arXiv:2302.05404v1 [stat.ML] 10 Feb 2023

极小极大工具变量回归：无需识别或闭性即可获得 L_2 收敛保证

Andrew Bennett * 1, Nathan Kallus[1], Xiaojie Mao[2], WhitneyNewey[3],
VasilisSyrgkanis 4, 以及 Masatoshi Uehara †1

[1]康奈尔大学[2]清华大学[3]麻省理
工学院[4]斯坦福大学

2023年2月13日

摘要

本文研究工具变量（IV）回归的非参数估计。近年来，许多灵活的机器学习方法已被用于工具变量估计。然而，这些方法至少存在以下一项局限：（1）将 IV 回归限制为具有唯一识别解；（2）只能用拟度量（例如投影范数）而非有效度量（例如 L_2 范数）表示估计误差收敛速率；或（3）施加所谓的闭性条件，要求某个条件期望算子足够平滑。本文提出首个能够规避上述全部三项局限、同时仍允许一般函数逼近的方法及其分析。具体而言，我们提出一种新的惩罚极小极大估计量，即使存在多个解，也能收敛到一个固定的工具变量解；并证明在宽松条件下，该估计量具有较强的 L_2 误差收敛速率。值得注意的是，这一保证只需要广泛使用的源条件和可实现性假设，而不需要所谓的闭性条件。我们指出，源条件与闭性条件本质上相互冲突，因此放宽后者显著改进了同时要求这两个条件的现有研究。我们的估计量能够实现这一改进，是因为它基于一种将工具变量估计问题表述为约束优化问题的新颖方法。

1 引言

工具变量（IV）估计在许多应用中都是一个重要问题。例子包括因果推断（Angrist and Imbens, 1995；Newey 和 Powell，2003；Deaner, 2018；Cui 等人，2020）、缺失数据问题（Wang 等人，2014；Miao 等人，2015）、资产定价模型（Chen 等人，2014；Christensen, 2017；Escanciano 等人，2020）、动态离散选择模型（Kalouptsidi 等人，2021）以及强化学习（Liao 等人，2021；Uehara 等人，2021）。

本文重点研究非参数工具变量回归（NPIV）的估计（Newey 和 Powell，2003）。该问题涉及三组变量 X、Y和 Z，它们分别取值于紧欧氏集合 D_X、 D_Y和 D_Z。在原始的工具变量估计问题中， X 表示内生变量， Y表示结果变量， Z表示外生工具变量。我们将 L_2（X） L_2（Z）定义为 L_2 关于 X的函数空间，{v58>

*字母顺序
†通讯作者：mu223@cornell.edu

分别根据它们的分布来定义。我们关注求解关于 $h \in L_2(Z)$的以下方程：

$$\mathbb{E}\left[Y - h(X) \mid Z\right] = 0.$$

该方程也可以改写为 $\mathcal{T}h = r_0$，其中 $r_0(Z) = \mathbb{E}[Y \mid Z]$，且 $\mathcal{T}: L_2(X) \to L_2(Z)$ 是一个有界线性算子，它将每个 $h \in L_2(X)$ 映射为 $\mathbb{E}\left[h(X) \mid Z\right] \in L_2(Z)$。这里，函数 r_0和算子 $\mathcal{T}$均未知。我们只能获取一组独立同分布观测 $\mathcal{D} := \{X_i, Y_i, Z_i\}_{i=1}^n$。

人们对非参数工具变量回归的兴趣日益高涨。许多经典研究提出了筛选估计量或基于核的估计量（例如，Carrasco 等，2007；Horowitz，2011；Newey，2013；Newey 和 Powell，2003；Chen，2007）。然而，非参数工具变量估计众所周知地难以处理，因为它是不适定逆问题。具体而言，非参数工具变量方程 $\mathcal{T}h = r_0$ 的解可能并不唯一；即使解唯一，也可能会随底层数据分布发生不连续变化（Carrasco 等，2007）。因此，现有研究通常假设非参数工具变量解唯一（Andrews，2017；Newey 和 Powell，2003 ）。即使并非如此，它们也会对线性算子 $\mathcal{T}$ 和非参数工具变量解施加限制（ Florens 等，2011；Chen，2021）。一种广泛采用的限制是源条件，它假设工具变量解属于由算子 $\mathcal{T}$定义的子空间（例如，Carrasco 等，2007；Cavalier，2011；Chen 和 Reiss，2011）。在这些条件下，这些经典文献提出的估计量能够提供有力的理论保证。然而，这些传统的非参数估计量无法整合现代、灵活的一般函数逼近方法，例如神经网络或基于树的方法。

为克服这一局限，近期研究提出了多种能够适用于一般函数逼近的算法。这些算法通常采用两个函数类，$\mathcal{H}$ 和 $\mathcal{G}$。具体而言，函数类 $\mathcal{H}$ 是非参数工具变量方程 $\mathcal{T}h = r_0$解的假设类。函数类 $\mathcal{G}$通常称为见证函数类或判别器类，用于检验每个给定函数在多大程度上h 违反非参数工具变量方程。随后，非参数工具变量估计量被定义为极小极大优化问题的解（Lewis and Syrgkanis，2018；Bennett 等人，2019；Dikkala 等人，2020；Liao 等人，2020；Muandet 等人，2020）：

$$\arg\min_{h \in \mathcal{H}} \max_{g \in \mathcal{G}} L(h, g)$$

其中，$L(h, g)$是一个从 $\mathcal{H} \times \mathcal{G}$ 到 $\mathbb{R}$的目标函数。

尽管极具灵活性，这些极小极大方法仍有若干局限。首先，它们通常假设非参数工具变量方程 $\mathcal{T}h = r_0$ 的解是唯一的。然而，如果工具变量不够强，这一假设很容易被违背（Andrews and Stock，2005年；Andrews 等人，2019），而在近端因果推断中，这一假设通常也不成立（Kallus 等人，2021年）。其次，极小极大估计量可能无法提供有力的L_2 误差收敛速率保证，而只能针对较弱的投影均方误差（MSE）给出误差收敛速率保证（Dikkala 等，2020）。然而，即使投影均方误差趋于零，极小极大估计量也可能不收敛到任何固定的工具变量解，因为投影均方误差是伪度量，而非 L_2 度量。第三，现有的极小极大估计量通常需要某种形式的闭性条件，例如对任意 $h \in \mathcal{H}$，$\mathcal{T}h \in \mathcal{G}$（Dikkala 等，2020；Liao 等，2020），或其他类似条件（Bennett 等人，2022）。然而，这一假设可能会对算子 $\mathcal{T}$施加严格限制；需要注意的是，为确保统计复杂度有界，$\mathcal{G}$ 必须是一个受限类。具体而言，闭性假设与广泛使用的源条件相矛盾，因为我们将证明，当 $\mathcal{T}$的谱衰减较慢时，闭性假设更为合理；而当谱衰减较快时，源条件则更为合理。

据我们所知，目前所有将一般函数逼近纳入工具变量问题的方法，都至少存在上述三项局限之一。本文提出了首个能够避开这三项局限的方法。具体而言，我们不假设非参数工具变量解唯一，而是以最小范数解 h_0为目标。这是处理具有非唯一解的逆问题的标准方法

表 1：采用一般函数逼近进行极小极大估计的现有文献汇总。我们的目标是针对未知的 $\mathcal{T}$和 r_0，求解关于 h 的 $\mathcal{T}h = r_0$。我们将其解集记为 $\mathcal{H}_0$，将最小范数解记为 h_0。估计量定义为某些极小极大优化的解：$\min_{h\in\mathcal{H}} \max_{g\in\mathcal{G}} L(h,g)$，其中 $\mathcal{H}$ 和 $\mathcal{G}$ 是假设类。为简单起见，我们重点比较 VC 类 $\mathcal{H}$ 和 $\mathcal{G}$ 的情形（尽管本文及这些论文中的结果都涉及一般函数类），并考虑指数为 1 的源条件。我们令 $\mathcal{T}^*$ 为 $\mathcal{T}$的伴随算子，$\mathcal{G}_0 = \{\bar{g}_0 : \mathcal{T}^*\bar{g}_0 = h_0\}$（在源条件下非空），$h_{0,\ \alpha} = \arg\min_h \|\mathcal{T}h - r_0\|_2^2 + \alpha\|h\|_2^2$，以及 $\alpha_0 > 0$ 为正数。请注意，我们的条件严格弱于 Bennett 等人（2022）的条件。

	主要假设	保证	Rate
Dikkala 等人(2020)	可实现性 $\mathcal{H}_0 \cap \mathcal{H} \neq \emptyset$，闭性 $\mathcal{T}\mathcal{H} \subset \mathcal{G} + r_0$	投影均方误差	$n^{-1/2}$
Liao 等人 (2020)	源 $h_0 \in \mathcal{R}(\mathcal{T}^*\mathcal{T})$，$h_0$的唯一性， 可实现性 $h_{0,\ \alpha} \in \mathcal{H}\ \forall\alpha \leq \alpha_0$，闭性 $\mathcal{T}\mathcal{H} \subset \mathcal{G} + r_0$	L_2 收敛速率	$n^{-1/6}$
Bennett 等人 （2022年）	源 $h_0 \in \mathcal{R}(\mathcal{T}^*\mathcal{T})$， 可实现性 $h_0 \in \mathcal{H}$，$\bar{\mathcal{G}}_0 \cap \mathcal{G} \neq \emptyset$ 和闭性 $\mathcal{T}^*\mathcal{G} \subset \mathcal{H}$	L_2 收敛速率	$n^{-1/4}$
这项工作	来源 $h_0 \in \mathcal{R}(\mathcal{T}^*\mathcal{T})$ 可实现性 $h_0 \in \mathcal{H}$，$\bar{\mathcal{G}}_0 \cap \mathcal{G} \neq \emptyset$	L_2 收敛速率	$n^{-1/4}$

（Florens 等，2011；Babii 和 Florens，2017；Chen，2021年；Bennett 等人，2022年）。我们证明，所提出的估计量可以收敛于最小范数工具变量解，并推导出其 L_2 误差收敛速率保证。这些理论保证仅需要相当标准的源条件和可实现性假设（即，$\mathcal{H}$ 和 $\mathcal{Q}$）。表1 总结了本文及相关研究中的假设和保证。

我们提出的估计量及其理论建立在一个新颖的洞见之上：寻找最小范数解h_0 到 $\mathcal{T}h = r_0$ 可以被视为一个约束优化问题。具体而言，我们证明，最小范数解可以唯一地被识别为拉格朗日函数极小极大优化的鞍点。尽管以往的极小极大估计量也利用了极小极大优化，但其内部最大化被用于近似投影均方误差 $\mathbb{E}[([\mathcal{T}h](Z) - r_0(Z)^2]$,，这需要闭性假设。相比之下，我们方法中的内部最大化源于拉格朗日乘子法，因此不需要闭性假设。有趣的是，我们证明，源条件是平稳拉格朗日乘子存在的充分必要条件，因此也是我们的极小极大优化问题存在鞍点的充分必要条件。这还揭示了广泛用于逆问题中的源条件的新作用。

本文结构安排如下。在第 2 节中，我们介绍工具变量估计的设定及当前研究在此设定下的局限性。在第 3 节中，我们将问题表述为约束优化问题，并据此介绍极小极大估计量。在第 4 节中，我们证明，在无限数据下，极小极大优化能够识别最小范数解。在第5节中，我们给出有限样本误差保证，即 L_2 收敛速率。在第 6 节中，我们将本文的估计量和理论与密切相关的研究进行比较。最后，我们在第 7 节结束本文。

1.1 相关工作

作为逆问题的一个子类，工具变量估计受到了广泛关注，详见Carrasco 等人(2007)；Cavalier (2011)；Newey(2013)；Ito 和 Jin (2014)。

即使算子 $\mathcal{T}$ 和响应变量 r_0 已知，非参数工具变量估计仍因其不适定性而面临重大困难。不适定性通常指存在以下一种或多种情况：（1）解不存在；（2）存在多个解；（3） $\mathcal{T}$ 的逆算子不连续。为应对这些挑战，人们提出了各种正则化技术，例如要求解空间具有紧性（Newey 和 Powell，2003）、Tikhonov 正则化以及 Landweber–Fridman 正则化（Carrasco 等，2007；Cavalier，2011）。在实际应用中，当 $\mathcal{T}$ 和 r_0 未知时，文献提出了多种估计量，包括基于级数的估计量（Ai 和 Chen，2003；Hall 和 Horowitz，2005；Blundell 等，2007；Chen 和 Reiss，2011；Darolles 等，2011；Chen 和 Pouzo，2012；Florens 等，2011；Chen，2021）、基于核的估计量（Hall 和 Horowitz，2005；Horowitz，2007）以及基于再生核希尔伯特空间的估计量（Singh 等，2019；Muandet 等，2020）。

近年来，人们对以统一方式将一般函数逼近技术（如深度神经网络和随机森林）应用于工具变量问题越来越感兴趣（Dikkala 等，2020；Lewis 和 Syrgkanis，2018；Bennett 等，2019；Zhang 等，2020）。在这些方法中，Dikkala 等人（2020）；Liao 等人（2020）；Bennett 等人（2022）给出了有限样本收敛速率保证。具体而言，Liao 等人（2020）在源条件假设下，通过将极小极大优化与 Tikhonov 正则化联系起来，建立了 L_2 收敛性。Bennett 等人（2022）则从不同角度，在源条件下建立了 L_2 收敛保证。值得注意的是，我们所需的假设严格弱于 Bennett 等人（2022）的假设。Dikkala 等人（2020）在不要求源条件的情况下，保证投影均方误差意义下的收敛；然而，当解不唯一时，这一保证不足以确定某个特定解。这些工作（Dikkala 等人，2020；Liao 等人，2020；Bennett 等人，2022）依赖所谓的闭性假设，该假设通过见证函数类对算子 $\mathcal{T}$ 的光滑性施加限制。在离线强化学习领域，这一假设引发了大量讨论，研究者一直在探索放宽该假设的方法（Chen 和 Jiang，2019；Uehara 等人，2020；Foster 等人，2021；Huang 和 Jiang，2022）。本文将在更一般的工具变量设定下考察如何放宽这一假设。这一点十分重要，因为源条件与闭性本质上相互冲突。

值得注意的是，将机器学习融入工具变量估计有多种替代方法（Hartford 等人，2017；Yu 等人，2018；Xu 等人，2020；Liu 等人，2020；Kato 等人，2021；Lu 等人，2021）。然而，据我们所知，这些方法在缺少唯一性假设时，并未提供 L_2 收敛速率保证。

2 问题设定

我们旨在求解关于 h的以下方程：

$$\mathcal{T}h = r_0 \tag{1}$$

其中，$r_0(Z) = \mathbb{E}[Y \mid Z] \in L_2(Z)$ 是一个未知函数，而 $\mathcal{T}\colon L_2(X) \to L_2(Z)$ 是一个未知的条件期望算子，它将任意 $h \in L_2(X)$ 映射为 $\mathbb{E}[h(X) \mid Z]$。注意，由詹森不等式可知，$\mathcal{T}$是一个有界算子，因为其范数上界为 1。此外，我们用 $\mathcal{T}^\star\colon L_2(Z) \to L_2(X)$ 表示 $\mathcal{T}$ 的伴随算子，即对于任意 $h \in L_2(X)$ $g \in L_2(Z)$，都有 $\langle g, \mathcal{T}h\rangle_{L_2(Z)} = \langle \mathcal{T}^\star g, h\rangle_{L_2(X)}$，其中 $\langle\cdot,\cdot\rangle_{L_2(X)}$ 和 $\langle\cdot,\cdot\rangle_{L_2(Z)}$ 分别是 $L_2(X)$ 和 $L_2(Z)$ 上的内积。已知对于任意 $g \in L_2(Z)$，$\mathcal{T}^*$ 可表示为 $[\mathcal{T}^*g](X) = \mathbb{E}[g(Z) \mid X]$（Carrasco 等，2007）。重要的是，这里我们不假设 $\mathcal{T}$ 具有紧性，因为只要 X, Z 包含共同变量，紧性就不成立；许多应用中都是如此

(Deaner，2018；Cui 等人，2020）。此外，我们用 $\mathcal{T}$ 的值域表示为 $\mathcal{R}(\mathcal{T})$，即 $\mathcal{R}(\mathcal{T}) = \{\mathcal{T}h : h \in L_2(X)\}$。

本文中，我们假设方程（1）存在一个解。

假设1（解的存在性）。我们有 $r_0 \in \mathcal{R}(\mathcal{T})$，即 $\mathcal{N}_{r_0}(\mathcal{T}) := \{h \in \mathcal{H} : \mathcal{T}h = r_0\} \neq \emptyset$。

现有文献大多还假设 $\mathcal{T}$ 是单射，且方程（1）的解唯一。然而，即使在这种情况下，方程（1）仍对应一个不适定逆问题，因为逆算子 $\mathcal{T}^{-1}$ 通常无界，因此 NIPV 解可能对数据分布 $\hat{}$ 的微小扰动都非常敏感。若无进一步限制，我们只能得到一个估计量 $\hat{h}$，其投影均方误差 $\mathbb{E}[\{\mathcal{T}\hat{h} - r_0\}^2(Z)] = \mathbb{E}[\{\mathcal{T}(\hat{h} - h)\}^2(Z)]$ 在 $h \in \mathcal{N}_{r_0}(\mathcal{T})$ 下具有收敛保证。然而，投影均方误差只是一种伪度量。因此，即使 $\mathbb{E}[\{\mathcal{T}(\hat{h} - h)\}^2(Z)]$ 趋于零，估计量 $\hat{h}$ 也未必收敛到一个固定点。此外，投影均方误差弱于有效度量，例如 L_2 度量。实际上，根据詹森不等式，我们有 $\mathbb{E}[\{\hat{h} - h)\}^2(X)] \geq \mathbb{E}[\{\mathcal{T}(\hat{h} - h)\}^2(Z)]$. 然而，反向关系通常并不成立。因此，即使 $\mathbb{E}[\{\mathcal{T}(\hat{h} - h)\}^2(Z)]$ 成立，$\mathbb{E}[\{\hat{h} - h)\}^2(X)]$ 也可能不会趋于零。

在许多问题中，L_2 收敛速率保证更可取，甚至是必要的（Hall 和 Horowitz，2005年；Chen 和 Reiss，2011；Kallus 等人，2021年；Uehara 等人，2021年）。为了实现 L_2 收敛，我们需要进一步限制 NPIV 问题的不适定性。一种常见的方法是限制不适定性度量 $\sup_{h \in \mathcal{H}} \frac{\mathbb{E}[\{h - h'\}^2(X)]}{\mathbb{E}[\{\mathcal{T}(h - h')\}^2(Z)]}$ 的大小，其中任意解为 $h' \in \mathcal{N}_{r_0}(\mathcal{T})$，而 $\mathcal{H}$ 是用于获得估计量的函数类（例如，Dikkala 等人，2020；Chen 和 Pouzo，2012）。这使我们能够在方程（1）具有唯一解的前提下，将投影均方误差保证转化为相应的 L_2 误差收敛速率。

然而，在本文中，我们不假设方程（1）存在唯一解，因为在许多实际情形下这一假设可能不成立。具体而言，当工具变量较弱时，唯一性会被破坏（Andrews andStock，2005年；Andrews 等人，2019）。例如，当空间 D_X 和 D_Z 为离散空间，且 D_Z 的基数超过 D_X 的基数时，通常不存在唯一解。此外，近端因果推断中通常也会出现唯一性被破坏的情况，正如 Kallus 等人 (2021年) 在多个例子中所展示的那样。当解不唯一时，方程（1）会变得更加不适定。在这种情况下，现有估计量仍可能具有投影均方误差保证，但要获得 L_2 收敛速率保证则困难得多。具体而言，不适定性度量通常为无穷大，因而无法提供有用信息。大多数现有估计量不一定会在 $\mathcal{N}_{r_0}(\mathcal{T})$ 度量下收敛到 L_2 中的任何特定解。

鉴于在 $\mathcal{N}_{r_0}(\mathcal{T})$ 中可能存在（无穷多个）解，我们拟选取一个具有最小范数的特定解，即

$$h_0 = \underset{h \in \mathcal{N}_{r_0}(\mathcal{T})}{\arg\min}\ 0.5\langle h, h\rangle_{L_2(X)}. \tag{2}$$

该最小范数解定义良好，因为它是将 $L_2(X)$ 中的原点投影到闭仿射空间 $\mathcal{N}_{r_0}(\mathcal{T}) \subseteq L_2(X)$ 上所得的投影。我们在下述引理中对此进行形式化表述。

引理 1. 假设 *1* 成立。则最小范数解 $h_0 \in \mathcal{N}_{r_0}(\mathcal{T})$ 唯一存在，且 $\{h_0\} = \mathcal{R}(\mathcal{T}^\star) \cap \mathcal{N}_{r_0}(\mathcal{T})$，其中 $\mathcal{R}(\mathcal{T}^\star)$ 是值域 $\mathcal{R}(\mathcal{T}^\star)$ 的闭包。

我们注意到，现有的一些文献也针对工具变量方程存在非唯一解时的最小范数解（Florens 等，2011；Santos，2011；Chen，2021年），但这些文献都聚焦于经典的筛选或基于核的估计量。唯一的例外是Bennett 等人（2022年），他们采用一般函数逼近，同时允许存在非唯一解。但正如我们在第6.3节所讨论的，他们的方法要求满足闭性假设，这对算子施加了严格限制 $\mathcal{T}$。本文提出了一种针对最小范数解的新估计量 h_0，并具有强有力的 L_2 收敛保证。重要的是，我们的估计量既支持一般函数逼近，又无需满足闭性假设，从而改进了现有文献。

3 惩罚极小极大工具变量回归

本节中，我们提出了针对方程（ h_0 2）中最小范数解 h_0的估计量。为此，我们首先对解 h_0进行重新表述。注意，

$$h_0 = \arg\min_{h \in L_2(X)} 0.5\langle h, h\rangle_{L_2(X)}, \text{ subject to } \mathcal{T}h = r_0.$$

这是一个定义在希尔伯特空间 $L_2(X)$ 上的约束优化问题。根据拉格朗日乘子法，我们可以考虑另一种极小极大优化：

$$h_0 = \arg\min_{h \in L_2(X)} \sup_{g \in L_2(Z)} L(h, g), \quad L(h, g) := 0.5\langle h, h\rangle_{L_2(X)} + \langle r_0 - \mathcal{T}h, g\rangle_{L_2(Z)}, \tag{3}$$

其中， g 对应一个拉格朗日乘子。

在方程（3）中，目标函数 $L(h, q)$ 是未知的，因为两个内积涉及未知函数 r_0、未知算子 $\mathcal{T}$以及 X 和 Z的未知分布。为了构造基于方程（3）的估计量，我们首先将内积改写为关于 X 和 Z的分布的期望：

$$\langle h, h\rangle_{L_2(X)} = \mathbb{E}\left[h^2(X)\right], \quad \langle r_0 - \mathcal{T}h, g\rangle_{L_2(Z)} = \mathbb{E}\left[(Y - h(X))\, g(Z)\right].$$

接下来，我们可以用经验均值替代未知期望，并将函数 h 和 g 限定在某些函数类 $\mathcal{H} \subset [D_X \to \mathbb{R}]$和 $\mathcal{G} \subset [D_Z \to \mathbb{R}]$中。由此得到以下估计量：

$$\hat{h}_{\text{mn}} \in \arg\min_{h \in \mathcal{H}} \max_{g \in \mathcal{G}} L_n(h, g), \quad L_n(h, g) := 0.5\mathbb{E}_n[h^2(X)] + \mathbb{E}_n\left[(Y - h(X))\, g(Z)\right], \tag{4}$$

其中， $\mathbb{E}_n[\cdot]$ 表示基于样本数据 $\mathcal{D} = \{X_i, Y_i, Z_i\}$的经验平均算子。例如，有 $\mathbb{E}_n[h^2(X)] = \frac{1}{n}\sum_{i=1}^n h^2(X_i)$。值得注意的是，方程（ $\mathbb{E}_n[h^2(X)]$ 中的项（4）可视为惩罚项，因此我们称该估计量为惩罚极小极大估计量。该惩罚项的作用将在定理 1中讨论。

ˆ方程（ h_{mn} ）4中的估计量具有极小极大优化表述。计算视角将在C节讨论。这与近期许多采用一般函数逼近的机器学习工具变量估计量一致（综述见第1.1节）。然而，我们在方程（4）中的极小极大优化是由拉格朗日乘子法启发的，而现有的极小极大估计量则基于截然不同的原理。因此，我们的目标函数 L_n（h， g）不同于现有极小极大估计量所用的目标函数。具体而言，我们的极小极大估计量所需的条件截然不同，详见第6节。

为了论证（4）中的目标函数，我们还需要进一步保证

$$h_0 = \arg\min_{h \in \mathcal{H}} \max_{g \in \mathcal{G}} L(h, g). \tag{5}$$

在第4节中，我们在相当宽松的条件下确立了方程（5）。基于此，我们进一步推导出所提出估计量 L_2ˆ的收敛速率 h_{mn}。

4 最小范数解的识别

在本节中，我们将证明，正如公式（h_0 所示，我们提出的极小极大形式确实能够识别最小范数解5）。我们首先介绍一项对我们的结果至关重要的假设，然后给出在该假设下的识别结果。

4.1 源条件

我们的识别结果在很大程度上取决于以下源条件。

假设 2（源条件）。函数 r_0 满足 $r_0 \in \mathcal{R}(\mathcal{T}\mathcal{T}^\star)$。

假设 2 进一步强化了假设 1，因为它将 r_0 限制在更小的子空间 $\mathcal{R}(\mathcal{T}\mathcal{T}^\star) \subseteq \mathcal{R}(\mathcal{T})$ 中。特别地，我们有 $\mathcal{R}(\mathcal{T}) = \mathcal{T}(\mathcal{R}(\mathcal{T}^\star))$[1] 和 $\overline{\mathcal{R}(\mathcal{T}\mathcal{T}^\star)} = \mathcal{T}(\mathcal{R}(\mathcal{T}^\star))$，因此 $\mathcal{R}(\mathcal{T}\mathcal{T}^\star)$ 通常是 $\mathcal{R}(\mathcal{T})$ 的真子集，除非 $\mathcal{R}(\mathcal{T}^\star)$ 是闭集。众所周知，对于不适定逆问题，算子 $\mathcal{T}^\star$ 通常没有闭值域（Carrasco 等，2007），因此，一般而言，假设 2 对逆问题的不适定性施加了非平凡的限制。在第6.4节中，我们提供了一个更具体的例子来说明这些限制。

源条件是逆问题文献中用于推导强收敛速率保证的常见假设。它们广泛应用于算子已知的逆问题（例如，Engl 等，1996；Ito 和 Jin，2014）以及算子未知的工具变量问题（例如，Florens 等，2011；Carrasco 等，2007；Liao 等人，2021年）。文献中的一个标准源条件是，解 h_0 满足 $h_0 \in \mathcal{R}((\mathcal{T}\mathcal{T}^\star)^{\beta/2})$，其中指数 $\beta > 0$ 为正。通过线性算子的谱理论（Cavalier，2011），可以证明我们的假设 2 中的源条件等价于 $h_0 \in \mathcal{R}((\mathcal{T}\mathcal{T}^\star)^{1/2})$。因此，我们的假设 2 是此类源条件，其源指数为 $\beta = 1$。

假设 2意味着存在 $\bar{g}_0 \in L_2(Z)$ 使得

$$r_0 = \mathcal{T}\mathcal{T}^\star \bar{g}_0. \tag{6}$$

事实上，任何满足 $\bar{g}_0$ 方程（6）的函数都与最小范数解 h_0密切相关 .

引理 2。 如果假设 2 成立，那么 $\bar{g}_0$ 满足 (6) 当且仅当 $\mathcal{T}^\star \bar{g}_0 = h_0$。

特别地，根据引理 2，满足方程（ $\bar{g}_0$ 6）的函数为：

$$\mathcal{N}_{h_0}(\mathcal{T}^\star) := \{g \in L_2(Z) : \mathcal{T}^\star g = h_0\}. \tag{7}$$

在下一小节中，我们将说明假设 2给出的源条件的重要性。特别是，该条件确保我们可以从鞍点中得到 h_0，这些鞍点属于 $L(h, g)$。

4.2 极小极大优化的鞍点

这里，我们刻画 L（h， g）在假设 2下的鞍点，如下所示：

引理 3. 设假设 2 成立，且令 h_0 为方程（2）中的最小范数解， $\mathcal{N}_{h_0}(\mathcal{T}^\star)$ 为方程（7）中给出的函数集合。则 $L(h, g)$ 在 $h \in L_2(X)$ $g \in L_2(Z)$ 上的鞍点集合，即满足以下条件的点 (h', g')：

$$L(h, g') \geq L(h', g') \geq L(h', g), \ \forall h \in L_2(X), \forall g \in L_2(Z),$$

由集合 $\{h_0\} \times \mathcal{N}_{h_0}(\mathcal{T}^\star) = \{(h_0, \bar{g}) : \bar{g} \in \mathcal{N}_{h_0}(\mathcal{T}^\star)\}$ 给出。

众所周知，(h', g') 是鞍点，当且仅当满足“强对偶性”条件 $\inf_{h \in L_2(X)} \sup_{g \in L_2(Z)} L(h, g) = \sup_{g \in L_2(Z)} \inf_{h \in L_2(X)} L(h, g)$，并且

$$h' \in \arg\min_{h \in L_2(X)} \sup_{g \in L_2(Z)} L(h, g), \quad g' \in \arg\max_{g \in L_2(Z)} \inf_{h \in L_2(X)} L(h, g).$$

我们在第 I 节中给出了这一点的形式化证明。根据鞍点的这一等价刻画，我们可以得到以下推论，见引理 3。

[1] 注意到 $L_2(X) = \mathcal{R}(\mathcal{T}^\star) \oplus \overline{\mathcal{R}(\mathcal{T}^\star)^\perp}$，且 $\mathcal{R}(\mathcal{T}^\star)^\perp = \mathcal{N}(\mathcal{T})$，因此 $\mathcal{T}(L_2(X)) = \mathcal{T}(\mathcal{R}(\mathcal{T}^\star)$).

推论1. 如果假设 2成立，则有

$$h_0 = \arg\min_{h\in L_2(X)} \sup_{g\in L_2(Z)} L(h,g), \quad \mathcal{N}_{h_0}(\mathcal{T}^\star) = \arg\max_{g\in L_2(Z)} \inf_{h\in L_2(X)} L(h,g). \tag{8}$$

值得注意的是，即使没有源条件，方程（8）中关于 h_0 的等式仍然成立。此外，即使不存在该源条件，强对偶性 $\inf_{h\in L_2(X)} \sup_{g\in L_2(Z)} L(h,g) = \sup_{g\in L_2(Z)} \inf_{h\in L_2(X)} L(h,g)$ 也成立。然而，源条件对于确立 $\arg\max_{g\in L_2(Z)} \inf_{h\in L_2(X)} L(h,g)$ 的存在性以及（8）中的第二个陈述至关重要。换言之，这表明源条件保证了方程（2）中的问题存在平稳拉格朗日乘子，且平稳拉格朗日乘子的集合为 $\mathcal{N}_{h_0}(\mathcal{T}^\star)$。

到目前为止，我们已经证明，假设 2是鞍点存在的充分条件。有趣的是，它也是鞍点存在的必要条件。

引理 4。 假设假设 *1* 成立 $r_0 \in \mathcal{R}(\mathcal{T})$ 那么，当且仅当假设 $L(h,g)$ 2成立时，存在一个鞍点。

上述引理的证明首先说明，当且仅当存在以下问题的解时，鞍点才存在：最小化变量$_{g\in L_2(Z)}$ $\|\mathcal{T}^* g - h_0\|_2^2$。接着，我们证明该优化问题存在当且仅当源条件（2）成立。我们的 引理 3 和 引理 4 表明，源条件与 h_0的约束优化表述中平稳拉格朗日乘子的存在密切相关。据我们所知，这种关系在文献中尚属首次。

引理 3刻画了不受限制的 $L_2(X)$ 和 $L_2(Z)$ 空间上的鞍点。然而，在实际估计中，我们只能使用统计复杂度有限的某些函数类 $\mathcal{H} \subset L_2(X)$ $\mathcal{G} \subset L_2(Z)$ 。为了使这两个受限函数类能够包含一些鞍点，它们需要满足以下可实现性假设。

假设 3（最小范数解的可实现性）。我们有 $h_0 \in \mathcal{H}$。

假设 4（平稳拉格朗日乘子的可实现性）。我们有 $\mathcal{N}_{h_0}(\mathcal{T}^\star) \cap \mathcal{G} \neq 0$。

上述可实现性假设要求函数类 $\mathcal{H}$ 和 $\mathcal{G}$ 定义良好，即其中至少包含一些真实的鞍点。特别地，假设 4 等价于 $h_0 \in \mathcal{T}^*\mathcal{G}$。在下述定理中，我们进一步将 h_0 在推论 1 中的鞍点刻画扩展到满足这些可实现性条件的受限函数类。

定理 1 （关键识别定理）。 假设 *2*、*3*和*4*成立，则

$$h_0 = \arg\min_{h\in\mathcal{H}} \max_{g\in\mathcal{G}} L(h,g).$$

定理 1 表明，在源条件和可实现性假设下，对函数类 $\mathcal{H}$、$\mathcal{G}$ 上的所提目标进行极小极大优化，可以找回这些函数类中的鞍点。从较高层面看，该定理的证明通过以下方式展开：(1) 原始函数类上的鞍点仍是受限函数类上的鞍点；(2) 受限函数类上任何额外的鞍点都是原始函数类上某个鞍点的最佳响应；(3) h_0 是任意 $\bar{g} \in \mathcal{N}_{h_0}(\mathcal{T}^\star)$ 的唯一最佳响应，这是因为$L(h, g)$ 在 h 中具有由 $\langle h, h\rangle_{L_2(X)}$ 所诱导的 附录 B 强凸性。详情参见该部分。

5 有限样本保证

如第 4 节所述，我们提出的极小极大优化表述在总体分布已知时能够识别目标最小范数解 h_0。本节将进一步证明，式$\hat{}h_{\mathrm{mn}}$ （4）中的有限样本估计量收敛至 h_0，并推导其 L_2 误差收敛速率。

定理 2 （L_2 收敛速率）。假设假设 2、3、4成立。则有

$$\|\hat{h}_{\mathrm{mn}} - h_0\|_2 \leq \sqrt{2 \sup_{h\in\mathcal{H}, g\in\mathcal{G}} \left|(\mathbb{E}_n - \mathbb{E})[(Y - h(X))\, g(Z) + 0.5h(X)^2]\right|}.$$

值得注意的是，定理 2 所需的假设与定理 1中的假设完全相同。特别是，这两个定理都只要求函数类 $\mathcal{H}$ 和 $\mathcal{Q}$满足可实现性条件假设 3、4。可实现性是统计学习理论中的一项基本假设。例如，在最小二乘回归问题中，可实现性是一项标准假设（例如， Wainwright， 2019）。据我们所知，现有的极小极大工具变量回归估计量还需要更强的条件，例如 $\mathcal{T}\mathcal{H} \subset \mathcal{G}$ 或 $\mathcal{T}^*\mathcal{G} \subset \mathcal{H}$。这些条件通常称为闭性条件，并对算子$\mathcal{T}$施加了额外限制。详细讨论见第 6 节。

接下来只需对定理 2 右侧的项进行界定。该项是一个经验过程项，对于任何具有有界统计复杂度的合理函数类 $\mathcal{H}, \mathcal{G}$，都可以利用标准统计学习理论轻松地给出上界。具体而言，我们可以使用标准的对称化论证，以 $\mathcal{H}, \mathcal{G}$的 Rademacher 复杂度来界定定理 2 的右侧。函数类 $\mathcal{H}$ 的 Rademacher 复杂度 $\mathfrak{R}_n(\mathcal{H})$定义为 $\mathfrak{R}_n(\mathcal{H}) = n^{-1}\mathbb{E}[\sup_{h\in\mathcal{H}} \sum_{i=1}^n \sigma_i h(X_i)]$，其中 $\{\sigma_1, \ldots, \sigma_n\}$ 是从 Rademacher 分布中独立抽取的随机变量。函数类 $\mathcal{G}$ 的 Rademacher 复杂度 $\mathfrak{R}_n(\mathcal{G})$也可以类似地定义。

推论 2. 假设 2、3、4 均成立。设 $\|Y\| \leq C_Y$， $\|h\|_\infty \leq C_{\mathcal{H}}$ 对任意 $h \in \mathcal{H}$ 以及 $\|g\|_\infty \leq C_{\mathcal{G}}$ 对$g \in \mathcal{G}$均成立。那么，存在一个通用正常数 c，使得以至少 $1 - \delta$的概率，有

$$\|\hat{h}_{\mathrm{mn}} - h_0\|_2 \leq c\sqrt{(C_{\mathcal{H}} + C_{\mathcal{G}})(\mathfrak{R}_n(\mathcal{G}) + \mathfrak{R}_n(\mathcal{H})) + (C_{\mathcal{G}} + C_{\mathcal{H}})C_{\mathcal{H}}\sqrt{\ln(1/\delta)/n}}$$

此外，对于给定的函数类 $\mathcal{H}$， $\mathcal{Q}$，通过代入 Rademacher 复杂度的现成结果，我们可以得到最终的 L_2 收敛速率。例如，将定理 2应用于有限函数类，可得以下推论。

推论 3. 当 $\mathcal{H}$、$\mathcal{G}$为有限函数类时，以至少 $1-\delta$的概率，有 $\|\hat{h}_{\mathrm{mn}}-h_0\|_2 = \mathrm{Poly}(C_{\mathcal{H}}, C_{\mathcal{G}}) \left(\frac{\ln(|\mathcal{H}||\mathcal{G}|/\delta)}{n}\right)^{1/}$，其中 $\mathrm{Poly}(C_{\mathcal{H}}, C_{\mathcal{G}})$ 是关于 $C_{\mathcal{H}}$ 和 $C_{\mathcal{G}}$ 的多项式项。

作为另一个例子，我们将定理 2 应用于更一般的非参数函数类，其复杂度由覆盖数刻画。

推论 4. 设 $M(\epsilon, \mathcal{H}, \|\cdot\|_\infty)$ 和 $M(\epsilon, \mathcal{G}, \|\cdot\|_\infty)$ 分别是 $\mathcal{H}, \mathcal{G}$ 关于 L^∞-范数的覆盖数。假设对于某个 $\beta > 0$，$\ln(M(\epsilon, \mathcal{H}, \|\cdot\|_\infty)) = O(\epsilon^{-\beta})$ 且 $\ln(M(\epsilon, \mathcal{G}, \|\cdot\|_\infty)) = O(\epsilon^{-\beta})$，并且推论 2 中的条件成立。那么，以至少 $1 - \delta$ 的概率，我们有

$$\|\hat{h}_{\mathrm{mn}} - h_0\|_2 = \begin{cases} \mathrm{Poly}(C_{\mathcal{H}}, C_{\mathcal{G}})\{n^{-1/4} + (\ln(1/\delta)/n)^{1/4}\}, & (\beta < 2) \\ \mathrm{Poly}(C_{\mathcal{H}}, C_{\mathcal{G}})\{n^{-1/4}\ln(n) + (\ln(1/\delta)/n)^{1/4}\}, & (\beta = 2) \\ \mathrm{Poly}(C_{\mathcal{H}}, C_{\mathcal{G}})\{n^{-1/(2\beta)} + (\ln(1/\delta)/n)^{1/4}\} & (\beta > 2). \end{cases}$$

如果将推论 4 应用于 Sobolev 球 $\mathcal{H}$， $\mathcal{G}$ 其中光滑度参数为 α，输入维度为 d，则有 $\beta = d/\alpha$，因此收敛速率变为 $O(n^{-1/4})$，当 $\alpha/d > 2$ 时；而当 $O(n^{-\alpha/(2d)})$，且 $\alpha/d \leq 2$时。这个收敛速率是否最优是一个值得探讨的问题。尽管 Chen 和 Reiss (2011) 推导了非参数工具变量回归估计的极小极大速率，但他们的结果要求 NIPV 方程有唯一解，并且对函数类施加了更强的条件，因此无法与我们的收敛速率直接比较。收敛速率最优性的深入研究留待未来工作。

最后，我们还考虑函数类 $\mathcal{H}$，$\mathcal{G}$ 存在设定错误、因而可能不满足可实现性假设的情形。当我们使用基于依赖样本的函数类 $\mathcal{H}$ 和 $\mathcal{G}$的筛估计量，来逼近某些函数空间时，这一结果很有用。例如，$\mathcal{H}$，$\mathcal{G}$ 可以是具有多项式基函数的线性模型，或维度不断增长的神经网络，它们能够逐步逼近 H¨older 空间或 Sobolev 球（Chen，2007）。

定理 3（错设条件下的有限样本结果）。假设 2 成立，且存在 $h^\dagger \in \mathcal{H}$和 $g^\dagger \in \mathcal{G}$，使得 $\|h^\dagger - h_0\|_2 \leq \epsilon_h$ 且 $\inf_{\bar{g}_0 \in \mathcal{N}_{h_0}(\mathcal{T}^\star)} \|g^\dagger - \bar{g}_0\|_2 \leq \epsilon_{g_0}$。那么

$$\|\hat{h}_{\mathrm{mn}} - h_0\|_2 \leq \sqrt{\{2C_{\mathcal{H}} + C_{\mathcal{G}}\}\epsilon_h + C_{\mathcal{H}}\epsilon_g + 2 \sup_{h\in\mathcal{H}, g\in\mathcal{G}} |(\mathbb{E}_n - \mathbb{E})[(Y - h(X))\, g(Z) + 0.5h(X)^2]|}.$$

与定理 2相比，定理 3 的上界还包含额外的错设误差 ϵ_h，ϵ_g 这是由于 $\mathcal{H}$、$\mathcal{G}$存在错设所致。定理 3 中的经验过程项同样可以用 Rademacher 复杂度进行界定。

6 讨论

本节将我们的方法与现有的极小极大 NPIV 估计量进行比较，重点讨论 Dikkala 等人 (2020）；Liao 等人(2020）；Bennett 等人 (2022年的研究，因为它们与本文最为相关。其他现有的极小极大估计量与之类似，因此我们仅在第 1.1 节中简要回顾。

6.1 与 Dikkala 等人（2020）

Dikkala 等人 （2020）考虑了以下极小极大估计量：

$$\hat{h}_{\mathrm{pro}} = \arg\min_{h\in\mathcal{H}} \max_{g\in\mathcal{G}} \tilde{L}_n(h, g), \quad \tilde{L}_n(h, g) := -0.5\mathbb{E}_n[g^2(Z)] + \mathbb{E}_n[(Y - h(X))\, g(Z)].$$

这里为简单起见，我们省略针对 h和 g的其他可能正则项。

Dikkala 等人 （2020）假设闭性条件 $\mathcal{T}(\mathcal{H} - h_0^\diamond) \subset \mathcal{G}$，其中 $h_0^\diamond$ 可以是 $\mathcal{T}h = r_0$ 的任意解（注意，该条件不受$h_0^\diamond$选择的影响）。在此条件下，令 $L(h,\ g)$ 为 $\tilde{L}_n\ h\ g$ $\max_{g\in\mathcal{G}} \mathbb{E}[\tilde{L}\ h\ g\ ] = 0.5\mathbb{E}[\ \mathcal{T}\ h_0^\diamond - h\ [Z]\ ^2] =$ (,) 的总体对应 $0.5\mathbb{E}[\ [\mathcal{T}h]\ Z\ - r_0\ Z\ \ ^2]$ 2020 量，可以证明max (,) (()) (() ())。换言之，Dikkala 等人 ( ) 中的极小极大目标用于在闭性条件下近似投影均方误差目标。相比之下，我们提出的极小极大目标受拉格朗日乘子法启发，无需满足闭性条件。

为了比较 Dikkala 等人（2020）的理论与我们的理论，为简单起见，我们考虑有限函数类 $\mathcal{H}, \mathcal{G}$。那么，Dikkala 等人（2020）的理论表明，若 $\mathcal{N}_{r_0}(\mathcal{T}) \cap \mathcal{H} \neq 0$ 且 $\mathcal{T}(\mathcal{H} - h_0^\diamond) \subset \mathcal{G}$ 对 $h_0^\diamond \in \mathcal{N}_{r_0}(\mathcal{T})$ 成立，则以概率 $1-\delta$ 有 $\mathbb{E}[\{\mathcal{T}(\hat{h}_{\mathrm{pro}} - h_0^\diamond)\}^2(Z)] = O\left(\left(\frac{\ln(|\mathcal{H}||\mathcal{G}|/\delta)}{n}\right)^{1/2}\right)$。

请注意，上述收敛速率 $O((\ln(|\mathcal{H}||\mathcal{G}|)/n)^{1/2})$ 比推论 3 中我们的收敛速率 $O((\ln(|\mathcal{H}||\mathcal{G}|)/n)^{1/4})$ 更快。然而，上述收敛速率针对的是弱投影均方误差，而推论 3 中的收敛速率针对的是更强的 L_2误差，因此二者无法比较。具体而言，若无进一步限制，投影均方误差收敛速率无法转化为 L_2 收敛速率。Dikkala 等人 （2020）考虑对病态性度量 $\sup_{h\in\mathcal{H}} \frac{\mathbb{E}[\{\ \ h-h\ \}^2\ X\ ]}{\mathbb{E}[\{\mathcal{T}\ h-h\ \}^2\ Z}$()()()施加限制。然而，这种病态性度量通常可能为无穷大；事实上，当 NPIV 问题的解不唯一时，它必定为无穷大，因此，利用它来获得 L_2 收敛速率往往会遇到问题。

备注 1（兼得两全之利）。这里我们注意到，Dikkala 等人（*2020*）中的估计量能够实现较快的投影均方误差收敛速率，而我们的估计量则实现较慢的 L_2 收敛速率。人们或许会想，是否有可能同时实现这两种保证。我们在附录 A中探讨了这一问题，并发现如果不考虑计算方面的考虑，这是可能的。

6.2 与 Liao 等人 （2020）

Liao 等人（2020）基于Dikkala 等人（2020）的方法，在极小极大优化中加入了额外的 Tikhonov 正则化：

$$\min_{h\in\mathcal{H}}\max_{g\in\mathcal{G}} -0.5\mathbb{E}_n[g^2(Z)] + \mathbb{E}_n[(Y-h(X))\,g(Z)] + \alpha\mathbb{E}_n[h^2(X)]. \tag{9}$$

Liao 等人 (2020) 同样需要满足Dikkala 等人 (2020)中的闭性假设，以及一项可实现性假设：Tikhonov 正则化解 $h_{0,\alpha}$ 在 $\mathcal{H}$ 中，且 α足够小时成立。此外，他们假设非参数工具变量解唯一，并满足指数为 $\beta\in$ 的源条件（$0,1]$, ，且正则化强度 α以适当的速率趋于 0，随着 $n\to\infty$。在这些条件下，他们可以推导出一个 L_2 收敛速率。具体而言，当函数类例如为有限类或 VC 类时，他们的 L_2 收敛速率为 $O(n^{-1/6})$，且 $\beta=1$。

我们提出的估计量和理论与 Liao 等人（2020）有显著差异。具体而言，我们的估计量不涉及 $\mathbb{E}_n[g^2(Z)]$ 项，且我们的正则化项 $\mathbb{E}_n[h^2(X)]$ 的系数为常数 0.5，而公式（9）需要一个趋于零的 α。此外，我们的理论允许解不唯一，并采用不同的可实现性假设。值得注意的是，在我们的源条件 $\beta=1$下，我们的收敛速率 $O(n^{-1/4})$快于 Liao 等人（2020）中的收敛速率$O(n^{-1/6})$。

6.3 与 Bennett 等人 (2022年)的比较

在相同的源条件下，Bennett 等人(2022年) [2] 将 L_2 的误差 h_0 表述为投影均方误差： $\mathbb{E}[\{h_0-h\}^2(X)] = \mathbb{E}[(\mathcal{T}^*\{\bar{g}_0-g\})^2(X)]$ 其中 $\mathcal{T}^*g=h$ 且 $\mathcal{T}^*\bar{g}_0=h_0$。首先，注意到对于任意固定的 $\bar{g}_0$ 满足$T^\star\bar{g}_0=h_0$，我们有

$$\mathcal{N}_{h_0}(\mathcal{T}^*) = \arg\min_{g\in\mathcal{G}}\mathbb{E}[(\mathcal{T}^*\{\bar{g}_0-g\})^2(X)] = \arg\min_{g\in\mathcal{G}} 0.5\mathbb{E}[(\mathcal{T}^*g)^2(X)] - \mathbb{E}[Yg(Z)].$$

那么，在闭性假设 $\mathcal{T}^*\mathcal{G}\subset\mathcal{H}$ 下，我们有

$$\mathcal{N}_{h_0}(\mathcal{T}^*) = \arg\max_{g\in\mathcal{G}}\min_{h\in\mathcal{H}} 0.5\mathbb{E}[h^2(X)] + \mathbb{E}[\{Y-h(X)\}g(Z)].$$

那么，注意到对于任意给定的 g，内部最小化解 h 满足 $\mathcal{T}^*g=h$，并回想我们的原始目标是找到满足 $\mathcal{T}^*\bar{g}_0=h_0$ 的 h_0，我们可以推导出 [3]

$$\{\bar{g}_0,h_0\} = \arg\max_{g\in\mathcal{G}}\arg\min_{h\in\mathcal{H}} 0.5\mathbb{E}[h^2(X)] + \mathbb{E}[\{Y-h(X)\}g(Z)].$$

ˆ 最后，他们提出的估计量 h_{fli} 是通过用经验均值替代期望得到的。

与我们提出的估计量相变量和最小化变量互换了。由于 $\mathcal{G}$、$\mathcal{H}$可能是非凸的，这两个估计量通常并不相同。事实上，这导致所需假设方面存在显著差异。在 h_{fli}中，主要假设是源条件、 $\bar{g}_0\in\mathcal{G}$和 $\mathcal{T}^*\mathcal{G}\subset\mathcal{H}$（注意， $h_0\in\mathcal{H}$ 由后两个条件隐含）。相反，在我们提出的估计量 h_{mn}中，主要假设是源条件以及 $\bar{g}_0\in\mathcal{G}$、 $h_0\in\mathcal{H}$。这一条件严格更弱，因为我们不再要求闭性。这一改进意义重大，因为源条件与闭性之间存在固有冲突，下一节将对此进行说明。

[2] 请注意，Bennett 等人（2022年）的主要关注点是以 L_2 误差率估计 Riesz 表示元（在他们的记号中为 $q^\dagger$）。不过，他们的论证很容易调整以适用于我们的情形。[3]在此，令损失函数为 $L^\star(h,g)$，方程（$\bar{g}_0,h_0$） $=$最大化变量$_g$最小化变量$_h$ $L(h,g)$ 表示 $\bar{g}_0=$最小化变量$_g$ $\max_h L(h,g)$ 和 $h_0=$最大化变量$_h$ $L(h,\bar{g}_0)$。

6.4 源条件与闭性之间的张力

在第 6.1 至 6.3 节中，现有估计量均要求某种闭性假设：要么是对任意解 $h_0^\diamond$，$\mathcal{T}(\mathcal{H}-h_0^\diamond)\subset\mathcal{G}$ 满足 $\mathcal{T}h=r_0$，要么是 $\mathcal{T}^*\mathcal{G}\subset\mathcal{H}$。相比之下，我们提出的估计量不需要任何闭性假设。本小节将说明闭性条件与源条件之间存在内在张力。这体现了摆脱源条件的好处。为简单起见，我们考虑一个紧线性算子 $\mathcal{T}$，它具有奇异值分解（SVD）$\{\sigma_i,u_i,v_i\}_{i=1}^\infty$，其中 $\{u_i\}_{i=1}^\infty$、$\{v_i\}_{i=1}^\infty$ 分别是希尔伯特空间 $L_2(Z)$ 和 $L_2(X)$ 中的标准正交基，而 $\sigma_1\geq\sigma_2\geq\cdots$ 是奇异值。由此可得 $\mathcal{T}v_i=\sigma_iu_i$、$\mathcal{T}^*u_i=\sigma_iv_i$，且 $\mathcal{T}\mathcal{T}^\star$ 具有奇异值分解 $\{\sigma_i^2,u_i,u_i\}_{i=1}^\infty$。这里我们仅为获得简单的可数奇异值分解而假设算子是紧的。非紧算子也可以类似处理，但会涉及更繁琐的记号（Cavalier，2011）。

为了理解假设 2 中的源条件，我们将函数 r_0 写作 $r_0=\sum_{i=1}^\infty\gamma_iu_i$，其中$\sum_{i=1}^\infty\gamma_i^2<\infty$。源条件 $r_0\in\mathcal{R}(\mathcal{T}\mathcal{T}^\star)$ 表示存在 $\bar{g}_0=\sum_{i=1}^\infty\beta_iu_i$，使得 $\sum_{i=1}^\infty\beta_i^2<\infty$，并且 $h_0=\mathcal{T}^*\mathcal{T}\bar{g}_0$。由 $\mathcal{T}\mathcal{T}^\star$ 的奇异值分解可知 $\gamma_i=\sigma_i^2\beta_i$。因此，源条件要求$\sum_{i=1}^\infty\gamma_i^2/\sigma_i^4<\infty$。这意味着函数 r_0 相对于 $\mathcal{T}$ 的谱必须足够平滑。显然，当 $\{\sigma_i\}_{i=1}^\infty$ 的衰减速率较慢时，源条件更容易满足；也就是说，算子 $\mathcal{T}$ 和 $\mathcal{T}^\star$ 较不平滑。相反，当 $\{\sigma_i\}_{i=1}^\infty$ 衰减较快且算子 $\mathcal{T}$ 和 $\mathcal{T}^\star$ 更平滑时，闭性条件通常更容易满足。

因此，我们观察到，源条件和闭性对算子 $\mathcal{T}$和 $\mathcal{T}^\star$的光滑性施加了相反的限制。

7 结论

本文研究采用一般函数逼近的非参数工具变量回归。我们提出了一种惩罚极小极大估计量，该估计量基于最小范数工具变量解的一种新颖的约束优化表述。我们证明，该估计量收敛到这一最小范数解，并推导出其 L_2 在源条件以及极小极大估计量所涉及的两个函数类均满足可实现性假设的条件下的收敛速率。值得注意的是，我们的估计量不要求非参数工具变量解具有唯一性，也避免了现有极小极大估计量通常假设的闭性条件。未来还有许多值得探索的研究方向。其中一个方向是将我们的工作推广到更一般的逆问题，包括非线性逆问题（Ito 和 Jin，2014）。另一个方向是将我们的工作推广到工具变量分位数回归（Chernozhukov 等人，2017）。

参考文献

Ai, C. 和 X. Chen（2003）。含有未知函数的条件矩限制模型的有效估计。《计量经济学》71(6)，1795–1843。Andrews, D. 和 J. H. Stock（2005）。弱工具变量下的推断。Andrews, D. W.（2017）。l2 完备分布和有界完备分布的例子。《计量经济学杂志》199(2)，213–220。Andrews, I.、J. H. Stock 和 L. Sun（2019）。工具变量回归中的弱工具变量：理论与实践。《经济学年度评论》11(1)，727–753。Angrist, J. 和 G. Imbens（1995）。局部平均处理效应的识别与估计。Babii, A. 和 J.-P. Florens（2017）。完备性是必要的吗？非识别线性模型中的估计。arXiv 预印本 arXiv:1709.03473。

Bennett, A.、N. Kallus、X. Mao、W. Newey、V. Syrgkanis 和 M. Uehara（2022）。对弱识别函数的强识别泛函进行推断。arXiv 电子预印本，arXiv–2208。Bennett, A.、N. Kallus 和 T. Schnabel（2019）。用于工具变量分析的深度广义矩估计法。《神经信息处理系统进展》32。Blundell, R.、X. Chen 和 D. Kristensen（2007）。形状不变恩格尔曲线的半非参数工具变量估计。《计量经济学》75(6)，1613–1669。Carrasco, M.、J.-P. Florens 和 E. Renault（2007）。结构计量经济学中的线性逆问题：基于谱分解和正则化的估计。《计量经济学手册》6，5633–5751。Cavalier, L.（2011）。统计学中的逆问题。载于《逆问题与高维估计》，第3–96页。施普林格。Chen, J. 和 N. Jiang（2019）。批量强化学习中的信息论考量。载于《国际机器学习会议》，第1042–1051页。PMLR。Chen, Q.（2021）。具有部分识别的部分线性工具变量模型的稳健最优估计。《计量经济学杂志》221(2)，368–380。Chen, X.（2007）。半非参数模型的大样本筛估计。《计量经济学手册》6，5549–5632。Chen, X.、V. Chernozhukov、S. Lee 和 W. K. Newey（2014）。非参数与半参数模型的局部识别。《计量经济学》82(2)，785–809。Chen, X. 和 D. Pouzo（2012）。具有可能非光滑广义残差的非参数条件矩模型估计。《计量经济学》80(1)，277–321。Chen, X. 和 M. Reiss（2011）。计量经济学中病态逆问题的收敛速率最优性。《计量经济理论》27(3)，497–521。Chernozhukov, V.、C. Hansen 和 K. W¨uthrich（2017）。工具变量分位数回归。Chapman and Hall/CRC。Christensen, T. M.（2017）。非参数随机贴现因子分解。《计量经济学》85(5)，1501–1536。Cui, Y.、H. Pu、X. Shi、W. Miao 和 E. T. Tchetgen（2020）。半参数近端因果推断。arXiv 预印本 arXiv:2011.08411。Darolles, S.、Y. Fan、J.-P. Florens 和 E. Renault（2011）。非参数工具变量回归。《计量经济学》79(5)，1541–1565。Daskalakis, C.、A. Ilyas、V. Syrgkanis 和 H. Zeng（2017）。以乐观主义训练生成对抗网络。arXiv 预印本 arXiv:1711.00141。Deaner, B.（2018）。代理控制变量与面板数据。arXiv 预印本 arXiv:1810.00283。Dikkala, N.、G. Lewis、L. Mackey 和 V. Syrgkanis（2020）。条件矩模型的极小极大估计。《神经信息处理系统进展》33，12248–12262。

Engl, H. W., M. Hanke，和 A．Neubauer（1996）．逆问题正则化，第375卷．施普林格科学与商业媒体出版社。Escanciano, J. C., S. Hoderlein, A. Lewbel, O. Linton 和 S. Srisuma（2020）。非参数欧拉方程识别与估计。《计量经济理论》。Florens, J.-P., J. Johannes 和 S. Van Bellegem（2011）。通过惩罚在非参数工具变量回归中进行识别与估计。《计量经济理论》27(3), 472–496。Foster, D. J., A. Krishnamurthy, D. Simchi-Levi 和 Y. Xu（2021）。离线强化学习：价值函数逼近的根本障碍。arXiv 预印本 arXiv:2111.10919。Hall, P. 和 J. L. Horowitz（2005）。存在工具变量时的非参数推断方法。《统计年鉴》33(6), 2904–2929。Hartford, J., G. Lewis, K. Leyton-Brown 和 M. Taddy（2017）。深度工具变量：一种用于反事实预测的灵活方法。载于《国际机器学习会议》，第1414–1423页。PMLR。Horowitz, J. L.（2007）。非参数工具变量估计量的渐近正态性。《国际经济评论》48(4), 1329–1349。Horowitz, J. L.（2011）。应用非参数工具变量估计。《计量经济学》79(2), 347–394。Huang, A. 和 N. Jiang（2022）。超越回报：在用户指定的误差测量分布下进行离策略函数估计。载于 NeurIPS。Ito, K. 和 B. Jin（2014）。逆问题：Tikhonov理论与算法，第22卷。World Scientific。Kallus, N., X. Mao 和 M. Uehara（2021）。利用负对照在未测量混杂下进行因果推断：一种极小极大学习方法。arXiv 预印本 arXiv:2103.14029。Kalouptsidi, M., P. T. Scott 和 E. Souza-Rodrigues（2021）。结构动态离散选择模型的线性工具变量回归估计量。《计量经济学杂志》222(1), 778–804。Kato, M., M. Imaizumi, K. McAlinn, S. Yasui 和 H. Kakehi（2021）。通过重要性加权从条件矩限制中学习因果模型。载于《国际学习表征会议》。Lewis, G. 和 V. Syrgkanis（2018）。对抗性广义矩估计法。arXiv 预印本 arXiv:1803.07164。Liao, L., Y.-L. Chen, Z. Yang, B. Dai, M. Kolar 和 Z. Wang（2020）。结构方程模型的神经估计具有可证明的效率：一种对抗性方法。载于《神经信息处理系统进展》，第33卷，第8947–8958页。Liao, L., Z. Fu, Z. Yang, Y. Wang, M. Kolar 和 Z. Wang（2021）。用于因果离线强化学习的工具变量价值迭代。arXiv 预印本 arXiv:2102.09907。Liu, R., Z. Shang 和 G. Cheng（2020）。关于深度工具变量估计。arXiv 预印本 arXiv:2004.14954。Lu, Y., H. Chen, J. Lu, L. Ying 和 J. Blanchet（2021）。椭圆型偏微分方程的机器学习：快速收敛率泛化界、神经缩放定律与极小极大最优性。arXiv 预印本 arXiv:2110.06897。Miao, W., L. Liu, E. T. Tchetgen 和 Z. Geng（2015）。利用影子变量对不可忽略缺失数据进行识别、双重稳健估计与半参数效率理论。arXiv 预印本 arXiv:1509.02556。

Muandet, K., A. Mehrjou, S. K. Lee 和 A. Raj（2020）。双重工具变量回归。《神经信息处理系统进展》33, 2710–2721。Newey, W. K.（2013）。非参数工具变量估计。《美国经济评论》103(3), 550–56。Newey, W. K. 和 J. L. Powell（2003）。非参数模型的工具变量估计。《计量经济学》71(5), 1565–1578。Santos, A.（2011）。用于恢复连续线性泛函的工具变量方法。《计量经济学杂志》161(2), 129–146。Singh, R., M. Sahani 和 A. Gretton（2019）。核工具变量回归。《神经信息处理系统进展》32。Uehara, M., J. Huang 和 N. Jiang（2020）。用于离策略评估的极小极大权重与q函数学习。载于《国际机器学习会议》，第9659–9668页。PMLR。Uehara, M., M. Imaizumi, N. Jiang, N. Kallus, W. Sun 和 T. Xie （ 2021 ）。极小极大离线强化学习的有限样本分析：完备性、快速收敛率与一阶效率。arXiv 预印本 arXiv:2102.02981。Wainwright, M. J.（2019）。高维统计学：一种非渐近视角，第48卷。剑桥大学出版社。Wang, S., J. Shao 和 J. K. Kim（2014）。一种用于不可忽略无响应下识别与估计的工具变量方法。中华统计学志，1097–1116。Xu, L., Y. Chen, S. Srinivasan, N. de Freitas, A. Doucet 和 A. Gretton（2020）。工具变量回归中的深层特征学习。arXiv 预印本 arXiv:2010.07154。Yu, B. 等（ 2018）。深度Ritz方法：一种用于求解变分问题的基于深度学习的数值算法。《数学与统计学通讯》6(1), 1–12。Zhang, R., M. Imaizumi, B. Sch¨olkopf 和 K. Muandet（2020）。工具变量回归的最大矩限制。arXiv 预印本 arXiv:2010.07684。

A 尽享两全之美

到目前为止，我们遇到了两种保证：较慢的 L_2 收敛速率和较快的投影均方误差。下一步是获得同时具备这两种特性的保证。若暂且不考虑计算效率问题，这实际上是可以实现的。估计量定义如下：

$$\hat{h}_{\text{both}} = \arg\min_{h\in\mathcal{H}_n} \max_{g\in\mathcal{G}} \tilde{L}_n(h,g)$$

其中

$$\mathcal{H}_n = \{h \in \mathcal{H}; \max_{g\in\mathcal{G}} L_n(h,g) - \min_{h\in\mathcal{H}} \max_{g\in\mathcal{G}} L_n(h,g) \le \mu_n\}.$$

这里， μ_n 是一个超参数。集合 $\mathcal{H}_n$ 的定义使得其中每个元素在源条件下都具有 L_2 收敛保证。

定理 4（快速投影均方误差 + 慢速 L_2 误差）。为简单起见，假设 $\mathcal{H},\mathcal{G}$ 有限。假设 $h_0 \in \mathcal{H}, \mathcal{T}(\mathcal{H}-h_0) \subset \mathcal{G}, \mathcal{N}_{h_0}(\mathcal{T}^\star) \cap \mathcal{G} \neq \emptyset$。那么，当我们取 $\mu_n = (C_{\mathcal{H}} + C_{\mathcal{G}})^2 \sqrt{\ln(|\mathcal{H}||\mathcal{G}|/\delta)/n}$ 时，以概率 $1-\delta$，有

$$\|\mathcal{T}(\hat{h}_{\text{both}} - h_0)\|_2 \le c(C_{\mathcal{H}} + C_{\mathcal{G}}) \sqrt{\frac{\ln(|\mathcal{H}||\mathcal{G}|/\delta)}{n}}, \quad \|\hat{h}_{\text{both}} - h_0\|_2 \le c(C_{\mathcal{H}} + C_{\mathcal{G}}) \left(\frac{\ln(|\mathcal{H}||\mathcal{G}|/\delta)}{n}\right)^{1/4}.$$

B 鞍点的一般刻画

首先，注意到

$$\{h_0\} = \min_{h\in L_2(X)} L(h, \bar{g}_0) \tag{10}$$

对任意 $\bar{g}_0 \in \mathcal{N}_{h_0}(\mathcal{T}^*)$ 都成立。换言之，对 $L(h,\bar{g}_0)$ 的最优响应唯一为 h_0。这源于以下两点：(1) 对于 $\mathcal{N}_{h_0}(\mathcal{T}^*)$ 中的任意元素 $\bar{g}_0$， h_0 都是最佳响应，因为根据引理 3，$(h_0, \bar{g}_0)$ 是一个鞍点；(2) 对每个 $\bar{g}_0$，其最佳响应都是唯一的，因为由于 $\langle h, h\rangle_{L_2}(x)$ 项， $L(h,\bar{g}_0)$ 关于 h 严格凸。

接下来，我们引用鞍点的一般刻画。这里，$(\tilde{x},\tilde{y}) \in$最小化变量$_{x\in\mathcal{X}'}$最大化变量$_{y'\in\mathcal{Y}'} f(x,y$) 表示 $\tilde{x} \in$最小化变量$_{x\in\mathcal{X}'}$max$_{y\in\mathcal{Y}'} f(x,y)$，且 $\tilde{y} \in$最大化变量$_{y\in\mathcal{Y}'} f(\tilde{x},y)$。

引理 5（约束集合上的鞍点刻画）。设 $\mathcal{Z}$ 是 $f(x,y)$˜在 $\mathcal{X},\mathcal{Y}$上的鞍点集合。设 $\mathcal{Z}_{\mathcal{X}} =$最大化变量 min$_{x\in\mathcal{X}}max_{y\in\mathcal{Y}} f(x,y)$，$(\cdot,\mathcal{Z}_{\mathcal{X}}) =$最大化变量 max$_{y\in\mathcal{Y}}$最小化变量 min$_{x\in\mathcal{X}} f(x,y)$。那么，对于 $\mathcal{X}' \subset \mathcal{X}, \mathcal{Y}' \subset \mathcal{Y}$，若 $\mathcal{Z} \cap (\mathcal{X}',\mathcal{Y}')$ 非空，则有

$$\mathcal{Z}_{\mathcal{X}} \cap \mathcal{X}' \subset \arg\min_{x\in\mathcal{X}'} \max_{y\in\mathcal{Y}'} f(x,y) \tag{11}$$

and

$$\arg\min_{x\in\mathcal{X}'} \max_{y\in\mathcal{Y}'} f(x,y) \subset \tilde{\mathcal{Z}}_{\mathcal{X}} \cap \mathcal{X}'. \tag{12}$$

在引理 5 中，主要假设 $\mathcal{Z} \cap (\mathcal{X}',\mathcal{Y}') \neq \emptyset$意味着， $\mathcal{X}',\mathcal{Y}'$中包含某个鞍点（相对于 $\mathcal{X},\mathcal{Y}$）。等式 (11) 表明，无约束函数类 $(\mathcal{X},\mathcal{Y})$上的任何鞍点 $(\mathcal{Z}_{\mathcal{X}} \cap \mathcal{X}')$ 仍然是有约束函数类 $(\mathcal{X}',\mathcal{Y}')$上的鞍点。等式 (12) 表明，有约束函数类 $(\mathcal{X}',\mathcal{Y}')$上的任何鞍点˜都包含于 $\mathcal{Z}_{\mathcal{X}}$中。

我们将上述对鞍点的刻画与引理 3 结合起来，令 $(\mathcal{X}',\mathcal{Y}') = (\mathcal{H},\mathcal{G})$，$(\mathcal{X},\mathcal{Y}) = (L_2(X)\ L_2(Z))$ $f(x,y) = L(h,g)$。直接可得，当 $h_0 \in \mathcal{H}, \mathcal{N}_{h_0}(\mathcal{T}^\star) \cap \mathcal{G} \neq \emptyset$（即鞍点包含在 $(\mathcal{H},\mathcal{G})$ 中）时，利用 (11)，有 $\{h_0\} \subset$ arg min$_{h\in\mathcal{H}}$max$_{g\in\mathcal{G}} L(h,g)$。接下来，利用 (12)，有 arg min$_{h\in\mathcal{H}}max_{g\in\mathcal{G}} L(h,g) \subset \{h_0\}$，因为根据 (10)，$(\cdot,h_0) =$arg max$_{g\in\mathcal{G}}$ arg min$_{h\in\mathcal{H}} L(h,g)$。

C 计算视角

为求解方程 (4) 中的优化问题，即使函数类 $\mathcal{H}$ 和 $\mathcal{G}$ 既非凸也非凹（例如神经网络类），我们也可以利用极小极大优化算法的最新进展（Daskalakis 等人，2017）。具体而言，将再生核希尔伯特空间（RKHS）球用作 $\mathcal{G}$ 尤其方便，因为此时方程 (4) 中的内部最大化问题有闭式解。具体来说，当 $\mathcal{G} = \{g : \|g\|_K \leq 1\}$ 对应于正定核 $K : D_Z \times D_Z \to \mathbb{R}$ 及其关联的再生核希尔伯特空间范数 $\|\cdot\|_K$时，方程 (4) 可化简为

$$\arg\min_{h\in\mathcal{H}} 0.5\mathbb{E}_n[h^2(X)] + \left(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(Y_i - h(X_i))\,K(Z_i, Z_j)\,(Y_j - h(X_j))\right)^{1/2}.$$

D 第 2 节中的证明

D.1 引理 1 的证明

这里，我们有 $\mathcal{N}_{r_0}(\mathcal{T}) = h_0 + \mathcal{N}(\mathcal{T})$。$h_0$ 是 $\mathcal{N}_{r_0}(\mathcal{T})$中的最小范数解，它是将 $\mathcal{N}_{r_0}(\mathcal{T})$中的任意元素投影到（闭子空间）$\mathcal{N}(\mathcal{T})^\perp$上的结果。因此，$\{h_0\} = \mathcal{N}(\mathcal{T})^\perp \cap \mathcal{N}_{r_0}(\mathcal{T}) = \mathcal{R}(\mathcal{T}^\star) \cap \mathcal{N}_{r_0}(\mathcal{T})$。这里，我们使用了 $\mathcal{N}(\mathcal{T})^\perp = \mathcal{R}(\mathcal{T}^\star)$。

E 第 4 节中的证明

E.1 引理 2 的证明

这可由引理 1看出。

E.2 引理3 的证明

证明如下。根据第 I 节，点 $(h', g') \in (L_2(X)\ L_2(Z))$ 是一个鞍点，当且仅当强对偶性成立，且 $h' \in$最小化变量$_{h\in L_2(X)} \sup_{g\in L_2(Z)} L(h, g)$ 并且 $g' \in$最大化变量$_{g\in L_2(Z)} \inf_{h\in L_2(X)} L(h, g)$。我们来验证这一条件。

因此，我们首先证明

$$\{h_0\} = \arg\min_{h\in L_2(X)} \sup_{g\in L_2(Z)} L(h, g), \quad 0.5\|h_0\|_2^2 = \min_{h\in L_2(X)} \sup_{g\in L_2(Z)} L(h, g) \tag{13}$$

首先，对于任意 $h \neq \mathcal{N}_{r_0}(\mathcal{T})$，我们有 $\sup_{g\in L_2(Z)} L(h, g) = \infty$。因此，解需要属于 $\mathcal{N}_{r_0}(\mathcal{T})$。由于对于任意 $h \in \mathcal{N}_{r_0}(\mathcal{T})$，$\sup_{g\in L_2(Z)} L(h, g) = 0.5\mathbb{E}[h^2(X)]$，利用引理 2 可知，由 h_0的定义，解为 h_0。

接下来，我们证明

$$\mathcal{N}_{h_0}(\mathcal{T}^\star) = \arg\max_{g\in L_2(Z)} \inf_{h\in L_2(X)} L(h, g), \quad 0.5\|h_0\|_2^2 = \max_{g\in L_2(Z)} \inf_{h\in L_2(X)} L(h, g). \tag{14}$$

我们先求解内层最小化问题。然后，

$$\begin{aligned}
\inf_{h\in L_2(X)} L(h, g) &= \inf_{h\in L_2(X)} 0.5\|h - \mathcal{T}^\star g\|_2^2 + \langle r_0, g\rangle_{L_2(Z)} - 0.5\langle \mathcal{T}^\star g, \mathcal{T}^\star g\rangle_{L_2(X)} \\
&= \langle r_0, g\rangle_{L_2(Z)} - 0.5\langle \mathcal{T}^\star g, \mathcal{T}^\star g\rangle_{L_2(X)} \\
&= \langle \mathcal{T}h_0, g\rangle_{L_2(Z)} - 0.5\langle \mathcal{T}^\star g, \mathcal{T}^\star g\rangle_{L_2(X)} && \text{(Use } r_0 = \mathcal{T}h_0\text{)} \\
&= -0.5\|\mathcal{T}^\star g - h_0\|_2^2 + 0.5\|h_0\|_2^2.
\end{aligned}$$

利用假设2，由于 $\mathcal{N}_{h_0}(\mathcal{T}^*)$ 非空，因此我们有

$$\mathcal{N}_{h_0}(\mathcal{T}^*) = \arg\max_{g\in L_2(Z)} \inf_{h\in L_2(X)} L(h,g).$$

最后，由于根据 (13) 和 (14)成立强对偶性，鞍点集合为 $(h_0, \mathcal{N}_{h_0}(\mathcal{T}^\star))$。

E.3 引理 4 的证明

回顾一下，当且仅当 最小化变量$h\in L_2(X)$ 上确界$g\in L_2(Z)$ $L(h,g)$ 和 最大化变量$g\in L_2(Z)$ 下确界$h\in L_2(X)$ $L(h,g)$存在且强对偶性成立时，鞍点才存在。我们已经证明，假设 2 足以保证鞍点存在。在本证明中，我们将证明假设 2 是保证鞍点存在的必要条件。

为保证鞍点存在，我们需要确保 $\arg\max_{g\in L_2(Z)} \inf_{h\in L_2(X)} L(h,g)$ 存在。该优化问题等价于

$$\arg\min_{g\in L_2(Z)} \|\mathcal{T}^\star g - h_0\|_2^2 \tag{15}$$

正如我们在引理 3的证明中所见。该解存在，当且仅当 $h_0 \in \mathcal{R}(\mathcal{T}^\star) + \mathcal{R}(\mathcal{T}^\star)^\perp$。为证明这一点，我们将投影算子定义为投影到 $\mathcal{R}(\mathcal{T}^\star)$ 上的 $P_{\mathcal{R}(\mathcal{T}^\star)}$。那么，(15) 的解存在，当且仅当 $P_{\mathcal{R}(\mathcal{T}^\star)}h_0 \in \mathcal{R}(\mathcal{T}^\star)$。这里，$P_{\mathcal{R}(\mathcal{T}^\star)}h_0 \in \mathcal{R}(\mathcal{T}^\star)$ 意味着

$$h_0 = P_{\overline{\mathcal{R}(\mathcal{T}^\star)}}h_0 + (I - P_{\overline{\mathcal{R}(\mathcal{T}^\star)}})h_0 \in \mathcal{R}(\mathcal{T}^\star) + \mathcal{R}(\mathcal{T}^\star)^\perp.$$

此外， $h_0 \in \mathcal{R}(\mathcal{T}^\star) + \mathcal{R}(\mathcal{T}^\star)^\perp$ 意味着 $P_{\overline{\mathcal{R}(\mathcal{T}^\star)}h_0} \in \mathcal{R}(\mathcal{T}^\star)$ 回忆 $\mathcal{H} = \mathcal{R}(\mathcal{T}^\star)\overline{\bigoplus}\mathcal{R}(\mathcal{T}^\star)^\perp$。这就完成了对（15）的解存在当且仅当 $h_0 \in \mathcal{R}(\mathcal{T}^\star) + \mathcal{R}(\mathcal{T}^\star)^\perp$。

最后，回忆 $h_0 \in \overline{\mathcal{R}(\mathcal{T}^\star)}$，利用 引理 1。因此， $h_0 \in \mathcal{R}(\mathcal{T}^\star) + \mathcal{R}(\mathcal{T}^\star)^\perp$ 意味着 $h_0 \in \mathcal{R}(\mathcal{T}^\star)$，因为若$h_0 = h_{0,\,2} + h_{0,\,3}$,$h_{0,\,2} \in \mathcal{R}(\mathcal{T}^\star)$$h_{0,\,3} \in \mathcal{R}(\mathcal{T}^\star)^\perp$，则有 $h_{0,\,3} = h_0 - h_{0,\,2} \in \mathcal{R}(\mathcal{T}^\star) \cap \overline{\mathcal{R}(\mathcal{T}^\star)^\perp} = \{0\}$。

根据 $h_0 \in \mathcal{R}(\mathcal{T}^\star)$ 蕴含 $r_0 \in \mathcal{R}(\mathcal{T}\mathcal{T}^\star)$ 这一事实，即可得出该结论。

E.4 引理 5 的证明

显然， $\mathcal{Z}\cap(\mathcal{X},\mathcal{Y})$ 中的每个元素都是关于 $\mathcal{X}'$、$\mathcal{Y}'$ 的鞍点，因为这是关于 $\mathcal{X}$、$\mathcal{Y}$ 的鞍点。因此，

$$\mathcal{Z}_{\mathcal{X}} \cap \mathcal{X}' \subset \arg\min_{x\in\mathcal{X}'} \max_{y\in\mathcal{Y}'} f(x,y).$$

现在，我们证明第二个陈述。设 (x_0,y_0) 是 $\mathcal{Z}\cap(\mathcal{X},\mathcal{Y})$ 中的一个元素（该元素存在且为鞍点）。于是，取:

$$\tilde{x} \in \arg\min_{x\in\mathcal{X}'} \sup_{y\in\mathcal{Y}'} f(x,y), \quad \tilde{y} \in \arg\max_{y\in\mathcal{Y}'} \inf_{x\in\mathcal{X}'} f(x,y).$$

由于 $(\tilde{x},\tilde{y})$ 是关于 $\mathcal{X}'$、$\mathcal{Y}'$ 的鞍点，我们有

$$f(x_0,y_0) \geq f(x_0,\tilde{y}) \geq f(\tilde{x},\tilde{y}) \geq f(\tilde{x},y_0) \geq f(x_0,y_0).$$

于是，上述不等式均为等式。因此，我们有

$$f(x_0,\tilde{y}) = f(\tilde{x},y_0), \quad f(x_0,y_0) = f(x_0,\tilde{y})$$

这意味着

$$\tilde{x} \in \mathcal{Z}'_{\mathcal{X}} \subset \mathcal{X}'.$$

回忆一下 $y_0 \in$最大化变量$y\in\mathcal{Y}$$\min_{x\in\mathcal{X}} f(x,y)$。

E.5 定理 1 的证明

我们给出两种证明。

第一种证明。我们使用引理 5。首先，利用（11），

$$\{h_0\} \subset \arg\min_{h\in\mathcal{H}} \max_{g\in\mathcal{G}} L(h,g).$$

其次，我们利用（12）。此处回顾引理 3的证明，可得

$$(\mathcal{N}_{h_0}(\mathcal{T}^*), h_0)) \in \arg\max_{g\in\mathcal{G}} \arg\min_{h\in\mathcal{H}} L(h,g).$$

因此，利用引理 3，可得

$$\arg\min_{h\in\mathcal{H}} \max_{g\in\mathcal{G}} L(h,g) \subset \{h_0\}.$$

因此，

$$\arg\min_{h\in\mathcal{H}} \max_{g\in\mathcal{G}} L(h,g) = \{h_0\}.$$

第二种证明。我们给出一个更直接的证明，以便稍后展示有限样本结果。

我们取 $\bar{g}_0$ 自 $\mathcal{N}_{h_0}(\mathcal{T}^*) \cap \mathcal{G}$中的某个元素。该元素满足 $\mathcal{T}^*\bar{g}_0 = h_0$。定义

$$L(h,g) := 0.5\mathbb{E}[h^2(X)] + \mathbb{E}[g(Z)\{Y - h(X)\}],$$
$$\hat{g}(h) := \arg\max_{g\in\mathcal{G}} L(h,g), \quad \hat{h} := \arg\min_{h\in\mathcal{H}} \sup_{g\in\mathcal{G}} L(h,g),$$

$\hat{}$和 $\hat{}g := \hat{}g(h)$。因此，对于任意 $h \in \mathcal{H}$,

$$\begin{aligned}
&L(h,\bar{g}_0) - L(h_0,\bar{g}_0) \\
&= 0.5\mathbb{E}[h^2(X)] + \mathbb{E}[\bar{g}_0(Z)\{Y - h(X)\}] - 0.5\mathbb{E}[\{h_0\}^2(X)]0\mathbb{E}[\bar{g}_0(Z)\{Y - h_0(X)\}] \\
&= 0.5\mathbb{E}[h^2(X)] + \mathbb{E}[\bar{g}_0(Z)\{h_0(X) - h(X)\}] - 0.5\mathbb{E}[\{h_0\}^2(X)] \\
&= 0.5\mathbb{E}[h^2(X)] + \mathbb{E}[h_0(X)\{h_0(X) - h(X)\}] - 0.5\mathbb{E}[\{h_0\}^2(X)] \qquad \text{(We use } \mathbb{E}[h_0(X) \mid Z] = \bar{g}_0(Z)) \\
&= 0.5\mathbb{E}[\{h(X) - h_0(X)\}^2].
\end{aligned}$$

因此，对于任意 $h \in \mathcal{H}$,

$$\mathbb{E}[\{h(X) - h_0(X)\}^2] = L(h,\bar{g}_0) - L(h_0,\bar{g}_0). \tag{16}$$

此外，

$$\begin{aligned}
L(\hat{h},\hat{g}) &\geq L(\hat{h},\bar{g}_0) && \text{（估计量的构造）} \\
&\geq L(h_0,\bar{g}_0) && \text{（鞍点性质）} \\
&\geq L(h_0,\hat{g}(h_0)). && \text{（鞍点性质）}
\end{aligned}$$

ˆ 根据定义，我们有 $L(h, \hat{g}) \leq L(h_0, \hat{g}(h_0))$，因此以上所有不等式均为等式。于是，我们有

$$L(\hat{h}, \bar{g}_0) - L(h_0, \bar{g}_0) = 0. \tag{17}$$

综上，将 (16) 与 (17) 结合可得

$$\mathbb{E}[\{\hat{h}(X) - h_0(X)\}^2] \leq L(\hat{h}, \bar{g}_0) - L(h_0, \bar{g}_0) = 0.$$

ˆ因此，$h(X) = h_0(X)$。

F 第5节的证明

F.1 定理 2的证明

我们从 $\mathcal{N}_{h_0}(\mathcal{T}^*) \cap \mathcal{G}$ 中取一个元素 $\bar{g}_0$。它满足 $\mathcal{T}^* \bar{g}_0 = h_0$。定义

$$\begin{aligned}
L(h, g) &:= 0.5\mathbb{E}[h^2(X)] + \mathbb{E}[g(Z)\{Y - h(X)\}], \\
L_n(h, g) &:= 0.5\mathbb{E}_n[h^2(X)] + \mathbb{E}_n[g(Z)\{Y - h(X)\}], \\
\hat{g}(h) &:= \arg\max_{g \in \mathcal{G}} 0.5\mathbb{E}_n[h^2(X)] + \mathbb{E}_n[g(Z)\{Y - h(X)\}], \\
M(\mathcal{H}, \mathcal{G}) &:= \sup_{h \in \mathcal{H}, g \in \mathcal{G}} |(\mathbb{E}_n - E)[\{Y - h(X)\}g(Z) + 0.5h(X)^2]|.
\end{aligned}$$

利用 (16)，回顾一下，对于 $h \in \mathcal{H}$，有

$$\mathbb{E}[\{h(X) - h_0(X)\}^2] = L(h, \bar{g}_0) - L(h_0, \bar{g}_0).$$

这里，我们有

$$\begin{aligned}
L_n(\hat{h}, \hat{g}(\hat{h})) &\geq L_n(\hat{h}, \bar{g}_0) && \text{(Construction of estimators)} \\
&\geq L(\hat{h}, \bar{g}_0) - M(\mathcal{H}, \mathcal{G}) \\
&\geq L(h_0, \bar{g}_0) - M(\mathcal{H}, \mathcal{G}) && \text{(Saddle point property)} \\
&\geq L(h_0, \hat{g}(h_0)) - M(\mathcal{H}, \mathcal{G}) && \text{（鞍点性质）} \\
&\geq L_n(h_0, \hat{g}(h_0)) - 2M(\mathcal{H}, \mathcal{G}) \\
&\geq L_n(\hat{h}, \hat{g}(\hat{h}) - 2M(\mathcal{H}, \mathcal{G}). && \text{(Construction of estimators.)}
\end{aligned}$$

因此，我们有

$$L(\hat{h}, \bar{g}_0) - L(h_0, \bar{g}_0) \leq 2M(\mathcal{H}, \mathcal{G}).$$

最后，我们得到

$$\mathbb{E}[\{\hat{h}(X) - h_0(X)\}^2] \leq L(\hat{h}, \bar{g}_0) - L(h_0, \bar{g}_0) \leq 2M(\mathcal{H}, \mathcal{G}).$$

F.2 推论2

我们计算以下经验过程项:

$$\sup_{h\in\mathcal{H},g\in\mathcal{G}} |(\mathbb{E}_n - \mathbb{E})[\{Y - h(X)\}g(Z) + 0.5h(X)^2]|.$$

因此，根据Wainwright (2019，定理4.10），该式的上界为

$$c\left\{\mathfrak{R}_n(\mathcal{A}_1) + \mathfrak{R}_n(\mathcal{A}_2) + \mathfrak{R}_n(\mathcal{A}_3) + (C_\mathcal{G} + C_\mathcal{H})C_\mathcal{H}\sqrt{\ln(1/\delta)/n}\right\}$$

其中

$$\mathcal{A}_1 = \{yg(z); g \in \mathcal{G}\}, \quad \mathcal{A}_2 = \{h(x)g(z); h \in \mathcal{H}, g \in \mathcal{G}\}, \quad \mathcal{A}_3 = \{0.5h(x)^2; h \in \mathcal{H}\}.$$

首先，我们有

$$\mathfrak{R}_n(\mathcal{A}_1) \lesssim C_\mathcal{G}\mathfrak{R}_n(\mathcal{G}).$$

其次，我们有

$$\mathfrak{R}_n(\mathcal{A}_2) \lesssim (C_\mathcal{H} + C_\mathcal{G})(\mathfrak{R}_n(\mathcal{G}) + \mathfrak{R}_n(\mathcal{H})).$$

这里，我们使用了Kallus 等人 (2021年，推论3的证明）。第三，我们有

$$\mathfrak{R}_n(\mathcal{A}_3) \lesssim 2C_\mathcal{H}\mathfrak{R}_n(\mathcal{H}).$$

综合所有结果，经验过程项可上界为

$$c\left\{(C_\mathcal{H} + C_\mathcal{G})(\mathfrak{R}_n(\mathcal{G}) + \mathfrak{R}_n(\mathcal{H})) + (C_\mathcal{G} + C_\mathcal{H})C_\mathcal{H}\sqrt{\ln(1/\delta)/n}\right\}.$$

F.3 推论的证明4

我们将 Dudley积分定理 5 与推论 2结合起来。

F.4 定理 3 的证明

我们取 $\mathcal{N}_{h_0}(\mathcal{T}^\star) \cap \mathcal{G}$ 中的某个元素 $\bar{g}_0$。该元素满足 $\mathcal{T}^*\bar{g}_0 = h_0$。

$$\begin{aligned}
L(h,g) &:= 0.5\mathbb{E}[h^2(X)] + \mathbb{E}[g(Z)\{Y - h(X)\}],\\
L_n(h,g) &:= 0.5\mathbb{E}_n[h^2(X)] + \mathbb{E}_n[g(Z)\{Y - h(X)\}],\\
\hat{g}(h) &:= \arg\max_{g\in\mathcal{G}} 0.5\mathbb{E}_n[h^2(X)] + \mathbb{E}_n[g(Z)\{Y - h(X)\}],\\
M(\mathcal{H},\mathcal{G}) &:= \sup_{h\in\mathcal{H},g\in\mathcal{G}} |(\mathbb{E}_n - E)[\{Y - h(X)\}g(Z) + 0.5h(X)^2]|.
\end{aligned}$$

ˆ且 ˆ$g = \hat{g}(h)$。回想一下，对于任意 $h \in \mathcal{H}$,

$$\mathbb{E}[\{h(X) - h_0(X)\}^2] \le L(h,\bar{g}_0) - L(h_0,\bar{g}_0).$$

此外，

$$
\begin{aligned}
&L(\hat{h}, \bar{g}_0) - L(h_0, \bar{g}_0) \\
&= \underbrace{-L(h_0, \bar{g}_0) + L(h^\dagger, \hat{g}(h^\dagger))}_{(a)} \underbrace{-L(h^\dagger, \hat{g}(h^\dagger)) + L(\hat{h}, g^\dagger)}_{(c)} \underbrace{-L(\hat{h}, g^\dagger) + L(\hat{h}, \bar{g}_0)}_{(f)}.
\end{aligned}
$$

项 (a) 的上界如下：

$$
\begin{aligned}
L(h^\dagger, \hat{g}(h^\dagger)) - L(h_0, \bar{g}_0) &\le 0.5\mathbb{E}[\{h^\dagger\}^2(X)] + \|\hat{g}(h^\dagger)\|_2 \|h_0 - h^\dagger\|_2 - 0.5\mathbb{E}[h_0^2(X)] \\
&\le 0.5\mathbb{E}[\{h^\dagger\}^2(X)] + \sup_g \|g\|_2 \|h_0 - h^\dagger\|_2 - 0.5\mathbb{E}[h_0^2(X)] \\
&\le 0.5\|h^\dagger + h_0\|_2 \|h^\dagger - h_0\|_2 + \{\sup_g \|g\|_2\} \|h_0 - h^\dagger\|_2 \\
&\le \{2C_{\mathcal{H}} + C_{\mathcal{G}}\} \|h_0 - h^\dagger\|_2.
\end{aligned}
$$

项 (c) 的上界如下：

$$
\begin{aligned}
-L(h^\dagger, \hat{g}(h^\dagger)) + L(\hat{h}, g^\dagger) &\le -L(h^\dagger, \hat{g}(h^\dagger)) + L_n(h^\dagger, \hat{g}(h^\dagger)) - L_n(h^\dagger, \hat{g}(h^\dagger)) + L_n(\hat{h}, g^\dagger) - L_n(\hat{h}, g^\dagger) + L(\hat{h}, g^\dagger) \\
&\le M(\mathcal{H}, \mathcal{G}) + (-L_n(h^\dagger, \hat{g}(h^\dagger)) + L_n(\hat{h}, \hat{g})) + M(\mathcal{H}, \mathcal{G}) \\
&\le 2M(\mathcal{H}, \mathcal{G}).
\end{aligned}
$$

项 (f) 的上界如下：

$$
L(\hat{h}, \bar{g}_0) - L(\hat{h}, g^\dagger) \le \|\hat{h}\|_2 \|g_0 - g^\dagger\|_2 \le C_{\mathcal{H}} \|g_0 - g^\dagger\|_2
$$

综上，我们有

$$
\mathbb{E}[\{h(X) - h_0(X)\}^2] \le L(\hat{h}, \bar{g}_0) - L(h_0, \bar{g}_0) \le \{2C_{\mathcal{H}} + C_{\mathcal{G}}\} \|h_0 - h^\dagger\|_2 + C_{\mathcal{H}} \|g_0 - g^\dagger\|_2 + 2M(\mathcal{H}, \mathcal{G}).
$$

G 第 6 节的证明第 6 节

G.1 第6.1 节中的收敛速率证明

回顾

$$
\hat{h}_{\text{pro}} = \arg\min_{h \in \mathcal{H}} \max_{g \in \mathcal{G}} \tilde{L}_n(h, g), \quad \tilde{L}_n(h, g) := -0.5\mathbb{E}_n[g^2(Z)] + \mathbb{E}_n[\{Y - h(X)\} g(Z)].
$$

Let

$$
\begin{aligned}
\hat{g}_h &:= \arg\max_{g \in \mathcal{G}} \tilde{L}_n(h, g), \quad g_h := \mathbb{E}[Y - h(X) \mid Z = \cdot], \\
\Gamma(h, g) &:= -0.5g^2(Z) + \{Y - h(X)\} g(Z), \quad \kappa(h, g) := \Gamma(h, g) - \Gamma(h, g_h).
\end{aligned}
$$

第一步。我们的目标是证明

$$
\forall h \in \mathcal{H}; |\mathbb{E}_n[\kappa(h, \hat{g}_h)]| \lesssim \frac{(C_{\mathcal{H}}^2 + C_{\mathcal{G}}^2) \ln(|\mathcal{G}|/\delta)}{n}. \tag{18}
$$

此后我们固定 h 。

这里，首先，我们有

$$\mathbb{E}[\kappa(h,\hat{g}_h)] = 0.5\mathbb{E}[(\hat{g}_h - g_h)^2(Z)].$$

于是，

$$\begin{aligned}\mathbb{E}[\kappa(h,\hat{g}_h)] &\leq \mathbb{E}_n[\kappa(h,\hat{g}_h)] + |(\mathbb{E}-\mathbb{E}_n)[\kappa(h,\hat{g}_h)]| \\ &\leq |(\mathbb{E}-\mathbb{E}_n)[\kappa(h,\hat{g}_h)]|.\end{aligned}$$

F从第一行到第二行，我们使用估计量的定义和 $g_h \in \mathcal{G}$。

现在，我们使用伯恩斯坦不等式。以 $1-\delta$ 的概率，有

$$\forall g \in \mathcal{G}, \forall h \in \mathcal{H}; (\mathbb{E}-\mathbb{E}_n)[\kappa(h,g)] \leq \sqrt{\frac{\mathrm{var}[\kappa(h,g)]\ln(|\mathcal{G}||\mathcal{H}|/\delta)}{n}} + \frac{(C_{\mathcal{H}}^2 + C_{\mathcal{G}}^2)\ln(|\mathcal{G}||\mathcal{H}|/\delta)}{n}.$$

下面，我们在该事件条件下进行讨论。于是，有

$$\mathbb{E}[\kappa(h,\hat{g}_h)] \lesssim \sqrt{\frac{\mathrm{var}[\kappa(h,\hat{g}_h)]\ln(|\mathcal{G}||\mathcal{H}|/\delta)}{n}} + \frac{(C_{\mathcal{H}}^2 + C_{\mathcal{G}}^2)\ln(|\mathcal{G}||\mathcal{H}|/\delta)}{n}. \tag{19}$$

这里，有

$$\begin{aligned}\mathrm{var}[\kappa(h,\hat{g}_h)] &= \mathbb{E}[\{\Gamma(h,g_h) - \Gamma(h,\hat{g}_h)\}^2] \\ &= \mathbb{E}[0.25\{\hat{g}_h(Z) - g_h(Z)\}^2\{\hat{g}_h(Z) + g_h(Z)\}^2] \\ &\leq C_{\mathcal{G}}^2\mathbb{E}[\{\hat{g}_h(Z) - g_h(Z)\}^2].\end{aligned}$$

因此，结合上述结果与 (19)，我们得到

$$\mathbb{E}[\{\hat{g}_h(Z) - g_h(Z)\}^2] \lesssim \frac{(C_{\mathcal{H}}^2 + C_{\mathcal{G}}^2)\ln(|\mathcal{G}||\mathcal{H}|/\delta)}{n}.$$

因此，

$$\begin{aligned}|\mathbb{E}_n[\kappa(h,\hat{g}_h)]| &\leq |\mathbb{E}[\kappa(h,\hat{g}_h)]| + |(\mathbb{E}_n - \mathbb{E})[\kappa(h,\hat{g}_h)]| \\ &= 0.5\mathbb{E}[\{\hat{g}_h(Z) - g_h(Z)\}^2] + |(\mathbb{E}_n - \mathbb{E})[\kappa(h,\hat{g}_h)]| \\ &\lesssim \frac{(C_{\mathcal{H}}^2 + C_{\mathcal{G}}^2)\ln(|\mathcal{G}||\mathcal{H}|/\delta)}{n} + |(\mathbb{E}_n - \mathbb{E})[\kappa(h,\hat{g}_h)]| \\ &\lesssim \frac{(C_{\mathcal{H}}^2 + C_{\mathcal{G}}^2)\ln(|\mathcal{G}||\mathcal{H}|/\delta)}{n}. \qquad \text{(Use (19))}\end{aligned}$$

第二步。我们定义

$$\Xi(h) := \Gamma(h,g_h) - \Gamma(h_0,g_{h_0}).$$

注意，$\Gamma(h_0,g_{h_0}) = 0$，因为 h_0。此外，

$$\mathbb{E}[\Xi(h)] = \mathbb{E}[g_h^2(Z)], \quad \mathbb{E}[\Xi^2(h)] \leq (C_{\mathcal{H}}^2 + C_{\mathcal{G}}^2)\mathbb{E}[g_h^2(Z)]. \tag{20}$$

那么，

$$\mathbb{E}[\Xi(\hat{h})] \leq \mathbb{E}_n[\Xi(\hat{h})] + |(\mathbb{E}-\mathbb{E}_n)[\Xi(\hat{h})]|$$

这里，利用第一个结论(18)，我们得到

$$\begin{aligned}\mathbb{E}_n[\Xi(\hat{h})] &= \mathbb{E}_n[\Gamma(\hat{h}, g_{\hat{h}}) - \Gamma(h_0, 0)] \\ &\leq \mathbb{E}_n[\Gamma(\hat{h}, \hat{g}_{\hat{h}}) - \Gamma(h_0, \hat{g}_{h_0})] + c\frac{(C_{\mathcal{H}}^2 + C_{\mathcal{G}}^2)\ln(|\mathcal{G}||\mathcal{H}|/\delta)}{n} \\ &\leq c\frac{(C_{\mathcal{H}}^2 + C_{\mathcal{G}}^2)\ln(|\mathcal{G}||\mathcal{H}|/\delta)}{n}.\end{aligned}$$

从第一行到第二行，我们使用 $h_0 \in \mathcal{H}$ 和 (18)。从第二行到第三行，我们使用估计量的构造。

因此，

$$\mathbb{E}[\Xi(\hat{h})] \leq c\frac{(C_{\mathcal{H}}^2 + C_{\mathcal{G}}^2)\ln(|\mathcal{G}||\mathcal{H}|/\delta)}{n} + |(\mathbb{E} - \mathbb{E}_n)[\Xi(\hat{h})]|.$$

这里，我们使用伯恩斯坦不等式。以概率 $1-\delta$，有

$$\forall h \in \mathcal{H}; |(\mathbb{E} - \mathbb{E}_n)[\Xi(h)]| \leq \sqrt{\frac{\mathrm{var}[\Xi(h)]\ln(|\mathcal{H}|/\delta)}{n}} + c\frac{(C_{\mathcal{H}}^2 + C_{\mathcal{G}}^2)\ln(|\mathcal{H}|/\delta)}{n}.$$

此后，我们以该事件为条件。因此，利用（20），有

$$\mathbb{E}[g_{\hat{h}}^2(Z)] \leq \sqrt{\frac{g_{\hat{h}}^2(Z)\ln(|\mathcal{H}|/\delta)}{n}} + c\frac{(C_{\mathcal{H}}^2 + C_{\mathcal{G}}^2)\ln(|\mathcal{H}||\mathcal{G}|/\delta)}{n}.$$

因此，经过一些代数运算，我们得到

$$\mathbb{E}[g_{\hat{h}}^2(Z)] \lesssim \frac{(C_{\mathcal{H}}^2 + C_{\mathcal{G}}^2)\ln(|\mathcal{H}||\mathcal{G}|/\delta)}{n}.$$

H 附录 附录 A

H.1 定理 4 的证明

我们采用定理 2中的记号。取 μ_n 使得 $2\mathcal{M}(\mathcal{H},\ \mathcal{G}) \leq \mu_n$ 以概率 $1-\delta$成立。我们以该事件为条件。

只要 h_0 以概率$\mathcal{H}_n$ 被包含在置信球中 $1-\delta$，按照第 6 节中的证明，投影均方误差的保证就显而易见。事实上，我们有

$$\begin{aligned}&L_n(h_0, \hat{g}(h_0)) - \min_h L_n(h, \hat{g}(h)) = L_n(h_0, \hat{g}(h_0)) - L_n(\hat{h}, \hat{g}(\hat{h})) \\ &\leq L_n(h_0, \hat{g}(h_0)) - L_n(\hat{h}, g(\hat{h})) \\ &= L_n(h_0, \hat{g}(h_0)) - L(h_0, \hat{g}(h_0))) + L(h_0, \hat{g}(h_0))) - L(\hat{h}, g(\hat{h})) + L(\hat{h}, g(\hat{h})) - L_n(\hat{h}, g(\hat{h})) \\ &\leq L_n(h_0, \hat{g}(h_0)) - L(h_0, \hat{g}(h_0))) + L(h_0, g(h_0))) - L(\hat{h}, g(\hat{h})) + L(\hat{h}, g(\hat{h})) - L_n(\hat{h}, g(\hat{h})) \\ &\leq 2M(\mathcal{H}, \mathcal{G}) \leq \mu_n.\end{aligned}$$

因此， $h_0 \in \mathcal{H}_n$。

ˆ 接下来，我们证明 L_2 收敛保证。这里，对于置信球 $\mathcal{H}_n$中的任意 h ，有

$$\begin{aligned}
L_n(\hat{h}, \hat{g}(\hat{h})) &\geq L_n(\hat{h}, \bar{g}_0) && \text{(Construction of estimators)} \\
&\geq L(\hat{h}, \bar{g}_0) - M(\mathcal{H}, \mathcal{G}) \\
&\geq L(h_0, \bar{g}_0) - M(\mathcal{H}, \mathcal{G}) && \text{(Saddle point property)} \\
&\geq L(h_0, \hat{g}(h_0)) - M(\mathcal{H}, \mathcal{G}) && \text{（鞍点性质）} \\
&\geq L_n(h_0, \hat{g}(h_0)) - 2M(\mathcal{H}, \mathcal{G}) \\
&\geq \min_h L_n(h, \hat{g}(h)) - 2M(\mathcal{H}, \mathcal{G}) \\
&\geq L_n(\hat{h}, \hat{g}(\hat{h})) - 2M(\mathcal{H}, \mathcal{G}) - \mu_n.
\end{aligned}$$

因此,

$$L_n(h_0, \hat{g}(h_0)) - \min_h L_n(h, \hat{g}(h)) \leq 2M(\mathcal{H}, \mathcal{G}) + \mu_n.$$

因此， L_2 收敛速率保证得以确保，因为

$$\mathbb{E}[\{\hat{h}(X) - h_0(X)\}^2] \leq L(\hat{h}, \bar{g}_0) - L(h_0, \bar{g}_0) \leq 2M(\mathcal{H}, \mathcal{G}) + \mu_n.$$

I 辅助引理

引理6. (x^*, y^*) 是 $f(x, y)$ 在 $(\mathcal{X}, \mathcal{Y})$ 上的一个鞍点，当且仅当强对偶性成立且

$$x^* \in \arg\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} f(x, y), \quad y^* \in \arg\max_{y \in \mathcal{Y}} \min_{x \in \mathcal{X}} f(x, y).$$

证明. 假设 (x^*, y^*) 是 $f(x, y)$ 在 $\mathcal{X}, \mathcal{Y}$ 上的一个鞍点。那么,

$$\inf_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} f(x, y) \leq \sup_{y \in \mathcal{Y}} f(x^*, y) \leq f(x^*, y^*) \leq \inf_{x \in \mathcal{X}} f(x, y^*) \leq \sup_{y \in \mathcal{Y}} \inf_{x \in \mathcal{X}} f(x, y).$$

因此，强对偶性成立。上述不等式实际上都是等式。因此 e,

$$\inf_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} f(x, y) = \sup_{y \in \mathcal{Y}} f(x^*, y) = f(x^*, y^*) = \inf_{x \in \mathcal{X}} f(x, y^*) = \sup_{y \in \mathcal{Y}} \inf_{x \in \mathcal{X}} f(x, y).$$

因此，我们有

$$x^* \in \arg\min_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} f(x, y), \quad y^* \in \arg\max_{y \in \mathcal{Y}} \inf_{x \in \mathcal{X}} f(x, y).$$

接下来，假设强对偶性成立，并且

$$x^* \in \arg\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} f(x, y), \quad y^* \in \arg\max_{y \in \mathcal{Y}} \inf_{x \in \mathcal{X}} f(x, y).$$

于是，我们有

$$\max_{y \in \mathcal{Y}} \inf_{x \in \mathcal{X}} f(x, y) = \inf_{x \in \mathcal{X}} f(x, y^*) \leq f(x^*, y^*) \leq \sup_{y \in \mathcal{Y}} f(x^*, y) = \min_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} f(x, y).$$

最后，利用强对偶性，上述关系实际上是等式。因此，

$$\inf_{x\in\mathcal{X}} f(x,y^*) = f(x^*,y^*), \quad \sup_{y\in\mathcal{Y}} f(x^*,y) = f(x^*,y^*).$$

这意味着（x^*， y^*）是一个鞍点，因为

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y}; f(x,y) \geq f(x^*,y^*) \geq f(x^*,y).$$

□

定理 5（Dudley积分）。考虑一个包含 f 的函数类 $\mathcal{F}$： $\mathcal{X} \to \mathbb{R}$。则有

$$\mathcal{R}_n(\mathcal{F}) \leq \inf_{\epsilon\geq 0} \left\{ 4\epsilon + 12 \int_{\epsilon}^{\|\mathcal{F}\|_\infty} \sqrt{\frac{\ln \mathcal{N}(\tau,\mathcal{F},\|\cdot\|_\infty)}{n}} \mathrm{d}\tau \right\}.$$