1. 反馈是强盗反馈，因此只能从选择的行动中学习。
2. 数据基于选择的记录策略存在偏差。

→ 一种抵消这种偏差的方法是使用倒数倾向性评分：

$$\arg\min_{\pi} \frac{1}{n} \sum_{i=1}^{n} -r_{\_i} \frac{\pi(a_{\_i}|x_{\_i})}{p_i}.$$

还可以使用方差正则化，从而产生反事实风险最小化（CRM）原则；

$$\arg\min_{\pi} \frac{1}{n} \sum_{i=1}^{n} -r_{\_i} \frac{\pi(a_{\_i}|x_{\_i})}{p_i} + \lambda \sqrt{Var(\pi,\, S)}$$

→ 受PAC分析启发，确保低泛化误差。

本研究：CRM的贝叶斯视角（受PAC-Bayes分析启发）。 得出了一个很好的泛化误差界限。

→ 应用于混合逻辑。

.....

3.3.13 每个决策选项计数 **[36]**

本次演讲由安娜·哈拉图尼安（Anna Haratunyan）进行。

动机：能够在长时间范围内进行推理的智能体。 但是：时间范围取决于γ的选择。

问：也许我们可以用选项[81]来做到这一点吗?

答：但是，它并不能真正处理不同的时间范围。

主要贡献：将选项框架推广到允许每个选项折扣。

→也就是说，添加每个决策选项的折扣(γ_r, γ_p)。

在这个框架中，他们确定了由这些新折扣捕捉到的偏差-方差权衡。

.....

3.3.14 RL中问题相关的遗憾界限 **[94]**

Andrea Zanette的演讲。

重点：在片段化的表格强化学习中进行探索。

→最先进的遗憾界限：

- •没有智能探索 $\tilde{O}(T)$
- •高效探索 $\tilde{O}(H\sqrt{\text{SAT}})$ $\overline{\text{· 下界：}\tilde{O}(\sqrt{\text{HSAT}})}$
- •HSAT) $\overline{SAT})$
- •下界以下：问题相关!

主要结果（需要价值的方差和奖励的缩放）

定理**7.**高概率下，可以实现遗憾为：

$$\tilde{O}(\sqrt{\mathbb{Q}^* SAT}),$$

其中 $\mathbb{Q}^*$ 是一个问题相关的项。

回答了Jiang和Agarwal 2018年（COLT）的一个开放猜想：任何算法在目标驱动的MDP中都必须遭受遗憾。

→还探讨了随机性对遗憾的影响。

.....

3.3.15 正则化MDP的理论 [29]

Matthieu Geist的演讲。

动机：许多深度强化学习算法使用正则化，但没有关于如何在强化学习中进行正则化的通用理论。

→这项工作：以两种方式推广了强化学习中的正则化：

1. 更大的正则化类别
2. 一般的修改策略迭代方案。

设 $\Omega : \Delta_A \to \mathbb{R}$为一个强凸函数。 凸共轭是一个平滑的最大值：

$$\forall_{q_s} \in \mathbb{R}^A, \Omega^*(q_s) = \max_{\pi_s \in \Delta_A} (\pi_s, q_s) - \Omega(\pi_s).$$

因此，我们可以对贝尔曼方程进行正则化：

$$T_{\pi,\Omega}[V] = T_\pi[V] - \Omega(V). \qquad (15)$$正则化贝尔曼算子具有与原始算子相同的性质：1) $T_{\pi,\Omega}$是仿射的，2) 单调性、分配性和 γ-收缩性。

→引入一种新的算法方案，正则化策略改进，并证明以下结果：

定理**8.**经过 k次**re***g-MPI*迭代，损失是有界的。

→还可以使用这个理论来描述RL中现有的正则化方法（如TRPO、DPP等）。

总结：

- 弥合了动态规划和优化之间的一些差距
- 引入了时间一致性方程，与熵一样
- 可以将现有的正则化方法推广到正则化策略梯度。

.....

3.3.16 通过最小化覆盖时间来发现探索选项 [43]

戴夫：现在，我们的论文！演讲由Yuu Jinnai进行。

目标：选择对探索有效的选项。

贡献：

1. 引入了一种称为覆盖时间的探索目标函数（见下文）。

2. 通过最小化覆盖时间的上界来发现选项的算法。

→基于MDP图的图诱导图计算*Fiedler*向量，使用它来最小化覆盖时间。

定义**14**（覆盖时间）：访问每个状态所需的预期步数。

定理**9.**覆盖时间的上界得到改进；

$$\mathbb{E}[C(G')] \leq \frac{n^2 \ln n}{\lambda_2 C(G')}$$

通过对不同类型选项的学习性能进行实证比较，进行了对比实验。

.....

3.3.17 策略证书**:** 迈向可追溯的**RL** [**20**]

Christoph Dann的演讲。

主要贡献：针对情节式表格MDPs的新算法，具有以下特点：

定理**10.** PAC界限；

$$\tilde{O}\left(\frac{SAH^2}{\varepsilon^2}\right),$$

以及相应的匹配后悔界限。

然而，动机是：问责制，而不一定是样本效率。

问：我的治疗效果如何？它是最好的吗？
→什么样的方法可以回答这些问题？

主要思想：引入策略证书以增加问责制。

定义**15**(策略证书)：最优和算法性能的置信区间

→模型基于乐观算法的自然扩展。

挑战：Q^π是随机的，因此很难计算其置信区间。但是！从乐观主义中，我们知道$Q^\pi \to Q^*$以已知速率。 因此，我们可以对这个量进行界定。

两个主要好处：

1. 通过准确的策略证书实现更可靠的算法。
2. 更好的探索奖励可以得到最小化的PAC和遗憾界限。

.....

3.3.18 行动鲁棒强化学习 [83]

陈特斯勒的演讲。

目标：在强化学习/马尔可夫决策过程中实现鲁棒性。考虑突然中断、高度随机的问题等情况。

→因此，研究鲁棒的MDP，其中；

$$\pi_\alpha(\pi, \pi') = \begin{cases} \pi & w.p. 1-\alpha \\ \pi' & w.p. \alpha \end{cases}$$

将算法引入作为两个玩家的对抗性游戏中的一方。 保证收敛到纳什均衡。

→基于这个鲁棒MDP算法提出了一种深度强化学习变体，在MuJoCo上进行实验。

.....

3.3.19 值函数多面体 [19]

罗伯特·达达什的演讲。

核心问题：我们能否描述给定MDP的可能价值函数空间的几何特征？

为什么要问这个问题？

- 策略空间和价值函数空间之间的关系
- 更好地理解现有算法的动态
- 强化学习中表示学习的新形式。

主要结果：给定MDP的值函数空间的几何结构是什么？

定理11。值函数的集合是（可能是非凸的）值多面体。

构建模块：1）线定理（类似策略混合的值函数描述值函数空间中的一条线），以及2）边界定理（Dave：错过了解释）。

未来的工作：RL的新表示学习方案，新的演员-评论家算法。

Dave：周二就到这里吧！

6月12日星期三：主会议

白天（实际上是下午）以多任务/终身学习的贡献演讲开始。

4.1 投稿演讲：多任务和终身学习

大多数演讲将再次是五分钟，有几个是20分钟。

4.1.1 领域无关学习与解缠表示[64]

Xingchao Peng的演讲。

思路：在领域发生变化时进行监督学习。
→在应用中有很好的动机（许多情况下，训练数据与真实任务不同）。

> 定义16(领域自适应):在一些源领域上进行训练，$P_S(X_S, Y_S)$，使用大量标记数据，然后在目标领域上进行测试，$P_T(X_T, Y_T)$。

例子：在经典的全彩图像上进行图像识别，然后转换为对草图的识别，其中一些可能是黑白的。

许多相关工作[75, 27, 85]。

新方法：深度对抗解缠自编码器（D**A**D**A**）。

- 类解缠：将特征解缠为与类别无关且领域不变的特征
- 领域解缠：将特征解缠为领域特定且领域不变的特征。

例子：类别不变性 vs. 领域不变性。
→领域不变性；给定两个图像，一个是真实汽车的图像，一个是汽车的绘画图 →将它们都输入到某个神经网络中会产生一些特征。 原则上，这些特征应该是领域不变的，因为它们在这两种汽车表现中都存在。

→类不变性：看看汽车的背景，这将产生类不变
特征，因为它们可以识别不同的类别。

类解缠：使用以下损失训练类别识别器：

$$L_{ce} = \mathbb{E}_{x,y}\left[\sum_{k=1}^{K} \mathbf{1}\{k = y\} \log C(f_D)\right]$$

为域解缠创建类似的损失（和架构部分）；

$$L_{vae} = ||\hat{f}_G - f_G||^2 + D_{\mathrm{KL}}\left(q(z \mid f_g) \mid\mid p(z)\right)。$$

实验：三个基准测试：1）5位数数据集，2）办公领域，3）DomainNet，每个数据集都包含各种领域和类别。

→发现：新模型（DAD）在这些数据集上的平均性能比SOTA提高了6%。

.....

4.1.2 强化学习中的值函数组合 [87]

Steve James的演讲。

> 核心问题：我们能否将来自不同任务的价值函数相结合，解决有趣的任务组合而无需进一步学习?

一般来说，考虑技能 Q_1和 Q_2。 通常，我们看到 $Q_1 \oplus Q_2 = \text{☹}$。

这项工作：考虑熵正则化强化学习：

$$r_{ent}(r,\, s) = r(s,\, a) - rD_{\mathrm{KL}}\left(\pi_s \mid\mid \bar{\pi}_s\right)。$$

通过熵正则化强化学习，我们证明了： $Q_1 \oplus Q_2 = \smile$

思路：OR任务组合：可以最优地组合 $Q(\square)$和 $Q(\circ)$来解决收集
$\square$ OR $\circ$的任务。

推论：在极限情况下，证明 $Q(\circ OR \square) = \max\{Q(\circ),\, Q(\square)\}$。

实验：代理应该拾取一些物体，在只需要拾取一种类型的任务中进行训练，然后在任意一种类型都可以的情况下进行测试。

总结：

1. 可以进行零射击组合，以确保找到OR任务的解决方案。
2. 在实验中表现良好!

.....

4.1.3 CAVIA: 快速上下文适应通过元学习 [95]

Luisa Zintgraf的演讲。

思路：元学习用于快速适应-学习如何将x映射到新任务上的y，快速且数据量少。

→Earl方法：MAML（参见元学习教程）。

新算法：CAVIA：通过元学习实现快速上下文适应- 1）不容易过拟合和2）可解释。

→许多任务和当前基准只需要任务识别。 许多参数和少量数据点可能导致过拟合。

实验**1**：正弦曲线实验。 任务由学习振幅和相位来定义。 MAML需要**1**500个参数，而CAVIA只需要2个上下文参数。

→上下文参数是可解释的，并且可以在任务之间重复使用。

实验**2**：**mini-im**ag**enet**实验。 随着网络变得更加复杂，MAML需要30000个参数，而cAVIA只需要几百个。

.....

4.1.4 基于梯度的元学习 [**45**]

Mikhail Khodak的演讲。

Q1：元学习算法可以利用哪些任务关系?

Q2：使用如此简单的方法是否限制了我们自己?

Q3：元学习与经典的多任务方法有什么关系?

这项工作：在凸情况下回答这些问题。答案：1. 如果最优任务参数彼此接近，

则每个任务的平均性能更好。

2. 在没有更强的任务相似性假设的情况下，GBML是我们能做的最好的。

3. GBML与正则化多任务学习之间存在自然联系

主要策略：与在线凸优化的联系。
→考虑一下标准的基于梯度的方法。 选择第一个初始化 $\phi_1 \in \Phi$，然后对于任务 $t = 1 \dots T$；1. 运行 m步的SGD。

2. 戴夫: 错过了。

然后可以从在线凸优化中导入保证，具体来说是从初始化距离编码的遗憾保证。

主要结果:

定理**12.** *GRBML*的平均遗憾为*:*

$$O\left(D + \frac{\log T}{T}\sqrt{m}\right),$$

其中 T是任务数量， D是半径，戴夫*:* 没听清 m，可能是样本数量。

.....

4.1.5 迈向理解知识蒸馏 [65]

玛丽・芳的演讲。

思路:知识蒸馏。 一位教师（经过训练的神经网络）提出/表示一个函数 $f_\theta : \mathcal{X} \to \mathcal{Y}$。 然后，希望将这个函数压缩成一个更小/更简单的网络（学生！）。

问题: 知识蒸馏有多有效？ 通过研究一个称为转移风险的量来具体化这个问题，该量度衡量蒸馏过程可能有多糟糕。

设置: 线性蒸馏: 1) 线性教师，2) 学生是一个深度神经网络。

结果：

1. 可以准确计算学生将学到的内容！如果教师是 $f_t(x) = w^\top x$，那么学生将会完全跟随教师（取决于数据量）。
2. 还可以限制转移风险：

$$n_data \geq d \implies risk = 0 \tag{16}$$

$$n_data < d \implies risk \leq \left(\frac{\log n_{data}}{n - data} \right)^\kappa, \tag{17}$$

$$\tag{18}$$

with $\kappa \in [0, inf\, ty)$ the easiness of the distr.

.....

4.1.6 可迁移的对抗训练 [53]

刘洪的演讲。

典型假设：训练数据和测试数据来自同一分布。

→如果我们改变这个假设呢?

∴这项工作：如何将学习器推广到不同的分布 P和 Q上。

目标：用训练误差（在分布 P上）来限制目标误差（在分布 Q上）。

→基于领域适应的现有理论。

之前的工作（对抗性特征适应）：

1. 最小化源风险
2. 最小化差异项：学习一些特征表示，使差异最小化

3. 定义为一个双人游戏：领域鉴别器试图区分源领域和目标领域，而特征提取器试图混淆它。

但是，进行领域适应的先决条件是通过 λ来量化适应性。 如果 λ很大，我们永远无法期望将学习器适应源领域。

问题：我们如何修正这个问题?

答案：一种方法是固定特征表示，防止适应性变差。

但是：现在我们需要弄清楚如何适应目标领域。

新模型：可转移对抗训练（TAT）。

- 主要思想：不再进行特征适应，而是将源领域和目标领域与可转移的示例相关联

- 生成可转移的示例来弥合两个领域之间的差距。
  →具体来说：训练一个分类器和一个领域鉴别器。
- 整体优化问题：
  1. 四个损失项，带有固定的特征表示。
  2. 无需特征适应（轻量级计算）
  3. 比对抗特征适应快一个数量级。

分析：考虑旋转的两个月亮问题。 目标领域与源领域相比旋转了30°。

实验：领域适应基准测试（Office031，Image-CLEF，Office-home，ViSDA）。

→发现：达到与SOTA相当的性能。

代码：https://github.com/thuml/Transferable-Adversarial-Training

.....

4.2 贡献演讲：强化学习理论

现在，更多强化学习理论。

4.2.1 仅从观察中可证明高效的模仿学习 [80]

Wen Sun的演讲。

之前的工作可以实现样本复杂度为：

$$\text{poly}(Horizon, |A|, |S|, 1/(1-\gamma))。$$

这个设置：从观察中进行模仿学习

•给定：观察的轨迹

•从观察中学习

•没有交互式专家，没有专家动作，没有重置，也没有成本信号

NE算法：前向对抗模仿学习（F**A**IL），无模型方法。

→思路：学习一系列策略，$\pi_1, \ldots, \pi_{T-1}$. 为了学习每个策略，将其视为一个双人博弈。

→为了解决这个博弈，将其视为一个极小极大博弈，将其作为最小化积分概率度量（IPM）来解决：

$$\max_{f \in \mathcal{F}} \mathbb{E}_{x \sim P}\left[f(x)\right] - \mathbb{E}_{x \sim Q}\left[f(x)\right],$$

其中 $\mathcal{F}$是一组判别器。π_0与动态模型一起定义一个代理（生成器），
而 π_1？作为专家分布。然后，希望：

$$\max_{\pi_0 \in \Pi} \max_{f \in \mathcal{F}} f(\pi_1^\star - f(\pi_0).$$

通过解决上述问题，得到下一个策略 π_1，并重复这个过程。

问：这种方法的样本复杂度是多少？

答：如果判别器非常强大，我们会过拟合。 但是如果它太弱，就无法很好地区分。

→因此，引入“内在贝尔曼误差”

> 定义**17**（内在贝尔曼误差）：函数类 $\mathcal{F}$的内在贝尔曼误差为：
>
> $$\Gamma^* f(x) \triangleq \mathbb{E}_{a \sim \pi^*, x \sim P(\cdot|s,a)}\left[f(x')\right]$$

失败的结果：

定理13。在可实现性假设下，所以 $\pi^* \in \Pi$和 $V^* \in \mathcal{F}$，为了学习一个接近最优的策略，我们需要：

$$poly(T,\, A,\, 1/\varepsilon,\, SC(\Pi),\, SC(\mathcal{F}))\,,$$

其中SC是函数族的统计复杂度。

但是，上述情况迫使我们忍受这种固有的贝尔曼误差。

问：在从观察中进行模仿学习的情况下，这种固有的贝尔曼误差是可以避免的吗？

答：是的！但是在基于模型的强化学习中。

- 从一个可实现的模型类开始，即 $P \in \mathcal{P}$。
- 然后存在一个算法，将 $\mathcal{P}$，F作为输入，输出上述样本界限下的 ε最优策略。

实验：在模拟结果（**fe**tch **re**ach和**re**ach**er**）上实施这种模仿学习算法，表现相当好（参数调整有限）。

要点：

- 仅通过专家的观察，我们可以学习到最优策略
- 接近最优的保证
- 类似于监督学习的样本复杂度
- 开箱即用的性能非常好。

.....

4.2.2 死胡同和安全探索 [25]

由Mehdi Fatemi发表的演讲。

问：什么是死胡同?

→如果一个终止状态是不希望的，它会阻止达到最大回报。

> 定义**18**（死胡同状态）：如果从状态 s开始的所有轨迹都以概率 1在某个有限（可能是随机的）步数内到达不希望的终止状态，则状态 s是死胡同。

问：为什么我们应该关心死胡同?

答：大多数算法只在无限（s, a）访问的假设下收敛 $\forall_{s,a}$。

→新的想法：

> 定义**19**（策略安全性）：如果对于任何 $\lambda \in [0, 1]$，策略都是安全的：
>
> $$\sum_{s'} T\ (s,\ a,\ s')\ \geq 1 - \lambda\ \implies\ \eta\ (s,\ a)\ \leq \lambda。$$

解决方案：创建一个名为“探索”MDP的新MDP，与原始MDP相似，但有一些细微的变化。结果：

定理14。设 η为任意策略，使得 $\eta(s,a) \leq 1 + Q^*(s,a)$，其中 $Q^*(s,a) = -1$至少对于一个动作。

那么， η是安全的。

（还要进行实验）。

.....

4.2.3 分布式强化学习中的统计和样本 [**74**]

马克・罗兰的演讲。

回顾：分布式强化学习旨在学习完整的回报分布：

$$Z^{\pi}(s,\, a) = \sum_{i=0}^{\infty} \gamma^t R_t。$$

主要贡献：新方法-直接学习回报分布的函数，如矩、尾部概率、期望等。

新框架！可以在新领域取得进展：

- 理论：从动态规划中可以学到回报分布的哪些属性

- 算法：近似学习回报分布统计的框架。

应用：期望分位数。 新的深度强化学习代理“期望回归DQN”（ER-DQN）相对于QR-DQN（分位数回归DQN）具有改进的平均性能。

.....

4.2.4 基于**Hessian**的策略梯度 [**78**]

沈泽邦的演讲。

思路：将策略优化形式化为轨迹的优化。

→旧方法：使用REINFORCE/SGD通过策略梯度缓慢改进策略。

新方法："无视"策略优化。 可以提高现有策略梯度在这种"无视"情况下的样本效率。

但是：在"无视"情况下，策略梯度算法会产生偏差。所以，需要进行修正。

总结：第一个能够通过策略优化将样本复杂度从$O(1/\varepsilon\text{^}4)$降低到$O(1/\varepsilon\text{^}3)$以达到$\varepsilon$最优的可证明方法。

.....

4.2.5 最大熵探索 [37]

Elad Hazan的演讲。

背景：智能体在没有奖励信号的马尔可夫决策过程中。如何进行探索?

→ 现有方法：任务无关的探索、好奇心和探索奖励。

主要问题：我们能否高效地解决没有奖励的探索问题?

背景设定：

- 每个π都会产生一个状态分布
- 给定一个策略类Π和一个凹函数H，作用于状态分布，我们能否找到：

$$\max_{\pi \in \Pi} H(d_\pi),$$

其中d_π是由$_\pi$引起的状态的稳态分布。

命题**1**.$H(d_\pi)$在$_\pi$中不是凹的。

新算法：最大熵算法。 采用策略的均匀混合的概念 $C = (\pi_1, \ldots, \pi_k)$。

1. 密度估计器
2.

定理**15.** 在有限（小）的迭代次数内，算法保证$H(d\mathbf{m}i_x) \leq \mathbf{m}a_{x\pi \in \Pi}\, H(d_\pi)$ - ε（有效地解决了这个问题）

.....

4.2.6 结合多个模型进行离线策略评估 [32]

Omer Gottesman的演讲。

> 定义**20**（离线策略评估）：假设我们有一批由某个我们无法控制的策略π收集的数据，并使用这些数据来评估其他策略π。

两种常见的方法：

1. 基于模型：使用数据来学习环境模型，然后使用该模型来评估新策略。

2. 重要性采样：根据策略差异重新加权回报/状态访问。
   → 问题：方差很大！（即使其他统计性质良好）。

问：如果我们有多个具有不同优势的模型，我们能否将它们组合以获得更好的估计？

答：是的，这是该工作的主要思想！

直观的例子：MDP的两个区域-1）顶部，建模良好，2）底部，建模不良。假设我们有能够进入顶部/底部区域的模型：我们知道顶部的轨迹会更好！因此，坚持使用在该区域表现良好的模型。

→想法：可能需要在短期和长期准确性之间进行权衡。模型误差的上界：

$$|g_T - \hat{g}_T| \leq L - t \sum_{t=0}^{T} \gamma^t \sum_{t'=0}^{T} L_t^{t'} \varepsilon_t (t - t' - 1) + \sum_{t=0}^{T} \gamma_r^{\varepsilon}(t),$$

其中 ε_r和 ε_t是误差界限， g_T是回报（而$\hat{g}_T$是估计值）， L_t是Lipschitz常数
戴夫：（关于转换，我想是这样的吧？）

想法：使用MCTS来最小化整个轨迹上的回报误差界限。

两种类型的模型：

1. 非参数模型：根据与邻居的相似性预测给定状态-动作对的动态。

   优势：在数据丰富的状态空间区域可以非常准确。

2. 参数模型：任何参数回归模型或手工编码模型，包含领域知识。

   优势：倾向于更好地推广到与数据观察不同的情况。

→ 这两种模型的好处是：它们具有不同的优势！所以，让我们将它们结合起来。

实验：医学模拟器（模拟癌细胞和HIV的生长）。对于不同的行为策略，进行离线策略评估。

总结：

- 提供一个将多个模型结合以改进离线策略评估的通用框架
- 通过个体模型、误差估计或多个模型的组合来改进

.....

4.2.7 使用线性特征的样本最优参数化 Q**-Learning** [**92**]

由林F.杨发表的演讲。

考虑：维度诅咒！最优样本复杂度为：

$$\tilde{\Theta}\left((1-\gamma)^{-3}|\mathcal{S}||\mathcal{A}|\right)$$

→状态和动作太多了。 那么，我们如何最优地减少游戏的维度?

A: 利用结构!

方法：基于特征的马尔可夫决策过程。分解转移模型：

$$P(s' \mid s,\, a) = \sum_{k\in[K]} \phi_k(s,\, a)^T \psi_k(s'),$$

这将MDP分解为一些因素，其中 ϕ是已知的， ψ是未知的。

玩具例子：股票价格！在一些金融模型中，我们可以将股票价格分解为一组代表性股票的线性组合。 如果是这种情况，我们可以使用模拟这种线性关系的 Q来参数化地解决RL问题。

→思路：我们能够从生成模型中采样（任意的s, a）。 用参数 $w \in \mathbb{R}^k$来表示 Q函数，因此：

$$Q_w = r(s,\, a)\gamma\phi(s,\, a)^t w$$

然后，使用改进的Q-Learning进行学习可以实现以下样本复杂度：

$$\tilde{O}\left(\frac{K}{\varepsilon(1-\gamma)^7}\right).$$

问：这是最优的吗?

答：不是，可以证明在“锚定条件”下，它会崩溃为：

$$\tilde{\Theta}\left(\frac{K}{\varepsilon^2(1-\gamma)^3}\right).$$

.....

4.2.8 策略搜索中的样本迁移 [**84**]

André Tirinzoni的演讲。

策略搜索：连续控制任务的有效强化学习技术。

→高样本复杂度仍然是一个主要限制。 从几个来源获得的样本被丢弃。

形式上：给定一些源任务（具有不同模型的MDPs），重用从这些问题中收集的数据在一个新的相关任务中。

贡献**1**：一种新的重要性采样器技术，可以实现更有效的梯度估计。

→估计器的良好特性

1. 无偏且有界的权重。
2. 易于与其他方差减少技术结合使用。
3. 有效样本大小 $\equiv$可转移的知识（自适应批量大小）。
4. 对负迁移具有可证明的鲁棒性。

问题！ P通常是未知的，因此无法计算重要性权重。

→解决方案：在线上界最小化。

实证结果：在已知和未知模型（cartpole，minigolg）上表现良好。
非常有效地从不同策略但相同环境中重复使用样本。

.....

4.2.9 探索意识强化学习再探

问：为什么要考虑探索意识?

答：为了学习一个好的策略，强化学习代理必须进行探索!

→新目标：在特定的策略下，找到最优策略，知道将进行探索。

主要方法：考虑一个固定的探索方案（ε贪婪等）。然后：1）选择贪婪动作，2）选择探索动作，3）执行，4）接收**r**，s'。

→使用关于探索意识问题的信息。

两种方法：

1. 预期：更新实际采取的动作的Q值，但预计代理可能在下一个状态中进行探索。. 因此，使用此方法进行自举（但是，计算此期望可能很困难）。
2. 替代方法：（主要贡献）直接将探索性添加到环境中。

结论：探索意识的强化学习，特别是替代方法，可以轻松改进各种强化学习算法。

.....

4.2.10 基于核的鲁棒MDP的强化学习 [51]

Shiau Hong Lim的演讲。

> 定义21（鲁棒MDP）：通过考虑模型不匹配和参数不确定性来扩展MDP。

对于状态聚合，通过鲁棒策略可以改进$||V_{R\pi}$ - $V_^*||$的性能界限：

$$O(\frac{1}{(1\text{ - }\gamma)^2} \rightarrow O(\frac{1}{(1-\gamma)}$$

贡献；

1. 通过扩展到核平均器设置来改进鲁棒性能界限。
2. 制定了一个实用的基于核的鲁棒算法，并在基准任务上进行了实证结果。

定理16。将状态聚合的值损失界限扩展到基于计数的设置。

→实用算法：使用核平均器来近似MDP模型，然后使用近似鲁棒贝尔曼算子解决近似模型。

结论：1）对于鲁棒基于核的强化学习提供性能保证，2）展示了这种方法的显著实证效益。

戴夫：周三就到这里了！

6月13日星期四：主会议

常规会议的最后一天！

5.1 贡献演讲：强化学习

我们开始最后的强化学习会议。

5.1.1 在约束条件下的批量策略学习 [49]

由Hoang M. Le发表的演讲。

动机：

1. 改变强化学习目标：考虑通常的强化学习目标：

$$\min_x C(\pi) = \mathbb{E}\left[\sum cost(s,\, a)\right]$$

   但是，在现实中，这并不总是可行的。→ 很难定义一个单一的成本函数！

   因此，指定额外的约束似乎是自然的。

2. 样本高效批强化学习：但是，如果我们改变成本目标，我们需要解决一个新的问题，这可能使强化学习变得更加困难。

   → 在批强化学习中，倾向于考虑拥有一些现有数据集（来自某个先前策略π_D，可能是次优的）。

   问题1：我们能否利用这些丰富的数据来学习一个更好的新策略？

   问题2：如果我们想要改变我们的约束条件怎么办？ 在这些新的约束条件下，我们仍然能够学习一个新的策略吗？

符号说明：

- 数据集包含n个数据元组，$_D$= $\{(s, a, r, s'), \ldots,$ 由 π D生成。
- 目标是找到一个π，使得$\min$ π $C(_\pi)$，满足某个约束 $G(\pi) \leq 0$。

例子：1) 反事实和安全策略学习（约束条件指定避免某些状态），或者2) 多准则价值约束（驾驶，最小化行驶时间，同时满足在特定车道行驶的约束）。

问题：拉格朗日函数：

$$L(\pi, \lambda) = C(\pi) + \lambda^\top G(\pi)$$

，其中原始问题为：

$$\min_\pi \max_{\lambda \geq 0} L(\pi,\, \lambda)$$

对偶是：

$$\max_{\lambda \geq 0} \min_\pi L(\pi,\, \lambda).$$

因此：扩展策略类以允许处理非凸成本的随机策略。
方法：通过降低到监督学习和在线学习。

算法概述：重复以下步骤：

1. $\pi \leftarrow$最佳响应(λ)，通常来自拟合Q迭代[23]。

2. $L_{max}=$ 评估 (对偶) 固定 π

3. $L_{min}=$ 评估 (原始) 固定 λ

4. 如果 $L_{max} - L_{min} \leq \omega$，停止。

5. 否则， $\lambda \leftarrow$在线算法(所有先前 π)。

为了处理评估步骤，依赖于离线策略评估。 也就是说：给定 D，希望估计
$\hat{C}(\pi) \approx C(\pi)$。

$\rightarrow$ 提出了一种新的无模型直接方法，用于离线策略评估，称为Fitted Q Evaluation（FQE）。概述如下：

1. 选择一个函数类。然后，进行k次迭代：

2. 求解Q；$(s, a) \rightarrow y = c = Q_p\text{rev}(s', \pi(s'))$

3. $Q_{prev} \leftarrow Q$.

定理**17.**FQE的端到端保证：对于$n = p\mathbf{o}ly(\mathbf{1}/\varepsilon, l\mathbf{o}g\,\mathbf{1}/\delta, l\mathbf{o}g\,K, \mathbf{...})$，以$\mathbf{Pr}\ \mathbf{1}$-$\delta$为概率：

$$|C(\pi) - \hat{C}(\pi)| \leq O(\omega + \sqrt{\beta}\varepsilon),$$

其中ω是FQE的停止条件。

实验：一个自上而下的驾驶任务，目标是最小化行驶时间，同时满足平稳驾驶和保持在车道中心等约束条件。 在一个忽略这些约束的π_D上进行训练。

$\rightarrow$发现：算法的输出策略可以满足新的约束条件，同时保持性能高（旅行时间）.

.....

5.1.2 量化强化学习中的泛化能力 [17]

Karl Cobbe的演讲.

注意：许多深度强化学习算法在训练和测试中使用相同的任务.

$\rightarrow$因此，这项工作引入了新的环境，专门用于在深度强化学习中明确研究泛化。更好的基准$\Longrightarrow$更好的研究，更好的算法.

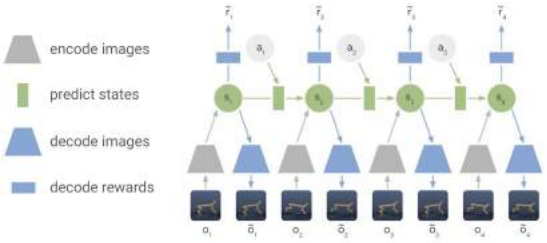


图7：PLaNET潜在动力学模型.

主要发现：事实上，许多深度强化学习算法在这个环境中过拟合（该环境明确要求代理能够很好地泛化）.

硬币收集领域：

1. 2D平台游戏，代理必须收集硬币
2. 根据难度生成关卡
3. 可以构建任意大小的训练集.

基准结果：在硬币收集领域的各种任务上运行各种架构（使用分割的训练/测试集）.

→发现1：过拟合实际上是一个重要问题.
→发现2：更深的架构往往更好地泛化（3个卷积层对比15个卷积层）。
→发现3：探索不同的正则化方案的效果，如 L_2、dropout、批归一化和数据增强。 它们在正则化方面都有类似的效果。

代码：https://github.com/openai/coinrun

.....

5.1.3 从像素中学习潜在动态规划 [**34**]

Danijar Hafner的演讲。

问题：当世界模型已知时（如AlphaGo），规划非常有效。但是，在复杂或未知的转换中，它已被证明是无效的。

→考虑视觉控制任务：MuJoCo，但输入是图像。

这项工作：PLaNet：1）大规模模型基于RL的规划方法，2）在潜在空间中进行高效规划，使用大批量数据，3）比模型无关方法更具样本效率。

实验：在MuJoCo任务的视觉变体上与无模型代理进行比较，以较少的样本获得竞争性表现（与无模型方法相比）。

因此：一种在图像空间中进行模型基于强化学习的可扩展方法。

代码：https://danijar.com/planet

.....

5.1.4 近似策略迭代的投影 [**3**]

Riad Akrour的演讲。

考虑：在强化学习中，使用熵正则化的演员-评论家方法很常见（参见：SA**C**，NA**C**，PPO，A3**C**，REPS，A**C**ER，A**C**KTR，TRPO）。

软约束：

$$max\pi_J(\pi) + \alpha H(\pi)。$$

其中 J是目标（回报）， H是熵。 相反，硬约束：

$$\max_{\pi} J(\pi) \quad \text{s.t.} \quad H(\pi) \geq \beta$$

→更难优化（硬约束），更容易调整。
贡献：1）将高斯或软最大化策略的香农熵的硬约束投影，2）对硬约束进行投影，得到深度强化学习的优化方案。

实验：比较不同的演员-评论家/策略优化算法与熵奖励
与它们的基于投影的方法相比。
→熵奖励在早期是无效的，但在严格的熵约束下，可以产生更好的探索和更好的策略。 简单的方案，可以融入任何强化学习算法。

.....

5.1.5 无意识学习结构化决策问题 [**39**]

Craig Innes的演讲。

重点：决策中的“无知”。

→例如：农业。代理人可能学到什么？ 粮食中蛋白质水平之间的动态关系，何时以及哪些粮食需要施肥等等。

问：但是如果在解决问题的过程中，我们发现新的状态变量和行动对于最佳性能至关重要怎么办？

答：如果我们不明确地对这种无知进行建模，代理人可能无法解决许多问题。

贡献：代理人通过领域试验和专家建议的证据逐步学习一个可解释的决策问题模型。

$\rightarrow$方法：如果代理在最近的k次试验中的表现低于真实策略π_+的阈值β，则说：
“在时间t，你应该选择a'而不是at。”代理可以从这个反馈中学到很多东西：
1）新的行动，2）a'比a_t更有价值，等等。

实验：一套随机生成的决策问题。

$\rightarrow$发现1：新的代理能够学习到真正的最优策略，尽管最初对状态变量和行动一无所知。$\rightarrow$发现2：专家对何时干预的容忍度不同，对代理的学习能力有深远影响。

.....

5.1.6 校准的基于模型的深度强化学习 [**56**]

Ali Malik的演讲。

考虑到：不确定性在现实世界中很普遍。 为了为现实世界设计有效的强化学习系统，我们必须直面这种不确定性。

$\rightarrow$因此，准确地建模不确定性非常重要。

问：在强化学习中，哪种类型的不确定性是重要的?

A: 转向统计学！例如考虑气象学家：采用“适当的评分规则”，这些规则在预测分布与真实分布匹配时成立。

定义**22**（适当的评分规则）：根据两个标准衡量模型捕捉不确定性的好坏程度：

1. 尖锐度：预测分布应该集中（方差较小）
2. 校准度：不确定性应该在经验上准确（如果我预测某事具有概率为p的倾向，那么在$p\%$的时间里我应该是正确的）。

论断：对于基于模型的强化学习来说，校准度尤为重要。

问：为什么?

A1：用于规划！校准的不确定性可以得到更好的期望估计。

定理**18**.在真实动力学下，MDP的π值等于在校准动力学模型T下的策略值。

A2：用于探索！许多探索/利用算法使用上界置信度（UCB）来指导选择，例如LinUCB：

$$\arg\max_{a\in\mathcal{A}}\left(x\theta_a=\alpha\sqrt{x\hat{\Sigma}_a x}\right).$$

校准自然地改善了UCB，从而实现更好的探索。

问：我们如何确保我们的系统具有校准的不确定性？（请注意，深度网络的不确定性通常是未校准的）

答：参考之前的工作方法：Kuleshov等人的重新校准方法[47]。

新算法：

- 探索：使用当前模型收集观察数据 T。
- 学习模型：使用新的观察数据重新训练转换模型。
- 学习校准：在保留的观察数据子集上学习重新校准器 R。
- 重新校准：使用重新校准器重新校准转换模型。

实验**1**：使用上下文强化学习来研究校准对探索的影响。

→发现：校准显著改善了上下文强化学习中的探索。

实验**2**：探索校准在MuJoCo连续控制任务中的影响。 将校准添加到Chua等人的算法中[16]（在MuJoCo上的最佳方法）。

→发现：添加校准可以改善连续控制任务的样本复杂度。

实验**3**：库存规划-校准贝叶斯**DenseNet**。

.....

5.1.7 可配置连续环境中的强化学习 [**59**]

Alberto Maria Metelli的演讲。

回顾：传统强化学习假设环境是一个固定的马尔可夫决策过程（MDP）。

→但是，在现实世界的场景中，我们可以对环境的某些方面进行部分控制。

例子：考虑赛车比赛。 离线时，我们可以修改汽车的某些方面，使比赛更有效。

定义**23**（可配置MDP）：代理寻求优化性能的环境配置ω下的策略参数θ，其中ω影响转移模型的方面。

假设（与先前的技术水平相比）：有限的状态-动作空间，对环境动态有完全了解。

算法：相对熵模型策略搜索（REMPS）：两个阶段-1）优化（在信任区域内找到新的稳态分布d'），和2）投影（找到策略π_θ'和配置分布p'_ω，以产生正确的稳态分布）。

实验：1）链式领域，2）倒立摆。

代码：https://github.comn/albertometelli/remps。

.....

5.1.8 基于目标的时差学习 [50]

Niao He的演讲。

回顾：在使用DQN类算法中，使用目标网络是普遍的。思路是在Q更新中使用一个单独的网络来进行引导，而不是进行引导。

$\rightarrow$ 但是！我们对于为什么目标网络有效并不了解太多。

问题：我们能否理解在使用目标变量时TD学习的收敛性？
答案：是的！研究了三种变体：1）平均（A-TD），2）双重（D-TD），3）周期性（P-TD），TD学习。

经典的TD学习：

$$\theta_{t+1} \leftarrow \theta_t - \alpha_t \nabla\theta L(\theta; \theta_t'; e), \tag{19}$$

然后进行目标变量更新：$\theta_t'_{+1} = \theta_{t+1}$每隔一段时间。

主要结果：

定理**19**。**1**）目标扩展$\mathbf{T}D$学习（A-$\mathbf{T}D$和D-$\mathbf{T}D$）渐近收敛到正确解，*2*）对这些方法的收敛速度进行样本分析。

实验：比较A-TD、D-TD和P-TD在模拟中的性能。

.....

5.1.9 线性化控制：稳定算法和复杂性保证[73]

Vincent Roulet的演讲。

问题：系统由一些非线性动力学描述的非线性控制问题。

Q1：ILQR（用于解决这些问题的经典算法）是否收敛？能否加速？
Q2：如何刻画非线性控制的复杂性？

贡献：

1. ILQR是高斯-牛顿方法：正则化ILQR收敛到一个稳定点
2. 通过外推步骤进行潜在加速：加速ILQR类似于Catalyst加速。
3. Oracle复杂性分析

代码：https://github.com/vroulet/ilqc。
Dave：现在，我们的其他选项论文。

.....

寻找最小化规划时间的选项 [42]
由Yuu Jinnai演讲。

贡献：寻找最小化规划时间的最优选项的问题是NP-Hard。

问：为什么要使用选项？
答：嗯，选项可以帮助强化学习和规划。

问：好的，但是我们如何获得好的选项？ 更一般地说：什么构成了一个好的选项？

答：规划！

贡献：

1. 通过价值迭代算法正式定义寻找规划的最优选项的问题。
2. 证明这个问题是NP-Hard。
3. 还证明这个问题很难近似解决：难以近似。 近似比的下界是：$2^{\log^{1-\varepsilon}(n)}$。
4. 在某些条件下确定计算最优选项的近似算法。
5. 对这些选项进行实验评估。

主要结论：只有当我们有结构、先验知识或假设时，选项发现对规划才有用。

.....

5.2 贡献演讲：深度学习理论

接下来是深度学习理论。

5.2.1 为什么更大的模型具有更好的泛化能力？[12]

由Alon Brutzkus发表的演讲。

目标：理解大型网络的泛化奥秘。

→ 经验观察：大型和小型模型都可以达到零训练误差，但大型模型的泛化能力更好。

两个主要挑战：

1. 证明泛化差距（需要超出样本复杂度的东西）
2. 分析神经网络上梯度下降的收敛性。

方法：分析一个简化的学习任务，该任务在经验上表现出相同的特征，并与实际问题共享特征。

主要结果：

- 研究了XOR问题的高维变体（XOR“检测”问题）
- 在某些假设下证明了过度参数化可以改善泛化能力

学习问题：地面真实卷积网络的二维滤波器。
→ 问题：当你尝试使用少量通道和大量通道学习这个函数时会发生什么？

实证发现：收敛时，大型网络能够准确找到真实函数，而小型网络（理论上也能找到真实函数）却找到了一个训练误差为0但测试误差不为0的函数。

→这项工作是关于解释上述内容的。

定义**24**（异或检测问题）：四个**2**维二进制模式，输入为 $x=（x_1，\ldots，x_d）\in\{-1，1\}^{2d}$。
异或检测器：

$$f^*(x)=\begin{cases}1 & \exists x_i 使得 x_i=_1 或 x_i=_3\\ -1 & 否则\end{cases}$$

考虑一个网络：三层CNN（卷积→最大池化→全连接），使用2维滤波器：

$$\sum_{i=1}^{n}\left[\max_j\left\{\sigma(w^i x_j)\right\}-最大_j\left\{\sigma(u^i x_j)\right\}\right].$$

使用：稍微修改的合页损失，随机梯度下降。

> 定义**25**(多样化训练集)：如果一个训练集 $\mathcal{S}$包含正负样本，则称其为多样化训练集，其中每个正/负样本都包含所有模式（对于正样本，p_1，p_2，p_3，p_4，对于负样本，p_2，p_4）。

训练；

- 假设有一个多样化训练集 $\mathcal{S}$。
- 对于 $k=2$（网络中的通道数），存在多个训练误差解（全局最小值）。→因此，只需检测到其中一个即可。

问题：

1. 对于 $k=2$，SGD是否收敛到全局最小值?
2. 如果是，是哪一个?
3. 当$k > 2$时会发生什么?

理论上说：对于一个小网络，SGD会收敛到一个不能恢复f^*的全局最小值，而一个大网络可以（以很高的概率）。→（也将上述内容翻译为PAC保证）。

实验：XOR检测。研究网络大小、收敛速度和探索效果的影响。

→发现：对于不同的训练规模，大网络的表现优于小网络。

.....

5.2.2 关于神经网络的谱偏差[69]

Nasim Rahaman的演讲。

问：当大规模神经网络可以学习随机标签时，为什么它们能够泛化？（参见Arpit等人的工作[6]）。

> A（本研究）：神经网络首先学习简单的函数（随着训练的进行逐渐构建更复杂的函数）。

问：但是如何量化简单性?

答：使用傅里叶谱！网络首先学习低频函数。

问：我为什么要在意呢?

答：嗯，考虑标签噪声。 我们通常认为是独立同分布的噪声（白噪声）。 但是，白噪声是具有平坦频谱的噪声的特例（形式在期望中是恒定的）。

→例如，高频噪声会导致验证分数大幅上升，而低频噪声则不会获得相同的好处。

.....

5.2.3 用于模块化深度学习的递归草图[**31**]

Joshua R. Wang的演讲

问题：物体识别。 训练一个神经网络，学习在图像中识别物体。

→想法：对这个任务进行改进：输出图像中物体的简洁表示/嵌入。

要点：可以使用这个模型来解决原始的物体识别问题。

> 定义**26**（模块化网络）：模块化网络由模块组成，这些模块是独立的神经网络组件，可以通过输出与其他模块的输入进行通信。

→使用模块化网络输出包含网络检测到的每个物体的草图。

可证明的属性：1）属性恢复，2）草图相似性，3）摘要统计，
4）优雅的擦除。

.....

5.2.4 深度网络中的零样本知识蒸馏[**60**]

Gaurav Kumar Nayak的演讲。

核心问题：我们能否在没有训练数据的情况下进行零样本知识蒸馏?

答案：是的！（这项工作）。

思路：伪数据合成。 通过Reddy Mopuri等人引入的技术，通过创建伪图像来构建类印象（CI）[71]。

→但是：CI有限，因为1）生成的样本不多样化，2）学生不具有很好的泛化能力。

这项工作：通过“数据印象”扩展类印象，克服这些问题（不是类特定的）。

→在MNIST和CIFAR-10上进行实验，发现数据印象增强了学习效果。

总结：基于数据印象的零样本知识蒸馏方法。

.....

5.2.5 深度学习的过参数化收敛理论[4]

考虑：给定 L层的深度网络，给定 n个训练点。

→假设网络是过度参数化的（因此参数的数量是训练数据的多项式）。

主要结果1：

定理**20.** 如果数据不是退化的，则使用SGD的过度参数化网络将在以下情况下找到全局最小值：

$$T = \frac{poly(n,\, L)}{\delta} \log \frac{1}{\varepsilon}.$$

关键信息：对于足够大的随机初始化邻域，该邻域几乎是凸的。 如果目标函数是光滑的，则可以通过正确的梯度步骤来降低目标值。

主要结果2：

定理**21.**如果神经元的数量足够大 $m \geq poly(n, L)$，则对于足够大的初始邻域，神经网络的行为类似于神经切向核（$N\mathbf{T}K$）。也就是说：

$$\text{过度参数化的深度网络} = NTK$$

因此：

$$\text{网络基本上是凸的和平滑的} \implies \text{训练很容易。}$$

.....

戴夫：出去吃午饭

.....

5.3 最佳论文奖：稀疏高斯过程回归的收敛速度

David Burt的演讲。

本文：获得良好近似的计算复杂度

背景：稀疏高斯过程（GP）回归。

- 优点：可以明确表示不确定性作为数据集大小的函数。
- 缺点：计算复杂度高：核方法，需要存储核矩阵 $O(N^2)$ 空间，并且需要求逆该矩阵，因此需要 $O(N^3)$ 时间。

伪观测近似：基于 $M \ll N$个引导点的近似。
→时间复杂度为 $O(N\,M^2 + M^3)$，内存复杂度为 $O(N\,M)$（远远小于之前的方法！）。

本研究：变分稀疏高斯推断。优点：

- 我们可以通过最大化ELBO来同时学习超参数。
- 在没有太多数据的区域，后验的近似仍然不确定。

> 核心问题：对于这个工作，我们真正需要多少个感应点（M）？

边界的起点：对于一个固定的数据集：

$$D_{\mathrm{KL}}\left(Q \mid\mid P\right) \qquad \leq \text{上界} - \mathrm{ELBO}。$$

随着数据集大小的增加，上界和下界在KL上趋于一致。 特别地，一个后验边界导致：

$$D_{\mathrm{KL}}\left(Q \mid\mid P\right) \quad \frac{\leq \text{降低矩阵近似的质量}}{2\sigma_n^2}\left(1 + \frac{||y||^2}{\sigma_n^2}\right).$$

为了证明先验边界，需要：1）选择感应点的高效过程，和2）关于训练数据分布的假设。

→近似核矩阵：

$$K_{ff} \approx K_{uf}^{\top} K_{uu}^{-1} K_{uf}$$

可以限制上述近似的偏差，但需要：1）了解特征值衰减的知识，以及2）由于初始化方案而产生的额外误差

目标：限制Q和P之间的KL散度。 现在他们有：1）消除了对诱导点位置的依赖，2）对训练数据的分布做出直观假设。

主要结果:

定理22。使用k-DPP初始化并假设$x_i \sim p(x)$：

$$D_{KL}\left(Q \mid\mid P\right) \quad \leq^{NM+Dave：有些我错过的项} \frac{}{2\sigma_n^2}\left(1+\frac{||y||^2}{\sigma_n^2}\right).$$

要点：

- GP回归的稀疏逼近快速收敛
- 平滑先验，密集输入数据$\Longrightarrow$非常稀疏的逼近可能。

戴夫：今天剩下的时间都是会议

6 6月14日星期五：研讨会

今天有研讨会！

6.1 研讨会：人工智能应对气候变化

首先，约翰·普拉特（对他几年前的精彩演讲的跟进！）

6.1.1 约翰·普拉特关于机器学习如何帮助应对气候变化

主要观点：能源研究不同于应对气候变化。

→气候危机：我们的碳预算即将耗尽！如果没有干预，我们最好的模型
预测

事实证明，我们无法停止能源基础设施（投资数万亿美元）。 我们能做的最好的办法是逐渐停止使用它们：可能实现3%的减少（能源使用），或者如果我们假设某种奇迹般的速率，
也许是10%。

→假设快速脱碳（10%），为了实现我们的目标（根据巴黎协定的定义），我们必须立即开始！

→好吧，但如果不是激进的呢？ 更保守的估计：3.3%。
测量指标：最终温度上升（°）。
需要发生什么？：

- 快速脱碳以避免温度上升超过2°。
- 零碳能源技术必须随时可用
  →可再生能源现在已经准备就绪，相对可扩展。
- 2040年后的零碳技术仍然有用。 →避免绝对最糟糕的气候变化的后备措施

  →每个人都能够实现和希望拥有充足能源的情景仍然可行和可取。

想法:假设我们想要部署可再生能源技术（太阳能、风能等）。

→问题:需求和供应波动剧烈，因此在我们有能源和需要能源之间可能会存在差距（考虑到多云天气！）。

有关可再生能源成本的可用代码/工作称为DOSCOE [66]。 代码:https://bit.ly/DOSCOE。

问题:机器学习能帮助解决这个问题吗?

- 最有价值的零碳源是可调度的（如水电）
- 也可以调度需求（在供应的碳强度低时开启）

•机器学习/人工智能可以进行需求响应吗？ →谷歌通过机器学习控制报告数据中心冷却能源消耗减少了40%。

两个注意事项:

1. 许多来源导致气候危机。 约翰的"悲伤饼图"中的细分:
工业: 21%，电力和热能生产: 25%，交通: 14%，建筑: 6%，
农业/林业和其他土地利用: 24%。

2. 需要"碳压力"：自由市场不鼓励这些做法!

- 目前没有需求响应的激励措施
- 提高效率会降低电力和化石燃料的价格
- 减少能源使用 →增加收入+经济增长 →增加能源。
- 效率使人们变得更好，即使存在杰文悖论（节省1焦耳的能源，导致 >消耗1焦耳的能源）。

想法：用机器学习来研究材料科学？ 1）制造混凝土的新方法？ 2）固定氮？ 3）更好的电池？ 4）中子抗性材料?

→其他想法：用机器学习检测甲烷泄漏吗？ 通过远程存在来减少交通（尤其是一个非常高的门槛）？优化货运?

机器学习可以做的其他事情：

1. 洪水预测：

- 洪水现在很严重：每年98亿美元的损失，每年影响250人
- 在更高的温度下会变得更糟!
- 目标：使用公共数据预测洪水影响。
→问题：数据通常太嘈杂，洪水通常发生在平坦地区。
- 一种解决方案：使用机器学习推导出地表图，可以更准确地进行预测。

2. 月球计划：在本世纪下半叶将二氧化碳排放出大气层。

- 利用碳捕获和封存的生物能源（已在IPCC路径中假设）。
- 通过植物增加土壤中的碳（土壤中的碳含量 >大气层中的 3×碳）
- 自由空气捕集（David Keith声称每吨成本为100美元）。
- 机器学习/人工智能能帮助其中任何一个吗?

总结：

1. 气候和能源是一个巨大的问题

2. 主要观点：多个时间尺度 1）机器学习/人工智能可以帮助长期技术（2040年后）2）但是，现在需要立即和彻底的碳化。

3. 温室气体有许多来源/汇。 但是，没有单一的万能方法，必须同时在许多方面努力。

4. 没有纯技术解决方案。

行动呼吁：

1. 找到一个机器学习应用（从最近的论文中获得灵感）
2. 与领域专家合作，3）开始工作并不停止！
3. 不要满足于节省1万吨二氧化碳。→这是一个巨大的问题：目标是百万吨！
   →此外：这是一个非常困难的问题。大多数方法都不起作用。

.....

6.1.2 杰克·凯利：为什么减缓气候变化很困难，如何做得更好

目标：帮助我们尽快减少排放。

核心问题：为什么减少碳排放很困难？

图示：工业在左侧，中间是一座“痛苦之山”，研究人员和初创公司在右侧。 这座山是什么，我们如何减轻它？

杰克最近创办了Open Climate Fix，这是一个非营利组织，基于两个直觉：

1. 开放科学，共享数据，帮助研究人员和公司在应对气候变化方面取得成功,
2. 开发更好的预测工具，以帮助预测可用太阳能的数量。
   →在全球开发一个太阳能板的开放平台。 通过更好的预测，减少旋转备用电力。

挑战：

1. 激励措施:世界上存在错误的激励措施-公司优化以最大化利润，科学家优化h指数，这两者都与气候挑战不太吻合。
2. 数据:这里有很多挑战:
   (a) 1) 组织不愿意共享（公司认为数据具有商业价值，数据是新的石油等）,(b) 大小
   (c) 未被诊断的系统
   (d) 质量: 经常非常嘈杂，缺少标签等。
   (e) 访问: 有时以不同/不寻常的物理形式存在，没有标准化。

(f) 没有完美的模拟器。

但是，一些解决方案: 1) 为数据工程预算充足的时间！2) 提早提问，3) 找到可以帮助的人，4) 保持灵活性，5) 使用简单的模型。

3. 知识产权：大部分事情都是闭门造车的。

→但是，有一些解决方案：1）开源一切，2）与商业化团队（大学/实验室）交流。

4. 行业不太可能阅读你的论文：你需要与行业人员坐下来，解决他们的问题。. 帮助行业使用你的算法！建立一个生产服务，提供咨询等。

5. 展示性能：没有标准数据集、度量标准或基准测试。

要点：

1. 尽可能多地与你所针对的行业人员交流。
2. 精简，构建一个最小可行产品，快速测试你的想法。 失败，并快速失败，然后继续下一个事情。

听众问题：你能提到为什么花时间与行业人员交流是如此有用吗？有没有具体的例子？

杰克A：当然！很好的观点。 考虑与云预测合作。 我们需要做的一件事是估计云的三维结构。 这很困难，但通过与气象学家的交流，我们可以从降雨和数值天气预报中估计体积信息。 与行业人员交流非常有价值，无法想象在线上做这件事。

听众问题：你认为碳税能更好地传达并付诸实施吗？

杰克：好问题！也许我们需要改变术语（人们不喜欢“税”这个词）。
找到更好的沟通方式非常重要。

.....

6.1.3 安德鲁・吴：通过合作应对人工智能在气候变化中的作用背景：我们已经看到气候变化带来了巨大的影响（参见：更多的火灾、洪水和其他灾害）。

→我们的许多孩子将在2100年还活着。 我们今天所做的决策将直接塑造他们生活的世界。

有很多事情要做：1）减缓（能源、工业、建筑、森林），2）适应（灾害预防、社会/生态），3）改变（封存），4）气候科学。

项目一：甲烷预测（第二重要的温室气体）。

- 甲烷的来源一半来自人为，一半来自自然。
- 甲烷的最大来源是湿地。
- 与二氧化碳不同：全球范围内探测甲烷的传感器数量有限。
  →安德鲁和同事参观了一个甲烷传感器：每隔几周，研究人员之一必须去收集数据。

- Ng等现在可以访问全球37个甲烷传感器。
  想法：利用这些数据来预测世界其他地方的甲烷排放！利用这些传感器的数据，现在可以预测全球范围内的甲烷。

项目二：风力涡轮机检测

- 世界正在向可再生能源转变
- 但是，存在一个问题：随着太阳能和风能发电场的建设，我们通常不知道它们的位置。

  →为了促进可再生能源的增长和整合，我们需要找到可再生能源的位置。
- 可以通过以下方式改进美国地质调查局的可再生能源位置数据库：1）自动整理，2）频繁更新。
- 为了找到风力涡轮机：
  - 数据包括10万张风力涡轮机位置的图像。
  - 在180万张图像上进行检测
  - 基准模型：DenseNet-121
  - 弱监督定位：GradCAM。
- 估计风能：通过这些模型，可以很好地估计风力发电机的位置
  和美国的风力估计（参见：http://hint.fm/wind/）。
  →结合他们的工作DeepWind，可以预测美国风力发电机的能量输出。

  其他项目：

- 用于建模大气对流的机器学习[61]
- 时空全球气候模型跟踪
- 减缓：优化可再生能源。 DeepMind提前36小时预测风力发电量。

- GasNet：使用深度学习从实时视频中自动检测天然气电厂的甲烷泄漏[89]。

- 经济影响：更好地了解气候变化对不同国家的经济影响[21]。

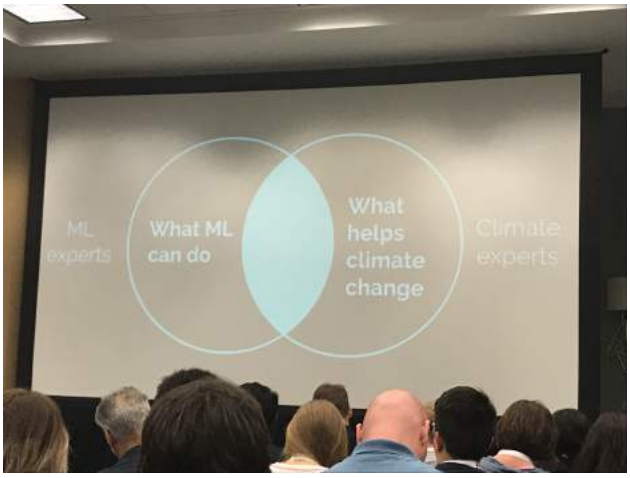


图8：呼吁机器学习和气候科学之间的合作。

回应杰克的观点：我们需要与气候科学家、生态学家、气象学家等合作！请参见图8。

总结思考：在技术颠覆的时代，合作很重要。 一起工作！
共享数据！这个问题比我们自己更大。

.....

6.2 研讨会：现实生活中的强化学习

我刚好赶上了小组讨论。

6.2.1 小组讨论

小组成员包括Craig Boutilier (CB)，Emma Brunskill (EB)，Chelsea Finn (CF)，MohammadGhavamzadeh (MG)，John Langford (JL)，David Silver (DS)，和Peter Stone (PS)。 我将用他们的首字母缩写。

由Alborz Geramifard (AG)主持。

问：哪些应用领域对强化学习最有前景？为什么？

JL：个性化是基本线，包括新闻报道、建议、布局等等。

CB：对我来说，我对能够代表用户行动的代理人感兴趣，但有很多问题（多智能体、偏好引导等）。 推荐系统有前景但也具有挑战性。 短期内我们看不到胜利，但长期而言，推荐系统在长时间内应该作为一个非常有价值的测试平台和压力测试强化学习算法的方式。

附言：我们需要一个有前途的定义来回答这个问题。 对于机器人和操作的强化学习的潜力感到兴奋，就像我的同事们一样，通过提示/影响人们的行为以有益的方式进行用户参与的一般类别（虽然也有隐秘的方式）。 帮助人们康复或服药；强化学习非常适合。我们必须考虑现实世界中哪些环境是安全/适合的。

CF：强化学习的最佳适应并不是社区关注的问题。例如，MuJoCo在一段时间内发挥了作用，但现在我们需要扩展到部分可观察性、迁移、连续状态/动作。 机器人学非常适合。

JL：超越目前部署的基本应用，我认为系统优化是强化学习应用的自然领域。 这些系统大部分都有很好的时间记录，因此有很多数据（不必担心不确定性）。

EB：个性化是一个原始的术语，世界上很多事情正在发生变化。 之前的应用（当我还在读研究生时）是机器人和游戏。 鉴于当前的世界，个性化现在有着巨大的责任和令人兴奋的机会。

DS：关于约翰在他的演讲中提到的一点——如果我们打好基础，将会有应用无处不在。 如果我们找到一个有效的算法，我们应该期望在各种场合看到强化学习的应用。 谁知道还有什么方法/方法论存在？ 让我们继续推动基础研究。 当然，我们应该继续努力并尝试部署它，但我认为我们也可以期待一些意外的发现。展望未来，一个开放的领域是个性化医疗保健，这里有巨大的潜力可以实现强化学习。 可以改变许多人的生活。

MG：运筹学应用无处不在——想想系统优化，对强化学习来说是巨大的机会。亚马逊的存储设施正在发生什么。 还有一个关于推荐系统的事情。 推荐系统无处不在，我们有大量的数据。 我们行动的后果是明确的，需要考虑安全性和鲁棒性。 尽管推荐系统的行动后果不会灾难性，但我们仍然重视正确性。

.....

AG Q: 这是一个很好的过渡到下一个问题。 我们应该如何在这些问题中进行探索？

数据科学：每个框架的一部分，甚至在模拟中也是如此。 我们总是需要处理探索（代理死亡，我们不会获得更多的奖励/数据）。 如果它有机会回来并学到更多，那么我们仍然可以从缺乏经验中学到更多。 在推荐系统中也很好，因为我们获得了“并行生命周期”，如果一段经验流停止了，我们仍然可以从中学习。这是有人停止使用系统的信号。

是的，除了在某些轨迹中死亡的人。 也许在推荐系统中不是这样，但在广泛的多用户系统中是如此。 在多用户系统中，您可以分配这些决策/探索。
可以以分布式方式从个别用户中学习，以确保安全性（没有个别行动会太糟糕）。在临床环境中，在推荐系统中，可以分布式地进行探索以避免这种情况

对任何个体的损害。

EB: 我认为这是一个非常有趣的问题，在许多情况下都是如此——这是使用策略梯度、模仿学习或具有单调改进方法的原因。一些研究表明，我们也可以是最优的（根据我们目前所拥有的信息贪婪地做得很好）。这些方法帮助我们避免了探索的主要负担。有很多工具和方向可以取得进展。

附言：对我来说，这是一个非常有趣的元学习问题。 当我们拥有一个庞大的用户群体时，我们应该选择一个小的子集并进行探索，还是应该在所有用户之间进行一些探索（或其他策略）？ 元目标：让人们对你的工作感到满意，那么我们应该如何实现这一点呢？

CB：快速回应。 如果我们有一个探索成本模型，所有这些都会得到解决。

JL：这是一个理论非常重要的地方。 如果我们只为效率进行优化，你会远离 ε-贪婪算法，如果我们想考虑安全性，我们可以使用界限来安全地进行决策。

MG：使探索更加保守的方法。 我们可以有多保守呢？ 嗯，这取决于个别任务/产品/环境。 如果我们了解用户/领域的信息，我们可以适当地设置探索。 另一个重要的问题是：如何填补模拟和现实世界之间的差距？非常困难。 但是拥有这些模拟器将帮助我们在现实世界中进行探索。

DS: 即使在现实世界中，有时用户离开，我们仍然可以从这些环境中继续学习。 如果我们做得对，一切都在框架中：我们应该如何最大化奖励？ 有时我们会采取行动来保持人们的参与度，有时不会，这是可以接受的。

EB: 我们一直在思考如何确保探索的安全性。 想法：策略证书。
可以明确承诺特定方法将如何探索/行为。 我们可以拥有工具，可以让我们在何时进行探索和何时不进行探索时进行暴露。

.....

AG-Q: 将RL引入现实世界的关键一般原则是什么？

JL: 第一个原则是问题的框架比任何其他东西都更重要。 第二个原则是：如果我给你一个比模拟世界更重要的事情的长长的清单。

CF: 安全性的一个重要原则是识别可以安全探索的情况。 也许你先买一个小机器人，然后将它学到的东西转移过来（就像孩子学习时摔倒是可以接受的）。 识别可以安全探索的情况，然后可以在新情况下进行探索。

附注：最大的教训是——模拟器永远不完美！它们非常吸引人，因为它们安全且我们可以获得大量数据。 我们应该使用模拟器，但我们永远无法使它们完美。 努力使模拟器完美是错误的，我们应该开发新的方法。

CF：完全同意！如果你能制作一个完美的模拟器，那么你已经解决了问题。 在现实世界中需要进行学习。

CB：我想提出另一个原则——奖励函数和目标函数的设计。 你指定的第一个奖励函数永远不是正确的！我们离开并说“这是最好的策略”，但显然它不基于行为，我们的奖励函数是错误的。 人在环路中的方法对于纠正/设计正确的目标/奖励函数非常重要。

EB：我想重申一下——我的实验室在教育和医疗保健方面做了很多工作。 将这项工作（和这些奖励函数的选择）与领域专家结合起来非常重要。 RL在教育领域的第一个已知应用是在1999年用于了解学生在问题上表现良好的情况。
例子：给学生提供问题，直到学生答对为止，然后一遍又一遍地给学生同样的问题（利用！）。

DS：在我的工作中，最重要的原则是可扩展性。 我们应该只追求能够随着资源的增加而扩展的方法（三个方面：1）计算能力，2）经验，3）内存）。如果我们能够实现这种可扩展性，我们只需要增加资源。如果我们设计的方法本身有一个上限（停止探索，停止学习，或其他限制），那个上限会对我们造成伤害。目标应该始终是可扩展性。如果你看一下深度学习的成功，我认为它之所以成功是因为它与这三个资源（计算能力，经验，内存）一起扩展。它在这些资源的支持下越来越好。所以我们应该寻找能够扩展的强化学习方法。

MG：第一个原则是确保强化学习是解决你问题的正确方法。 第二，确保我们有可用于强化学习的数据。想想一下监督学习，如果你的一半数据是错误的，你仍然可以做得很好。 但是对于强化学习来说，我们做不好。 重申克雷格所说的：奖励函数非常重要。 奖励函数的一部分可能是公司的利润（而不是顾客的幸福感）。

JL: 我想提到我同意David的观点。教育人们了解强化学习的现实过程总是很有趣的。

PS: Cogitai平台的设计原则是一个两部分的假设：1）强化学习已经准备好应对现实世界的问题（也就是说，我们在这个房间里已经准备好解决现实世界的问题），2）现实世界已经准备好接受强化学习。这是John提到的第二部分。我们知道强化学习已经准备好解决世界上很多问题，但是与RL/ML以外的人合作，帮助他们识别哪些问题是强化学习问题，并帮助他们使用强化学习工具来解决问题，这是一个很大的挑战。第一个方面很容易（我们找到强化学习问题），而第二个方面很困难（说服他人）。David提到的重要观点是：我们将会感到惊讶！制造智能手机的不是那些建造应用程序的人。我们构建技术和工具，使得这个房间外的人能够使用这些想法。这就是目标。使其足够稳定，以至于我们可以感到惊讶。

CB: RL的广泛接受最终是一个重要目标。几年前我们可能在讨论工业中的机器学习时也是这样。不久之前，让机器学习在大规模生产系统中被接受还是很困难的，但现在我们已经实现了。

.....

观众提问：（由PS改述）我们能否将监督学习和强化学习的方法应用于实际问题进行比较?

PS：嗯，我认为这个测试并不合适。 强化学习问题和监督学习问题是不同的。

JL：考虑推荐系统。 有大量的行动可供选择。 将强化学习应用于大量事物通常不起作用。 如果你有一个不受相同样本复杂性问题困扰的监督信号源。 可以在其上应用强化学习以获得个性化，而这是无法从监督学习中获得的。强化学习没有表示限制，可以适用于监督学习的表示方法也适用于强化学习。我们已经在这两者之间进行了转化，并且取得了成功。

CB: 你可以将强化学习问题视为推荐系统中的监督学习问题。 进行预测，存储结果。不同之处在于我们是短视还是着眼于长远。
问题是，如果你学习用户在长期和短期内的行为表示，是否会有一些好处？我不知道。

DS: 关于A-B测试的评论。 你提到它好像它是强化学习之外的，但你可以将其视为一种糟糕的强化学习算法。 做一件事并坚持下去，做另一件事，然后比较试验后它们的表现。这是强化学习的一部分，但我们也可以像上下文臂带一样做得更好。

.....

观众提问：关于时间跨度/奖励设计。 未来，算法可能更加利用人类的弱点，很多机器学习算法的行为可能是由于它们的短视性质。
能够提前多个时间步骤进行展望可能会有更好的奖励函数。你有什么想法?
其次，当我们在RL中使用人类参与的问题时，我们实际上可以使用的最长视野是多久。

JL: 目前，通常视野相对较短（几小时、几分钟、几天）。 随着视野的增加，数据变得更难获得。如果我们在RL方面取得成功，没有理由它不能持续一生。

EB: 这取决于您使用的粒度级别：分层可以在这方面有所帮助！这两种观点绝对是可能的。 我们正在开发的技术是双重用途和放大器。 将受到人们的限制，而不是技术本身。

CB: 完全同意John和Emma的观点。我仍然认为我们手头没有技术来进行多尺度规划，但我们会实现的。 关于为用户做正确的事情。

一个重要的事情是：我们确实没有好的方法来理解用户的偏好和效用，以及他们的行为如何揭示这些。 应用RL在长期视野上的真正限制因素是开发能够快速了解用户效用函数的技术。RL也涉及规划，但我们离实现这一点还有很长的路要走。

CF: 关于奖励设计和David关于可扩展性的评论 - 我们应该考虑可扩展的算法，但如果我们处于现实世界中，奖励函数监督不是可扩展的反馈来源。 在这些情况下，重要的是考虑如何构建问题，使其可扩展。 不应该害怕改变问题陈述。 例如，如果你想要一个可扩展的算法，设计一个能够与所涉及的资源一起扩展的环境和算法。 如果你的问题与经典的强化学习陈述不符，不要受到限制。

.....

观众提问：首先，何时可以安全地部署策略?

PS: 诱人但错误的做法：在最优时部署！这肯定取决于领域。 测试应该是经验性的 - 我们能比人类驾驶得更好吗?

EB: 通过观察验证（在强化学习和深度学习中）可以学到很多东西。 对于这些例子，有时人们无法参与验证。 可以利用这些来避免最坏情况的发生（例如：使飞机避免碰撞）。一些安全保证。

JL: 实用的答案是：将问题框架化，使得没有特别糟糕/灾难性的行动。 然后，在此基础上进行强化学习。
PS: 强化学习有一些是艺术的组成部分，也有一些是科学的组成部分。我们一直试图使其100%科学化，但我们需要接受其中的一部分是艺术。

.....

李宏杰在闭幕辞中说：

- 研讨会的目标：汇集来自工业界和学术界的研究人员和实践者，共同解决将强化学习应用于实际场景中的实际和/或理论问题。
- 现实生活中的强化学习：游戏、机器人技术、交通运输、资源分配。
- 挑战：部分可观测性、多智能体系统、随机性、大动作集、模拟器、评估、泛化。
- 通过slack进行研讨会后的讨论（在研讨会网站上：https://sites.google.com/view/RL4RealLife）。

.....

接下来是一个相关研讨会的下午会议。

6.3 研讨会：现实世界的顺序决策

研讨会以Emma Brunskill的邀请演讲开始。

6.3.1 Emma Brunskill 关于数据成本高昂时的高效强化学习

问题：教学生编程的最佳方法是什么？ 这个病人应该如何治疗癌症？ 职业培训会增加未来的收入吗？

→我们希望以基于证据的方式回答这些问题！参见：干预科学。

问题：目前我们进行干预科学的方式存在缺陷。 例子：开办一个集训营。 看看哪种干预措施导致更高的毕业率。 一个想法：动态调整我们将学生分配到条件的方式。⟹使用强化学习！

→可以使用强化学习来理解干预科学。

历史上的测试平台：视频游戏和机器人。 然而，现在有了新的测试平台。

最近的变化：大规模计算、大规模传感器和与人动态互动的能力。
∴出现了应用强化学习的机会！这也有很多负面影响，但也有很多机会。

目标：学会更快地做出好决策（以统计效率的方式），造福更多人。

问：我们如何创建能够帮助学生的人工智能？

答：项目-强化学习辅助可以实现$1.3\times$更快的学习、更长的学习时间和更多的学习。

关键技术：

1. 探索！
2. 反事实/批量离线强化学习（今天的重点）。

   →美丽的数学与影响社会之间的紧张关系。 但是！这个领域可以兼顾两者。

   →对因果推断和机器学习的兴趣日益增长。参见：《为什么的书》（Pearl和Mac*k*enzie [63]著）。

问：我们如何从历史序列数据中学习？

答：这真的很难！许多挑战：数据生成过程本质上不同。也就是说：

$$\text{不同的策略} = \quad \Rightarrow \text{ 不同的行动} = \quad \Rightarrow \text{ 不同的状态}$$

也就是说：决策复合。

→此外：我们甚至不知道行为策略。

经典挑战：批量离线策略评估。我们有一组由某个策略 π D收集的轨迹数据集$_D$，我们想要计算一些新策略 π'的价值。

$\rightarrow$ 一个方面：批量离线策略优化。 只想知道如何根据先前的数据行动。

例如：我们应该何时治疗患者？ 实际上，我们如何制定一个个性化策略来干预（例如，拔掉呼吸机）。

$\rightarrow$ 一种特定的干预策略：绝大多数决策都是“无干预”，然后突然我们不得不干预。 让我们称这样的决策为何时治疗策略。

问题：何时开始HIV治疗？ 何时停止某种治疗？

$\rightarrow$ 这些情况可以通过与相关背景条件化的参数化策略很好地捕捉到。

两种解决方案：重要性加权（也称为“倒数倾向性加权”）。 目标是估计新何时治疗策略的预期值。

$\rightarrow$ 优势：简单一致 $\rightarrow$但是：只使用与数据匹配的轨迹，不够稳健，方差较高。

新思路：优势分解：τ_π是某个过滤器下的停止时间根据策略 π决定何时开始治疗。然后：

$$\mu\text{现在}(S_{1:t}) = \mathbb{E}\left[\text{回报}\right] \tag{20}$$
$$\mu\text{下一步}(S_{1:t}) = \mathbb{E}\left[\text{回报}\right] \tag{21}$$

如果我们可以确定$\Delta_\pi := V^\pi - V^{\pi_0}$，其中 π_0是从不治疗的策略，我们可以确定特定治疗的优势。

$\rightarrow$ 具体来说：想要从数据中估计Δ_π，即使没有 π。

在何时治疗的背景下，得到了优势双重稳健$(\mathbf{ADR})$评估器。 也得到了遗憾界限。更多差异可在图9中可视化。
实验：保持健康度指标大于0，在状态随时间以随机方式演变
（具有布朗运动）。 使用治疗方法来尝试改善结果。

$\rightarrow$ 根据临床观察数据，希望找到选择有效治疗方法的方式。

另一个重要问题：我们如何估计 V^*？ （也许，特别是为了帮助人机协同）

问题：如果存在一个期望值超过阈值 b的线性阈值策略，我们能否通过$< O(d)$个样本来确定？

答案：对于某些上下文分布，是的！这是在我们能够返回任何策略之前。

.....

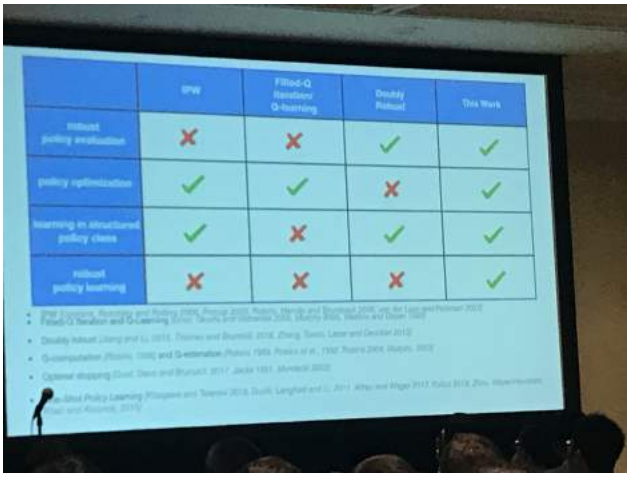


图9：不同离线策略评估类型的比较。

6.3.2 Miro Dudik：通过收缩进行双重稳健的离线策略评估

问题设置：上下文多臂赌博机。 在从交互数据中学习时出现。

> 定义**27**（上下文多臂赌博机）：一种交互协议，用户访问服务，服务向用户呈现对象，用户根据给定的对象做出一些决策。
>
> 也就是说，我们遵循以下步骤：*1*）观察上下文 x，*2*）选择动作 a，*3*）接收奖励 r，然后独立同分布地采样一个新的上下文。

→有很多上下文强化学习的应用。

运行示例：将新闻推荐视为上下文强化学习的典型实例。算法反复感知上下文（用户数据），向用户推荐一些文章，并观察一些数据来指示推荐文章的质量（点击率）。

→具有挑战性，因为奖励模型通常是在先前数据上训练的，$(x_1, a_1, r_1), \ldots\ldots\ldots, (x_n, a_n, r_n)$（通常对应不同的策略）。 此外，策略可能是随机的。
基本任务：离线策略评估。 如果我们遵循其他策略，点击率会是多少?

标准方法：重要性加权估计。 根据以下方式重新加权先前的经验：

$$\frac{\pi_{\text{新}}\ (a_i \mid x_i)}{\pi_{\text{旧}}\ (a_i \mid x_i)}.$$

→无偏，但方差较高。

关键问题：我们能否利用奖励估计器$\hat{r}$来改进这些方法?

两个技巧：

1. 双重稳健估计：将直接估计用作基线，对离基线的偏差应用重要性加权估计。

   →仍然无偏！

2. 权重收缩（见下面的理论依据）。 思路是使用参数 λ来剪裁估计器：

$$\hat{w}(x,\, a) = \begin{cases} w(x,\, a) & txif\ w(x,\, a) < \lambda \\ \lambda & \text{if} w(x,\, a) \geq \lambda \end{cases}$$

   允许 λ明确控制偏差-方差权衡，同时确保权重较小。

> 定义**28**（双重稳健估计）：双重稳健估计如下所示：
>
> $$\hat{V}_{direct} + \frac{1}{n} w(x,\, a)(r - \hat{r}(x,\, a))^2.$$

关键是：双重稳健估计在渐近情况下是最优的，但在有限样本中，其他估计量往往更准确。

→关键是：

均方误差 = 偏差^2+ 方差。

但是：IPS和DR是无偏的，所以误差完全是方差。它们的方差是：

$$IPS : \frac{1}{n}\mathbb{E}[w^2(x,\, a)r^2] + O(1/n) \tag{22}$$

$$DR : \frac{1}{n}\mathbb{E}[w^2(x,\, a)(r - \hat{r}(x,\, a))^2] + O(1/n) \tag{23}$$

$$\tag{24}$$

即使$\hat{r}\,(x, a) = \mathbb{E}[r\ |x, a]$，所以$\hat{V}_{\text{direct}}$是无偏的，但如果奖励本身有噪声，我们会受到较大的权重影响。 因此：权重收缩（见上文）。

问：有很多种方式可以收缩权重，我应该选择哪一种?

答1：悲观收缩！只假设

$$|r - \hat{r}(x,\, a)| \leq 1$$

这会引入偏差 $\mathbb{E}[(\hat{w}(x,\, a) - w(x,\, a))(r - \hat{r}(x,\, a))]$.

新的方差 $\leq \mathbb{E}[\hat{w}(x,\, a)]$.

∴可以提出正确的权重收缩方法。

答2：乐观收缩！假设 $\hat{r}$ 是通过优化加权平方误差来拟合的，其中加权函数为 $z(x, a)$:

$$\frac{1}{n}\sum_{i=1}^{n} z(x_i, a_i)(r_i - \hat{r}(x, a))^2$$

然后我们可以再次限制偏差和方差（使用类似的帕累托优化技巧）。 基于 λ的新估计器：

$$\hat{w}(x, a) = \frac{\lambda}{w^2(x, a) + \lambda} w(x, a).$$

当 $\lambda \to \infty$时，主导系数变为1，恢复DR估计器；当 $\lambda \to 0$时，恢复Dave：错过了。

实验1：我们是否需要乐观和悲观收缩？ 在108个条件下，每个条件更好的频率是多少？ （对于三种不同的奖励预测器）。

→发现：在所有三个奖励预测器下，悲观收缩往往比乐观收缩更好（但也有条件下乐观收缩更好或两者并列）。
实验2：我们是否需要所有这些奖励预测器？

→发现：再次出现了一些情况，每个预测器都优于其他预测器（使用直接估计器时）。
→当使用DR时，几乎总是在不同估计器之间达成平局。 →此外，当使用DR与收缩时，其中一个奖励估计器（$z = w^2$）突然变得占主导地位。. 所以：如果使用收缩，可能需要重新思考如何使用奖励估计器。

总结：

- 为了获得最佳的有限样本性能： →重要考虑奖励预测器和权重收缩 →不同的奖励预测器适用于不同的情境 →模型选择很重要，需要进一步研究

- 下一步：结构化动作，连续动作。

.....

戴夫：就这样结束了！遗憾的是我明天要错过研讨会（以及今天剩下的会议）。 很快就要离开长滩了。

编辑

非常感谢以下个人提供有用的编辑/捕捉到的拼写错误：

- Zeno Zantner 捕捉到的拼写错误。
- Brandon Amos 捕捉到的拼写错误。

参考文献

[1] Alessandro Achille 和 Stefano Soatto. 信息丢失：通过嘈杂计算学习最佳表示。I*EEE*模式分析与机器智能交易，40(12):2897–2905, 2018.

[2] Tameem Adel 和 Adrian Weller. Tibgm：一种可转移和基于信息的图模型方法用于强化学习。在国际机器学习大会上，页码71–81, 2019.

[3] Riad Akrour, Joni Pajarinen, Jan Peters, 和 Gerhard Neumann. 近似策略迭代算法的投影。 在国际机器学习大会上，页码181–190,2019.

[4] Zeyuan Allen-Zhu, Yuanzhi Li, 和 Zhao Song. 通过超参数化的深度学习收敛理论。ar*X*iv预印本 ar*X*iv*:1811.03962*, 2018.

[5] Amiran Ambroladze, Emilio Parrado-Hern´andez, 和 John S Shawe-taylor. 更紧密的PAC-Bayes界限。 在神经信息处理系统进展, 2007年，第9-16页。

[6] Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio,Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio等。深度网络中记忆化的更深入研究。 在第*34*届国际机器学习大会-第*70*卷, 2*0*1*7*年，第2*33*-2*4*2页。JM*L*R. org。

[7] Kavosh Asadi 和 Michael L Littman. 强化学习的另一种softmax运算符。 在第*34*届国际机器学习大会-第7*0*卷, 2*0*17年，第2*43*-252页。JMLR. org。

[8] Andr´e Barreto, Will Dabney, R´emi Munos, Jonathan J Hunt, Tom Schaul, Hado P van Hasselt,and David Silver. 强化学习中的传输后继特征。 在神经信息处理系统进展, 2017年，第4055-4065页。

[9] Mikhail Belkin, Daniel J Hsu, and Partha Mitra. 过拟合还是完美拟合？ 分类和回归规则的风险界限，插值。 在神经信息处理系统进展, 2018年，第2300-2311页。

[10] Marc G Bellemare, Will Dabney, and R´emi Munos. 强化学习的分布视角。ar*X*iv预印本 ar*X*iv*:*1*7*07.0*6887*, 201*7*年。

[11] Olivier Bousquet and Andr´e Elisseeff. 稳定性和泛化。机器学习研究杂志, 2(Mar):499–526, 2002年。

[12] Alon Brutzkus和Amir Globerson。为什么更大的模型更好地泛化？ 通过异或问题的理论视角。 在2019年国际机器学习大会上，页
822-830。

[13] Olivier Catoni。Pac-Bayesian监督分类：统计学习的热力学。ar*X*iv预印本ar*X*iv:*0*7*1*2.*0248*，*2007*年。

[14] Yash Chandak，Georgios Theocharous，James Kostas，Scott Jordan和Philip S Thomas。学习强化学习的动作表示。ar*Xiv*预印本ar*Xiv:1902.00183*，2019年。

[15] Ching-An Cheng，Xinyan Yan，Nathan Ratliff和Byron Boots。预测器-校正器策略优化。a*rXiv*预印本a*rXiv:1810.06509*，2*018*年。

[16] Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. 使用概率动力学模型在少数试验中进行深度强化学习。 在*Advances in NeuralInformation Processing Systems*, 2018年的第4754-4765页。

[17] Karl Cobbe, Oleg Klimov, Chris Hesse, Taehoon Kim, and John Schulman. 量化强化学习中的泛化能力。ar*X*iv预印本ar*Xiv:1812.02341*, *2018*年。

[18] Imre Csisz´ar. 概率分布的I-散度几何和最小化问题。概率论年鉴, 1975年的第146-158页。

[19] Robert Dadashi, Adrien Ali Ta¨ıga, Nicolas Le Roux, Dale Schuurmans, and Marc G Bellemare. 强化学习中的价值函数多面体。ar*X*iv预印本ar*X*iv:1901.11*524*, *2019*年。

[20] Christoph Dann, Lihong Li, Wei Wei, and Emma Brunskill. 策略证书：迈向可追溯的强化学习。ar*Xiv*预印本ar*Xiv:1811.03056*，2*018*年。

[21] Noah S Diffenbaugh和Marshall Burke. 全球变暖导致全球经济不平等加剧。美国国家科学院院刊，116(20):9808–9813，2019年。

[22] Gintare Karolina Dziugaite和Daniel M Roy. 熵-SGD优化PAC-Bayes界的先验：熵-SGD和数据相关先验的泛化性质。 ar*Xiv*预印本ar*Xiv:1712.09376*，2*017*年。

[23] Damien Ernst, Pierre Geurts和Louis Wehenkel. 基于树的批量模式强化学习。机器学习研究杂志，6(Apr):503–556，2005年。

[24] Mahdi M Fard和Joelle Pineau. 强化学习的Pac-Bayesian模型选择。 在*Advances in Neural Information Processing Systems*, 2010年的第1624-1632页。

[25] Mehdi Fatemi，Shikhar Sharma，Harm Van Seijen和Samira Ebrahimi Kahou。强化学习中的死胡同和安全探索。 在国际机器学习大会, 2019年的第1873-1881页。

[26] Scott Fujimoto，David Meger和Doina Precup。无需探索的离线深度强化学习。ar*X*iv预印本arXiv*:1812.02900*, 2*018*年。

[27] Yaroslav Ganin，Evgeniya Ustinova，Hana Ajakan，Pascal Germain，Hugo Larochelle，Fran¸coisLaviolette，Mario Marchand和Victor Lempitsky。神经网络的领域对抗训练。机器学习研究杂志, 17(1):2096-2030, 2016年。

[28] Marta Garnelo, Dan Rosenbaum, Chris J Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo J Rezende, and SM Eslami. 条件神经过程。*arXiv*预印本*arXiv:1807.01613*，2018年。

[29] Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. 正则化马尔可夫决策过程的理论。ar*X*iv预印本ar*X*iv*:1901.11275*，2*019*年。

[30] Carles Gelada, Saurabh Kumar, Jacob Buckman, Ofir Nachum, and Marc G Bellemare. 深度-mdp: 学习连续潜在空间模型进行表示学习。 ar*X*i*v*预印本*arXiv:1906.02736*，2019年。

[31] Badih Ghazi, Rina Panigrahy, and Joshua R Wang. 递归草图用于模块化深度学习。ar*X*i*v*预印本ar*X*i*v:1905.12730*，*2019*年。

[32] Omer Gottesman, Yao Liu, Scott Sussex, Emma Brunskill, and Finale Doshi-Velez. 组合参数化和非参数化模型进行离线策略评估。 ar*X*iv预印本*arXiv:1905.05787*，2019年。

[33] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 软性演员-评论家：具有随机演员的离线最大熵深度强化学习。ar*X*iv预印本*arXiv:1801.01290*，2018年。

[34] Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. 从像素中学习潜在动态规划。 ar*X*iv预印本*arXiv:1811.04551*，2018年。

[35] Josiah Hanna, Scott Niekum, and Peter Stone. 使用估计的行为策略进行重要性采样策略评估。ar*X*iv预印本 ar*X*iv*:1806.01347*, 2*018*年。

[36] Anna Harutyunyan, Peter Vrancx, Philippe Hamel, Ann Nowe, and Doina Precup. 每个决策的选项折扣。 在国际机器学习大会上, 2019年，第2644-2652页。

[37] Elad Hazan, Sham M Kakade, Karan Singh, and Abby Van Soest. 可证明高效的最大熵探索。ar*X*iv预印本 ar*X*iv*:1812.02690*, *2018*年。

[38] Jonathan J Hunt, Andre Barreto, Timothy P Lillicrap, and Nicolas Heess. 使用差异校正组合熵策略。2018年。

[39] Craig Innes and Alex Lascarides. 学习具有无意识的结构化决策问题。 在国际机器学习大会上, 2019年，第2941-2950页。

[40] Alexis Jacq, Matthieu Geist, Ana Paiva, and Olivier Pietquin. 从学习者中学习。 在2019年国际机器学习大会上, 第2990-2999页。

[41] 郑耀江和罗山。神经逻辑强化学习。 ar*X*i*v*预印本*arXiv:1904.10729*, 2019年。

[42] Yuu Jinnai, David Abel, Michael Littman, and George Konidaris. 寻找最小化计划时间的选项。 在2019年国际机器学习大会上, 2019年。

[43] Yuu Jinnai, Jee Won Park, David Abel, and George Konidaris. 通过最小化覆盖时间来发现探索选项。 在2019年国际机器学习大会上, 2019年。

[44] Christos Kaplanis, Murray Shanahan, and Claudia Clopath. 持续强化学习的策略整合。ar*X*i*v*预印本ar*X*i*v:1902.00255*, *2019*年。

[45] Mikhail Khodak, Maria Florina-Balcan, and Ameet Talwalkar. 自适应基于梯度的元学习方法。 ar*X*iv预印本 ar*X*iv*:1906.02717*, *2019*年。

[46] Taesup Kim, Jaesik Yoon, Ousmane Dia, Sungwoong Kim, Yoshua Bengio, and Sungjin Ahn. 贝叶斯模型无关元学习。ar*X*i*v*预印本 ar*X*i*v:180*6.*0383*6, 2*018*年。

[47] Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. 使用校准回归进行深度学习的准确不确定性。ar*X*iv预印本 ar*X*iv*:180*7*.00263*, *2018*年。

[48] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. 通过概率编程归纳实现人类级概念学习。Science, 350(6266):1332–1338, 2015年。

[49] Hoang M Le, Cameron Voloshin, and Yisong Yue. 在约束条件下的批量策略学习。 *arXiv*预印本 *arXiv:1903.08738*, 2019年。

[50] Donghwan Lee 和 Niao He. 基于目标的时序差异学习。 ar*X*i*v*预印本 *arXiv:1904.10945*, 2019年。

[51] Shiau Hong Lim 和 Arnaud Autef. 基于核的强化学习在鲁棒马尔可夫决策过程中的应用。 在国际机器学习大会上，第3973-3981页，2019年。

[52] Michael L Littman. 马尔可夫博弈作为多智能体强化学习的框架。 在*1994*年机器学习会议论文集上，第*1*57-*1*63页。Elsevier，*1994*年。

[53] Hong Liu, Mingsheng Long, Jianmin Wang, 和 Michael Jordan. 可迁移的对抗训练：一种适应深度分类器的通用方法。 在国际机器学习大会上，第4013-4022页，2019年。

[54] Qiang Liu和Dilin Wang。Stein变分梯度下降：一种通用的贝叶斯推断算法。 在*Advances In Neural Information Processing Systems*, 第2378-2386页，2016年。

[55] Ben London和Ted Sandler。贝叶斯反事实风险最小化。 ar*X*i*v*预印本 *arXiv:1806.11500*, 2018年。

[56] Ali Malik，Volodymyr Kuleshov，Jiaming Song，Danny Nemer，Harlan Seymour和StefanoErmon。校准的基于模型的深度强化学习。 在国际机器学习大会, 第4314-4323页，2019年。

[57] Borislav Mavrin，Shangtong Zhang，Hengshuai Yao，Linglong Kong，Kaiwen Wu和Yaoliang Yu。分布式强化学习用于高效探索。 ar*X*i*v*预印本 *arXiv:1905.06125*, 2019年。

[58] David A McAllester. 一些PAC-Bayesian定理。机器学习，37(3):355–363，1999年。

[59] Alberto Maria Metelli, Emanuele Ghelfi和Marcello Restelli. 可配置连续环境中的强化学习。 在国际机器学习大会上，页码4546–4555，2019年。

[60] Gaurav Kumar Nayak, Konda Reddy Mopuri, Vaisakh Shaj, R Venkatesh Babu和Anirban Chakraborty. 深度网络中的零样本知识蒸馏。 ar*X*iv预印本ar*X*iv*:1905.08114*，2*019*年。

[61] Paul A O'Gorman和John G Dwyer. 使用机器学习参数化湿对流：在气候、气候变化和极端事件建模方面的潜力。 地球系统建模进展杂志，10(10):2548–2563，2018年。

[62] Matteo Papini, Alberto Maria Metelli, Lorenzo Lupo, and Marcello Restelli. 通过多重重要性采样的乐观策略优化。 在2019年国际机器学习大会上，页码为4989-4999。

[63] Judea Pearl和Dana Mackenzie。为什么的书：因果关系的新科学。Ba*s*ic B*ook*s，2018年。

[64] Xingchao Peng，Zijun Huang，Ximeng Sun和Kate Saenko。具有解缠表示的领域不可知学习。arXi*v*预印本arXi*v:190*4.*12347*，*2019*年。

[65] Mary Phuong和Christoph Lampert。走向理解知识蒸馏。 在2019年国际机器学习大会上，页码为5142-5151。

[66] John Platt，J Pritchard和Drew Bryant。使用doscoe分析能源技术和政策。2017年。

[67] Chao Qu, Shie Mannor, and Huan Xu. 非线性分布梯度时差学习。arXi*v*预印本arXi*v:1805.0773*2，*2018*年。

[68] Goran Radanovic, Rati Devidze, David Parkes, and Adish Singla. 在马尔可夫决策过程中学习合作。ar*X*iv预印本ar*X*iv*:1901.08029*，*2019*年。

[69] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. 关于神经网络的谱偏差。2018年。

[70] Sachin Ravi and Alex Beatson. 分摊贝叶斯元学习。2018年。

[71] Konda Reddy Mopuri, Phani Krishna Uppala, and R Venkatesh Babu. 提问、获取和攻击：使用类印象生成无数据的UAP。 在欧洲计算机视觉会议（*ECC*V）论文集中，第19-34页，2018年。

[72] Joshua Romoff, Peter Henderson, Ahmed Touati, Yann Ollivier, Emma Brunskill, and Joelle Pineau. 在时间尺度上分离价值函数。ar*X*iv预印本ar*X*iv*:1902.01883*，*2019*年。

[73] Vincent Roulet, Dmitriy Drusvyatskiy, Siddhartha Srinivasa, and Zaid Harchaoui. 迭代线性化控制：稳定算法和复杂性保证。 在国际机器学习大会上，页码5518-5527，2019年。

[74] Mark Rowland, Robert Dadashi, Saurabh Kumar, R´ emi Munos, Marc G Bellemare, and Will Dabney. 分布式强化学习中的统计和样本。 ar*X*i*v*预印本 *arXiv:1902.08102*，2019年。

[75] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. 无监督领域自适应的最大分类器差异。 在I*EEE*计算机视觉和模式识别会议上，页码3723-3732，2018年。

[76] Lior Shani, Yonathan Efroni, and Shie Mannor. 探索意识强化学习再探讨。 在2019年国际机器学习大会上，第5680-5689页。

[77] John Shawe-Taylor和Robert C Williamson. 贝叶斯估计器的*P*AC分析。 在第十届计算学习理论年会上的计算学习理论年会研讨会，第6卷，第2-9页，1997年。

[78] Zebang Shen, Alejandro Ribeiro, Hamed Hassani, Hui Qian和Chao Mi. Hessian辅助策略梯度。在2019年国际机器学习大会上，第5729-5738页。

[79] Zhao Song, Ron Parr和Lawrence Carin. 重新审视Softmax Bellman算子：新的好处和新的视角。 在2019年国际机器学习大会上，第5916-5925页。

[80] Wen Sun, Anirudh Vemula, Byron Boots, and J Andrew Bagnell. 仅通过观察进行的可证明高效的模仿学习。ar*X*iv预印本 ar*X*iv*:1905.10948*, 2019年。

[81] Richard S Sutton, Doina Precup, and Satinder Singh. 在强化学习中介于MDPs和半MDPs之间的时间抽象框架。人工智能, 112(1-2):181–211, 1999年。

[82] Corentin Tallec, L´eonard Blier, and Yann Ollivier. 使深度Q学习方法对时间离散化具有鲁棒性。ar*X*iv预印本 ar*X*iv*:1901.09732*, 2019年。

[83] Chen Tessler, Yonathan Efroni, and Shie Mannor. 动作鲁棒性强化学习及在连续控制中的应用。ar*X*i*v*预印本 ar*X*i*v:1901.09184*, 2019年。

[84] Andrea Tirinzoni, Mattia Salvini, and Marcello Restelli. 通过多重重要性采样在策略搜索中传输样本。 在2019年国际机器学习大会上，页码为6264-6274。

[85] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. 对抗性鉴别领域适应。 在2017年IEEE计算机视觉和模式识别大会上，页码为7167-7176。

[86] Leslie G Valiant. 可学习性理论。 在1984年A*CM*理论计算年会上，页码为436-445。

[87] Benjamin Van Niekerk, Steven James, Adam Earle, and Benjamin Rosman. 强化学习中的价值函数组合。 在2019年国际机器学习大会上，页码为6401-6409。

[88] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra等。匹配网络用于一次性学习。 在神经信息处理系统进展, 第3630-3638页，2016年。

[89] Jingfan Wang, Lyne P Tchapmi, Arvind P Ravikumara, Mike McGuire, Clay S Bell, Daniel Zimmerle, Silvio Savarese和Adam R Brandt。使用红外相机的天然气甲烷排放检测的机器视觉。ar*X*i*v*预印本ar*X*i*v:1904.08500*, 2019年。

[90] Ruohan Wang, Carlo Ciliberto, Pierluigi Amadori和Yiannis Demiris。随机经验-专家蒸馏：通过专家策略支持估计进行模仿学习。 ar*X*iv预印本 *arXiv:1905.06750*, 2019年。

[91] Kelvin Xu, Ellis Ratner, Anca Dragan, Sergey Levine, and Chelsea Finn. 通过元逆强化学习学习先验意图。arXiv预印本 arXiv*:*1*805*.1*2573*, *20*1*8*年。

[92] Lin Yang和Mengdi Wang。使用线性可加特征的参数化Q学习的样本最优解。 在国际机器学习大会上, 2019年，第6995-7004页。

[93] Lantao Yu, Jiaming Song和Stefano Ermon。多智能体对抗逆强化学习。 在国际机器学习大会上, 2019年，第7194-7201页。

[94] Andrea Zanette和Emma Brunskill。在没有领域知识的情况下使用值函数界限在强化学习中获得更紧密的问题相关遗憾界限。 ar*X*iv预印本*arXiv:1901.00210*, 2019年。

[95] Luisa Zintgraf, Kyriacos Shiarli, Vitaly Kurin, Katja Hofmann, 和 Shimon Whiteson. 通过元学习实现快速上下文适应。 在2019年国际机器学习大会上，页码为7693-7702。