arXiv:2611.33069v1 [econ.EM] 24 Nov 2026

实用型深度去偏机器学习

作者：Victor Chernozhukov[1] 和 Chengchun Shi[2]

[1]麻省理工学院经济学系与统计与数据科学中心

[2]伦敦政治经济学院统计系

去偏机器学习（DML）将拟合的结果变量回归与加权残差相结合，在估计平均处理效应等参数时消除一阶正则化偏差；然而，只要回归误差与加权误差相互一致，二阶偏差仍会存在。对于线性回归泛函，我们通过训练强学习器，根据多个协变量预测第一阶段残差来针对这一二阶偏差，从而得到一个一维（或低维）工作模型，并在独立样本上估计其偏差贡献。为补充这一针对性搜索，我们还分析了一般的基于核的校正，包括协变量分区和平滑器。我们证明，深度去偏机器学习在代数上等同于采用事后回归更新进行评估的常规双重机器学习；该更新利用其余评估观测中的核加权残差，对每个得分观测进行调整。在我们的实证与模拟研究中，强大的初始双重机器学习学习器通常会使二阶校正相对于抽样不确定性保持较小；在这种情况下，该校正可作为诊断检查，并与敏感性界结合，用于评估任何剩余的模型设定错误偏差。

1. 引言。 普通的去偏机器学习（DML）[5]消除了将拟合的结果回归代入目标参数时产生的一阶正则化偏差。对于该回归的任何线性泛函（例如平均处理效应），剩余偏差都是二阶的。它等于结果回归误差与估计的里斯权重误差（例如逆倾向权重）内积的负值。由于这一剩余偏差是两个误差的乘积，因此通常较小；然而，只要两个学习器都遗漏了相同的协变量模式，其误差就会对齐，混合偏差很容易超过抽样不确定性。我们研究了一种针对这种对齐现象的单独校正项。

假设我们正在估计家庭的平均处理效应。在某个收入组内，估计的处理权重可能无法再现该组在总体中的真实占比，而拟合的结果回归则系统性地高估或低估同一组中的结果变量。尽管无法观测到任一真实干扰函数，但可以直接在独立评估样本上估计这两种偏差。对同一批观测值取样本均值的乘积会引入自身观测相关性（即一个对角协方差项），因此，我们将一个观测值的权重偏差与另一个不同观测值的结果残差配对。这样一来，偏差校正项就表现为对所有不同观测值对取平均，这种统计量称为 U统计量。在训练数据上选定的核指定了我们应用此校正项的组（或更一般地说，方向）。

选择用于偏差校正的基函数可能比为初始回归选择学习器更困难。预先指定的字典迫使分析者事先猜测拟合回归将遗漏哪些协变量模式，而投影到大型字典上可能会增加

2026年11月10日（修订于2026年11月24日）。我们与 GPT Astra 合作完成了本文。我们对本文内容承担全部责任。

相当大的方差。我们根据第一阶段残差学习用于偏差校正的字典。将这些残差对协变量进行灵活回归，可以生成一个预测初始拟合误差所在位置的单一特征。随后，在独立评估样本上，我们将结果残差与沿该学习特征的加权差异配对，以估计其对目标量偏差的贡献。如果学习特征能够在尺度意义上捕捉回归误差，那么投影到这一单一方向上即可消除全部混合偏差，同时使工作空间保持一维。第3.2节展示了如何利用包含不断增加的预测变量的残差回归，足够准确地学习这一方向，从而使剩余偏差小于抽样误差。

一旦我们找到了回归出错之处，除了直接校正目标量之外，一个自然的替代方案是沿着学习得到的方向更新回归，再将其代入常规双重机器学习。定理1表明，当我们使用其他评估观测值的残差和相同的核，对每个观测分别更新回归时，这两种方法会给出完全相同的估计值。对于单个学习得到的方向，更新会将该方向加入初始回归，并用留出当前得分观测值后的数据估计斜率。虽然每个观测值在更新自身时都被留出，但更新后的得分会共享数据。成对表示使我们能够确定，这种事后校正的双重机器学习估计量消除了多少偏差，以及更新会付出多少抽样方差代价。

对训练选定的工作空间进行精确投影，可以消除该空间内所有混合偏差，只留下空间外误差分量的内积（命题 2）。两个真实的干扰函数都不必属于该工作空间，因为该空间仅用于指定要消除哪些误差方向。该投影也保留了通常的最坏情形乘积界，不过，当空间内外的误差恰好相互抵消时，单次拟合的实际偏差可能会增大（例4）。

成对校正可能同时改变一阶影响函数以及各成对样本贡献的抽样噪声。精确的Hoeffding分解将这两个组成部分分离开来，从而得到完整条件方差的闭式有限样本无偏估计量（命题 9），并给出明确条件，说明校正估计量何时仍保有常规双重机器学习的渐近分布和标准误（命题 8）。

即使经过第二步去偏，工作空间之外仍可能存在一些模型设定错误和干扰函数学习偏差。我们采用 Chernozhukov、Cinelli、Newey、Sharma 和 Syrgkanis [6] 的双因素敏感性框架来控制剩余的模型设定错误偏差，将对残余干扰误差的明确容许度转化为偏差感知置信区间。当初始双重机器学习学习器已经较为强大时，进一步校正通常只会使点估计发生很小变化。在这种情况下，校正可作为沿学习得到的方向进行的诊断检查，而敏感性界则量化结论对剩余方向上的模型设定错误偏差的容忍程度。

我们的实证分析说明，初始学习器的强弱决定了第二步去偏的价值。在 401(k)资格数据集中，以线性双重机器学习为起点时，随机森林学得的单一方向会使资格对上限净金融资产的估计效应增加 1,849 美元，而 CausalPFN [1]学得的方向则会使其增加 2,808 美元。对于同一批家庭，以随机森林双重机器学习为起点时，这两种校正项对估计值的影响仅为 54 美元和 −113 美元。在真实处理效应已知的 ACIC 模拟中，森林学得的校正项使均方根误差在线性初始拟合后降低 38–43%，在梯度提升后降低不到 2%，在 CausalPFN 结果拟合后降低约 2%。在采用提升干扰函数的 ACIC设定29中，常规双重机器学习和校正后的 95%置信区间在 84.5% 的抽样中均覆盖真实效应。若按 2% 的敏感性预算对剩余偏差作统一调整，常规双重机器学习的覆盖率升至 95.5%，双重去偏机器学习的覆盖率则升至 94.5%。

我们还考察变量选择未能捕捉弥散混杂信号的情形。在 Chetverikov、Sørensen 和 Tsyvinski 的高维线性设计基础上加以扩展 [8],

我们将一个共同的混杂成分分散到 500 个控制变量中的 400 个弱控制变量上。后 Lasso 会对每个干扰回归所选出的变量进行重拟合，但许多个体较弱的控制变量仍未被选中，却依然可能朝同一方向推动处理效应系数。对处理残差进行岭回归，可以将全部 500 个控制变量合并为一个学习特征；由此得到的秩一校正，在我们研究的两种尾部强度和两种相关系数水平下，将均方根误差降低了 7–22%。我们还报告了原始几何设计的结果；在该设计中，仅靠后置 Lasso 重拟合几乎没有留下让三重 Lasso 得分进一步改善估计值的空间。

在这些实验中，当初始拟合遗漏了系统性模式时，单个学习得到的方向能够降低偏差和均方根误差。初始拟合效果较好时，估计校正项相对于处理效应估计的标准误通常较小。随后，我们利用它检查沿学习得到的方向的拟合情况，并报告结论能够容忍多大的残余偏差。

与文献的关系。 基于投影的偏差校正建立在 Robins、Li、Tchetgen 和 van der Vaart [21] 以及 Robins、Li、Mukherjee、Tchetgen Tchetgen 和 van der Vaart [20], 提出的高阶影响函数之上。他们将函数投影到预先指定的基函数张成空间上，并利用成对观测值估计该投影所捕获的偏差。Liu、Mukherjee 和 Robins [18] 同样使用投影偏差的估计值，来检验传统置信区间能否达到其名义覆盖率。

多种情境下都出现了相关的二阶校正。Díaz、Carone 和 van der Laan [9] 针对缺失情形下的均值构造了二阶估计量；Shi、Wan、Chernozhukov 和 Song [22] 则将逐步去偏应用于离策略评估中的策略价值估计，并将这一方法称为“深度去偏”。Bonhomme、Jochmans、Newey 和 Weidner [3] 为干扰参数维数有限的模型构造了高阶正交矩；Chetverikov、Sørensen 和 Tsyvinski [8] 则引入二阶正交得分，以校正高维线性回归中的正则化偏差。

互补的事后策略应对常规双重机器学习推断的不同失效模式。Van der Laan、Luedtke 和 Carone [24] 对标量结果预测和里斯权重应用保序校准，从而在一个干扰参数收敛到错误极限、而另一个干扰参数足够准确时，实现双重稳健推断。Zheng、Bonvini 和 Guo [26] 则通过在扰动下重新拟合 Lasso 干扰参数，并围绕接近原始 DML拟合的估计值取 Wald 区间的并集，构造扰动 DML 区间。相比之下，我们的方法针对两个干扰函数之间相关的协变量层面的误差：它沿学习到的协变量方向减去估计的二阶混合偏差后，将单个区间重新居中，并使用双因素敏感性预算来考虑任何残余偏差。

我们独立且同时开展了本文的研究，时间上与 Escanciano 和 Ter-schuur [13]同期。对于仿射矩和有限字典，他们的命题 6 与我们的命题 2 给出了相同的剩余混合偏差公式。他们的分析还涵盖非线性矩，并刻画了实现精确二阶正交性所需的最小方差代价。与 Robins 等人 [21, 20], 的投影方法一样，他们的构造从给定字典 $p(X)$ 出发，并使用训练数据在其张成空间内选择或收缩方向。相反，我们关注于从第一阶段残差中学习字典本身：森林、岭回归拟合或基础模型将许多协变量压缩为一个干扰误差预测器，从而保留一个方向，并在独立样本上估计该方向的偏差贡献。在这一针对性校正之外，我们还展示了，关于两个学习误差的给定界如何控制学习空间之外剩余的模型设定错误偏差。

本文其余部分安排如下。第2节和第3节介绍成对校正项，并说明实用的工作空间。第4节和第5节讨论方差估计，

渐近推断和偏差感知敏感性区间。第6节汇总实施步骤和区间选择规则。第7节比较401(k)、ACIC 和三重 Lasso 的结果。第8节阐述校正项与高阶正交性及影响函数的关系，第9节则讨论选择和评估校正项的意义。所有证明均汇总于附录 B。

2. 成对校正及其偏差几何。我们使用一个大小为 n 的独立同分布评估样本 $W_1, \ldots, W_n$，该样本独立于其 σ-代数为 $\mathcal{T}$ 的训练样本。对于成对估计量，我们取 $n \geq 2$。所有干扰函数学习器 $\widehat{g}$ 和 $\widehat{\alpha}$，以及校正核 K，都仅在训练样本上拟合，并在评估期间保持固定。除非另有说明，第2–5节中涉及已拟合对象的所有期望、方差和范数，均是在给定 $\mathcal{T}$ 的条件下计算的。第 4.5 节说明了这些拆分样本结果何时可推广至标准的 K 折交叉拟合。

2.1. **目标泛函、动机示例与常规双重机器学习**。令 W 表示一个包含协变量 $X \subset W$ 和标量结果变量 Y 的完整观测，并记 $g_0(x) := \mathbb{E}[Y \mid X = x]$ 为真实的条件均值函数。我们的目标参数是已知线性泛函对 g_0 的期望，

$$g_0(x) := \mathbb{E}[Y \mid X = x], \qquad \theta_0 := \mathbb{E}[m(W, g_0)], \tag{1}$$

其中，映射 $h \mapsto m(W, h)$ 关于 h 是线性的。我们假设该泛函在均方意义下连续，从而保证存在一个平方可积的 Riesz 表示元 $\alpha_0(X)$，满足

$$\mathbb{E}[m(W, h)] = \mathbb{E}[\alpha_0(X) h(X)] \tag{2}$$

对于包含下文拟合对象的线性定义域中的每个函数 h，在标准可积性条件下，期望可以通过 m。我们用 $\langle f, h \rangle := \mathbb{E}[f(X)h(X)]$ 表示 $L^2(P_X)$ 内积，用 $\|f\|_2^2 := \langle f, f \rangle$ 表示相应的平方范数。向得分中加上一个可积的、仅依赖于数据的项 $r(W)$，并向目标量中加上其期望 $\mathbb{E}[r(W)]$，不会改变下文的任何偏差公式。

我们用这一符号表示两个熟知的计量经济学目标量。

- 平均处理效应。对于二元处理 D 和基线协变量 Z，令 $W = (Y, D, Z)$、$X = (D, Z)$ 和 $e_0(Z) := \Pr(D = 1 \mid Z)$。协变量调整对比为 $m(W, g) = g(1, Z) - g(0, Z)$。在满足重叠性和平方可积条件时，由迭代期望可知，逆倾向得分权重 $\alpha_0(D, Z) = D/e_0(Z) - (1 - D)/(1 - e_0(Z))$ 满足里斯恒等式 (2)。一致性和条件可交换性随后赋予这一对比因果解释，即平均处理效应。
- 部分线性回归。假设 $Y = D\beta_0 + g_0(X) + \epsilon$ 且 $D = \gamma_0(X) + \nu$，其中 $\mathbb{E}[\epsilon \mid D, X] = 0$ 且 $\mathbb{E}[\nu \mid X] = 0$。对于任意候选值 β，定义残差化结果 $Z_\beta := Y - \beta D$ 及其条件均值 $g_\beta(X) := \mathbb{E}[Z_\beta \mid X]$。Robinson 正交矩 $\mathbb{E}[\{Z_\beta - g_\beta(X)\}\{D - \gamma_0(X)\}]$ 在真实系数 β_0 处为零。固定 β 后，将乘积展开，即可将该矩写成线性泛函形式 (1)–(2)，其中 $m(W, f) = -Df(X)$、Riesz 表示元 $\alpha_0(X) = -\gamma_0(X)$，以及可观测加项 DZ_β。第 3.3 节详细介绍高维线性特化。

给定训练估计值 $\widehat{g}$ 和 $\widehat{\alpha}$，普通双重机器学习得分和估计量 [5]将代入项与加权结果变量残差相结合，

$$s(W) := m(W, \widehat{g}) + \widehat{\alpha}(X)\{Y - \widehat{g}(X)\}, \qquad \widehat{\theta}_{\mathrm{DB}} := \mathbb{P}_n s, \tag{3}$$

其中，$\mathbb{P}_n f := n^{-1}\sum_{i=1}^n f(W_i)$ 表示评估样本均值。用 $\delta_g := \widehat{g} - g_0$和 $\delta_\alpha := \widehat{\alpha} - \alpha_0$ 表示两个干扰参数估计误差。由于 $\mathbb{E}[Y - \widehat{g}(X) \mid X] = -\delta_g(X)$，里斯恒等式（2）使一阶代入误差 $\mathbb{E}[m(W,\delta_g)] = \langle\alpha_0、\delta_g\rangle$与一阶残差项 $\mathbb{E}[\alpha_0(X)\{Y - \widehat{g}(X)\}] = -\langle\alpha_0、\delta_g\rangle$相互抵消。最终只剩下两个误差的乘积，从而得到熟知的混合偏差恒等式

$$\mathbb{E}[s(W)] - \theta_0 = -\langle\delta_\alpha, \delta_g\rangle. \tag{4}$$

在平均处理效应的示例中，(3) 是增广逆概率加权（AIPW）估计量，而 (4) 表明，其偏差等于逆概率加权权重误差与结果回归误差的内积的相反数。常规双重机器学习依赖于该乘积小于抽样尺度，而我们的校正项则直接估计并消除 (4) 中可检测的部分。

2.2. 将可观测矩条件违背转化为成对校正项。 尽管既未观测到 g_0，也未观测到 α_0，评估样本却揭示了它们的误差沿任意由训练集选定的测试函数 $h(X)$ 的投影。应用里斯恒等式 (2) 和条件均值恒等式 $\mathbb{E}[Y - \widehat{g}(X) \mid X] = -\delta_g(X)$，可得到两个可观测的总体诊断量：一个是里斯方程（协变量平衡）违背，另一个是回归残差可预测性违背：

$$\begin{aligned}\mathcal{R}_\alpha(h) &:= \mathbb{E}[\widehat{\alpha}(X)h(X) - m(W,h)] = \langle\delta_\alpha, h\rangle,\\ \mathcal{R}_g(h) &:= \mathbb{E}[h(X)\{Y - \widehat{g}(X)\}] = -\langle\delta_g, h\rangle.\end{aligned} \tag{5}$$

对于平均处理效应，令 $h(D, Z) := \mathbf{1}\{D = 1, Z \in C\}$ 表示某个单位接受了处理并属于协变量单元 C。那么 $\mathcal{R}_\alpha(h) = \mathbb{E}[(D/\widehat{e}(Z) - 1)\mathbf{1}\{Z \in C\}]$ 用于检验拟合的逆概率加权权重是否恢复了该单元的总体占比，而 $\mathcal{R}_g(h) = \mathbb{E}[\mathbf{1}\{D = 1, Z \in C\}(Y - \widehat{g}(1, Z))]$ 用于检验结果回归在 C 内是否留下非零的平均结果残差。这两个矩都可以直接根据评估数据进行估计。然而，如果我们简单地对同一组观测值计算它们的样本均值并相乘，每个观测值 i 都会与自身配对，从而引入一个对角协方差项；即使总体偏差为零，该项也可能非零。为了在估计两个总体误差乘积时避免这种对角偏差，我们将一个观测值的平衡违背与另一个不同观测值的结果残差进行配对。

为了一次性记录多个方向上的这些违背集选定的核，其各列属于 m 的定义域，并令 T 表示其有界积分算子，

$$(Tf)(x) := \mathbb{E}[\widehat{K}(x, X)f(X)]. \tag{6}$$

$\widehat{\ }$核 K 不必是对称或正定的。构造稳定投影时，对称性很方便；而对于一般核，伴随算子 T^* 则会出现在偏差公式中。

使用此核，我们将观测 W 在协变量点 x'处的 Riesz 差异计算为

$$A(W; x') := \widehat{\alpha}(X)\widehat{K}(X, x') - m(W, \widehat{K}(\cdot, x')),$$

并将其与来自另一条观测 W'的结果残差 $\varepsilon_g(W') := Y' - \widehat{g}(X')$ 配对，构成成对核 $H(W, W') := A(W; X')\varepsilon_g(W')$。对所有不同观测值组成的有序对取平均，即可得到 U统计量偏差校正项和深度去偏机器学习（*DDML*）估计量，

$$\widehat{B}_K := \frac{1}{n(n-1)}\sum_{i\neq j} H(W_i, W_j), \qquad \widehat{\theta}_K := \widehat{\theta}_{\mathrm{DB}} - \widehat{B}_K. \tag{7}$$

由于排除了对角项 $i=j$，观测值 W_i 和 W_j 在给定 $\mathcal{T}$ 的条件下相互独立^，因此 B_K 是混合偏差中目标分量的无偏估计量。

2.3. 将双重去偏机器学习作为事后回归更新。 进入式（7）校正项的同一组残差，也可以用于更新结果回归本身。为了计算观测 i 的得分，我们将其他评估残差的核加权平均值加到初始拟合上。将留一回归更新定义为

$$\widetilde{g}_{-i}(x) := \widehat{g}(x) + \frac{1}{n-1}\sum_{j\neq i}\widehat{K}(x, X_j)\{Y_j - \widehat{g}(X_j)\}. \tag{8}$$

核保留了训练数据上选定的归一化。更新后的回归是 x 的函数，因此在平均处理效应示例中，我们可以针对同一个家庭，在处理组和对照组下分别对其进行评估。

定理 1（双重去偏机器学习与事后校正的双重机器学习完全等价）。设 $n \geq 2$、^，并固定训练拟合 $\widehat{g}$、$\widehat{\alpha}$ 和 K。假设 $m(w,\cdot)$ 在一个包含 $\widehat{g}$^及核列 $K(\cdot, X_j)$的定义域上线性。对于每个评估样本，式(7)中的双重去偏机器学习估计量等于使用式 (8) 中的更新后回归和原始 *Riesz* 权重计算的普通双重机器学习得分的平均值，

$$\widehat{\theta}_K = \frac{1}{n}\sum_{i=1}^{n}\left[m(W_i, \widetilde{g}_{-i}) + \widehat{\alpha}(X_i)\{Y_i - \widetilde{g}_{-i}(X_i)\}\right]. \tag{9}$$

^对于单个学习得到的方向 f 及其训练归一化 $\widehat{\nu}_f > 0$，核 $K(x,\ x') = f(x)f(x')/\widehat{\nu}_f$给出的更新为 $\widetilde{g}_{-i} = \widehat{g} + \widehat{c}_{-i}f$，其中

$$\widehat{c}_{-i} := \frac{1}{(n-1)\widehat{\nu}_f}\sum_{j\neq i} f(X_j)\{Y_j - \widehat{g}(X_j)\}.$$

当我们留出观测值 i时，只需更新一个系数。在训练数据上拟合一次的系数通常会给出不同的估计值，正如第 7.3 节中的匹配比较所示。在（9）中，每个更新后的回归在给定训练集条件下都与其自身的评估观测值独立，但不同的得分使用了相互重叠的残差。第 2.4 节和第 4 节利用成对表示推导这一事后更新的偏差和方差。

里斯权重的留一更新。 双重去偏机器学习有两种对称的解释：既可以将其理解为在固定 $\widehat{\alpha}$ 的同时更新结果回归 $\widehat{g}$（定理1），也可以将其理解为在固定 $\widehat{g}$ 的同时更新里斯权重 $\widehat{\alpha}$。先对衡量权重失衡的观测值求和同样的成对项，可得到这一伴随表示。定义留一平衡调整

$$\widehat{d}_{-j}(x) := \frac{1}{n-1}\sum_{i\neq j}\{\widehat{\alpha}(X_i)\widehat{K}(X_i, x) - m(W_i, \widehat{K}(\cdot, x))\}, \tag{10}$$

$$\widehat{\theta}_K = \frac{1}{n}\sum_{j=1}^{n}\left[m(W_j, \widehat{g}) + \{\widehat{\alpha}(X_j) - \widehat{d}_{-j}(X_j)\}\{Y_j - \widehat{g}(X_j)\}\right].$$

这样看来，双重去偏机器学习保留结果回归 $\widehat{g}$ 不变，并通过减去 d_{-j}（X_j）来更新每个观测 j 的 Riesz^权重。在总体中，$\mathbb{E}[d_{-j}$（x）$]=$（$T^*\delta_\alpha$）（x），因此，精确投影会从权重中移除 δ_α 的投影分量。由于留一法权重在不同索引 j之间共享评估观测，第4节中的方差分析明确考虑了它们的依赖性。

注1（有限秩核的快速计算）。虽然对所有不同的配对计算一般核需要 $O(n^2)$ 次运算，但实践中最常使用的有限秩核计算速度快得多。设 $K(x,x') = b(x)'Mb(x')$ 表示特征向量 $b(x) \in \mathbb{R}^k$，并令 $Z_\alpha(W) := \widehat{\alpha}(X)b(X) - m(W,b)$，以及 $Z_g(W) := b(X)\{Y - \widehat{g}(X)\}$。先对所有配对 (i,j) 求和，再减去 n 对角项后，双重求和可分解为样本总和的单个双线性型：

$$\widehat{B}_K = \frac{1}{n(n-1)} \left[\left(\sum_{i=1}^{n} Z_\alpha(W_i) \right)' M \left(\sum_{j=1}^{n} Z_g(W_j) \right) - \sum_{i=1}^{n} Z_\alpha(W_i)' M Z_g(W_i) \right].$$

一旦特征和拟合矩阵 M 构建完成，聚合乘积的计算成本为 $O(nk + k^2)$ 次运算。直接计算 n 对角双线性型，对于一般稠密的 M，成本为 $O(nk^2)$；当 M 为对角矩阵时，成本为 $O(nk)$。对于平均处理效应单元核，每个观测仅属于固定数量的处理组与单元条目，因此一旦分配好单元归属，计算单元总量并减去对角项只需 $O(n+k)$ 次运算。

2.4. 精确条件偏差与工作模型稳健性。 θ_K 的一个核心优势是，其条件偏差满足一个精确的有限样本恒等式，而不要求 K 在无限维函数空间中近似该恒等式。

命题 2（精确剩余偏差）。 在 （1）–（6）及所述可积性条件下，

$$\mathbb{E}[\widehat{B}_K] = -\langle \delta_\alpha, T\delta_g \rangle, \qquad b_K := \mathbb{E}[\widehat{\theta}_K] - \theta_0 = -\langle \delta_\alpha, (I-T)\delta_g \rangle. \tag{11}$$

等价地， $b_K = -\langle (I - T^*)\, \delta_\alpha,\ \delta_g \rangle$。

命题 2 澄清了，偏差消除具有方向性。若要使偏差相对于根-n 尺度可忽略，只需满足

$$\sqrt{n}\, \left| \langle \delta_\alpha, (I-T)\delta_g \rangle \right| = o_p(1). \tag{12}$$

如果算子 T 能捕捉结果回归误差 δ_g 的大部分，即满足 $\|(I-T)\delta_g\|_2 \le r_K \|\delta_g\|_2$，那么根据柯西–施瓦茨不等式可得 $|b_K| \le r_K \|\delta_\alpha\|_2 \|\delta_g\|_2$；只要 $r_K = o_p(1)$，这就严格放宽了标准的双重机器学习乘积速率。对于任何平方可积核，由于其积分算子是紧的，要求 $\|I - T\|_{\mathrm{op}} = o_p(1)$ 在一个无限维空间中对所有函数一致成立是不可能的。偏差条件 (12) 只涉及 $I - T$ 对两个拟合误差的作用。

工作模型分解。为了用熟悉的回归术语理解这一几何结构，设 $\mathcal{H}_k$ 为训练集选定的、维数为 k 的工作空间，例如收入与处理组合单元的指示变量、浅层回归树的叶节点，或单个学习得到的残差特征。令 Π 表示投影到 $\mathcal{H}_k$ 上的总体 $L^2(P_X)$ 正交投影，它是弗里施–沃–洛弗尔回归投影的总体类比。将两个干扰误差都分解为位于 $\mathcal{H}_k$ 内部和外部的分量，

$$\delta_\alpha^{\perp} := (I - \Pi)\delta_\alpha, \qquad \delta_g^{\perp} := (I - \Pi)\delta_g,$$

正交性将常规双重机器学习的偏差分解为两部分，

$$\mathbb{E}[\widehat{\theta}_{\mathrm{DB}}] - \theta_0 = -\langle \Pi\delta_\alpha, \Pi\delta_g \rangle - \langle \delta_\alpha^{\perp}, \delta_g^{\perp} \rangle. \tag{13}$$

当 $T = \Pi$ 是对 $\mathcal{H}_k$ 的精确投影时，双重去偏机器学习会完全消除第一项，只留下正交余项的内积，

$$b_\Pi = -\langle \delta_\alpha^{\perp}, \delta_g^{\perp} \rangle. \tag{14}$$

在训练数据上估计格拉姆矩阵或单元概率，会得到一个近似投影 $T = \Pi T \Pi$，其值域仍位于 $\mathcal{H}_k$中。其剩余偏差可分为两个易于解释的部分，即 $\mathcal{H}_k$ 之外的遗漏方向偏差 和 $\mathcal{H}_k$之内的*Gram*复现误差：

$$b_K = -\langle \delta_\alpha^\perp, \delta_g^\perp \rangle - \langle \Pi \delta_\alpha, (I - T)\Pi \delta_g \rangle. \tag{15}$$

g_0 和 α_0 都不必属于 $\mathcal{H}_k$，因为工作空间只是定义了我们要消除哪些误差方向，而不是对数据生成过程强加参数模型。

为了明确双重去偏机器学习在全局误差界和残差界下所能保证的性质，假设拟合干扰函数满足全局 L^2 界 $\|\delta_\alpha\|_2 \le \epsilon_\alpha$ 和 $\|\delta_g\|_2 \le \epsilon_g$，并且在 $\mathcal{H}_k$之外还满足可能更严格的残差界 $\|\delta_\alpha^\perp\|_2 \le \rho_\alpha \le \epsilon_\alpha$ 和 $\|\delta_g^\perp\|_2 \le \rho_g \le \epsilon_g$。给定训练样本，定义误差类

$$\mathcal{U}_\Pi := \left\{ (a, v) \in L^2(P_X)^2 : \begin{array}{ll} \|a\|_2 \le \epsilon_\alpha, & \|v\|_2 \le \epsilon_g, \\ \|(I-\Pi)a\|_2 \le \rho_\alpha, & \|(I-\Pi)v\|_2 \le \rho_g \end{array} \right\}. \tag{16}$$

命题3（误差球上的精确偏差界）。若 $\mathcal{H}_k$ 的正交补非平凡，则

$$\sup_{(a,v)\in\mathcal{U}_\Pi} |\langle a, (I-\Pi)v \rangle| = \rho_\alpha \rho_g. \tag{17}$$

特别是，仅有全局界（$\rho_\alpha = \epsilon_\alpha$ 和 $\rho_g = \epsilon_g$）时，上确界等于 $\epsilon_\alpha \epsilon_g$。若 $T = \Pi T \Pi$ 和 $\|\Pi - T\|_{\mathrm{op}} \le \omega$，则对于实际的干扰误差，

$$|b_K| \le \|\delta_\alpha^\perp\|_2 \|\delta_g^\perp\|_2 + \omega \|\Pi \delta_\alpha\|_2 \|\Pi \delta_g\|_2. \tag{18}$$

命题3中的上确界揭示了一个重要边界：仅凭全局误差范数，投影所能保证的程度是有限的。如果我们只知道 $\|\delta_\alpha\|_2 \le \epsilon_\alpha$ 和 $\|\delta_g\|_2 \le \epsilon_g$，那么所有干扰误差原则上都可能隐藏在 $\mathcal{H}_k$ 之外，因此最坏情况下的偏差仍为 $\epsilon_\alpha \epsilon_g$。要得到更紧的界，需要了解 $\mathcal{H}_k$ 之外剩余误差的信息。第5节根据结果残差和拟合权重，表示了这些误差的大小。这个确定性计算补充了 Jin 和 Syrgkanis [17] 的与结构无关的极小极大下界。

与此同时，任何一般核都满足算子范数备用条件

$$|b_K| \le (1 + \|T\|_{\mathrm{op}}) \|\delta_\alpha\|_2 \|\delta_g\|_2. \tag{19}$$

具有 $\|T\|_{\mathrm{op}} = O_p(1)$ 的稳定算子会保留偏差通常的乘积速率界，但有效推断还取决于第4节分析的一阶方差项和成对方差项。然而，即使是精确投影，当两个项的符号相反并相互部分抵消时，也可能增加某个特定训练拟合的绝对偏差。

例4（校正项可以消除抵消效应）。对于标准正交向量 e_1 和e_2，令 $\delta_\alpha = t_n(e_1 + e_2)$、$\delta_g = t_n(e_1 - e_2)$ 和 $T = \Pi_{e_1}$。在常规双重机器学习中，两个坐标产生大小相等、符号相反的偏差，因此它们的内积完全抵消，得到 $-\langle \delta_\alpha, \delta_g \rangle = 0$。投影到 e_1 上只会移除第一项贡献，而第二项仍然保留，因而产生大小为 $b_K = t_n^2$的校正偏差。当 $t_n = n^{-1/5}$时，该校正偏差等于 $n^{-2/5}$，超过了 $n^{-1/2}$ 抽样尺度；尽管两个全局误差范数的乘积为 $2n^{-2/5}$，因此常规双重机器学习的乘积速率条件在此也不成立。

备注5（一致性与推断的双重稳健性）。命题 2 立即推出总体矩具有双重稳健性，因为只要 $b_K = 0$ 在$\delta_g = 0$ 或 $\delta_\alpha = 0$成立时就成立。若一个干扰参数是一致的、另一个误差有界、算子 T 稳定且随机波动趋于零，则一致性成立。当一个干扰参数设定错误时，要实现双重稳健的置信区间 则更难，因为一致干扰参数的估计误差$\widehat{\ }$以及 B_K 的一阶投影仍会产生影响 [12]。

3. 在实践中选择校正核。由于剩余偏差等于$\backsim b_K = -\langle \delta_\alpha,\ (I - T)\delta_g \rangle$，核 K的选择决定了要消除哪些误差方向，以及会增加多少抽样方差。本节介绍实用的核构造方法，从 Robins 等人 [21, 20] 提出的经典有限基投影出发，进而讨论数据自适应残差方向、协变量分区和平滑核（总结见表1）。第3.2节给出学习得到的方向的偏差率。第 3.3 节说明有限基构造如何应用于具有任意惩罚学习器的高维线性回归，并推广了 Chetverikov、Sørensen 和 Tsyvinski [8]提出的三重 Lasso。

3.1. 实用的核构造方法。

经典起点：有限基投影。高阶影响函数文献中由 Robins 等人 [21, 20] 开创的经典投影方法，从一个预先指定的基函数字典 $b(x) = (b_1(x) \dots, b_k(x))'$出发，例如样条、多项式或交互项。令 $\Sigma := \mathbb{E}[b(X)b(X)']$ 为其总体格拉姆矩阵，并令 $M \approx \Sigma^{-1}$ 为基于训练协变量计算的逆格拉姆矩阵估计。设定$\widehat{\ }$为 $K(x, x') = b(x)'Mb(x')$，即可定义一个有限秩校正核，其值域包含于$\mathcal{H}_k =$张成空间$\{b_1, \dots, b_k\}$中。当 $M = \Sigma^{-1}$时，所诱导的算子是到 $\mathcal{H}_k$上的精确正交投影 Π，其算子范数为 $\|\Pi\|_{\mathrm{op}} = 1$，平方希尔伯特–施密特范数为$\|K_\Pi\|^2_{\mathrm{HS}} = k$。正如第 3.3 节所示，这种构造为采用任何第一阶段惩罚学习器的高维回归提供了非对角偏差校正，是对三重 Lasso 的推广。然而，在中等或高维情形下，选择一个能捕捉初始机器学习学习器出错之处的预先指定基可能并不容易；扩大 k 还可能同时增加格拉姆矩阵估计误差和成对方差。由此引出了下文的定向构造和数据自适应构造。

1. 定向秩一（或低秩）残差方向。我们无需投影到宽泛的预先指定字典上，而可以让第一阶段残差本身告诉我们应该校正哪个方向。在给定固定的第一阶段拟合 $\widehat{g}$ 的条件下，结果残差 $Y - \widehat{g}(X)$ 对 X 的总体回归等于 $-\delta_g(X) = g_0(X) - \widehat{g}(X)$。将这些残差对 X 进行回归，可得到一个用于检验初始拟合的方向。实证示例使用了森林、岭回归，以及 CausalPFN [1] 与初始结果预测之间的差值。由此得到的预测器 $f(X)$ 提供了一个单函数字典，其核为秩一核

$$K_f(x, x') := \frac{f(x)f(x')}{\|f\|_2^2}. \tag{20}$$

采用总体归一化并令 $\|f\|_2 > 0$，该核诱导出到 f 张成空间上的精确正交投影 Π_f。其希尔伯特–施密特范数的平方为 1，这是非零正交投影所能达到的最小值。剩余偏差满足

$$|b_{\Pi_f}| = |\langle (I - \Pi_f)\delta_\alpha, (I - \Pi_f)\delta_g \rangle| \le \|(I - \Pi_f)\delta_\alpha\|_2 \, \|(I - \Pi_f)\delta_g\|_2.$$

因此，只要任一误差与 f成比例，校正项就会消除全部混合偏差；近似成比例则通过该界控制绝对余项，不过

消除的比例也取决于原始偏差中可能存在的抵消（例4）。在训练集（或单独的无标签校准子集）上估计 $\|f\|_2^2$，可得到近似投影，同时确保评估样本独立，以便进行推断。当特征或其校准二阶矩为零时，我们使用 $K_f = 0$ 并恢复常规双重机器学习。第3.2节量化了学习和归一化该方向所产生的误差。对于多个学习得到的方向，同样的构造会用到它们的格拉姆矩阵。

由于双向划分将所有训练操作与评估分开，命题2和命题7会对用于构建干扰函数和核的每一步进行条件化。若要学习特定初始拟合 $\widehat{g}$ 的误差，可以固定该拟合，在新的训练观测上构造残差标签，再用 X 对这些标签进行回归。如果残差标签来自训练集内的折外预测，那么每一折针对的都是其自身拟合的误差，但精确的偏差恒等式仍适用于学习得到的方向，以及评估中使用的最终重新拟合。令 $\bar{g}$ 表示按各留出折大小加权的标签生成拟合的平均值，取 $c = 1$ 并应用三角不等式，可得

$$\inf_c \|(g_0 - \widehat{g}) - cf\|_2 \le \|f - (g_0 - \bar{g})\|_2 + \|\widehat{g} - \bar{g}\|_2.$$

这里，第一项衡量学习特征对折平均拟合误差的预测效果；而要将这一速率迁移到最终重新拟合 $\widehat{g}$，还需要控制稳定性差距 $\|\widehat{g} - \bar{g}\|_2$。

2. 协变量划分与残差拟合回归树。一种直接的非参数替代方案是利用训练数据将协变量空间划分为具有正总体概率的 k 互不相交的单元 $\mathcal{C}$（在平均处理效应问题中，再与处理组交叉）。这些单元可以来自领域选定的分位数，如我们的 401(k) 划分分析；也可以来自一棵拟合于 Y 和 D的标准化训练残差的浅层回归树，如我们的 ACIC 划分实验。总体单元投影核为

$$K_{\mathcal{C}}(x, x') := \sum_{C \in \mathcal{C}} \frac{\mathbf{1}\{x \in C\}\mathbf{1}\{x' \in C\}}{P_X(C)}. \tag{21}$$

其积分算子在每个单元内对函数取平均，算子范数为1，其希尔伯特–施密特范数的平方为 $\|K_{\mathcal{C}}\|_{\mathrm{HS}}^2 := \mathbb{E}[K_{\mathcal{C}}(X,\ X')^2] = k$。将 $P_X(C)$ 替换为训练单元频率 $\widehat{p}_C$（通过最小叶节点大小和可选的伪计数进行稳定化），即可根据 $O(n + k)$ 中的单元总量直接计算校正项，耗时为常数（注1）。稀有单元会放大 $\widehat{p}_C^{-1}$ 并引入 Gram 复现误差，因此在实践中设置最小叶节点大小很重要。使用总体归一化时，对多棵树的核取平均也会产生正的收缩。

3. 局部平滑器与再生核希尔伯特空间岭核。当预期存在光滑的非线性误差，且没有硬单元边界时，我们可以对相邻协变量值进行平滑。局部核平滑器使用

$$K_h(x, x') := \frac{\kappa_h(x - x')}{\mathbb{E}[\kappa_h(x - X)]}, \tag{22}$$

其分母通过训练协变量估计（参见 [9] 二阶目标估计）。或者，对于再生核希尔伯特空间（RKHS） $\mathcal{H}$，其有界核为 k、特征映射为 $k_x := k(x,\ \cdot)$、协方差算子为 $C := \mathbb{E}[k_X \otimes k_X]$，岭惩罚为$\lambda > 0$，其等价核为

$$K_\lambda(x, x') := \langle k_x, (C + \lambda I)^{-1} k_{x'} \rangle_{\mathcal{H}}. \tag{23}$$

若 T_0 表示 k 在 $L^2(P_X)$ 上的积分算子，其特征对为 $(\mu_\ell,\ e_\ell)$，则 T_λ 会按照以下方式对每个特征方向进行平滑收缩

$$T_\lambda := T_0(T_0+\lambda I)^{-1}, \qquad T_\lambda e_\ell = \frac{\mu_\ell}{\mu_\ell+\lambda} e_\ell, \tag{24}$$

有效自由度为 $\|K_\lambda\|_{\mathrm{HS}}^2 = \sum_\ell \{\mu_\ell/(\mu_\ell+\lambda)\}^2$。更一般地，对于任何具有完备正交归一特征基（包括零空间）的自伴平方可积核，其特征对为 $(s_\ell,\ e_\ell)$，误差投影为 $a_\ell := \langle \delta_\alpha,\ e_\ell\rangle$ 和 $b_\ell := \langle \delta_g,\ e_\ell\rangle$，其谱展开

$$b_K = -\sum_\ell (1-s_\ell) a_\ell b_\ell, \qquad \|\widehat{K}\|_{\mathrm{HS}}^2 = \sum_\ell s_\ell^2 \tag{25}$$

展示了不同谱方向上的偏差–方差权衡。如果两个干扰误差都属于 $\mathcal{H}$，则谱演算给出显式的偏差界

$$|b_{K_\lambda}| = |\langle \delta_\alpha, (I-T_\lambda)\delta_g\rangle| \le \lambda \|\delta_\alpha\|_{\mathcal{H}} \|\delta_g\|_{\mathcal{H}}. \tag{26}$$

减小 λ 会使该偏差界更紧，但代价是增大 $\|K_\lambda\|_{\mathrm{HS}}^2$ 以及 $\widehat{C}$ 中的估计噪声。在 λ 固定时，剩余偏差仍为二次项。

最后，任何核 $\widehat{K}$ 也可以通过替换而向常规双重机器学习收缩 K 与 ρ $\widehat{}$ K对于训练集选择的因子 $\rho \in [0,1]$，该因子将成对方差缩放为 ρ^2，同时保留所表示混合偏差的 $1-\rho$（在 $\rho=0$时恢复常规双重机器学习）。

3.2. 学习得到的方向的偏差率条件。 随着初始回归的改进，即使是零预测器在预测残差均值时也能实现渐近消失的 L^2 误差，因此，对于偏差校正而言，关键在于初始回归误差中有多少在投影到学习特征后仍然存在。给定一个训练集选择的特征 f，满足$\nu_f := \mathbb{E}[f(X)^2] > 0$，以及一个为正的训练估计值 $\widehat{\nu}_f$，定义

$$\Delta_f := \inf_{a\in\mathbb{R}} \|\delta_g - a f\|_2 = \|(I-\Pi_f)\delta_g\|_2, \qquad \omega_f := |\nu_f/\widehat{\nu}_f - 1|.$$

这里，Δ_f 衡量 f 在忽略尺度和符号的情况下对回归误差形状的捕捉程度，而ω_f 衡量核所使用的逆二阶矩中的相对误差。

C推论 6（秩一校正后的剩余偏差）。在命题 *2*的条件下，核 $K_f(x,x') = f(x)f(x')/\widehat{\nu}_f$ 满足

$$\begin{aligned} |b_{K_f}| &\le \|(I-\Pi_f)\delta_\alpha\|_2\, \Delta_f + \omega_f \|\Pi_f \delta_\alpha\|_2 \|\Pi_f \delta_g\|_2 \\ &\le \|\delta_\alpha\|_2 \big(\Delta_f + \omega_f \|\delta_g\|_2\big). \end{aligned} \tag{27}$$

如果任一上界为 $o_p(n^{-1/2})$，则在满足其得分和抽样变异条件时，命题 *8* 的预言机展开成立。

只要 f 捕捉到任一干扰误差，（27）中的第一个界就可能很小；而第二个界则为仅基于结果残差训练的学习器提供了直接判据。固定初始的干扰项拟合，并使用 m 个新的训练观测值，通过最小二乘法将 $Y-\widehat{g}(X)$ 回归到由 p 个特征$\xi_p(X)$ 构成的固定字典上，其中回归目标为 $v := g_0 - \widehat{g} = -\delta_g$。令 a_p 为一个确定性上界；该上界以趋于 1 的概率成立，适用于逼近误差 $\inf_\beta \|v - \xi_p'\beta\|_2$，其中所有 p 系数都可以非零。然后，我们再使用ℓ 个独立的协变量观测来计算 $\widehat{\nu}_f := \ell^{-1}\sum_{j=1}^{\ell} f(X_j)^2$,个观测值留作评估，并令 $m,\ \ell,\ n\to\infty$。

TABLE 1应用工作中校正核的实用比较

核族	校正方向	训练数据上的选择 训练数据	评估 cost	示例
基底投影 / 高维回归	张成空间 $b(X)$（样条，交互项、协变量）	基函数维数 k 或稀疏性	$O(nk^2)$ 用于稠密的 M	典范 HOIF [21]，部分线性模型。
定向秩一 (20)	单个学习得到的残差预测器 $f(X)$	f 的学习器（仅用于训练）	$O(n)$	一个已学习的残差方向含有许多协变量。
单元划分 (21)	单元内／叶节点平均值（与 D）	叶节点数量 k 和最小叶节点大小	$O(n+k)$，用于单元总量	ATE 与分位数单元或残差-树叶节点。
局部平滑器 (22)	平滑局部协变量邻域	带宽 h	$O(n^2)$	Low-维度的连续 X 与光滑非线性误差。
RKHS岭回归 (23)	主导谱模态核 k	岭惩罚 λ 和核宽度	$O(n^2)$ 核评估	光滑非参数度量校正项没有硬单元边界。

假设 $\xi_p(X)$ 中的特征一致有界，其总体 Gram 矩阵的特征值有上界且远离零，并且 $p = o(\sqrt{m})$，其中 $\mathbb{E}[(Y - \widehat{g}(X))^2 | \widehat{g}, \widehat{\alpha}] = O_p(1)$。为校准非零的 f，进一步假设 $\mathbb{E}[f(X)^4]/\nu_f^2 = O_p(1)$ 给定拟合特征成立（例如，独立的对称 $\{-1, 1\}$ 特征具有单位 Gram 矩阵，并且任意非零线性组合的四阶矩比至多为 3）。若对于 $j \in \{g, \alpha\}$，$\|\delta_j\|_2 = O_p(r_j)$，则

$$|b_{K_f}| = O_p\Big(r_\alpha \left[a_p + \sqrt{p/m} + r_g/\sqrt{\ell}\right]\Big). \tag{28}$$

方括号中的前两项体现了近似和拟合残差回归的成本，而最后一项则体现了一维投影的校准成本。

为说明这一优势，令 m，$\ell \asymp n$ 且 $r_g = r_\alpha = n^{-1/5}$。当两个误差都等于 $n^{-1/5}u$（对于单位函数 u）时，常规双重机器学习的绝对偏差为 $n^{-2/5}$。在对称符号设计中，$u = p^{-1/2}\sum_{j=1}^{p}\xi_{pj}$ 对每个特征都赋予权重，并得到 $a_p = 0$。更一般地，只要 $a_p = o(n^{-3/10})$ 且 $p = o(n^{2/5})$，式（28）就给出 $|b_{K_f}| = o_p(n^{-1/2})$。这样一来，$p$ 可以在残差学习阶段增大，而评估样本只需估计一个偏差贡献。

在这个例子中，常规双重机器学习也可以通过在训练样本上更新回归来达到可忽略偏差：对于 $\widetilde{g} := \widehat{g} + f$，回归误差为 $f - v$，因此式（4）将绝对偏差界定为 $\|\delta_\alpha\|_2 \|f - v\|_2 = O_p\{r_\alpha (a_p + \sqrt{p/m})\}$。这种训练更新依赖于 f 的拟合幅度，而精确的秩一投影仅取决于其方向，因为 Π_f 对非零缩放保持不变；当 f 经过正则化，或以任意尺度学习时，这一点尤为重要。第 7.3 节比较了沿同一学习得到的方向进行的训练更新和评估更新。

3.3. 特殊情形：高维回归。有限基构造可直接应用于采用任意第一阶段线性学习器的高维部分线性回归。考虑如下模型

$$Y = D\beta_0 + X'\vartheta_0 + \epsilon, \qquad \mathbb{E}[\epsilon \mid D, X] = 0, \qquad D = X'\gamma_0 + \nu, \quad \mathbb{E}[X\nu] = 0, \tag{29}$$

并将 $\varphi_0 := \vartheta_0 + \beta_0\gamma_0$ 记为约化式结果变量的系数。在训练样本上，使用任意惩罚型或后正则化线性学习器（例如岭回归 [16], Lasso [23], 后 Lasso [2], 或 lava 和后 lava [7], ）估计 γ_0 和 φ_0，并构造第一阶段残差

$$v := D - X'\widehat{\gamma}, \qquad u := Y - X'\widehat{\varphi}, \qquad a_\beta := u - \beta v.$$

常规双重机器学习（Robinson）得分是样本均值 $\mathbb{P}_n[a_\beta v]$。给定精度矩阵 $\Sigma^{-1} = \mathbb{E}[XX']^{-1}$ 的训练估计值 $\widehat{\Theta}$，非对角U统计量校正为

$$\widehat{M}_U(\beta) := \mathbb{P}_n[a_\beta v] - \frac{1}{n(n-1)} \sum_{i \neq j} a_{\beta,i} X_i' \widehat{\Theta} X_j v_j. \tag{30}$$

要看出 (30) 是 (7) 的一个精确实例，令 $Z_\beta := Y - \beta D$，将注意力限制在线性张成空间 X 上，并定义

$$\widehat{g}_\beta(x) := x'(\widehat{\varphi} - \beta\widehat{\gamma}), \qquad m(W, f) := -Df(X), \qquad \widehat{\alpha}(x) := -x'\widehat{\gamma},$$

并将 DZ_β 加到得分中，将 $\mathbb{E}[DZ_\beta]$ 加到目标量中。在这一替换下，普通双重机器学习得分可简化为 $DZ_\beta - D\widehat{g}_\beta(X) + \widehat{\alpha}(X)\{Z_\beta - \widehat{g}_\beta(X)\} = va_\beta$，其沿任意线性函数 f 的 Riesz 差异为 $vf(X)$。选择核 $\widehat{K}(x, \widetilde{x}) = x^\top \widehat{\Theta}^\top \widetilde{x}$ 并交换求和指标，即可复现（30）；这里保留转置，是因为稀疏化的逆Gram估计量不一定是对称的。线性投影在此已足够，因为 $\mathbb{E}[X\{Z_\beta - X'(\varphi_0 - \beta\gamma_0)\}] = 0$ 以及 $\mathbb{E}[X(D - X'\gamma_0)] = 0$。这些是线性核所检验的唯一残差矩。

将两个干扰误差写成 $e_\gamma := \widehat{\gamma} - \gamma_0$ 和 $e_\varphi := \widehat{\varphi} - \varphi_0 - \beta_0 e_\gamma$，在 β_0 处计算期望得分，得到

$$\mathbb{E}[\widehat{M}_U(\beta_0)] = e_\varphi'(\Sigma - \Sigma\widehat{\Theta}\Sigma)e_\gamma, \qquad \Sigma := \mathbb{E}[XX'], \tag{31}$$

这就是矩阵形式的命题 2。当精度矩阵是精确的（$\widehat{\Theta} = \Sigma^{-1}$）时，$\widehat{\ }$矩阵因子 $\Sigma - \Sigma\Theta\Sigma$ 消失，且校正得分在 β_0 处具有零均值，无论 $(\widehat{\gamma}, \widehat{\varphi})$ 来自岭回归、Lasso、后 Lasso、lava 还是后 lava。这扩展了 Chetverikov、Sørensen 和 Tsyvinski [8], 提出的三重 Lasso，他们专门关注第一阶段的 Lasso 拟合。

*U*统计量与*V*统计量的实现。Chetverikov、Sørensen 和 Tsyvinski [8] 将三重 Lasso 实现为样本均值的乘积（V统计量），

$$\widehat{M}_V(\beta) = \mathbb{P}_n[a_\beta v] - \mathbb{P}_n[a_\beta X']\widehat{\Theta}\mathbb{P}_n[Xv], \tag{32}$$

其中包含对角项 $i = j$。对于任意成对数组 F_{ij}，V统计量与U统计量的平均值之间存在如下精确恒等式

$$\frac{1}{n^2}\sum_{i,j} F_{ij} = \frac{n-1}{n}\frac{1}{n(n-1)}\sum_{i\neq j} F_{ij} + \frac{1}{n^2}\sum_i F_{ii}. \tag{33}$$

在式（30）中排除对角项可消除自身观测偏差，并使我们能够直接应用第4节的无偏方差公式；而对于式（32），随着协变量维度增大，必须单独跟踪对角项$\widehat{\ }$ $\widehat{\ }$ $\widehat{\ }$。为了对 β 的根进行推断 $M_U(\beta) = 0$，通常的局部线性化会将 β_0 处的得分方差除以总体雅可比矩阵的平方，并要求该雅可比矩阵非零。均值为零的得分不一定对应无偏的根。

4. 抽样变异、标准误与交叉拟合。 减去成对校正项$\widehat{\ }$ B_K 会改变估计量的条件均值，也可能改变其抽样方差。本节将这种抽样变异分解为两个精确部分，构造完整条件方差的有限样本无偏估计量，说明常规DML标准误和 K折交叉拟合在何种情况下仍然有效，并提醒注意刀切方差估计量中重复计算的陷阱。

4.1. 抽样变异的两个来源：*Hoeffding*分解。由于 B_K 中的非对角求和遍历所有有序对 $i \neq j$，因此我们也可以对无序对 $i < j$ 上的对称化成对核取平均。将式（7）中的 $H(w,\ w')$ 对称化可得

$$h(w,w') := \tfrac{1}{2}\{H(w,w') + H(w',w)\}, \qquad \mu_h := \mathbb{E}[h(W,W')], \tag{34}$$

因此 $\widehat{B}_K = \mathbb{U}_n h$，其中 *U*-平均 $\mathbb{U}_n h := \binom{n}{2}^{-1}\sum_{i<j} h(W_i,\ W_j)$ 是对所有评估观测的无序对上的对称核取平均。根据Hoeffding正交分解 [15], ，任何对称成对平均都可以分解为单个观测值（线性）平均与纯成对（退化）余项之和。定义

$$L(w) := 2\{\mathbb{E}[h(w,W')] - \mu_h\}, \tag{35}$$

$$d(w,w') := h(w,w') - \mu_h - \tfrac{1}{2}\{L(w) + L(w')\},$$

其中 $\mathbb{E}[L(W)] = 0$ 和 $\mathbb{E}[d(W,\ W') \mid W] = 0$。线性项 $L(w)$ 衡量观测值 w 平均而言在所有独立搭档 W' 中对成对校正产生多大影响，而 $d(w,\ w')$ 则是在剔除单个观测值的影响后剩余的纯成对交互。将线性项与原始DML得分结合，定义校正估计量的有效影响函数，

$$\phi_K(w) := s(w) - \mathbb{E}[s(W)] - L(w). \tag{36}$$

命题 7（精确条件展开与方差）。若 $s \in L^2(P)$ 且 $h \in L^2(P \otimes P)$，则

$$\widehat{\theta}_K - \theta_0 = b_K + \mathbb{P}_n\phi_K - \mathbb{U}_n d, \tag{37}$$

$$\tau_n^2 := \mathrm{Var}(\widehat{\theta}_K \mid \mathcal{T}) = \frac{\sigma_{1,n}^2}{n} + \frac{2\sigma_{2,n}^2}{n(n-1)}, \qquad \sigma_{1,n}^2 := \mathbb{E}[\phi_K^2], \quad \sigma_{2,n}^2 := \mathbb{E}[d^2]. \tag{38}$$

特别是，$\sigma_{1,\ n}^2 = \mathrm{Var}(s) + \mathrm{Var}(L) - 2\mathrm{Cov}(s,\ L)$。

命题 7 中的协方差项 $-2\mathrm{Cov}(s,\ L)$ 解释了为什么简单地将校正项 B_K 的方差加到常规双重机器学习方差上会得到错误的结果。当 $2\mathrm{Cov}(s,\ L) > \mathrm{Var}(L)$ 时，一阶标准误恰好会下降。在包含两个或四个收入单元格的 401(k) 森林拟合中，估计的一阶标准误会下降（表16）。为了解释干扰误差如何进入 L，写作 $f: = T\delta_g$ 以及 $a: = T^*\delta_\alpha$。对 H 中的第二个观测进行积分可得

$$L(W) = m(W,f) - \widehat{\alpha}(X)f(X) + a(X)\{Y - \widehat{g}(X)\} + 2\langle\delta_\alpha, T\delta_g\rangle. \tag{39}$$

因此，L 直接取决于投影后的干扰误差（f、a）以及得分和结果残差的矩。下面命题 8 之后的连续性条件说明了该公式如何界定 $\|L\|_2$。

4.2. *DDML* 何时达到根-n 预言机极限。令 $\psi_0(W): = m(W、g_0) + \alpha_0(X)\{Y - g_0(X)\}$ 表示预言机影响得分，并令 $\phi_0: = \psi_0 - \theta_0$。应用工作中的一个核心问题是，双重去偏机器学习估计量 θ_K 何时具有根-n 渐近正态性，并以该预言机影响函数 ϕ_0为影响函数，从而标准的双重机器学习置信区间可以直接应用于 θ_K，而无需加入成对方差项。命题 8 同时涵盖了两种不同情形。第一，当全局乘积 $\|\delta_\alpha\|_2\|\delta_g\|_2$ 过大而导致常规双重机器学习失效时，只要核消除足够多的相关误差，使 $b_K = o_p(n^{-1/2})$，同时保持 L 和成对项可忽略，双重去偏机器学习仍可达到根-n 预言机极限。第二，当常规双重机器学习满足其通常的乘积速率条件时，算子稳定性会控制剩余偏差。如果校正项的线性波动和成对波动均可忽略，则同样可以得到该预言机极限。

命题 8（根-n oracle 推断的双重去偏机器学习）。假设 $0 < \mathbb{E}[\phi_0^2] < \infty$ 且

$$\|s - \psi_0\|_{L^2(P)} = o_p(1), \qquad \|L\|_{L^2(P)} = o_p(1), \qquad \mathbb{E}[d^2] = o_p(n), \qquad b_K = o_p(n^{-1/2}). \tag{40}$$

则双重去偏机器学习估计量满足

$$\sqrt{n}(\widehat{\theta}_K - \theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_0(W_i) + o_p(1) \rightsquigarrow N(0, \mathbb{E}[\phi_0^2]). \tag{41}$$

偏差条件 $b_K = o_p(n^{-1/2})$ 要么由方向性偏差消除（12）成立，如残差回归速率（28）所示；要么由 $\|T\|_{\mathrm{op}} = O_p$（1）结合标准的双重机器学习乘积速率 $\|\delta_\alpha\|_2\|\delta_g\|_2 = o_p(n^{-1/2})$ 成立。

双重去偏机器学习核的允许复杂度。成对核的二阶矩决定了在（40）下工作空间维数 k 可以有多大。假设以趋近于一的概率，$\|\widehat{\alpha}\|_\infty \le C$、$\mathbb{E}[\varepsilon_g^2 \mid X] \le C$，且该泛函是 L^2-连续的，

$$\mathbb{E}[m(W,f)^2] \le C\|f\|_2^2, \tag{42}$$

在一致重叠性下，平均处理效应满足这一条件，但导数泛函可能不满足。如果两个干扰函数学习器都是 L^2-一致的，且 $\|T\|_{\mathrm{op}} = O_p(1)$，那么（39）立即推出 $\|L\|_2 = o_p(1)$。在给定 X' 的条件下，成对二阶矩由该核的希尔伯特–施密特范数的平方界定，

$$\mathbb{E}[d^2] \le \mathbb{E}[h^2] \le \mathbb{E}[H^2] \lesssim \|\widehat{K}\|_{\mathrm{HS}}^2 := \mathbb{E}[\widehat{K}(X,X')^2]. \tag{43}$$

对于秩为k 的正交投影，$\|T\|_{\mathrm{op}} = 1$ 和 $\|\widehat{K}\|_{\mathrm{HS}}^2 = k$。由于退化项是在 $\binom{n}{2} \asymp n^2$ 个不相关对上取平均，其条件方差为 $O_p(k/n^2)$，其

标准差为 $O_p(\sqrt{k/n})$。选择 $\bar{k}=o(n)$，可使双重去偏机器学习的成对噪声相对于 $n^{-1/2}$ 抽样尺度渐近可忽略。当$\sigma_{1,n}^2 \le C_1$ 且$\sigma_{2,n}^2 \le C_2 k$时，结合式（38）与命题3，可得条件均方根误差界

$$\left\{\mathbb{E}[(\widehat{\theta}_{\Pi}-\theta_0)^2 \mid \mathcal{T}]\right\}^{1/2} \le \rho_\alpha \rho_g + \sqrt{\frac{C_1}{n}+\frac{2C_2 k}{n(n-1)}}. \tag{44}$$

仅凭这些方差界和全局误差半径，$k \lesssim n$ 的均方根误差至多为 $\epsilon_\alpha \epsilon_g + n^{-1/2}$ 量级。只要 $k=o(n)$，且工作空间之外的残差误差乘积满足 $\rho_\alpha \rho_g = o(n^{-1/2})$，剩余偏差和成对噪声都会以 $n^{-1/2}$ 的尺度趋于零，因此即使常规双重机器学习失效，得分条件和（40）中的线性项条件仍能给出均方根-n 预言机极限。当 k 足够大（或干扰误差足够大），以致 L 或成对项 $\mathbb{U}_n d$ 不再可忽略时，完整方差估计量可在满足方差一致性条件和第4.3节和第4.4节的中心极限定条件下进行高斯推断。

4.3. 实际计算完整标准误。为同时估计一阶方差分量和成对方差分量，我们将普通双重机器学习得分与校正项结合在一个单一的对称核中，

$$q(w,w') := \tfrac{1}{2}\{s(w)+s(w')\} - h(w,w'), \qquad \widehat{\theta}_K = \mathbb{U}_n q. \tag{45}$$

$\widehat{\ }$将 θ_K 写成单个 U统计量 $\mathbb{U}_n q$ 意味着其 Hoeffding方差会自动纳入方差(s)、方差(L)、 -2协方差(s,L)以及方差$(\mathbb{U}_n d)$。令 $q_{ij} :=q(W_i,\ W_j)$ 并定义三个总体矩

$$A := \mathbb{E}[q_{12}^2], \qquad B := \mathbb{E}[q_{12}q_{13}], \qquad C := \mathbb{E}[q_{12}q_{34}] = (\mathbb{E}q_{12})^2, \tag{46}$$

其满足 $\sigma_{1,\ n}^2 = 4(B-C)$ 以及 $\sigma_{2,\ n}^2 = A-2B+C$。下降阶乘 ($n_{\ r} = n\ n-1\ \cdots\ n-r+1$ r $n \ge 4$) : ( ) ( ) 表示不同评估索引的有序 -元组数量。对于 ，对相应的成对、三元组和四元组取平均可得

$$\widehat{A} := \frac{\sum_{i\neq j} q_{ij}^2}{(n)_2}, \qquad \widehat{B} := \frac{\sum_{i,j,k \text{ distinct}} q_{ij}q_{ik}}{(n)_3}, \qquad \widehat{C} := \frac{\sum_{i,j,k,l \text{ distinct}} q_{ij}q_{kl}}{(n)_4}. \tag{47}$$

命题 9（总体方差的无偏估计量）。在命题 7 的二阶矩条件下，

$$\widehat{V} := \underbrace{\frac{4}{n}(\widehat{B}-\widehat{C})}_{\widehat{V}_1 \text{ (first-order)}} + \underbrace{\frac{2}{n(n-1)}(\widehat{A}-2\widehat{B}+\widehat{C})}_{\widehat{V}_2 \text{ (pairwise)}} \tag{48}$$

满足 $\mathbb{E}[\widehat{V} \mid \mathcal{T}] = \tau_n^2$。

该构造将无偏 U统计量方差估计专门化为 [25]。计算（47）无需遍历三元组或四元组，因为行和与平方和已足够。定义

$$r_i := \sum_{j\neq i} q_{ij}, \qquad S := \sum_i r_i, \qquad Q := \sum_{i\neq j} q_{ij}^2, \qquad R := \sum_i r_i^2.$$

将相等指标与互异指标分开，可将（47）简化为 $O(n^2)$ 次运算，

$$\widehat{A} = \frac{Q}{(n)_2}, \qquad \widehat{B} = \frac{R-Q}{(n)_3}, \qquad \widehat{C} = \frac{S^2-4R+2Q}{(n)_4}. \tag{49}$$

将每个非对角线元素 q_{ij} 减去 $\mathbb{U}_n q$ 进行中心化，不会改变（48），并能提高数值稳定性。对于秩为 k 的核，从特征总和计算所有行和需要 $O(nk)$ 次运算，对角线为 M 时则需要 $O(nk^2)$ 次运算；对于一般的稠密 M，则需要 $\widehat{\ }$ $\widehat{\ }$ $\widehat{\ }$（备注 1）。当 $V>0$时，比值 V_2/V可估计成对占比。

学生化推断的渐近有效性要求比率一致性 $\widehat{V}/\tau_n^2 \to_p 1$，而满足以下条件即可：

$$|\widehat{A}-A| = o_p(n^2\tau_n^2), \qquad |\widehat{B}-B| + |\widehat{C}-C| = o_p(n\tau_n^2). \tag{50}$$

$\widehat{\ }$由于 V 是矩估计的无偏线性组合，而非平方和，因此当核复杂度相对于 n 过高时，在有限样本中偶尔会变为非正值（附录 C.2 的 1,200 次 ACIC 抽样中出现了 2 次）。非正的 V警示我们应检查核复杂度和稀有单元，而不是盲目地将其截断为零。

行投影捷径与重复计数陷阱。当成对噪声可忽略不计时（$\mathbb{E}[d^2]/n = o_p(1)$ 且 $\sigma_{1,\ n}^2$ 远离零），经验行投影（U 统计量刀切法）给出一阶方差的非负估计量：

$$\widehat{\phi}_i := 2\left\{ \frac{1}{n-1}\sum_{j\neq i} q_{ij} - \mathbb{U}_n q \right\}, \qquad \widehat{V}_{\mathrm{lin}} := \frac{1}{n^2}\sum_i \widehat{\phi}_i^2. \tag{51}$$

在 ϕ_K^2 一致可积且成对噪声可忽略不计的条件下，$\widehat{V}_{\mathrm{lin}}$ 一致估计有效一阶方差，同时捕捉 s 和 L。然而，在 V_{lin}之上再单独估计成对方差，会重复计算成对变异，因为经验$\widehat{\ }$行 ϕ_i 已将其纳入其中。在纯退化情形下（$q = d_q$ 均值为零），直接代数推导可得：

$$\mathbb{E}[\widehat{V}_{\mathrm{lin}} \mid \mathcal{T}] = \frac{4(n-2)}{n^2(n-1)}\mathbb{E}[d_q^2], \qquad \mathrm{Var}(\mathbb{U}_n q \mid \mathcal{T}) = \frac{2}{n(n-1)}\mathbb{E}[d_q^2], \tag{52}$$

$\widehat{\ }$因此，V_{lin} 对退化方差的高估幅度趋近于2倍。只要成对噪声不可忽略，就应改用精确的无偏分解式（48）。

4.4. 一般推断情形。当成对项对 τ_n^2的贡献显著时，有效的高斯推断需要一个同时涵盖线性部分和$\widehat{\ }$ $\widehat{\ }$退化部分的联合中心极限定理。为将 $\theta_K - \mathbb{E}[\theta_K \mid \mathcal{T}]$ 表示为鞅差之和，令 $\mathcal{F}_i := \sigma(\mathcal{T}, W_1, \ldots, W_i)$ 并定义

$$\Delta_i := \frac{\phi_K(W_i)}{n} - \frac{2}{n(n-1)}\sum_{j<i} d(W_i, W_j). \tag{53}$$

给定 $\mathcal{F}_{i-1}$，$\phi_K(W_i)$ 以及每个退化项 $d(W_i, W_j)$ 对于 $j<i$ 的条件均值均为$\widehat{\ }$ $\widehat{\ }$零，因此 $\mathbb{E}[\Delta_i \mid \mathcal{F}_{i-1}] = 0$ 和 $\sum_{i=1}^n \Delta_i = \theta_K - \mathbb{E}[\theta_K \mid \mathcal{T}]$。

命题 10（允许额外方差的推断）。假设 $\tau_n > 0$ 其概率趋于 *1*，且

$$\tau_n^{-2}\sum_i \mathbb{E}[\Delta_i^2 \mid \mathcal{F}_{i-1}] \to_p 1, \tag{54}$$

$$\tau_n^{-2}\sum_i \mathbb{E}[\Delta_i^2 \mathbf{1}\{|\Delta_i| > \epsilon\tau_n\} \mid \mathcal{F}_{i-1}] \to_p 0 \quad \textit{for every } \epsilon > 0.$$

若 $\widehat{V}/\tau_n^2 \to_p 1$，则中心化后的估计量满足

$$\frac{\widehat{\theta}_K - \mathbb{E}[\widehat{\theta}_K \mid \mathcal{T}]}{\sqrt{\widehat{V}}} \rightsquigarrow N(0,1). \tag{55}$$

此外，若 $b_K = o_p(\tau_n)$，则

$$\frac{\widehat{\theta}_K - \theta_0}{\sqrt{\widehat{V}}} \rightsquigarrow N(0,1). \tag{56}$$

当 $n\tau_n^2$ 以概率有界且远离零时，所得置信区间的宽度为根-n 阶；否则，其宽度随 τ_n 缩放。

在实践中，当线性分量 $\mathbb{P}_n\phi_K$ 占主导地位，或退化成对分量 $\mathbb{U}_n d$ 的有效秩不断增大（如 Hall [14] 中所示）时，条件方差条件 (54) 成立；而如果固定秩退化分量占主导地位，其极限为非高斯加权卡方-$\widehat{}$平方分布。由于 $\mathbb{E}[\theta_K \mid \mathcal{T}] = \theta_0 + b_K$，全方差定律将评估样本间的变异与训练样本间条件偏差的变异区分开来，

$$\mathrm{Var}(\widehat{\theta}_K) = \mathbb{E}[\tau_n^2] + \mathrm{Var}(b_K). \tag{57}$$

由于 $\widehat{V}$ 估计条件评估方差 τ_n^2，速率条件 $b_K = o_p(\tau_n)$ 使训练集中的随机波动对于学生化极限而言可以忽略。当 b_K 相对于 τ_n 不可忽略时，则需要采用第 5 节中的偏差感知区间。

4.5. 交叉拟合。 校正项可能使偏差移动超过一个标准误，同时仍使标准误在一阶上保持不变。为说明原因，考虑一个评估折 I_k，其大小为 n_k，折内平均值为 $\mathbb{P}_k$；条件化于用于训练其干扰函数和核的数据。式 (35) 中的分解将估计校正项拆分为三部分：

$$\widehat{B}_k = \underbrace{\mu_k}_{\text{平均校正项}} + \underbrace{\mathbb{P}_k L_k}_{\text{单个观测的波动}} + \underbrace{\mathbb{U}_{n_k} d_k}_{\text{成对波动}} .$$

$\widehat{}$从常规双重机器学习中减去 B_k，会使其条件均值移动 $-\mu_k$（足以消除 $n_k^{-2/5}$ 阶的偏差，如第3.2节所示），而在训练拟合固定时，只有剩余的两项会随评估样本而波动。

假设命题 8 在每个折内均成立。由于中心化函数 L_k 的 L^2 范数为 $o_p(1)$，其折内平均值为 $o_p(n_k^{-1/2})$；根据式 (38) 和 $\mathbb{E}[d_k^2] = o_p(n_k)$，退化成对余项同样为 $o_p(n_k^{-1/2})$（在给定训练数据的条件下，对独立的评估样本对取期望）。将这些界与拟合得分向神谕得分的 L^2 收敛，以及剩余的 $o_p(n_k^{-1/2})$ 偏差相结合，得到折内展开：

$$\widehat{\theta}_k - \theta_0 = \frac{1}{n_k}\sum_{i \in I_k} \phi_0(W_i) + o_p(n_k^{-1/2}).$$

虽然每一折都使用各自拟合的干扰函数和校正项，但其主导项都是同一个中心化预言机双重机器学习得分 ϕ_0 的平均值。因此，按评估样本量对各折估计值加权后，即使训练样本彼此重叠，每个观测值仍具有相同的权重，并且可以直接看出哪些得分构成了标准误。

推论 11（采用普通双重机器学习标准误的交叉拟合双重去偏机器学习）。将N 个独立同分布观测值划分为固定数量 K 的折，且划分独立于观测值，并满足 $n_k := \lvert I_k\rvert \asymp N$ 对每个 k 均成立。对于每一折，仅使用 I_k之外的数据拟合干扰函数和校正核。假设命题 *8* 的条件对每一折均成立，评估样本量为 n_k。那么，交叉拟合估计量 $\widehat{\theta}_{\mathrm{CF}} := \sum_{k=1}^{K} (n_k/N)\, \widehat{\theta}_k$ 满足

$$\sqrt{N}(\widehat{\theta}_{\mathrm{CF}} - \theta_0) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \phi_0(W_i) + o_p(1). \tag{58}$$

对于 $i \in I_k$，令 $\widehat{s}_i := s_{-k}(W_i)$ 表示在 I_k 之外训练的干扰项上评估的普通双重机器学习得分 (3) 。其样本方差 $\widehat{\sigma}^2 := (N-1)^{-1} \sum_{i=1}^{N} (\widehat{s}_i - s)^2$，其中$s := N^{-1} \sum_{i=1}^{N} \widehat{s}_i$，满足

$$\widehat{\sigma}^2 \to_p \mathbb{E}[\phi_0^2], \qquad \frac{\sqrt{N}(\widehat{\theta}_{\mathrm{CF}} - \theta_0)}{\widehat{\sigma}} \rightsquigarrow N(0,1).$$

在此情形下，普通的折外得分为以校正后的估计值为中心的区间提供标准误 $\widehat{\sigma}/\sqrt{N}$ 。附录 B 中的证明利用了每个评估折与其自身训练拟合之间的独立性，以及各折共用的神谕得分。

当校正项改变一阶影响函数或产生显著的成对变异时，如何估计合并交叉拟合估计量的方差在此仍是一个悬而未决的问题。命题 9 给出了单个独立训练与评估划分的完整条件方差估计量。对于存在重叠训练样本的情形，合并方差的计算还必须考虑各折之间可能存在的主要协方差。

5. 偏差感知推断与设定错误敏感性分析。在训练样本与评估样本相互独立的情况下，校正项估计两个干扰误差沿所选方向的对齐程度。为了探究校正项可能遗漏了多少误差，我们采用 Chernozhukov、Cinelli、Newey、Sharma 和 Syrgkanis [6] 的双因素敏感性分析。分析者指定两个剩余误差的合理大小。它们的乘积决定了观测数据目标的区间需要扩大多少。

5.1. 通过短目标与长目标敏感性控制模型设定错误偏差。Chernozhukov、Cinelli、Newey、Sharma 和 Syrgkanis [6]将“长”目标量与“短”回归目标之间的差异表示为遗漏结果变量与 Riesz 分量的内积。尽管该框架通常用于介绍未观测混杂，但完全相同的希尔伯特空间几何也支配着双重去偏机器学习中的泛函设定错误和剩余学习偏差。

为以最简单的形式了解这一联系，固定总体投影 Π 到训练选定的工作空间 $\mathcal{H}_k$ ，并定义与 $\mathcal{H}_k$相关联的短目标如下：

$$g_{\mathcal{H}} := \Pi g_0, \qquad \alpha_{\mathcal{H}} := \Pi \alpha_0, \qquad \theta_{\mathcal{H}} := \mathbb{E}[m(W, g_{\mathcal{H}})]. \tag{59}$$

这里， $g_{\mathcal{H}}$ 是在 $\mathcal{H}_k$内对真实回归 g_0 的最佳 $L^2(P_X)$ 近似，而 $\alpha_{\mathcal{H}}$ 是限制在 $\mathcal{H}_k$上的目标泛函的 Riesz 表示元，这意味着对于每个 $h \in \mathcal{H}_k$，都有 $\mathbb{E}[m(W,\ h)] = \langle \alpha_{\mathcal{H}},\ h \rangle$ 。

命题12（短目标与校正中心）。在满足里斯恒等式和平方可积条件的情况下，

$$\theta_0 - \theta_{\mathcal{H}} = \langle g_0 - g_{\mathcal{H}}, \alpha_0 - \alpha_{\mathcal{H}} \rangle. \tag{60}$$

如果 *bo* 拟合干扰函数 $\widehat{g}$ 和 $\widehat{\alpha}$ 属于 $\mathcal{H}_k$，则精确投影校正满足

$$\mathbb{E}[\widehat{\theta}_{\Pi} \mid \mathcal{T}] = \theta_{\mathcal{H}}, \tag{61}$$

而无需要求任一拟合干扰函数等于其总体投影。

当两个拟合干扰函数都位于 $\mathcal{H}_k$ 内时，命题12表明，精确投影^校正会消除 $\mathcal{H}_k$ 内的全部估计偏差，并将 θ_{Π} 居中于短目标 $\theta_{\mathcal{H}}$，仅留下函数错设差距（60）。将该差距按短模型尺度 $\sigma^2_{\mathcal{H}} := \mathbb{E}(Y - g_{\mathcal{H}})^2 > 0$ 和 $\nu^2_{\mathcal{H}} := \|\alpha_{\mathcal{H}}\|_2^2 > 0$ 进行标准化，定义

$$C_{Y,\mathcal{H}} := \frac{\|g_0 - g_{\mathcal{H}}\|_2}{\sigma_{\mathcal{H}}}, \qquad C_{\alpha,\mathcal{H}} := \frac{\|\alpha_0 - \alpha_{\mathcal{H}}\|_2}{\nu_{\mathcal{H}}}, \qquad S_{\mathcal{H}} := \sigma_{\mathcal{H}}\nu_{\mathcal{H}}, \tag{62}$$

并令 $\lambda_{\mathcal{H}} \in [-1, 1]$ 为 $g_0 - g_{\mathcal{H}}$ 与 $\alpha_0 - \alpha_{\mathcal{H}}$ 之间的余弦（若任一余项为零，则设为零）。于是，错设差距可分解为

$$\theta_0 - \theta_{\mathcal{H}} = \lambda_{\mathcal{H}} C_{Y,\mathcal{H}} C_{\alpha,\mathcal{H}} S_{\mathcal{H}}, \tag{63}$$

这是定理 2 在 [6] 中的工作空间类比。这里 $C^2_{Y,\ \mathcal{H}} \leq 1$ 自动成立，因为 $\mathbb{E}(Y - g_{\mathcal{H}})^2 = \mathbb{E}(Y - g_0)^2 + \|g_0 - g_{\mathcal{H}}\|_2^2$；而 $C_{\alpha,\ \mathcal{H}}$ 等于被省略部分 Riesz 表示元的均方根除以其投影的均方根。例如，若 $e \perp \mathcal{H}_k$ 是一个单位向量，且 $g_0 = g_{\mathcal{H}} + ue$ 和 $\alpha_0 = \alpha_{\mathcal{H}} + ve$，则 $\theta_0 - \theta_{\mathcal{H}} = uv$。改进 $\mathcal{H}_k$ 内的系数估计无法消除这一差距，而将 e 加入 $\mathcal{H}_k$ 则可以消除该差距，其方差变化由（38）给出。

扩展到任意机器学习拟合。拟合得到的森林、提升回归或 CausalPFN 预测可能位于工作空间 $\mathcal{H}_k$ 之外。在这种情况下，相同的范数和对齐恒等式适用于留在 $\mathcal{H}_k$之外的残差估计误差。定义

$$r_g := (I - \Pi)\delta_g, \qquad r_\alpha := (I - \Pi)\delta_\alpha, \qquad \overline{\theta}_{\Pi} := \mathbb{E}[\widehat{\theta}_{\Pi} \mid \mathcal{T}], \tag{64}$$

因此命题 2 给出

$$\theta_0 - \overline{\theta}_{\Pi} = \langle r_\alpha, r_g \rangle. \tag{65}$$

为了理解为什么这与短模型与长模型错设差距的表现完全相同，定义总体精化干扰函数

$$g^\dagger := \widehat{g} - \Pi\delta_g = g_0 + r_g, \qquad \alpha^\dagger := \widehat{\alpha} - \Pi\delta_\alpha = \alpha_0 + r_\alpha. \tag{66}$$

这些函数将 g_0 和 α_0 投影到仿射空间 $\widehat{g} + \mathcal{H}_k$ 和 $\widehat{\alpha} + \mathcal{H}_k$ 上，并在$(g^\dagger,\ \alpha^\dagger)$ 处计算普通双重机器学习得分，可得到精确的修正中心：

$$\mathbb{E}\Big[m(W, g^\dagger) + \alpha^\dagger(X)\{Y - g^\dagger(X)\}\Big] = \theta_0 - \langle r_\alpha, r_g \rangle = \overline{\theta}_{\Pi}. \tag{67}$$

将 $R_\alpha := \|r_\alpha\|_2^2$ 和 $R_g := \|r_g\|_2^2$ 分别记作两个残差均方，并记其 L^2 余弦（当 $1 \in \mathcal{H}_k$ 且两个残差方差均为正时，该余弦等于它们的相关系数）

$$\lambda_\perp := \frac{\langle r_\alpha, r_g \rangle}{\sqrt{R_\alpha R_g}} \quad (\text{or } 0 \text{ if } R_\alpha R_g = 0), \tag{68}$$

我们可以将 (65) 改写为

$$\theta_0 - \overline{\theta}_{\Pi} = \lambda_\perp \sqrt{R_\alpha R_g}. \tag{69}$$

拟合校正项的双因素预算。 为将同样的敏感性分析应用于估计投影或正则化核，我们还要考虑其归一化在工作空间内留下的误差。假设拟合算子 T 有界且自伴，这与我们所有的实证比较一致。记 $A:=I-T$ 并定义

(70) $$R_{j,T}:=\langle\delta_j,|A|\delta_j\rangle,\qquad j\in\{g,\alpha\},$$

其中，$|A|$ A A $=$ 取各个 $|A|^{1/2}$ A $|A|^{1/2}$ 11 谱值的绝对值。由于 sgn()和符号算子的范数至多为一，( ) 中的柯西–施瓦茨不等式给出 $|b_K|\le\sqrt{R_{g,T}R_{\alpha,T}}$。对于精确投影，$R_{j,\Pi}=\|(I-\Pi)\delta_j\|_2^2$；而对于秩一校正，$T=c\Pi_f$ 沿 f 方向的平方误差以权重 $|1-c|$ 计入，而与 f 正交的平方误差以权重 1计入。常规双重机器学习对应于 $T=0$，并使用每次拟合的完整平方误差。

为了用特定回归拟合的单位表示结果变量误差预算，我们使用其训练残差均方 $s_Y^2:=\mathbb{P}_{\mathrm{tr}}[(Y-\widehat{g}(X))^2]$, 其中，$\mathbb{P}_{\mathrm{tr}}$ 表示训练平均值。对于加权误差，自然的参照量是总体二阶矩$\mathbb{E}[\alpha_0(X)^2]$, ，它对于所有针对同一泛函的方法都是共同的。遵循 Chernozhukov 等人的正交得分，[6],任何候选表示元 a 均满足

$$\mathbb{E}[2m(W,a)-a(X)^2]=\mathbb{E}[\alpha_0(X)^2]-\|a-\alpha_0\|_2^2$$

根据里斯恒等式 (2)，最小化期望里斯损失 $\mathbb{E}[a(X)^2-2m(W,a)]$，即可选出均方意义下最接近 α_0 的候选项。

我们将训练样本划分为若干折 J_1，……，J_V，并为每个训练划分估计一个共享参考尺度。对于每一折 v，我们根据其补集上的交叉验证里斯损失选择一个参考学习器，并在该补集上重新拟合，得到 $\widehat{a}_{-v}$，从而使 J_v 中的每个观测值都由一个未使用该观测值进行选择和拟合的学习器计算得分。对于 n_{tr} 个训练观测，得到的去偏二阶矩估计为

(71) $$\widehat{\nu}_\alpha:=\frac{1}{n_{\mathrm{tr}}}\sum_{v=1}^{V}\sum_{i\in J_v}\left\{2m(W_i,\widehat{a}_{-v})-\widehat{a}_{-v}(X_i)^2\right\}.$$

每当 $\widehat{\nu}_\alpha>0$ 时，我们都会在该划分的所有方法中固定 $s_\alpha:=\sqrt{\widehat{\nu}_\alpha}$：每种方法在估计方程中仍使用各自的权重，但共享 s_α 会使每个加权误差预算都处于共同尺度上，同时允许 s_Y 随结果拟合而变化。

对于共同的敏感性参数 $0\le q<1$，考虑以下预算

(72) $$R_{g,T}\le q\,s_Y^2,\qquad R_{\alpha,T}\le\frac{q}{1-q}\,s_\alpha^2\quad\Longrightarrow\quad|b_K|\le B_q:=\frac{q}{\sqrt{1-q}}\,s_Y s_\alpha.$$

第一个不等式将结果误差设定为训练残差均方的一定比例，第二个则使用对 [6], 进行双因子里斯校准所得到的优势比 $q/(1-q)$；取几何平均数后，可得到适用于对齐任一方向的偏差界 B_q。单独设定预算 $(q_Y,\ q_\alpha)$ 可得 $s_Y s_\alpha\sqrt{q_Y q_\alpha}/(1-q_\alpha)$，我们设定 $q_Y=q_\alpha=q$，以便在单一尺度上报告每项实证比较。以这些由训练集确定的单位计，$q=1\%$ 表示均方根结果误差最高可达 (70) 中几何结构下 s_Y 的十分之一。未知的干扰误差是否确实满足 (72)，仍属于敏感性假设；而几乎对训练结果变量进行插值的学习器会产生较小的 s_Y，因此在给定 q 时，对结果误差的假设要求更为严格。

注13（扩展至非对称核）。对于有界的非对称核算子，同一界 $|b_K|\le\sqrt{R_{g,T}R_{\alpha,T}}$ 可由极分解 $A=U|A|$ 推出：由于 $A=|A^*|^{1/2}U|A|^{1/2}$ 且 $\|U\|_{\mathrm{op}}\le1$，只需对结果误差令 $R_{g,T}:=\langle\delta_g,\ |A|\delta_g\rangle$，并对加权误差令 $R_{\alpha,T}:=\langle\delta_\alpha,\ |A^*|\delta_\alpha\rangle$。

使用嵌套工作空间进行残差误差基准测试。为评估所提预算的合理性，我们可以比较两个由训练集选择、相互嵌套且对应投影 Π_1 和 Π_2 的工作空间。令 $D := \Pi_2 - \Pi_1$，即可将较粗工作空间的剩余偏差分解为

$$\theta_0 - \overline{\theta}_{\Pi_1} = \langle D\delta_\alpha, D\delta_g \rangle + \theta_0 - \overline{\theta}_{\Pi_2}. \tag{73}$$

两个校正估计值之间的差异衡量了新增方向的偏差贡献。其大小为分析者判断 B_q 提供了参考。若要推断更丰富空间之外部分的上界，则需要一个外推假设。

备注 14（区分模型设定错误偏差与未观测混杂）。本节中的敏感性界控制了观测数据泛函 $\theta_0 = \theta_{\text{obs}}$ 的剩余模型设定错误偏差和学习偏差。如果因果识别也存在疑问，那么允许存在未观测混杂因素会在因果参数 θ_{causal} 与 θ_{obs} 之间引入一个单独的差距，

$$\theta_{\text{causal}} - \overline{\theta}_{\Pi} = \underbrace{\theta_{\text{causal}} - \theta_{\text{obs}}}_{\text{omitted-confounder discrepancy}} + \underbrace{\theta_{\text{obs}} - \overline{\theta}_{\Pi}}_{\text{remaining learning bias}}. \tag{74}$$

更强的学习器或更丰富的工作空间只能缩小第二项。如果分析者分别对 $|\theta_{\text{causal}} - \theta_{\text{obs}}| \le B_{\text{conf}}$ 设定界限，那么将有效的偏差感知区间在每一侧扩大 B_{conf}，即可分别展示这两个不确定性来源。

5.2. 偏差感知置信区间。在评估之前选择 q，并假设（72）中的两个不等式在训练样本上以趋近于一的概率成立。令 $\widehat{\text{se}}_K$ 为条件标准误的相对一致估计量。在命题10的中心化正态近似（55）下，

$$\frac{\widehat{\theta}_K - \mathbb{E}[\widehat{\theta}_K \mid \mathcal{T}]}{\widehat{\text{se}}_K} \rightsquigarrow N(0,1). \tag{75}$$

对于名义覆盖率 $1-\gamma$，将 $z := z_{1-\gamma/2}$ 记为标准正态分位数。去偏差区间为

$$\mathcal{I}_n(q) := [\widehat{\theta}_K - z\widehat{\text{se}}_K - B_q,\ \widehat{\theta}_K + z\widehat{\text{se}}_K + B_q]. \tag{76}$$

只要以中心为基准的 Wald 区间包含 $\mathbb{E}[\theta_K \mid \mathcal{T}]$ 和 $|b_K| \le B_q$，三角不等式便给出 $\theta_0 \in \mathcal{I}_n(q)$。因此，$\liminf_{n\to\infty} \Pr\{\theta_0 \in \mathcal{I}_n(q)\} \ge 1-\gamma$。在此陈述中，预算以实际训练尺度为单位。以 $\mathcal{T}$ 为条件，可固定选定的参考拟合和 s_α，因此，估计共同尺度不会给 B_q 增加任何评估样本不确定性。抽样不确定性仅通过中心体现。若科学问题指定的是以未知总体尺度表示的界，则对其端点进行推断时，还必须考虑该尺度的估计及其与中心的协方差，如 [6] 的定理 4 所示。

对于正的效应估计值，报告下端点达到零时对应的预算很有用。假设 $q=0$ Wald 下端点 $L_0 := \theta_K - z\widehat{\text{se}}_K$ 和尺度乘积 $s_Y s_\alpha$ 均为正。令 $B_q = L_0$，并对 q 求解，可得

$$t := \frac{L_0}{s_Y s_\alpha}, \qquad q^* := \frac{2t}{\sqrt{t^2+4}+t}. \tag{77}$$

该区间支持 $0 \le q < q^*$ 存在正向效应。如果 $L_0 \le 0$，那么即使点估计为正且为 $q=0$，区间也已包含零。

6. 区间的实施与选择。在实证研究中应用双重去偏机器学习，首先要选择目标泛函和工作空间。在训练数据上确定干扰函数和校正核后，评估样本同时承担两项任务：估计偏差校正，并揭示相较于常规双重机器学习，校正项会在多大程度上改变抽样变异。

6.1. 实现。

1. 选择目标泛函，并将样本划分为训练集和评估样本。对于平均处理效应（ATE），请预先明确结果变量、处理对比以及任何倾向得分截尾，并在合并交叉拟合估计量之前参阅第4.5节。2. 在训练数据上，拟合结果回归和 Riesz 表示元，学习方向，并估计其二阶矩（或格拉姆矩阵）。在评估之前，固定这三项拟合结果及归一化。内部划分可用于提供新的残差标签或校准协变量。第7节说明了各项实验采用的分配方式。^3. 在评估样本上，计算常规双重机器学习估计值 θ_{DB}、非对角^U统计量校正项 B_K （见式（7）），以及校正后的估计值 θ_K。并列报告两个点估计，便于直观呈现中心位置的变化。^4. 计算原始 DML 标准误、校正后的一阶方差估计 V_1、^ ^ ^以及式（48）中的完整无偏方差估计 $V = V_1 + V_2$。当两种方差均为正时，报告完整标准误相对于常规双重机器学习的百分比变化^ ^，以及成对占比 V_2/V；若方差估计为非正，则应标记出来，提示检查核复杂度或稀有单元。5. 使用训练尺度和选定的预算 q （见式（72）），以效应量为单位表示容许量 B_q。式（73）中的嵌套工作空间可作为其大小的基准。6. 报告表2中抽样假设和偏差假设所支持的置信区间。当残差学习误差作为敏感性参数纳入时，在模拟中同时报告区间宽度和覆盖率；在应用中则针对不同误差预算报告区间端点。如果还在研究因果识别，可依据备注 14另行加入混杂容许量。

6.2. 选择区间。成对方差占比较小，并不意味着校正项不会改变一阶影响函数。在包含16个收入单元格的401(k)分析中，更新一阶影响函数会使随机森林的标准误从$987升至$1,015。计入成对方差后，标准误升至$1,018。尽管成对项仅占方差的0.55%（表16），整体标准误仍上升了3.1%。这些诊断描述了拟合估计量的表现，而表2汇总的理论条件则决定了哪种置信区间具有形式覆盖保证。

7. 实证应用与蒙特卡罗证据。在 401(k) 应用中，我们使用相同的家庭，比较不同学习器下估计效应及其置信区间的变化；而 ACIC 模拟提供了已知的处理效应，因此我们可以衡量实证覆盖率，以及为容纳残余偏差所需的区间宽度。两项分析均依赖独立的训练与评估划分，而高维线性模拟则比较五折交叉拟合估计量在匹配抽样上的表现。表3总结了四种设定中使用的内部样本分割分配。

表2 2为原始目标选择区间

条件	标准误和区间
普通影响函数和 可忽略偏差	在（40）下，使用普通双重机器学习标准误和 Wald 区间 。条件包括得分收敛、可忽略的 L、可忽略的 成对噪声，以及 $b_K = o_p(n^{-1/2})$。对于固定折交叉 拟合，应用推论 11。
变更后的影响函数或 显著成对噪声	在独立评估、中心化中心极限定理、比率相合的 完整方差估计以及 $b_K = o_p(\tau_n)$ 的条件下，使用 Wald 区间， 并采用 $\sqrt{\widehat{V}}$。如果只需考虑一阶项，则一致的基于行的 标准误就足够了。合并限制请参见第4.5节。
残差学习误差满足 双因素预算	在中心化中心极限定理和一致学生化条件下，加入训练集 误差项 B_q 至每个半宽度中，如式（76）所示。对于 观测数据目标量，在指定的预算下，覆盖率成立。

表 3经验研究与模拟研究中的样本分割分配

实验	总量 / 评估	基线拟合	方向学习和范数 校准
401(k) 主体（第7.1节）	9,915 / 4,958	全部 4,957 个训练 集	森林：前一折上的二折残差 3,717。PFN上下文：前1 ,800. 范数：后1 ,240.
ACIC线性或提升 （第7.2节）	4,800 / 2 ,400	全部 ,400 个训练 2 集	森林：前 1,800 个的二折残差 PFN上下文：前 1,800 个。 范数：最后 600 个。
ACIC PFN基线 （第7.2节）	4,800 / 2 ,400	PFN 1 ,800, 约 2,400	基于残差的森林方向 在接下来的 400 个样本上校准范数 最终 200。
匹配更新 （第7.3节）	401(k) 4,958, ACIC 2,400	401(k) 3,717, ACIC 1,800	斜率和 范数的校准划分：1,240（401(k)）或 600 （ACIC），PFN上下文为 1,800。
高维线性 （第 7.4–7.5 节）	1,000（5 折）	800 个训练观测值 / 200 eval	在相同的 800 个训练观测值上使用残差学习器 ，在 400 个训练观测值上计算范数 独立未标记 协变量。

7.1. *401(k)*资格与家庭金融资产。我们使用来自 1991年收入与项目参与调查 [5, 11] 的 9,915 个家庭，估计 401(k)资格对家庭净金融资产的影响。全文将净金融资产限制在–$50,000 和 $250,000，并采用标准的九个家庭协变量，将数据划分为 4,957 个训练家庭和 4,958 个评估家庭。要对调整后的对比进行因果解释，需要满足条件可交换性、一致性和重叠性。我们比较两种基准双重机器学习设定：一种是简单的参数化设定，在每个资格组内使用线性结果回归，并采用逻辑倾向得分模型；另一种是灵活的设定，使用随机森林处理两个干扰函数。

按照表3中的分配，我们在全部4,957个训练家庭上进行基线干扰函数拟合。为学习森林方向，我们将前3,717个训练家庭平分为两组，在每组之外拟合基线以形成折外残差，并在合并后的残差上训练组别特定森林。对于CausalPFN方向，我们在前1,800个训练家庭上拟合处理组均值并在

其上减去全训练基线预测；两种方向都在剩余的1,240个训练家庭上估计二阶矩（这些家庭也会纳入最终的基线拟合）。随后，我们使用全训练干扰函数在留出的4,958个家庭上进行一次评估，并加入第三种比较：将相同的CausalPFN处理组均值直接代入使用原始倾向得分的常规双重机器学习中。

表格 4*401(k)*资格对封顶资产的影响，含名义 *95%* 区间

估计量	估计值	SE	95%置信区间	变化	变化的标准误	标准误缩减（%）
线性结果变量，逻辑倾向得分						
常规双重机器学习	5,914	1,090	[3,778, 8,050]	—	—	—
双重机器学习，直接PFN结果变量	8,680	1,046	[6,630, 10,730]	+2,766	438	4.0
双重去偏机器学习，PFN方向	8,722	1,020	[6,722, 10,722]	+2,808	561	6.4
双重去偏机器学习，森林方向	7,763	1,049	[5,708, 9,818]	+1,849	321	3.8
森林结果与倾向评分						
常规双重机器学习	7,559	987	[5,625, 9,493]	—	—	—
双重机器学习，直接PFN结果	7,446	983	[5,518, 9,373]	-113	232	0.3
双重去偏机器学习，PFN方向	7,446	983	[5,520, 9,372]	-113	224	0.4
双重去偏机器学习，森林方向	7,613	998	[5,657, 9,569]	+54	54	-1.2

除百分比外，所有数值均以1991年美元计。每个面板内的所有估计量都使用相同的倾向得分和4,958个评估家庭。变化和标准误缩减均相对于该面板的原始DML拟合值计算。Wald区间采用DML的普通得分标准误和DDML的完整条件标准误，并假设剩余偏差可以忽略。

在固定逻辑倾向得分的情况下，普通线性DML的估计值为$5,914，而直接CausalPFN将估计值提高至$8,680，秩一PFN校正则给出$8,722，且区间几乎相同（表4）；森林方向使线性估计值提高了$1,849。相比之下，在使用随机森林初始拟合时，最大的变化仅为$113，而普通标准误为$987。尽管所有以零为中心的区间都不包含零，但这里的真实效应未知，因此仅凭这些变化无法证实偏差有所降低。

森林倾向得分在全部三个训练折中均取得较低的验证 Riesz 损失，因此得到共享参考尺度 $s_\alpha = 2.443$。在 $q = 1\%$ 时，线性结果拟合在区间两侧各给出 740 美元的容许幅度。应用PFN方向后，下端点从 3,038 美元升至 5,983 美元（表5）。在 $q = 2\%$ 时，校正区间为 5,235 至 12,209 美元。对于常规线性双重机器学习，下端点在预算为 5.00% 时达到零；对于PFN方向和直接 PFN 结果替代，分别在预算为 8.73% 和 9.08% 时达到零。使用森林初始拟合时，常规双重机器学习及两个学习得到的方向的阈值介于 7.81% 至 8.00% 之间。直接 PFN 结果替代在两个面板中的容许幅度相同，因为其结果拟合和 Riesz 参考尺度都是共享的。

表 *5401(k)*效应的剩余偏差容许值和区间

	$q = 1\%$		$q = 2\%$		
估计量	B_q	95%区间	B_q	95% 区间	q^* (%)
线性结果变量，逻辑倾向得分					
常规双重机器学习	740	[3,038, 8,790]	1,487	[2,291, 9,537]	5.00
双重机器学习，直接PFN结果变量	700	[5,931, 11,429]	1,406	[5,224, 12,136]	9.08
双重去偏机器学习，PFN方向	740	[5,983, 11,462]	1,487	[5,235, 12,209]	8.73
双重去偏机器学习，森林方向	740	[4,968, 10,558]	1,487	[4,221, 11,305]	7.46
森林结果与倾向评分					
常规双重机器学习	682	[4,943, 10,175]	1,370	[4,255, 10,863]	7.96
双重机器学习，直接PFN结果	700	[4,819, 10,073]	1,406	[4,112, 10,779]	7.62
双重去偏机器学习，PFN方向	682	[4,839, 10,054]	1,370	[4,150, 10,743]	7.81
双重去偏机器学习，森林方向	682	[4,975, 10,251]	1,370	[4,286, 10,940]	8.00

容许量 B_q 和区间端点均以1991年美元计。(72) 中两个干扰误差预算均适用相同的 q。所有行均采用去偏 Riesz 尺度，并使用各自结果拟合的残差尺度以及表4中的标准误。对于 $q < q^*$，区间不包含零（阈值经过四舍五入）。

7.2. 初始拟合较强时的 *ACIC* 模拟。在 ACIC 设置 27 和 29 中，四个固定响应面分别规定了 4,802 条协变量记录上的条件结果均值和处理概率，并在重复实验中保持不变。每次重复都会抽取相互独立的训练样本和评估样本，各包含 2,400 个观测值，并采用表3汇总的内部训练划分。对于线性或提升基线，森林方向和 PFN方向遵循与 401(k) 应用相同的流程；而在 CausalPFN 提供初始结果回归时，我们将其在 1,800 个观测值上的拟合固定下来，利用接下来的 400 个观测值上的残差学习森林方向，并在最后 200 个观测值上校准其范数，同时使用全部 2,400 个训练观测拟合倾向得分。为构造式 (71) 中的共享敏感性尺度，我们将训练集分成三折，并在每个由两折组成的补集上，根据双折 Riesz 损失选择倾向评分学习器（在 ACIC 中比较逻辑回归与提升法，在 401(k) 数据中则比较逻辑回归与森林）。评估样本仅用于点估计和方差估计；每个秩一双重去偏机器学习区间均使用完整方差估计（48），在中心化高斯近似和可忽略的剩余偏差下，这可实现名义 Wald 覆盖率。

表 6*ACIC* 上秩一森林校正前后的均方根误差和标准误

设定	结果变量 / 倾向得分	抽样	DML	DDML	均方根误差缩减（%）	标准误缩减（%）
27	CausalPFN / 提升法	20	0.02976	0.02921	1.9	0.7
	CausalPFN / 逻辑回归	20	0.03075	0.03023	1.7	0.8
	提升法 / 提升法	200	0.02222	0.02187	1.6	1.3
	线性 / 逻辑回归	200	0.06681	0.03781	43.4	22.4
29	CausalPFN / 提升法	20	0.02050	0.01997	2.6	0.8
	CausalPFN / 逻辑回归	20	0.02682	0.02629	2.0	0.3
	提升法 / 提升法	200	0.02832	0.02790	1.5	1.8
	线性 / 逻辑回归	200	0.09756	0.06086	37.6	22.9

DML 和 DDML 列报告的是以边际结果标准差表示的均方根误差。两项百分比降幅均以常规双重机器学习为基准。对于标准误，我们计算每个配对抽样中估计标准误的百分比降幅，然后对各次抽样取平均值。每种设置使用四个固定响应面；对于 CausalPFN，每个响应面抽取五次，对于其他学习器则抽取五十次。

使用线性/逻辑干扰函数时，森林方向在设置27中使均方根误差降低43.4%，在设置29中降低37.6%（表6）。使用提升干扰函数时，两种设置下的降幅均低于2%。表7考察区间是否覆盖真实效应。在每一对方法中，双重机器学习和双重去偏机器学习具有相同的初始结果拟合，并共享训练尺度 s_Y 和 s_α。因此，在式（72）中的每个预算 q下，我们都为每种方法各自的Wald半宽度加上相同的 B_q。对于常规双重机器学习，预算适用于完整干扰误差（$T=0$）；对于双重去偏机器学习，预算适用于按 $|I-T|$加权的误差。宽度和容许量均以各曲面上的边际结果标准差为单位。

表7*ACIC* 上常规双重机器学习与秩一*DDML*的覆盖率和平均区间宽度

			$q=0$		$q=1\%$		$q=2\%$	
初始拟合 / 方向	抽样	方法覆盖率		宽度	Cov.	宽度	Cov.	宽度
设置*27*								
线性回归、逻辑回归／森林	200	DML	74.0	0.143	79.0	0.166	83.5	0.189
		DDML	88.0	0.111	93.5	0.133	96.0	0.156
提升 , 提升树 / 森林	200	DML	93.0	0.081	96.0	0.088	99.0	0.096
		DDML	94.0	0.080	96.5	0.087	99.0	0.095
线性回归、逻辑回归／PFN	20	DML	85.0	0.137	85.0	0.159	90.0	0.182
		DDML	80.0	0.088	90.0	0.111	95.0	0.134
PFN、逻辑回归／森林	20	DML	85.0	0.084	85.0	0.092	90.0	0.099
		DDML	85.0	0.084	85.0	0.091	90.0	0.099
PFN, 提升树 / 森林	20	DML	90.0	0.082	90.0	0.090	90.0	0.097
		DDML	90.0	0.082	90.0	0.089	90.0	0.097
设置*29*								
线性回归、逻辑回归／随机森林	200	DML	52.5	0.152	63.5	0.175	71.0	0.198
		DDML	67.0	0.118	74.0	0.141	86.5	0.164
提升 , 提升树 / 森林	200	DML	84.5	0.083	91.5	0.094	95.5	0.104
		DDML	84.5	0.082	91.0	0.092	94.5	0.103
线性、逻辑回归 / PFN	20	DML	65.0	0.145	70.0	0.167	70.0	0.189
		DDML	95.0	0.095	100.0	0.117	100.0	0.139
PFN、逻辑回归 / 森林	20	DML	95.0	0.083	95.0	0.090	100.0	0.097
		DDML	95.0	0.082	100.0	0.090	100.0	0.097
PFN, 提升树 / 森林	20	DML	95.0	0.080	95.0	0.088	95.0	0.095
		DDML	95.0	0.079	95.0	0.087	95.0	0.094

覆盖率是区间包含真实效应的配对抽样所占百分比。宽度是以边际结果标准差为单位的完整区间平均宽度。在每次抽样中，两种方法都会在各自的 Wald 半宽度上加上相同的 B_q。第一列中的方向由双重去偏机器学习使用。不可用的区间计为未覆盖，且不纳入平均宽度的计算。表 19 报告了 B_q 覆盖双重去偏机器学习绝对条件偏差的频率。

在线性／逻辑回归拟合下，森林校正将设置27中的覆盖率从74%提高到88%，将设置29中的覆盖率从52.5%提高到67%，同时缩短了两个设置中的区间。在$q=2\%$水平下，常规双重机器学习在设置27的83.5%抽样中实现覆盖，在设置29的71.0%抽样中实现覆盖。在相同的容许偏差下，双重去偏机器学习分别达到96.0%和86.5%的覆盖率，且平均区间更短（设置27中为0.156，而非0.189；设置29中为0.164，而非0.198）。PFN 试点还允许对其20次抽样进行匹配比较。在设置29且采用逻辑倾向得分时，将PFN方向应用于线性拟合，以及直接 PFN 结果替代，两者都在20次抽样中的19次实现覆盖，平均宽度分别为0.095和0.083。每个表面只有五次抽样，因此这些试点比较仅供描述。

在设置29采用提升的初始拟合时，两个未调整区间都在84.5%的抽样中实现覆盖。在 $q = 2\%$水平下，常规双重机器学习的覆盖率为95.5%，平均宽度为0.104；双重去偏机器学习的覆盖率为94.5%，平均宽度为0.103。这里覆盖率的提升来自区间变宽，而在这些抽样中，常规双重机器学习至少同样受益。对于双重去偏机器学习，用每个拟合的估计误差减去其真实条件偏差后，所得标准差为0.0216，接近0.0217的估计标准误均方根。未调整的双重去偏机器学习区间在94%的抽样中覆盖其条件均值，尽管2%的容许偏差仅在40.5%的拟合中界定了真实绝对条件偏差。因此，接近名义水平的目标量覆盖率可能与频繁违反所假定的偏差界同时出现。附录 C.2 报告了每个秩一拟合以及32叶校正的这项诊断。

7.3. 在训练数据或评估数据上更新回归。定理1确定了能够复现双重去偏机器学习的留一回归更新，这促使我们将其与一种更简单的替代方法进行比较，即在评估前拟合单个斜率。利用与两种回归拟合均分离的校准观测，我们将 $Y - \widehat{g}(X)$ 对 $f(X)$ 进行回归：$= \widehat{g}_{\mathrm{PFN}}(X) - \widehat{g}(X)$，不含截距，并在 $\widehat{g} + \widehat{c} f$ 处使用原始权重评估常规双重机器学习（设置 $\widehat{c} = 1$ 可恢复直接PFN替代）。双重去偏机器学习使用相同的$\mathbb{E}[f(X)^2]$ 校准估计，并在评估数据上更新斜率，同时对每个得分观测执行留一处理。

为进行匹配比较，我们在拟合任一基线干扰函数（表3）之前预留校准划分，因此每个面板都在相同的评估观测上比较相同的学习得到的方向和权重拟合。基于训练集拟合的替代方法使用常规条件得分标准误，而双重去偏机器学习使用第4.3节的完整条件标准误。

表8 *401(k)*效应的匹配回归更新

初始拟合 / 更新	估计值	SE	95% Wald区间	区间，$q = 2\%$
线性/逻辑回归				
初始双重机器学习	5,876	1,072	[3,774, 7,977]	[2,286, 9,465]
直接 PFN	8,607	1,022	[6,603, 10,611]	[5,197, 12,017]
训练斜率	7,534	1,022	[5,531, 9,537]	[4,111, 10,957]
DDML	8,617	993	[6,672, 10,563]	[5,184, 12,051]
森林 / 森林				
初始双重机器学习	7,389	980	[5,468, 9,310]	[4,059, 10,718]
直接 PFN	7,697	966	[5,804, 9,590]	[4,397, 10,996]
训练斜率	7,649	965	[5,758, 9,540]	[4,358, 10,940]
DDML	7,702	965	[5,810, 9,593]	[4,402, 11,002]

所有金额均以1991年美元计。斜率和方向范数使用相同的1,240个校准家庭。$q = 2\%$区间使用共享的Riesz尺度，以及每个结果拟合的训练残差尺度。

采用线性401(k)拟合时，估计的校准斜率为0.61。所得估计值介于初始双重机器学习与直接PFN替代之间，而双重去偏机器学习几乎复现了直接替代的结果。采用森林拟合时，斜率为0.84，训练集斜率估计值与双重去偏机器学习估计值仅相差53美元（表8）。即使在2%预算下，每个面板中的四种方法给出的下端点也都是正值。

表格 9*ACIC* 上的匹配回归更新

更新	均方根误差	均值 $\lvert b\rvert$	覆盖率	覆盖率，$q=2\%$	宽度，$q=2\%$	
设置*27*，线性/逻辑回归						
初始双重机器学习	0.0609	0.0428	75	0.150	95	0.196
直接 PFN	0.0312	0.0113	90	0.088	90	0.103
训练斜率 0.0314		0.0103	85	0.088	90	0.103
DDML	0.0320	0.0077	80	0.090	100	0.137
设置*27*，提升／提升						
初始双重机器学习	0.0291	0.0070	90	0.084	90	0.102
直接 PFN	0.0337	0.0051	85	0.085	90	0.100
训练斜率 0.0298		0.0056	85	0.079	90	0.093
DDML	0.0299	0.0054	85	0.079	90	0.096
设置*29*，线性/逻辑回归						
初始双重机器学习	0.1118	0.0774	60	0.178	65	0.223
直接 PFN	0.0317	0.0200	90	0.092	100	0.106
训练斜率 0.0305		0.0187	100	0.092	100	0.106
DDML	0.0254	0.0127	90	0.089	100	0.134
设置*29*，提升／提升						
初始双重机器学习	0.0264	0.0153	90	0.092	100	0.113
直接 PFN	0.0217	0.0115	90	0.077	95	0.092
训练斜率 0.0223		0.0118	90	0.077	95	0.092
DDML	0.0199	0.0101	90	0.074	100	0.096

每种设置重复 20 次，每个固定响应面重复 5 次。均方根误差、平均绝对条件偏差和区间宽度均以边际结果标准差为单位。覆盖率是包含真实效应的名义 95% 区间所占的百分比。前两列覆盖率和宽度数据采用零偏差余量。

在 ACIC 设置27中，无论采用哪种基线，训练集斜率更新与双重去偏机器学习的均方根误差都相近；而在设置29中，双重去偏机器学习在两种基线下的均方根误差都更低：采用提升拟合时，均方根误差从0.0223降至0.0199，而两种未调整区间在20次重复中均有18次覆盖（表9）。与其根据这些仅有20次抽样的面板数据得出普遍的排名，不如说这一比较在方向固定后，揭示了在训练数据还是评估数据上估计斜率所产生的影响。

7.4.学习校正项以处理弱混杂因素的稠密尾部。后 Lasso 等稀疏变量选择方法可能会漏掉许多个体影响较弱、但会使处理和结果变量朝同一方向变化的控制变量。为了衡量弱混杂因素的稠密尾部会产生多大偏差，以及单个学习得到的方向可以消除多少偏差，我们扩展了 Chetverikov、Sørensen 和 Tsyvinski [8] 的高斯线性设计。每次抽样包含 $n=1{,}000$ 个观测值和 500 个具有托普利兹协方差 $\mathbb{E}[X_jX_k]=\rho^{|j-k|}$ 的标准化高斯控制变量。主要稀疏成分 $S(X):=\sum_{j=1}^{500}2^{1-j}X_j$ 对前几个控制变量赋予较大权重；而 $U(X)$ 等于最后 400 个控制变量之和除以其总体标准差，因此在任何相关系数水平下，稠密尾部的方差均为单位方差。我们生成

$$D=S(X)+aU(X)+\nu,\qquad Y=D+S(X)+aU(X)+\varepsilon,$$

误差项为相互独立的标准正态变量 ν 和 ε，且与 X独立，因此真实处理效应系数为 1。即使在 $\rho=0$ 和 $a=0.50$ 的情况下，每个尾部控制变量在每个干扰方程中的系数也只有 0.025，但 400 个尾部控制变量合起来贡献了一个共同信号 $aU(X)$，其标准差为 0.50。我们评估尾部强度 $a\in\{0.25,0.50\}$和相关系数 $\rho\in\{0,0.8\}$。

在五个交叉拟合折中，我们分别对处理变量和约化式结果变量拟合 Lasso回归，并用最小二乘法重新拟合所选变量。随后，我们比较后 Lasso双重机器学习、三重后套索（使用相同的重新拟合结果评估所选行的三重得分）以及我们的非对角秩一得分（30）。由于该模型中的 Riesz 表示元是处理变量回归的相反数，因此，在训练数据上预测处理残差可以学习到 Riesz更新的方向（相差一个符号）。我们首先通过对全部 500 个控制变量进行岭回归来实现这一步；其次使用 CausalPFN（提供与训练数据独立的人工二元处理臂标签，并对两个处理臂的预测取平均，同时在估计方程中保留实际的连续处理变量）。对于每个学习得到的方向，我们使用每折 400 个独立无标签校准协变量估计其平方范数。在每一折中，我们使用与后 Lasso 相同的 800 个观测值拟合残差学习器，并在另外 200 个观测值上进行评估。我们在留出折上求解校正矩条件，并对五个根取平均。表 10 衡量这些根在独立抽样中的离散程度。第 4.5节给出了汇总折估计值以进行推断的条件。

表10当 *400* 个弱控制变量进入两个干扰函数时的均方根误差

	独立控制变量		相关控制变量（$\rho = 0.8$）	
尾部强度 a	0.25	0.50	0.25	0.50
A. 对 *400* 次配对抽样进行岭回归比较				
Post-Lasso双重机器学习	0.0663	0.1945	0.0619	0.0748
三重后套索	0.0663	0.1961	0.0639	0.0881
双重去偏机器学习，岭方向	0.0617	0.1568	0.0482	0.0637
均方根误差降低（%）	6.8	19.4	22.1	14.8
抽样标准差降低（%）	0.5	1.9	0.2	2.8
B. 前 *12* 次抽样中的 *CausalPFN*比较				
Post-Lasso双重机器学习	0.0607	0.1991	0.0595	0.0743
三重后套索	0.0569	0.1964	0.0555	0.0750
双重去偏机器学习，岭方向	0.0550	0.1597	0.0407	0.0582
DDML, CausalPFN方向	0.0595	0.1865	0.0403	0.0549
均方根误差降低（%）	1.9	6.3	32.3	26.1
抽样标准差降低（%）	-3.4	-7.3	-1.6	-2.9

真实处理系数为 1。每个面板中的所有方法都使用相同的抽样、五折划分和后套索干扰项拟合。每个百分比均表示所选方向相对于 Post-Lasso双重机器学习的变化。抽样标准差是处理效应估计在各次抽样中的样本标准差。负向降幅表示离散程度更大。

在 400 次配对抽样中，沿单个岭回归学习得到的方向进行校正，在全部四种稠密尾部设计中将均方根误差降低了 6.8–22.1%，同时略微降低了抽样离散度（表 10，面板 A）。对于独立控制变量和 $a = 0.50$，处理变量套索平均仅选出 400 个尾部变量中的 0.17 个，使得 Post-Lasso双重机器学习的偏差达到 0.192。由于岭方向与真实尾部的留出样本均方相关性达到 0.242，投影到该单一方向可将偏差降至 0.153，并将均方根误差从 0.1945 降至 0.1568；而三重后套索的均方根误差为 0.1961。

面板 B 比较了前 12 次抽样中的岭回归方向与 CausalPFN方向。在采用独立控制变量且 $a = 0.50$ 时，岭回归带来的均方根误差降低幅度更大；而在采用相关控制变量（$\rho = 0.8$）时，CausalPFN 进一步略微降低了均方根误差：从 0.0743 降至 0.0549，而岭回归对应的数值为 0.0582。在 $a = 0.25$ 时，CausalPFN方向在独立控制变量下将均方根误差降低了 1.9%，在相关控制变量下则降低了 32.3%。在这 12 次抽样的全部四种设计中，CausalPFN校正项使抽样标准差相较于

Post-Lasso双重机器学习增加了 1.6–7.3%，这证实其均方根误差的改善源于偏差降低。

7.5. 几何设计与后 *Lasso* 重拟合的价值。当混杂信号近似稀疏而非弥散时，通过对选定变量进行重拟合，已经可以消除 Lasso 的大部分收缩偏差。设定 $a = 0$ 可恢复 Chetverikov 等人 [8], 的原始几何设计，其中所有 500 个系数均非零，并按几何级数衰减为 2^{1-j}。表11在相同的 400 次配对抽样上，将干扰参数拟合方式（惩罚 Lasso 与后置 Lasso 重拟合）和得分类型（双重与三重）进行交叉比较。比较结果对五折的根取平均值，并将合并后估计影响值的标准差除以 $\sqrt{1,\ 000}$，以计算所报告的 Wald 区间。此处的覆盖率是根据多次重复抽样中已知的系数衡量的。

表 11对几何衰减系数进行重拟合并使用三重 *Lasso* 得分

	独立控制变量			相关控制变量（$\rho = 0.8$）		
干扰项拟合和得分	Bias	RMSE	覆盖率（%）	Bias	RMSE	覆盖率（%）
后 Lasso ，双重	0.00184	0.0313	96.75	-0.00019	0.0315	96.25
后置 Lasso、三重	0.00172	0.0315	96.25	0.00017	0.0323	95.00
Lasso ，双重	0.04570	0.0542	67.50	0.01674	0.0348	93.00
Lasso、三重	0.01496	0.0353	93.00	-0.00680	0.0336	95.75

新增尾部的强度为 $a = 0$。500 个几何系数仍保留在两个结构性干扰项中。所有估计量均使用相同的 400 次配对抽样和五折交叉验证。覆盖率指名义 95% 区间。

在独立控制变量情形下，对惩罚 Lasso 拟合应用三重 Lasso，可将偏差从 0.04570 降至 0.01496，并使名义 95% 覆盖率提高 25.5 个百分点（从 67.50% 升至 93.00%）；代价是估计标准误平均增加 4.3%（相关控制变量情形下增加 8.1%）。然而，改用双重得分的后 Lasso 干扰项拟合，已能将偏差降至 0.00184、覆盖率提高至 96.75%，因此改用三重得分仅使覆盖率变化 -0.5 个百分点，并使标准误增加 0.1%（相关控制变量情形下增加 1.8%）。在 $\rho = 0.8$ 时，两种后 Lasso 估计量的绝对偏差均低于 0.0002。在几何设计中，仅后 Lasso 重拟合就几乎获得了消除收缩偏差所带来的全部实际收益；而表 10 显示，当存在弱混杂因素的稠密尾部时，在选定控制变量之外学习一个方向，仍能大幅降低偏差。

对选定控制变量进行得分校正之所以能够紧密追踪最小二乘重新拟合，有一个简单的代数原因。在共同样本上，令 z 收集一组共享的选定控制变量，其样本 Gram 矩阵 $G_n = \mathbb{P}_n[zz']$ 可逆，并令 Π_n 表示在 z 上的样本最小二乘投影。对于第 3.3 节中的残差 v 和 $a_\beta = u - \beta v$，正规方程意味着

$$\mathbb{P}_n[a_\beta v] - \mathbb{P}_n[a_\beta z']G_n^{-1}\mathbb{P}_n[zv] = \mathbb{P}_n[(a_\beta - \Pi_n a_\beta)(v - \Pi_n v)].$$

只要两个初始线性拟合都位于 z 的张成空间内，且残差化处理变量的样本方差为正，求解右侧的估计方程就会得到与在 z上重新拟合两个干扰函数所得完全相同的处理效应系数。在实践中，三重 Lasso 得分使用训练折逆格拉姆估计中的选定行，并结合留出折上的残差矩；而后 Lasso 则在每个回归各自的训练支持集上分别强制满足正规方程，这就是为什么在某次抽样中这两个估计量仍可能略有不同。

8. 高阶正交性与影响函数的联系。 命题 2 给出了任意成对核修正所留下偏差的精确公式。为将该公式置于二阶去偏这一更广泛的计量经济学理论背景中，我们首先区分实践中常被混为一谈的四种不同“高阶”概念，然后说明工作空间投影如何在 $\mathcal{H}_k$ 内实现精确的二阶 Neyman 正交性，同时在其外部保留二次敏感性。

8.1. 区分“高阶”的四种含义。二阶去偏的文献中，“阶”有四种不同含义，概述于表 12。对于在 ℓ 个独立数据副本上评估的矩函数 ψ，令 $\Psi(\theta,\eta):=\mathbb{E}[\psi(W_1,\ldots,W_\ell;\theta,\eta)]$ 表示其在真实数据生成规律下的期望。若矩函数以真值为中心，即 $\Psi(\theta_0,\eta_0)=0$，且关于干扰参数 η 的方向导数直至 q 阶均为零，则称该矩函数为 阶 ***Neyman*正交**，

$$D_\eta^j\Psi(\theta_0,\eta_0)[h_1,\ldots,h_j]=0,\qquad j=1,\ldots,q, \tag{78}$$

对于所有可容许的干扰参数方向均如此。当第 阶导数在局部有界时，泰勒展开留下的偏差余项阶数为 $\|\eta-\eta_0\|^{q+1}$；对于 $q=1$（常规双重机器学习），它是二次的；对于 $q=2$，则是三次的。

表12去偏估计中四种“高阶”概念的区别

Term	所描述的性质
Neyman正交阶数 q	在固定的数据分布下，沿指定干扰参数空间的 q 阶导数均为零
高阶影响函	固定的数据分布。 核表示目标量随数据
tion	分布会发生变化。
阶数为 ℓ 的U统计量	估计量核将 ℓ 个不同的观测值作为参数。
高阶偏差缩减	剩余偏差严格更小，针对实际实现的干扰参数 估计误差而言 .

仅使用二阶成对U统计量，并不能据此确立二阶正交性，也不能保证更快的偏差收敛率。在 Bonhomme、Jochmans、Newey 和 Weidner 的矩条件框架 [3]（第2.3节）中，干扰参数向量明确包含一个逆雅可比矩阵；正交性则通过独立副本，对原始干扰参数及其逆矩阵的混合导数进行联合约束（他们的公式（9）采用U统计量的形式）。相比之下，Bonhomme、Jochmans 和 Weidner 的似然构造 [4]（第5节）使用指定的条件密度 f，将单个模型观测值（如面板数据或网络）投影到广义得分 $f^{-1}\partial_\eta^j f$（$j=1$、……、q）上；而完全正交性 要求期望矩在所有干扰参数取值下均为零。

有限维线性对比示例。一个简单的线性回归示例说明，当模型为有限维时，双重去偏机器学习如何实现二阶 Neyman 正交性。考虑 $g_0(x)=b(x)'\beta_0$，其格拉姆矩阵 $\Sigma:=\mathbb{E}[b(X)b(X)']$可逆，线性对比为 $\theta_0=c'\beta_0$，泛函为 $m(W,\ b'\beta):=c'\beta$。给定对称的训练估计值 $M\approx\Sigma^{-1}$，选择

$$\widehat{g}=b'\widehat{\beta},\qquad \widehat{\alpha}=b'Mc,\qquad \widehat{K}(x,x')=b(x)'Mb(x') \tag{79}$$

在（7）中得到校正估计量

$$\widehat{\theta}_K=c'\widehat{\beta}+\frac{1}{n}\sum_j c'\{2M-M\widehat{\Sigma}_{-j}M\}b(X_j)\{Y_j-b(X_j)'\widehat{\beta}\}, \tag{80}$$

其中 $\widehat{\Sigma}_{-j} := (n-1)^{-1}\sum_{i\neq j} b(X_i)\,b(X_i)'$。将常规双重机器学习估计量$\widehat{\ }$与 θ_K进行比较，它们的条件偏差为

(81) $$\mathbb{E}[\widehat{\theta}_{\mathrm{DB}}]-\theta_0 = c'(I-M\Sigma)(\widehat{\beta}-\beta_0), \qquad \mathbb{E}[\widehat{\theta}_K]-\theta_0 = c'(I-M\Sigma)^2(\widehat{\beta}-\beta_0).$$

由于 $I-M\Sigma=-(M-\Sigma^{-1})\Sigma$，校正偏差关于 $(\widehat{\beta}-\beta_0, M-\Sigma^{-1})$ 联合呈三次，因此实现了精确的二阶 Neyman 正交性。在式（35）中令 $r(W,\beta)=b(X)\{Y-b(X)'\beta\}$并取 $\Lambda=-M$，可得得分

$$c'\widehat{\beta}+2c'Mb(X_2)\varepsilon_g(W_2)-c'Mb(X_3)b(X_3)'Mb(X_2)\varepsilon_g(W_2),$$

其在不同观测值上的平均恰好重现（80）。在这个例子中只需两份副本，因为线性对比向量 c 与数据无关，而随机目标导数或非线性泛函则需要额外的结构。

8.2. 工作空间内的二阶正交性与空间外的偏差。回到一般线性泛函（1），固定一个工作空间 $\mathcal{H}_k=\mathrm{span}\{b_1,\dots,b_k\}$，其格拉姆矩阵 $\Sigma=\mathbb{E}[b(X)b(X)']$ 为正定矩阵，核为 $K_M(x,x') := b(x)'Mb(x')$。令 $\mathcal{M}_M(g,\alpha)$ 表示在候选函数（g,α）处求值的双重去偏机器学习估计量（7）的总体均值，并定义投影误差矩 $u_g := \mathbb{E}[b(X)\delta_g(X)]$ 和$u_\alpha := \mathbb{E}[b(X)\delta_\alpha(X)]$。

命题 15（工作空间正交性与逼近偏差）。在命题 *2* 的条件及前述基函数假设下，令 Π 为投影到 $\mathcal{H}_k$ 上的总体正交投影。对于任意平方可积的候选误差，

(82) $$\mathcal{M}_M(g,\alpha)-\theta_0 = -\langle (I-\Pi)\delta_\alpha, (I-\Pi)\delta_g\rangle + u_\alpha'(M-\Sigma^{-1})u_g.$$

通过 (59) 中的 $g_{\mathcal{H}}=b'\beta_{\mathcal{H}}$ 和 $\alpha_{\mathcal{H}}=b'\gamma_{\mathcal{H}}$ 定义 $\beta_{\mathcal{H}}$，$\gamma_{\mathcal{H}}\in\mathbb{R}^k$。若候选项属于 $\mathcal{H}_k$，则

(83) $$\mathcal{M}_M(b'\beta,b'\gamma)-\theta_{\mathcal{H}} = (\gamma-\gamma_{\mathcal{H}})'\Sigma(M-\Sigma^{-1})\Sigma(\beta-\beta_{\mathcal{H}}).$$

因此，$\mathcal{M}_M(b'\beta,\ b'\gamma)-\theta_{\mathcal{H}}$ 在 $(\beta,\ \gamma,\ M)$ 处关于 $(\beta_{\mathcal{H}},\ \gamma_{\mathcal{H}},\ \Sigma^{-1})$ 联合满足二阶 *Neyman* 正交性。若 $g_0,\ \alpha_0\in\mathcal{H}_k$，则该性质对 θ_0 本身也成立。

在 k 固定时，(83) 中的三次余项有界于 $\|\Sigma\|_{\mathrm{op}}^2\|\gamma-\gamma_{\mathcal{H}}\|\|M-\Sigma^{-1}\|_{\mathrm{op}}\|\beta-\beta_{\mathcal{H}}\|$；而随着 k 增大，仍须明确控制算子范数、收敛速率和抽样方差。当 $\mathcal{H}_k$ 设定错误时，短目标 $\theta_{\mathcal{H}}$ 的精确二阶正交性与逼近偏差 $\theta_0-\theta_{\mathcal{H}}$ 并存。对于不受限机器学习拟合，(82) 保留了 $\mathcal{H}_k$ 之外的残差内积，而这正是第5节敏感性界所控制的确切项。

这一分解还解释了为什么不存在固定的平方可积核，能够在 L^2所有（P_X）上实现精确二阶正交性。在固定 K 的情况下，命题 2 给出了混合二阶导数

(84) $$D_\alpha D_g\mathcal{M}_K(g_0,\alpha_0)[a,h] = -\langle a,(I-T)h\rangle.$$

若要使该导数对所有 a、$h\in L^2(P_X)$ 都为零，算子 T就必须等于恒等算子 I，而这对于无限维$L^2(P_X)$ 空间上的紧积分算子是不可能的。精确的工作投影可在 $\mathcal{H}_k$ 内实现二阶正交性，同时控制其 $\mathcal{H}_k$之外剩余的二次敏感性。

8.3. 高阶影响函数与重复校正。 我们的成对估计量与高阶影响函数（HOIFs） [21, 20],密切相关；后者同样利用独立观测副本估计干扰误差的乘积。此外，它也与稳定化高阶影响函数 [19], 相关，后者旨在解决估计所得 Gram逆矩阵的不稳定性。在 Shi等人 [22], 的马尔可夫离策略设定中，Bellman残差和条件访问权重分别扮演回归残差和协变量核的算子角色，不过，时间依赖需要单独进行渐近分析。

为了应用校正项 r 次，将每个独立观测值 W_j 与$\widehat{}$更新算子 $(\mathcal{D}_j g)(x) := g(x) + K(x, X_j)\{Y_j - g(X_j)\}$ 相关联，其期望满足$\mathbb{E}[\mathcal{D}_j g] - g_0 = (I - T)(g - g_0)$。在不同观测值上组合 r 次更新，在另一个不同的观测值上计算普通双重机器学习得分，并对所有不同的 $(r+1)$ 元组取平均，得到

$$\mathbb{E}[\widehat{\theta}^{(r)}] - \theta_0 = -\langle \delta_\alpha, (I - T)^r \delta_g \rangle, \qquad 0 \le r \le n-1, \tag{85}$$

前提是迭代得分可积。这里， $r = 0$ 恢复普通双重机器学习， $r = 1$给出定理 1 的留一回归更新。然而，对于精确投影 Π，幂等性给出 $(I - \Pi)^r = I - \Pi$，对每个 $r \ge 1$ 均成立。重复使用固定投影无法消除 $\mathcal{H}_k$ 之外的遗漏方向偏差，同时会引入更高阶的 Hoeffding 项，而我们的 $r = 1$ 方差公式 (48) 未涵盖这些项。

9. 结论。 深度去偏机器学习为常规双重机器学习与完整高阶影响函数理论之间提供了一个务实的折中方案。由于不存在能够在所有 $L^2(P_X)$ 上实现精确二阶正交性的平方可积核，任何实用的二阶校正都必须选择要针对的误差方向。直接从第一阶段残差中学习低维工作空间，可以让灵活的学习器将高维混杂模式压缩为一个或少数几个方向，并在独立样本上估计其混合偏差贡献。定理1表明，这种成对U统计量校正在代数上等同于使用留一事后回归（或Riesz权重）更新来评估常规双重机器学习；与此同时，Hoeffding分解提供了精确的有限样本方差核算，并明确说明了常规双重机器学习标准误和交叉拟合保持有效的条件。

在我们的实证研究和模拟研究中，第二步去偏的实际作用取决于初始学习器的强度。当初始拟合遗漏了系统性非线性或一批分散的弱混杂因素时，单个学习得到的方向能够显著降低偏差和均方根误差。当初始学习器已经较强时，估计校正项相对于抽样不确定性通常很小。在这种情况下，双重去偏机器学习主要作为沿所搜索方向进行的诊断检查，而第5节的双因素敏感性框架则量化了实质性结论能够容忍多少剩余模型设定错误偏差（包括学习空间之外的偏差，以及由归一化误差沿该空间产生的偏差）。

未来工作有两个主要方向值得关注。首先，我们目前的残差学习器分别针对结果误差或 Riesz误差，而混合偏差取决于二者的对齐 $\langle \delta_\alpha, \delta_g \rangle$。设计能够直接寻找使两种可观测矩条件违背相互对齐的协变量方向、同时又不会增大方差的正则化训练目标，或可进一步改进偏差消除。其次，尽管对于静态线性泛函，单次成对校正已充分发挥低秩投影带来的收益，但将定向残差方向学习和工作空间敏感性界扩展至非线性矩条件及动态序贯场景（例如马尔可夫离策略评估，并与 Shi等人 [22]的多步深度去偏相衔接），同时将交叉拟合方差估计拓展至预言机方差情形之外，仍是一项重要议题。

参考文献

[1] Balazadeh, V., H. Kamkari, V. Thomas, B. Li, J. Ma, J. C. Cresswell, and R. G. Krishnan（2025）。CausalPFN：通过上下文学习摊销因果效应估计。神经信息处理系统进展。https://arxiv.org/abs/2506.07918。[2] Belloni, A., and V. Chernozhukov（2013）。高维稀疏模型中模型选择后的最小二乘法。伯努利 19(2), 521–547。https://doi.org/10.3150/11-BEJ410。[3] Bonhomme, S., K. Jochmans, W. K. Newey, and M. Weidner（2026）。矩条件模型中的高阶Neyman正交性。工作论文，第3版，11月5日。https://arxiv.org/abs/2605. 10842v3。[4] Bonhomme, S., K. Jochmans, and M. Weidner（2026）。似然模型中的正交矩。TSE 工作论文 26-1774，9月25日。https://www.tse-fr.eu/sites/default/files/TSE/documents/doc/ document/wp_tse_1774.pdf。[5] Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins（2018）。用于处理效应和结构参数的双重／去偏机器学习。计量经济学期刊 21, C1–C68。https://doi.org/10.1111/ectj.12097。[6] Chernozhukov, V., C. Cinelli, W. K. Newey, A. Sharma, and V. Syrgkanis（2026）。长话短说：因果机器学习中的遗漏变量偏差。《经济学与统计学评论》。https://doi.org/10. 1162/REST.a.1705。扩展版本：https://arxiv.org/abs/2112.13398v6。[7] Chernozhukov, V., C. Hansen, and Y. Liao（2017）。以 lava 攻克稠密与稀疏信号之和的恢复问题。《统计年鉴》45(1), 39–76。https://doi.org/10.1214/16-AOS1434。[8] Chetverikov, D., J. R.-V. Sørensen, and A. Tsyvinski（2026）。三重／双重去偏Lasso。工作论文。https://arxiv.org/abs/2603.20134。[9] Díaz, I., M. Carone, and M. J. van der Laan（2016）。对随机缺失变量均值的二阶推断。《国际生物统计学杂志》12, 333–349。https://arxiv.org/abs/1511. 08369。[10] Dorie, V., J. Hill, U. Shalit, M. Scott, and D. Cervone（2019）。自动化方法与自行操作方法在因果推断中的比较：从数据分析竞赛中汲取的经验。《统计科学》34, 43–68。https://arxiv.org/abs/1707.02641。[11] DoubleML（访问日期：2026年11月9日）。Python：401(k) 对金融财富的影响。https: //docs.doubleml.org/stable/examples/py_double_ml_pension.html。数据：https://github.com/VC2015/DMLonGitHub/blob/master/sipp1991.dta。[12] Dukes, O., S. Vansteelandt, and D. Whitney（2024）。半参数回归中双重机器学习的双重稳健推断。《机器学习研究杂志》25(279), 1–46。https://www.jmlr. org/papers/v25/22-1233.html。[13] Escanciano, J. C., and J. R. Terschuur（2026）。二阶正交性的统计代价。工作论文，11月19日。https://arxiv.org/abs/2611.25702。[14] Hall, P.（1984）。多元非参数密度估计量积分平方误差的中心极限定理。《多元分析杂志》14, 1–16。https://doi.org/10.1016/0047-259X(84)90044-7。[15] Hoeffding, W.（1948）。一类具有渐近正态分布的统计量。《数理统计年鉴》19, 293–325。https://doi.org/10.1214/aoms/1177730196。[16] Hoerl, A. E., and R. W. Kennard（1970）。岭回归：非正交问题中的有偏估计。《技术计量学》12(1), 55–67。https://doi.org/10.1080/00401706.1970.10488634。[17] Jin, J., and V. Syrgkanis（2026）。一般线性泛函估计的尖锐、结构无关下界。工作论文，第2版，1月5日。https://arxiv.org/abs/2512.17341v2。[18] Liu, L., R. Mukherjee, and J. M. Robins（2020）。关于对机器学习估计的因果参数标称置信区间覆盖率进行近乎无假设检验。《统计科学》35, 518–539。https://doi.org/10.1214/20-STS786。[19] Liu, N., C. Li, Y. Gu, and L. Liu（2026）。稳定化高阶影响函数：一类双线性型的统计理论。工作论文，第3版。https://arxiv.org/abs/2607.04743v3。[20] Robins, J., L. Li, R. Mukherjee, E. Tchetgen Tchetgen, and A. van der Vaart（2017）。高维模型的高阶估计方程。《统计年鉴》45(5), 1951–1987。https://doi.org/10. 1214/16-AOS1515。[21] Robins, J., L. Li, E. Tchetgen, and A. van der Vaart（2008）。高阶影响函数与非线性泛函的极小极大估计。《概率与统计：纪念David A. Freedman文集》，335–421。https://arxiv.org/abs/0805.3040。[22] Shi, C., R. Wan, V. Chernozhukov, and R. Song（2021）。深度去偏的离策略区间估计。机器学习研究会议录 139, 9580–9591。https://proceedings.mlr.press/v139/shi21d. html。

[23] Tibshirani, R.（1996）。通过 Lasso 进行回归收缩与选择。《皇家统计学会会刊：*B*辑》58(1)，267–288。https://doi.org/10.1111/j.2517-6161.1996.tb02080.x。[24] van der Laan, L.、A. Luedtke 和 M. Carone（2026）。通过校准进行双重稳健推断。工作论文，第3版，8月20日。https://arxiv.org/abs/2411.02771v3。[25] Wang, Q. 和 B. Lindsay（2014）。一般 U统计量的方差估计及其在交叉验证中的应用。《中华统计学志》24，1117–1141。https://doi.org/10.5705/ss.2012.215。[26] Zheng, M.、M. Bonvini 和 Z. Guo（2026）。扰动双重机器学习：超越参数长度的非标准推断。工作论文，第2版，2月21日。https://arxiv.org/abs/2511. 01222v2。

附录 A：符号说明

表 13 汇总了全文使用的主要符号。给定训练样本 $\mathcal{T}$ 条件下，所有期望、方差、内积和 $L^2(P_X)$ 范数均取自独立评估分布 P，除非另有说明。

表 13主要符号汇总

符号	含义
W, X, Y	具有协变量 $X \subset W$ 和标量结果 Y 的观测 W
$g_0(x)$ $\alpha_0(X)$ θ_0	真实回归 $\mathbb{E}[Y \mid X = x]$、Riesz 表示元和目标量 $\mathbb{E}[m(W,\ g_0)]$
$\mathcal{T}$, $\mathbb{P}_n$	训练样本 σ-代数和评估样本均值 $n^{-1}\sum_{i=}^{n}$ 1
$\mathbb{U}_n h$	U-平均 $\binom{n}{2}^{-1}\sum_{i\ <j} h(W_i,\ W_j)$ 对称情形下的无序对核 h
$(n)_r$	互异指标的有序 r 元组数量，$n(n-1)\cdots(n-r+1)$。在特别地，$(n)_3 = n(n-1)(n-2)$。
$\widehat{g}$、$\widehat{\alpha}$、δ_g、δ_α	训练拟合和干扰误差 $\delta_g := \widehat{g} - g_0$，$\delta_\alpha := \widehat{\alpha} - \alpha_0$
$\langle f, h\rangle$、$\|f\|_2$	内积 $\mathbb{E}[f(X)h(X)]$ 和 $L^2(P_X)$ 范数 $\langle f,\ f\rangle^{1/2}$
$s(W)$, $\widehat{\theta}_{\mathrm{DB}}$	普通双重机器学习得分 $m(W,\ \widehat{g}) + \widehat{\alpha}(X)\{Y - \widehat{g}(X)\}$ 和估计量 $\mathbb{P}_n s$
$\widehat{K}(x,\ x')$ T, T^*	训练选定的核、积分算子 $(Tf)(x) := \mathbb{E}[\ \widehat{K}(x,\ X)f(X)]$，以及伴随算子
$\mathcal{H}_k$、Π、$\delta_j^{\perp}$	k维工作空间、投影到 $\mathcal{H}_k$的 $L^2(P_X)$，以及$(I-\Pi)\delta_j$
$A(W;\ x')$ $\varepsilon_g(W)$ H	Riesz 差异，结果残差 $Y - \widehat{g}(X)$，and 成对核 $A(W;\ X')\varepsilon_g(W')$
$\widehat{B}_K$, $\widehat{\theta}_K$, b_K	U统计量校正项，双重去偏机器学习估计量 $\widehat{\theta}_{\mathrm{DB}} - \widehat{B}_K$，以及偏差 $\mathbb{E}\widehat{\theta}_K] - \theta_0$
h, L, d, ϕ_K	对称化核、Hoeffding 线性部分与退化部分，以及有效影响函数 $s - \mathbb{E}[s] - L$
$q(w,\ w')$ τ_n^2	组合配对核 $\frac{1}{2}\{s(w)+s(w')\} - h(w,\ w')$ 和条件方差 $\mathrm{Var}(\widehat{\theta}_K \mid \mathcal{T})$
$\widehat{V}$, $\widehat{V}_1$, $\widehat{V}_2$, $\widehat{V}_{\mathrm{lin}}$	无偏全方差估计量 ($\widehat{V}_1 + \widehat{V}_2$)和刀切线性方差
$g_{\mathcal{H}}$, $\alpha_{\mathcal{H}}$, $\theta_{\mathcal{H}}$	短模型投影 Πg_0, $\Pi\alpha_0$和短目标 $\mathbb{E}[m(W,\ g_{\mathcal{H}})]$
r_g, r_α, $\lambda_\perp$	残差误差（$I-\Pi$）δ_g、（$I-\Pi$）δ_α及其 L^2 余弦值
ϵ_j, ρ_j, B_q	全局误差半径、残差误差半径以及双因素 q-预算偏差补贴
$R_{j,T}$, s_Y, s_α	加权均方干扰误差、训练结果残差均方根，以及共享去偏Riesz尺度

附录 B：证明

在本附录中，除非另有说明，所有涉及训练拟合对象（$\widehat{g}$、$\widehat{\alpha}$、K、T、Π 和 $\mathcal{H}_k$）的期望、方差、协方差和 L^2（P_X）范数均为给定训练σ-域 $\mathcal{T}$ 条件下的量。

定理1的证明。固定训练拟合 $(\widehat{g}, \widehat{\alpha}, K)$ 和评估样本$W_1, \ldots, W_n$。对每个 $i \in \{1, \ldots, n\}$，将式（8）中的留一回归增量写为

$$u_i(x) := \widetilde{g}_{-i}(x) - \widehat{g}(x) = \frac{1}{n-1}\sum_{j \neq i} \widehat{K}(x, X_j)\varepsilon_g(W_j),$$

其中 $\varepsilon_g(W_j) = Y_j - \widehat{g}(X_j)$。由于 $\widetilde{g}_{-i} = \widehat{g} + u_i$，且映射 $g \mapsto m(W_i, g)$ 在包含 $\widehat{g}$ 和每个核列 $K(\cdot, X_j)$的$\widehat{\ }$上是线性的，将 $\widetilde{g}_{-i}$ 代入观测 i 的普通双重机器学习得分（3），可得

$$\begin{aligned} m(W_i, \widetilde{g}_{-i}) + \widehat{\alpha}(X_i)\{Y_i - \widetilde{g}_{-i}(X_i)\} &= m(W_i, \widehat{g}) + \widehat{\alpha}(X_i)\{Y_i - \widehat{g}(X_i)\} \\ &\quad + m(W_i, u_i) - \widehat{\alpha}(X_i)u_i(X_i) \\ &= s(W_i) - \big\{\widehat{\alpha}(X_i)u_i(X_i) - m(W_i, u_i)\big\}. \end{aligned}$$

利用 u_i 的定义以及 m（W_i，$\cdot$）的线性性质展开方括号内的调整项，得到

$$\begin{aligned} &\widehat{\alpha}(X_i)u_i(X_i) - m(W_i, u_i) \\ &\quad = \frac{1}{n-1}\sum_{j \neq i} \big\{\widehat{\alpha}(X_i)\widehat{K}(X_i, X_j) - m(W_i, \widehat{K}(\cdot, X_j))\big\}\varepsilon_g(W_j) \\ &= \frac{1}{n-1}\sum_{j \neq i} A(W_i; X_j)\varepsilon_g(W_j) = \frac{1}{n-1}\sum_{j \neq i} H(W_i, W_j), \end{aligned}$$

根据 (7) 前面 A（W_i；x'）和 H（W_i，W_j）的定义。对 $i = 1$、.....、n 求平均，得到

$$\frac{1}{n}\sum_{i=1}^{n}\big[m(W_i, \widetilde{g}_{-i}) + \widehat{\alpha}(X_i)\{Y_i - \widetilde{g}_{-i}(X_i)\}\big] = \mathbb{P}_n s - \frac{1}{n(n-1)}\sum_{i \neq j} H(W_i, W_j) = \widehat{\theta}_{\mathrm{DB}} - \widehat{B}_K,$$

$\widehat{\ }$这等于 θ_K，并由此得到 (9)。 □

命题 2 的证明。设 W 和 W' 是从给定 $\mathcal{T}$ 条件下的评估$\widehat{\ }$分布中独立抽取的样本，协变量分量为 X 和 X'。由于 B_K 是对所有满足 n（$n-1$）的有序对（i，j）取平均，且每一对（W_i，W_j）与（W，W'）具有相同的联合分布，因此由期望的线性性质可得 $\mathbb{E}[B_K] = \mathbb{E}[H$（$W$，$W'$）]。

我们分两步评估 $\mathbb{E}[H$（W，W'）]，首先对残差观测 W' 积分消去。根据条件均值 g_0（x'）$= \mathbb{E}[Y' \mid X' = x']$的定义，结果残差$\varepsilon_g$（$W'$）$= Y' - \widehat{g}$（$X'$）满足

$$\mathbb{E}[\varepsilon_g(W') \mid W, X'] = \mathbb{E}[Y' - \widehat{g}(X') \mid X'] = g_0(X') - \widehat{g}(X') = -\delta_g(X'),$$

其中，第一个等号利用了在给定 $\mathcal{T}$ 的条件下，W' 与 W 的独立性。将两边乘以$\frown$ $K(x, X')$，并对 X' 取期望，可得对于每个固定的 x，

$$\mathbb{E}[\widehat{K}(x, X')\varepsilon_g(W') \mid W] = -\mathbb{E}[\widehat{K}(x, X')\delta_g(X')] = -(T\delta_g)(x).$$

现在将此恒等式代入 $H(W, W') = A(W; X')\varepsilon_g(W')$。根据 (7) 前面对$A(W; x')$的定义，结合 $m(W, \cdot)$的线性，以及允许将关于 W' 的期望移入 m内的可积性条件，我们得到

$$\begin{aligned} \mathbb{E}[H(W, W') \mid W] &= \mathbb{E}[\{\widehat{\alpha}(X)\widehat{K}(X, X') - m(W, \widehat{K}(\cdot, X'))\}\varepsilon_g(W') \mid W] \\ &= -\widehat{\alpha}(X)(T\delta_g)(X) + m(W, T\delta_g). \end{aligned}$$

对 W 取外层期望，并应用式 (2) 中的 Riesz表示恒等式$\mathbb{E}[m(W, T\delta_g)] = \langle \alpha_0, T\delta_g \rangle$ 可得

$$\mathbb{E}[\widehat{B}_K] = \mathbb{E}[H(W, W')] = -\mathbb{E}[\widehat{\alpha}(X)(T\delta_g)(X)] + \mathbb{E}[\alpha_0(X)(T\delta_g)(X)] = -\langle \delta_\alpha, T\delta_g \rangle.$$

^最后，将此表达式与式 (4) 中的常规双重机器学习偏差恒等式 $\mathbb{E}[\theta_{\mathrm{DB}}] - \theta_0 = -\langle \delta_\alpha, \delta_g \rangle$ 结合可得

$$b_K = \mathbb{E}[\widehat{\theta}_{\mathrm{DB}} - \widehat{B}_K] - \theta_0 = -\langle \delta_\alpha, \delta_g \rangle + \langle \delta_\alpha, T\delta_g \rangle = -\langle \delta_\alpha, (I-T)\delta_g \rangle.$$

等价形式 $b_K = -\langle (I - T^*)\, \delta_\alpha,\ \delta_g \rangle$ 可直接由$L^2(P_X)$ 伴随算子 T^*的定义得到。□

命题3的证明。由于 Π 是到 $\mathcal{H}_k$上的 $L^2(P_X)$ 正交投影，补算子 $I - \Pi$ 也同样是自伴且幂等的，即 $(I-\Pi)^* = I - \Pi$ 以及 $(I-\Pi)^2 = I - \Pi$。对于任意一对 $(a,\ v) \in \mathcal{U}_\Pi$，结合这两个性质与柯西–施瓦茨不等式可得

$$\begin{aligned} |\langle a, (I-\Pi)v \rangle| &= |\langle (I-\Pi)a, (I-\Pi)v \rangle| \\ &\le \|(I-\Pi)a\|_2 \|(I-\Pi)v\|_2 \le \rho_\alpha \rho_g. \end{aligned}$$

为证明该上界是紧的，假设正交补 $\mathcal{H}_k^\perp$非平凡，因此可以选取一个单位向量 $e \in \mathcal{H}_k^\perp$，满足 $\|e\|_2 = 1$ 且 $(I-\Pi)\, e = e$。令 $a := \rho_\alpha e$ 以及$v := \rho_g e$，则有 $\|a\|_2 = \|(I-\Pi)\, a\|_2 = \rho_\alpha \le \epsilon_\alpha$和 $\|v\|_2 = \|(I-\Pi)\, v\|_2 = \rho_g \le \epsilon_g$，因此$(a,\ v) \in \mathcal{U}_\Pi$。此外，$|\langle a,\ (I-\Pi)\, v \rangle| = \rho_\alpha \rho_g \langle e,\ e \rangle = \rho_\alpha \rho_g$，从而得到（17）。令$\rho_\alpha = \epsilon_\alpha$ 并令 $\rho_g = \epsilon_g$，可得全局球上确界 $\epsilon_\alpha \epsilon_g$。

对于近似投影界（18），假设 $T = \Pi T \Pi$ 和 $\|\Pi - T\|_{\mathrm{op}} \le \omega$。由于 $\Pi^2 = \Pi$，我们有 $(I-T)\Pi\delta_g = (\Pi - T)\Pi\delta_g$。对精确分解（15）应用三角不等式和柯西–施瓦茨不等式可得

$$\begin{aligned} |b_K| &\le |\langle \delta_\alpha^\perp, \delta_g^\perp \rangle| + |\langle \Pi\delta_\alpha, (\Pi - T)\Pi\delta_g \rangle| \\ &\le \|\delta_\alpha^\perp\|_2 \|\delta_g^\perp\|_2 + \|\Pi\delta_\alpha\|_2 \|(\Pi - T)\Pi\delta_g\|_2 \\ &\le \|\delta_\alpha^\perp\|_2 \|\delta_g^\perp\|_2 + \omega \|\Pi\delta_\alpha\|_2 \|\Pi\delta_g\|_2, \end{aligned}$$

由此完成证明。□

推论6的证明。在给定训练样本 $\mathcal{T}$ 条件下，由 $K_f(x,x') = f(x)f(x')/\widehat{\nu}_f$诱导的积分算子 T作用于任意 $u \in L^2(P_X)$时为

$$(Tu)(x) = \mathbb{E}\left[\frac{f(x)f(X)}{\widehat{\nu}_f} u(X) \right] = \frac{\nu_f}{\widehat{\nu}_f} \cdot \frac{\langle f, u \rangle}{\nu_f} f(x) = \frac{\nu_f}{\widehat{\nu}_f} (\Pi_f u)(x),$$

其中，Π_f 是将 $L^2(P_X)$ 正交投影到一维工作空间$\mathcal{H}_1 := \mathrm{span}\{f\}$上的投影。因此，$T = \Pi_f T \Pi_f$并且

$$\|\Pi_f - T\|_{\mathrm{op}} = \left| 1 - \frac{\nu_f}{\widehat{\nu}_f} \right| \|\Pi_f\|_{\mathrm{op}} = \omega_f.$$

将命题3中的近似投影界（18）与 $\Pi = \Pi_f$和 $\omega = \omega_f$ 一起应用，并代入 $\|(I - \Pi_f)\, \delta_g\|_2 = \Delta_f$，得到（27）中的第一个不等式：

$$|b_{K_f}| \le \|(I - \Pi_f)\delta_\alpha\|_2 \Delta_f + \omega_f \|\Pi_f \delta_\alpha\|_2 \|\Pi_f \delta_g\|_2.$$

由于 Π_f和 $I - \Pi_f$都是正交投影，因此它们是压缩映射（$\|\Pi_f\|_{\mathrm{op}} \le 1$和 $\|I - \Pi_f\|_{\mathrm{op}} \le 1$），所以 $\|(I - \Pi_f)\delta_\alpha\|_2 \le \|\delta_\alpha\|_2$、$\|\Pi_f \delta_\alpha\|_2 \le \|\delta_\alpha\|_2$和 $\|\Pi_f \delta_g\|_2 \le \|\delta_g\|_2$。代入这些界即可得到式（27）中的第二个不等式。如果任一上界为$o_p(n^{-1/2})$，则 $b_{K_f} = o_p(n^{-1/2})$，这验证了式（40）中的偏差条件，因此命题 8 适用。□

残差回归速率（28）的证明。我们分四步建立该速率：首先以初始干扰项拟合 $(\widehat{g},\widehat{\alpha})$ 为条件，对形状误差 Δ_f 进行界定；然后以学习特征 f 为条件，对校准误差 ω_f 进行界定。

步骤 *1*：总体投影与残差正交性。给定$(\widehat{g},\ \widehat{\alpha})$，并记$Z:=Y-\widehat{g}(X)$、$\xi:=\xi_p(X)$以及 $G:=\mathbb{E}[\xi\xi']$。令 $\xi'\beta_p$ 为将 $v=g_0-\widehat{g}=-\delta_g$ 投影到张成空间$\{\xi_p\}$上的总体 $L^2(P_X)$ 最小二乘投影，并令 $e:=Z-\xi'\beta_p$。根据 v的定义，我们有 $\mathbb{E}[Z\mid X]=v(X)$，因此总体正规方程给出

$$\mathbb{E}[\xi e]=\mathbb{E}[\xi(v-\xi'\beta_p)]=0,$$

并且，由 $v-\xi'\beta_p$ 和 $\xi'\beta_p$ 的正交性（以及 $\mathbb{E}[(Z-v)\xi'\beta_p]=0$），

$$\mathbb{E}[e^2]=\mathbb{E}[(Z-v)^2]+\mathbb{E}[(v-\xi'\beta_p)^2]\le\mathbb{E}[(Z-v)^2]+\mathbb{E}[v^2]=\mathbb{E}[Z^2]=O_p(1).$$

步骤 2： Δ_f的 Gram 矩阵集中与估计误差。令$\widehat{G}:=m^{-1}\sum_{i=1}^m\xi_i\xi_i'$为在 m 个新鲜的残差学习观测值上计算的样本 Gram 矩阵，其中 $\xi_i:=\widehat{\xi}_p(X_i)$。由于 ξ 的各元素均一致有界， $G-G$ 的每个元素都是 m 个独立同分布、均值为零且有界的随机变量的平均值，其方差为 $O(m^{-1})$。对全部 p^2个元素求和可得$\mathbb{E}[\|\widehat{G}-G\|_{\mathrm{F}}^2]=O(p^2/m)$，因此在 $p=o(\sqrt{m})$条件下，$\|\widehat{G}-G\|_{\mathrm{op}}\le\|\widehat{G}-G\|_{\mathrm{F}}=O_p(p/\sqrt{m})=o_p(1)$。由于最小特征值 $\lambda_{\min}(G)$有正的下界，Weyl 不等式意味着 $\lambda_{\min}(\widehat{G})\ge\frac{1}{2}\lambda_{\min}(G)>0$ 以趋于 1 的概率成立。

$\widehat{\ }$ $\widehat{\ }$ $\widehat{\ }$在该事件上，最小二乘估计量 β 满足样本正规方程 $G(\beta-\beta_p)=m^{-1}\sum_{i=1}^m\xi_ie_i$，其中 $e_i:=Z_i-\xi_i'\beta_p$。给定 $(\widehat{g},\widehat{\alpha})$，各加项 ξ_ie_i 独立同分布且均值为零，因此

$$\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^m\xi_ie_i\right\|^2\,\middle|\,\widehat{g},\widehat{\alpha}\right]=\frac{\mathbb{E}[\|\xi\|^2e^2\mid\widehat{g},\widehat{\alpha}]}{m}\lesssim\frac{p\,\mathbb{E}[e^2\mid\widehat{g},\widehat{\alpha}]}{m}=O_p(p/m),$$

其中 $\|\cdot\|$ 表示欧几里得范数，并且我们利用 ξ 各坐标的一致有界性来界定 $\|\xi\|^2\lesssim p$。将其与 $\underline{\lambda_{\min}}(\widehat{G})$ 的下界结合，可得 $\|\widehat{\beta}-\beta_p\|=O_p(\sqrt{p/m})$。由于 G 的最大特征值有上界，$\widehat{f}=\xi'\beta$ 的预测误差满足

$$\|\xi'(\widehat{\beta}-\beta_p)\|_2=\left\{(\widehat{\beta}-\beta_p)'G(\widehat{\beta}-\beta_p)\right\}^{1/2}\lesssim\|\widehat{\beta}-\beta_p\|=O_p(\sqrt{p/m}).$$

在 Δ_f 的定义中取 $a=-1$，并将 $v=-\delta_g$与三角不等式结合使用，因此得到

$$\Delta_f\le\|\delta_g+f\|_2=\|v-f\|_2\le\|v-\xi'\beta_p\|_2+\|\xi'(\widehat{\beta}-\beta_p)\|_2\le a_p+O_p(\sqrt{p/m}).$$

步骤 *3*：校准比率 ω_f。 现在，给定 $(\widehat{g},\ \widehat{\alpha})$ 以及拟合特征 f。每当 $\nu_f=\mathbb{E}[f(X)^2]>0$ 时，校准估计量 $\widehat{\nu}_f=\ell^{-1}\sum_{j=1}^\ell f(X_j)^2$ 就是 ℓ 个独立同分布抽样的均值为 ν_f 的平均值，因此比率 $\widehat{\nu}_f/\nu_f$的条件均值为1，条件方差为

$$\mathrm{Var}\left(\frac{\widehat{\nu}_f}{\nu_f}\,\middle|\,\widehat{g},\widehat{\alpha},f\right)=\frac{\mathbb{E}[f(X)^4]/\nu_f^2-1}{\ell}=O_p(\ell^{-1}).$$

由条件切比雪夫不等式，$\widehat{\nu}_f/\nu_f=1+O_p(\ell^{-1/2})$。特别地，在事件 $\nu_f>0$ 上，$\Pr(\widehat{\nu}_f=0\mid\widehat{g},\widehat{\alpha},f)\to_p0$，取倒数可得 $\omega_f=|\nu_f/\widehat{\nu}_f-1|=O_p(\ell^{-1/2})$。（对于独立的对称 $\{-1,1\}$ 特征，展开 $(\xi'\beta)^4$ 并去掉奇次幂后，叶节点为 $\mathbb{E}[(\xi'\beta)^4]=\sum_j\beta_j^4+6\sum_{j<k}\beta_j^2\beta_k^2=3\|\beta\|^4-2\sum_j\beta_j^4\le3\nu_f^2$，从而验证了四阶矩条件。）

第4步：合并各个界。 在事件 $\nu_f > 0$ 上，将 $\Delta_f = O_p(a_p + \sqrt{p/m})$、$\omega_f = O_p(\ell^{-1/2})$ 和 $\|\delta_j\|_2 = O_p$) 代入（27）的第二个界，即可直接得到（28）。最后，在事件 $f = 0$ 上， P_X 几乎必然成立；零核约定（$K_f = 0$）会恢复常规双重机器学习，其中 $|b_{K_f}| = |\langle \delta_\alpha, \delta_g \rangle| \le \|\delta_\alpha\|_2 \|v\|_2$，并且由于 $f = 0$ 几乎必然成立， $\|v\|_2 = \|v - f\|_2 \le a_p + O_p(\sqrt{p/m})$，因此相同的收敛速率（28）仍然成立。 □

命题 7 的证明。我们首先验证 Hoeffding展开 $\mathbb{U}_n h = \mu_h + \mathbb{P}_n L + \mathbb{U}_n d$。根据（35）中 d 的定义，每个满足 $1 \le i < j \le n$ 的无序对（i， j） 都满足（W_i， W_j） $= \mu_h + \frac{1}{2}\{L（W_i） + L（W_j）\} + d（W_i$， W_j）。对所有 $\binom{n}{2} = n（n-1）/2$ 无序对求和时，每个指标 $i \in \{1, \ldots, n\}$ 恰好出现在 $n-1$ 个对中，因此

$$\binom{n}{2}^{-1} \sum_{1 \le i < j \le n} \frac{L(W_i) + L(W_j)}{2} = \frac{2}{n(n-1)} \cdot \frac{n-1}{2} \sum_{i=1}^{n} L(W_i) = \mathbb{P}_n L.$$

因此 $\mathbb{U}_n h = \mu_h + \mathbb{P}_n L + \mathbb{U}_n d$。从 $\mathbb{P}_n s$ 中减去该展开式，再加上并减去 $\mathbb{E}[s（W）]$，得到

$$\widehat{\theta}_K - \theta_0 = \{\mathbb{E}[s(W)] - \mu_h - \theta_0\} + \mathbb{P}_n\{s - \mathbb{E}[s(W)] - L\} - \mathbb{U}_n d = b_K + \mathbb{P}_n \phi_K - \mathbb{U}_n d,$$

其中，第二个等式用到了命题 2（$\mathbb{E}[s（W）] - \mu_h - \theta_0 = b_K$）以及（36）中 ϕ_K 的定义。这就证明了（37）。

接下来我们计算条件方差 $\tau_n^2 = \mathrm{Var}(\mathbb{P}_n \phi_K - \mathbb{U}_n d \mid \mathcal{T})$。注意到 $\mathbb{E}[\phi_K(W)] = 0$，并且 d 是对称且退化的，这意味着 $\mathbb{E}[d(W, W') \mid W] = 0$ 以及 $\mathbb{E}[d(W, W') \mid W'] = 0$。我们检验两个正交关系。首先，对于任意索引 r 和任意一对 $i < j$，要么 $r \notin \{i, j\}$（因此 W_r 独立于 (W_i, W_j)，且 $\mathbb{E}[\phi_K(W_r) d(W_i, W_j)] = \mathbb{E}[\phi_K(W_r)] \mathbb{E}[d(W_i, W_j)] = 0$），要么 $r \in \{i, j\}$（设 $r = i$，此时以 W_i 为条件可得 $\mathbb{E}[\phi_K(W_i) \mathbb{E}[d(W_i, W_j) \mid W_i]] = 0$）。因此 $\mathrm{Cov}(\mathbb{P}_n \phi_K, \mathbb{U}_n d \mid \mathcal{T}) = 0$。其次，对于任意两个不同的无序对 $\{i, j\} \ne \{k, l\}$，这些对要么互不相交（由独立性可得协方差为零），要么恰好共享一个索引（设 $i = k$ 与 $j \ne l$，此时以 (W_i, W_j) 为条件可得 $\mathbb{E}[d(W_i, W_j) \mathbb{E}[d(W_i, W_l) \mid W_i]] = 0$）。因此， $\mathbb{U}_n d$ 中的所有 $\binom{n}{2}$ 项都两两不相关，且具有共同方差 $\sigma_{2,n}^2 = \mathbb{E}[d^2]$，从而得到

$$\mathrm{Var}(\mathbb{U}_n d \mid \mathcal{T}) = \binom{n}{2}^{-2} \cdot \binom{n}{2} \sigma_{2,n}^2 = \frac{2\sigma_{2,n}^2}{n(n-1)}.$$

由于给定 $\mathcal{T}$ 时， $\mathbb{P}_n \phi_K$ 的 n 项独立同分布，且均值为零、方差为 $\sigma_{1,n}^2 = \mathbb{E}[\phi_K^2]$ $\tau_n^2 = \sigma_{1n}^2/n + 2\sigma_{2n}^2/\{n\ n-1\}$ 38 $\sigma_{1n}^2 =$ $s - L|$，因此我们得到 . . ()，从 $\mathcal{T} =$ $s +$ $L -2$ s L 而证明了 ()。最后， . 方差() 方差() 方差() 协方差(,)。 □

命题 8 的证明。以命题 7 中的精确展开式 (37) 为起点，加上并减去中心化预言平均值 $\mathbb{P}_n \phi_0$（注意到 $\mathbb{E}[\psi_0] = \theta_0$），再乘以 $\sqrt{n}$，可得

$$\sqrt{n}(\widehat{\theta}_K - \theta_0) = \sqrt{n}\,\mathbb{P}_n \phi_0 + \sqrt{n}\, b_K + \sqrt{n}\,\mathbb{P}_n\big\{(s - \psi_0) - \mathbb{E}[s - \psi_0]\big\} - \sqrt{n}\,\mathbb{P}_n L - \sqrt{n}\,\mathbb{U}_n d.$$

我们证明最后四项为 $o_p(1)$。首先，根据假设，$\sqrt{n}\, b_K = o_p(1)$ 直接成立。其次，给定 $\mathcal{T}$ 时，中心化经验平均值 $R_1 := \sqrt{n}\,\mathbb{P}_n\{(s - \psi_0) - \mathbb{E}[s - \psi_0]\}$ 的均值为零，条件方差为

$$\mathrm{Var}(R_1 \mid \mathcal{T}) = \mathrm{Var}(s - \psi_0 \mid \mathcal{T}) \le \mathbb{E}[(s - \psi_0)^2 \mid \mathcal{T}] = \|s - \psi_0\|_{L^2(P)}^2 = o_p(1).$$

根据条件马尔可夫不等式，对于每个 $\varepsilon > 0$，我们都有 $\Pr(|R_1| > \varepsilon \mid \mathcal{T}) \le \min\{1, \mathrm{Var}(R_1 \mid \mathcal{T})/\varepsilon^2\} \to_p 0$，而有界收敛定理意味着

$\Pr(|R_1| > \varepsilon) \to 0$，因此 $R_1 = o_p(1)$。第三，对 $R_2 := \sqrt{n}\,\mathbb{P}_n L$进行完全相同的计算可得 $\mathrm{Var}(R_2 \mid \mathcal{T}) = \|L\|^2_{L^2(P)} = o_p(1)$，由此 $R_2 = o_p(1)$。第四，命题 7 给出

$$\mathrm{Var}(\sqrt{n}\,\mathbb{U}_n d \mid \mathcal{T}) = \frac{2n\mathbb{E}[d^2]}{n(n-1)} = \frac{2\mathbb{E}[d^2]}{n-1} = o_p(1)$$

在 $\mathbb{E}[d^2] = o_p$（n）下，因此 $\sqrt{n}\,\mathbb{U}_n d = o_p$（1）由相同的条件马尔可夫论证可得 .

汇总这些界得到 $\sqrt{n}$（$\widehat{\theta}_K$ − θ_0） = $\overline{n}^{-1/2}\sum_{i=1}^n \phi_0$（$W_i$） + o_p（1）。由于 ϕ_0（W_1）…… ϕ_0（W_n）是均值为零且方差为正常数的独立同分布变量 $\mathbb{E}[\phi_0^2]$，林德伯格–列维中心极限定理给出 (41)。$b_K = o_p$（$n^{-1/2}$）的两个充分条件直接由 (12) 和算子范数备用条件 (19) 得出。

□

命题 9 的证明。我们首先将 (46) 中的三个总体矩 （A、B、C）与 $\widehat{\theta}_K = \mathbb{U}_n q$ 的 Hoeffding 方差 （$\sigma^2_{1,\,n}$、$\sigma^2_{2,\,n}$）联系起来。令 μ_q：$= \mathbb{E}[q_{12}]$，对称核 q 在 (45) 中的 Hoeffding 分解为

$$q_{ij} = \mu_q + \tfrac{1}{2}\{\phi_K(W_i) + \phi_K(W_j)\} - d(W_i, W_j),$$

其中 $\phi_K(W_i)$ 的均值为零，方差为 $\sigma^2_{1,n}$； $d(W_i, W_j)$ 是退化的，方差为$\sigma^2_{2,n}$，且三个分量两两正交。对于给定 $\mathcal{T}$ 的独立评估抽样W_1, W_2, W_3, W_4，利用正交性直接展开可得

$$\begin{aligned}
A &= \mathbb{E}[q_{12}^2] = \mu_q^2 + \tfrac{1}{4}(\sigma^2_{1,n} + \sigma^2_{1,n}) + \sigma^2_{2,n} = \mu_q^2 + \tfrac{1}{2}\sigma^2_{1,n} + \sigma^2_{2,n},\\
B &= \mathbb{E}[q_{12}q_{13}] = \mu_q^2 + \tfrac{1}{4}\mathbb{E}[\phi_K(W_1)^2] = \mu_q^2 + \tfrac{1}{4}\sigma^2_{1,n},\\
C &= \mathbb{E}[q_{12}q_{34}] = \mu_q^2.
\end{aligned}$$

解该线性方程组可得 $\sigma^2_{1,n}$ 和 $\sigma^2_{2,n}$：

$$\sigma^2_{1,n} = 4(B - C), \qquad \sigma^2_{2,n} = A - 2B + C.$$

接下来，由于在给定 $\mathcal{T}$ 的条件下， $W_1, \ldots, W_n$ 独立同分布，因此每个有序不同二元组 (i,j) 都满足$\mathbb{E}[q_{ij}^2] = A$，每个有序不同三元组 (i,j,k) 都满足 $\mathbb{E}[q_{ij}q_{ik}] = B$，每个有序不同四元组 (i,j,k,l) 都满足 $\mathbb{E}[q_{ij}q_{kl}] = C$。由于分别有 $(n)_2, (n)_3$, 和 $(n)_4$ 个这样的$\widehat{\quad}$ $\widehat{\quad}$有序元组，式 (47) 中的估计量满足 $\mathbb{E}[A \mid \mathcal{T}] = A$、 $\mathbb{E}[B \mid \mathcal{T}] = B$ 和$\frown\mathbb{E}[C \mid \mathcal{T}] = C$。将这些期望代入式 (48)，并应用命题 7，可得

$$\mathbb{E}[\widehat{V} \mid \mathcal{T}] = \frac{4}{n}(B - C) + \frac{2}{n(n-1)}(A - 2B + C) = \frac{\sigma^2_{1,n}}{n} + \frac{2\sigma^2_{2,n}}{n(n-1)} = \tau_n^2.$$

□

命题 10 的证明。首先检查（53）中的鞅差 Δ_i 之和等于中心化估计量。对 $i = 1,$ ……， n 求和，并利用 d 的对称性，得到

$$\sum_{i=1}^n \Delta_i = \frac{1}{n}\sum_{i=1}^n \phi_K(W_i) - \frac{2}{n(n-1)}\sum_{1 \le j < i \le n} d(W_i, W_j) = \mathbb{P}_n\phi_K - \mathbb{U}_n d = \widehat{\theta}_K - \mathbb{E}[\widehat{\theta}_K \mid \mathcal{T}].$$

对于每个 $i \ge 1$，观测 W_i 独立于 $\mathcal{F}_{i-1} = \sigma$（$\mathcal{T}$、$W_1$、…、$W_{i-1}$）。因此， $\mathbb{E}[\phi_K$（W_i） $\mid \mathcal{F}_{i-1}] = \mathbb{E}[\phi_K$（$W_i$） $\mid \mathcal{T}] = 0$；并且对于每个 $j < i$， d 的退化性给出 $\mathbb{E}[d$（W_i、 W_j） $\mid \mathcal{F}_{i-1}] = \mathbb{E}[d$（$W_i$、 W_j） $\mid \mathcal{T}$， $W_j] = 0$。于是 $\mathbb{E}[\Delta_i \mid \mathcal{F}_{i-1}] = 0$，所以 $\{\Delta_i/\tau_n\}_{i=1}^n$ 是关于 $\{\mathcal{F}_i\}$ 的鞅差数组，因为 τ_n 是 $\mathcal{T}$-可测的（因而也是 $\mathcal{F}_0$-可测的）。

在满足条件方差收敛性和林德伯格条件（54）的情况下，鞅中心极限定理（参见 Hall [14] 关于变化的退化 U 统计量核）意味着

$$\frac{\widehat{\theta}_K - \mathbb{E}[\widehat{\theta}_K \mid \mathcal{T}]}{\tau_n} = \sum_{i=1}^{n} \frac{\Delta_i}{\tau_n} \rightsquigarrow N(0,1).$$

由于 $\widehat{V}/\tau_n^2 \to_p 1$ 蕴含 $\sqrt{\tau_n^2/\widehat{V}} \to_p 1$，斯卢茨基定理立即给出中心化的学生化极限 (55)。当 $b_K = o_p(\tau_n)$ 也成立时，重写

$$\frac{\widehat{\theta}_K - \theta_0}{\sqrt{\widehat{V}}} = \sqrt{\frac{\tau_n^2}{\widehat{V}}} \left(\frac{\widehat{\theta}_K - \mathbb{E}[\widehat{\theta}_K \mid \mathcal{T}]}{\tau_n} + \frac{b_K}{\tau_n} \right)$$

并再次应用斯卢茨基定理可得 (56)。Wald 置信区间$\widehat{\theta}_K \pm z_{1-\gamma/2}\sqrt{\widehat{V}}$的宽度为 $2z_{1-\gamma/2}\sqrt{\widehat{V}} = 2z_{1-\gamma/2}\tau_n\{1 + o_p(1)\}$，当 $n\tau_n^2$以概率有界且远离零时，其尺度为 $n^{-1/2}$；否则为 τ_n。 □

推论 11 的证明。我们先建立合并线性展开 (58)，然后验证合并样本外得分方差 $\hat{\sigma}^2$的一致性。

第 1 部分：合并的渐近线性表示。 对于每个折 $k \in \{1, \dots, K\}$，令 $\mathcal{T}_k$ 表示由折划分、 I_k之外的所有观测值以及训练期间使用的任何独立随机性生成的 σ-代数。由于该划分与数据独立，给定 $\mathcal{T}_k$ 后， I_k 中的 n_k 个观测值来自 P，且独立同分布。将命题 8 应用于折 k 可得

$$\widehat{\theta}_k - \theta_0 = \mathbb{P}_k\phi_0 + R_k, \qquad \sqrt{n_k}\, R_k = o_p(1),$$

其中 $\mathbb{P}_k\phi_0 := n_k^{-1}\sum_{i \in I_k} \phi_0(W_i)$。乘以 n_k/N并对 $k = 1$、.....、 K 求和后，领先的预言机平均值中的每个观测值权重均为 $1/N$，

$$\sum_{k=1}^{K} \frac{n_k}{N} \mathbb{P}_k\phi_0 = \frac{1}{N}\sum_{i=1}^{N} \phi_0(W_i),$$

因此按 $\sqrt{N}$ 缩放后得到

$$\sqrt{N}(\widehat{\theta}_{\mathrm{CF}} - \theta_0) - \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \phi_0(W_i) = \sum_{k=1}^{K} \sqrt{\frac{n_k}{N}}\, (\sqrt{n_k}\, R_k).$$

由于 $n_k/N \le 1$，有限和中的每一项都是 $o_p(1)$，而由有限个 K 项组成且每项均为 $o_p(1)$ 的和，无论各折之间是否存在依赖性，仍为 $o_p(1)$。这就证明了（58）。

第 2 部分：方差一致性与学生化极限。 对于每个折 k，定义留出集平均平方得分误差 $D_k := \mathbb{P}_k[(s_{-k} - \psi_0)^2] \ge 0$。给定 $\mathcal{T}_k$后， I_k中的观测值来自 P，且独立同分布，因此根据（40）中的得分收敛假设，有 $\mathbb{E}[D_k \mid \mathcal{T}_k] = \|s_{-k} - \psi_0\|_{L^2(P)}^2 = o_p(1)$。对于任意 $t > 0$，条件马尔可夫不等式给出

$$\Pr(D_k > t \mid \mathcal{T}_k) \le \min\left\{1,\ \frac{\|s_{-k} - \psi_0\|_{L^2(P)}^2}{t}\right\} \to_p 0.$$

由于右侧被1界定，有界收敛定理推出$\Pr(D_k > t) = \mathbb{E}[\Pr(D_k > t \mid \mathcal{T}_k)] \to 0$，因此对每个 k， $D_k = o_p(1)$ 均成立。由于 K 是固定的，合并后的均方得分误差也趋于零：

$$\frac{1}{N}\sum_{i=1}^{N} \{\widehat{s}_i - \psi_0(W_i)\}^2 = \sum_{k=1}^{K} \frac{n_k}{N} D_k = o_p(1).$$

现在令 $\widehat{\mathbf{s}}:=(\widehat{s}_1,\ldots,\widehat{s}_N)'$，且 $\boldsymbol{\psi}_0:=(\psi_0(W_1)\ldots,\psi_0(W_N))'$，并令 $M_N:=I_N-N^{-1}\mathbf{1}\mathbf{1}'$ 为样本中心化投影矩阵。将 $\widetilde{\sigma}:=(N-1)^{-1/2}\|M_N\boldsymbol{\psi}_0\|$ 记为预言机得分的样本标准差，将 $\widehat{\sigma}=(N-1)^{-1/2}\|M_N\widehat{\mathbf{s}}\|$ 记为……，由反三角不等式和压缩性质 $\|M_N\|_{\mathrm{op}}=1$，可得

$$|\widehat{\sigma}-\widetilde{\sigma}|\le\frac{\|M_N(\widehat{\mathbf{s}}-\boldsymbol{\psi}_0)\|}{\sqrt{N-1}}\le\left(\frac{1}{N-1}\sum_{i=1}^{N}\{\widehat{s}_i-\psi_0(W_i)\}^2\right)^{1/2}=o_p(1).$$

根据独立同分布且平方可积的得分 $\psi_0(W_i)$的大数弱定律，我们有$\widetilde{\sigma}^2\to_p\mathrm{Var}(\psi_0)=\mathbb{E}[\phi_0^2]>0$，因此有 $\widehat{\sigma}^2\to_p\mathbb{E}[\phi_0^2]$。将林德伯格–列维中心极限定理应用于（58），再结合斯卢茨基定理，可得 $\sqrt{N}(\widehat{\theta}_{\mathrm{CF}}-\theta_0)/\widehat{\sigma}\rightsquigarrow N(0,1)$。

□

命题12的证明。由泛函 m 的线性性以及 Riesz 表示（2），

$$\theta_0-\theta_{\mathcal{H}}=\mathbb{E}[m(W,g_0)]-\mathbb{E}[m(W,\Pi g_0)]=\mathbb{E}[m(W,g_0-\Pi g_0)]=\langle\alpha_0,(I-\Pi)g_0\rangle.$$

由于 $I-\Pi$ 在 $L^2(P_X)$上是自伴且幂等的，我们可以将一个 $I-\Pi$移到 α_0上，从而得到

$$\theta_0-\theta_{\mathcal{H}}=\langle(I-\Pi)\alpha_0,(I-\Pi)g_0\rangle=\langle\alpha_0-\alpha_{\mathcal{H}},g_0-g_{\mathcal{H}}\rangle,$$

这就证明了（60）。

现在假设两个拟合干扰函数 $\widehat{g}$和 $\widehat{\alpha}$都属于 $\mathcal{H}_k$。则 $(I-\Pi)\widehat{g}=0$ 且$(I-\Pi)\widehat{\alpha}=0$，因此它们的正交误差分量可化简为

$$\delta_g^{\perp}=(I-\Pi)(\widehat{g}-g_0)=-(g_0-g_{\mathcal{H}}),\qquad\delta_\alpha^{\perp}=(I-\Pi)(\widehat{\alpha}-\alpha_0)=-(\alpha_0-\alpha_{\mathcal{H}}).$$

将这两个恒等式代入精确投影偏差公式 (14)（来自命题 2），得到

$$\mathbb{E}[\widehat{\theta}_\Pi\mid\mathcal{T}]-\theta_0=-\langle\delta_\alpha^{\perp},\delta_g^{\perp}\rangle=-\langle\alpha_0-\alpha_{\mathcal{H}},g_0-g_{\mathcal{H}}\rangle=-(\theta_0-\theta_{\mathcal{H}}).$$

在两边加上 θ_0，得到 $\mathbb{E}[\theta_\Pi\mid\mathcal{T}]=\theta_{\mathcal{H}}$，从而证明了 (61)。 □

命题 15 的证明。回顾，$K_M(x,x')=b(x)'Mb(x')$，$\Sigma=\mathbb{E}[b(X)b(X)']$，$u_g=\mathbb{E}[b(X)\delta_g(X)]$, 以及 $u_\alpha=\mathbb{E}[b(X)\delta_\alpha(X)]$。对于任意平方可积的候选误差 δ_g，由 K_M 诱导的积分算子 T_M 作用如下：

$$(T_M\delta_g)(x)=\mathbb{E}[b(x)'Mb(X)\delta_g(X)]=b(x)'M\mathbb{E}[b(X)\delta_g(X)]=b(x)'Mu_g.$$

与 δ_α 取内积，得到

$$\langle\delta_\alpha,T_M\delta_g\rangle=\mathbb{E}[\delta_\alpha(X)b(X)']Mu_g=u_\alpha'Mu_g.$$

在特殊情形 $M=\Sigma^{-1}$下，$T_{\Sigma^{-1}}$是将 Π 正交投影到 $\mathcal{H}_k$上的精确 $L^2(P_X)$，因此 $\Pi\delta_g=b'\Sigma^{-1}u_g$ 且 $\langle\Pi\delta_\alpha,\ \Pi\delta_g\rangle=u_\alpha'\Sigma^{-1}u_g$。现在将命题 2 与 $-\langle\delta_\alpha,\ \delta_g\rangle$的正交分解 (13) 结合起来：

$$\begin{aligned}\mathcal{M}_M(g,\alpha)-\theta_0&=-\langle\delta_\alpha,\delta_g\rangle+\langle\delta_\alpha,T_M\delta_g\rangle\\&=-\langle(I-\Pi)\delta_\alpha,(I-\Pi)\delta_g\rangle-\langle\Pi\delta_\alpha,\Pi\delta_g\rangle+\langle\delta_\alpha,T_M\delta_g\rangle\\&=-\langle(I-\Pi)\delta_\alpha,(I-\Pi)\delta_g\rangle+u_\alpha'(M-\Sigma^{-1})u_g,\end{aligned}$$

由此证明了（82）。

接下来，假设候选项位于 $\mathcal{H}_k$，因此 $g = b'\beta$ 和 $\alpha = b'\gamma$。根据命题12，（82）右侧的第一项等于 $\theta_{\mathcal{H}} - \theta_0$。此外，由于 $g_0 - b'\beta_{\mathcal{H}} = (I - \Pi)g_0$ 与 $b(X)$ 的每个坐标都正交，因此我们有

$$u_g = \mathbb{E}[b(X)\{b(X)'\beta - g_0(X)\}] = \mathbb{E}[b(X)b(X)'](\beta - \beta_{\mathcal{H}}) = \Sigma(\beta - \beta_{\mathcal{H}}),$$

且恒等于 $u_\alpha = \Sigma(\gamma - \gamma_{\mathcal{H}})$。将这些表达式代入（82），并将 $\theta_{\mathcal{H}} - \theta_0$ 移至等式左侧，得到

$$\mathcal{M}_M(b'\beta, b'\gamma) - \theta_{\mathcal{H}} = (\gamma - \gamma_{\mathcal{H}})'\Sigma(M - \Sigma^{-1})\Sigma(\beta - \beta_{\mathcal{H}}),$$

这就是（83）。令 $\Delta\beta := \beta - \beta_{\mathcal{H}}$、$\Delta\gamma := \gamma - \gamma_{\mathcal{H}}$ 和 $\Delta M := M - \Sigma^{-1}$。由于右侧是关于 $(\Delta\beta, \Delta\gamma, \Delta M)$ 的三次齐次三线性型 $\Delta\gamma'\Sigma(\Delta M)\Sigma\Delta\beta$，因此它的值、一阶方向导数和二阶方向导数在 $(\Delta\beta, \Delta\gamma, \Delta M) = (0, 0, 0)$ 处均恒等为零。因此，$\mathcal{M}_M(b'\beta, b'\gamma) - \theta_{\mathcal{H}}$ 在 $(\beta_{\mathcal{H}}, \gamma_{\mathcal{H}}, \Sigma^{-1})$ 处关于 (β, γ, M) 联合满足二阶 Neyman正交性。若 $g_0, \alpha_0 \in \mathcal{H}_k$，则 $\theta_{\mathcal{H}} = \theta_0$，因此 θ_0 本身也满足相同的正交性。 □

附录C：关于划分与敏感性的补充实证结果

C.1. 401(k)划分与方差。我们分析1991年收入与项目参与调查中的公开401(k)资格数据集 [5, 11], ，该数据集涵盖9,915户家庭。令 D 表示401(k)资格，Z 表示九个标准家庭协变量，且 $X = (D, Z)$。目标量为协变量调整后的ATE泛函，

$$m(W, g) = g(1, Z) - g(0, Z), \qquad \alpha_0(D, Z) = \frac{D}{e_0(Z)} - \frac{1 - D}{1 - e_0(Z)},$$

其因果解释需要满足一致性、条件可交换性和重叠性。我们将观测值视为独立同分布，且不使用调查权重。

样本设计与基线估计。我们以随机方式（种子20261109）将样本划分为4,957个训练观测和4,958个评估观测，并比较两种干扰参数设定：（i）按资格组进行线性结果回归，并使用逻辑倾向得分；（ii）随机森林（500棵树，结果变量的最小叶节点大小为20，倾向得分的最小叶节点大小为40）。倾向得分被截断至 $[0.1, 0.9]$，以稳定得分，但这并不能确保重叠性；所有干扰项拟合和核均在训练数据上固定。

表14比较了原始净金融资产与封顶资产，

$$Y^c := \min\{250{,}000, \max\{-50{,}000, Y\}\},$$

结果预测采用相同的截断方式。封顶会改变105个家庭的结果变量，并定义一个不同的估计目标。所有货币金额均以1991年美元计。对于封顶资产，普通去偏将森林插件估计值从7,285美元提高到7,559美元，并将线性插件估计值从6,440美元降至5,914美元。将常规双重机器学习拟合从线性切换为森林后，封顶资产的估计标准误降低了9.5%，原始资产则降低了41.3%。

表 14从回归调整到普通去偏（*401(k)* 数据）

	原始资产		封顶资产	
方法	估计值	SE	估计值	SE
未调整对比	20,485	1,945	18,006	1,249
线性回归插件	3,626	—	6,440	—
线性增广逆概率加权	5,191	2,712	5,914	1,090
森林回归插件	6,289	—	7,285	—
森林增强逆概率加权	8,406	1,593	7,559	987

金额以1991年美元计。未调整行估计的是描述性组间对比。增广逆概率加权采用通常的去偏平均处理效应得分。报告的标准误反映中心化抽样变异，而要对平均处理效应进行有效推断，还需要学习偏差可以忽略不计，或设置偏差余量。未报告插件标准误，因为将拟合回归视为固定并不能解决其一阶估计偏差。

校正项及其方差。仅使用训练数据，我们将家庭收入划分为 $J \in \{2,4,8,16\}$ 个分位数单元，并将其与资格 D 交叉组合，形成 $2J$ 个单元指标 $b(D、\ Z)$，并构造单元投影核

$$\widehat{K}(x,x') := \sum_{\ell=1}^{2J} \frac{b_\ell(x) b_\ell(x')}{\widehat{p}_\ell},$$

其中，$\widehat{p}_\ell$ 表示单元格 ℓ的训练频数。表15报告了新增校正项 $-B_K$、更新估计值及其完整估计标准误。将收入划分为四个单元格后，线性增强逆概率加权估计值增加了979美元（从5,914美元增至6,893美元），而森林估计值仅增加了83美元。线性双重机器学习使完整标准误下降了1.3%，森林双重机器学习则使其下降了0.5%。随后，两种模型对资产增益的估计变得更加接近，且精度基本相同。若要量化已消除的偏差，就需要知道真实效应。

表15 15增加一个校正项，然后扩大其工作空间（*401(k)* 数据）

Fit	收入单元	新增校正	估计值	完整标准误	标准误缩减（%）
线性	2	284	6,198	1,071	1.7
线性	4	979	6,893	1,075	1.3
线性	8	883	6,797	1,065	2.2
线性	16	956	6,870	1,064	2.3
森林	2	-3	7,556	974	1.2
森林	4	83	7,642	982	0.5
森林	8	38	7,597	989	-0.2
森林	16	149	7,708	1,018	-3.1

封顶资产。新增校正为 $-\widehat{B}_K$，因此正值会提高常规估计值。收入截点和单元格归一化基于训练数据计算。工作空间的秩为 $2J$，其中每个收入单元格和资格状态分别对应一个指示变量。最后一列给出了相对于采用相同初始拟合的常规双重机器学习，完整标准误的百分比降幅。

在表16的四种森林设定中，成对项仅占估计总方差的 0.08% 至 0.55%。更新一阶影响函数对标准误的影响大于增加成对项，这解释了为何表2将这两个组成部分分开列出。

表格 16将影响函数的变化与成对方差分开（*401(k)* 森林）

收入单元	原始标准误	校正后的一阶标准误	完整标准误	成对占比
2	987	974	974	0.08%
4	987	981	982	0.17%
8	987	988	989	0.31%
16	987	1,015	1,018	0.55%

森林拟合、封顶资产。中间的标准误为 $\sqrt{\widehat{V}_1}$，完整标准误为 $\sqrt{\widehat{V}_1+\widehat{V}_2}$。最后一列为 $100\widehat{V}_2/(\widehat{V}_1+\widehat{V}_2)$。这两个组成部分均来自无偏的不同索引方差公式，在这些计算中未经截断时均为正值。

C.2. ACIC 划分、覆盖率与敏感性。 2016 年 ACIC 竞赛生成器 [10] 为 4,802 条协变量记录提供真实处理概率和两个条件结果均值，并以这些记录的平均处理效应作为目标量。在每次重复中，我们有放回地均匀抽取记录，按照真实处理概率分配处理，并在分配的均值上加入独立标准正态噪声。由于总体分布已知，我们可以直接评估每个训练拟合的条件偏差。

模拟设计与工作空间选择。我们使用设置27和设置29。这两个设置具有相同的多项式处理分配、阶跃函数结果均值、完全重叠性和中等效应异质性，但处理模型与结果模型中的协变量项共享程度不同（设置27的对齐参数为0.25，设置29为0.75）。在每个设置中固定前四个响应面后，我们针对每个响应面抽取50组相互独立的训练样本和评估样本，每个设置共抽取200次，$n=2,400$ 个样本用于每次划分。

常规双重机器学习按处理组拟合岭回归结果模型，并基于主效应拟合带惩罚项的逻辑回归倾向得分模型，将倾向得分截断至 $[0.05,0.95]$。为在训练数据 $\mathcal{H}_k$ 上构建数据自适应工作空间，我们对标准化的处理残差和结果残差拟合一棵回归树，再将其叶节点与处理组交叉；要求每个叶节点至少包含12个训练观测，并在估计 $\widehat{p}_c$时，为每个处理组-叶节点单元增加一个伪计数。我们报告所有三个预先指定的叶节点上限（4、32和128个叶节点），并与常规双重机器学习进行比较。

表 17*ACIC* 响应面上的偏差、均方根误差和名义 *95% Wald* 覆盖率

	设置27，对齐 0.25			设置29，对齐 0.75		
方法	Bias	RMSE	覆盖率	Bias	RMSE	覆盖率
常规双重机器学习	0.0372	0.0668	74.0%	0.0719	0.0976	52.5%
双重去偏机器学习，最多 4 个叶节点	0.0190	0.0401	85.5%	0.0409	0.0582	67.5%
双重去偏机器学习，最多 32 个叶节点	0.0131	0.0333	94.0%	0.0258	0.0498	83.5%
双重去偏机器学习，最多 128 个叶节点	0.0166	0.0496	92.5%	0.0222	0.0669	89.0%

偏差是各训练拟合中绝对条件偏差的均值，通过在已知分布下对拟合得分和校正项进行积分计算。偏差和均方根误差以每个曲面上的边际结果标准差为单位。覆盖率是每种设置下 200 次抽样中名义 95% Wald区间包含目标量的比例。方差估计非正则计为未覆盖。在设置29中，叶节点上限为 4 和 32 时各发生过一次这种情况。

在使用32个叶节点时，校正项使八个响应面中的每一个的均方根误差均有所降低，并且在两个设置中都大致将其减半（表17），使Wald 覆盖率从74.0%提高到94.0%，在

设置27中从52.5%提高到83.5%，在设置29中也是如此。对于两种方法的标准误均有效的抽样，32叶校正平均使估计标准误分别降低16.4%和15.5%。尽管如此，在设置27中，33.5%的个体拟合的绝对条件偏差有所增加，这反映了例4中的抵消机制。

更精细的划分会在更小的组内估计校正项。将叶节点上限从32提高到128，会使两种设置下的均方根误差都上升。在设置29中，更精细的划分将平均绝对条件偏差从0.0258降至0.0222，但平均估计标准误上升了40.4%，从0.0326增至0.0458。我们报告了所有三个预先设定的上限，因此在32个叶节点处的比较仍纯属描述性。

对划分残余误差的敏感性。对于真实概率为 p_c 且训练集样本量为 $\widehat{p}_c$的单元，算子 T 在单元指示函数空间内作用，其特征值为 $p_c/\widehat{p}_c$。(70)中的定义对沿该指示函数方向的平方误差赋予权重 $|1-p_c/\widehat{p}_c|$。单元空间之外的误差权重为1。表18将(72)应用于由此得到的32叶校正项。作为比较，若按真实绝对条件偏差扩大区间（在该生成机制下可观测），设置27和设置29的覆盖率分别为98.5%和95.0%。

表18 18在假设残余误差预算和*32*叶上限下的覆盖率与宽度（*ACIC*）

	设置27，对齐度0.25			设置29，对齐度0.75		
预算 q	覆盖率	宽度增幅	$\lvert b_K\rvert \le B_q$	覆盖率	宽度增加	$\lvert b_K\rvert \le B_q$
0%	94.0%	0.0%	0.0%	83.5%	0.0%	0.0%
1%	97.0%	19.1%	52.0%	89.0%	17.9%	22.0%
2%	98.5%	38.3%	83.0%	95.5%	36.1%	53.5%

宽度增幅是指在估计方差为正的抽样中，平均区间宽度相对于Wald区间的百分比增幅。标题为 $|b_K| \le B_q$ 的列给出了偏差余量覆盖真实绝对条件偏差的拟合占比。只有这项诊断使用已知偏差。

1%预算使覆盖率提高了3.0和5.5个百分点，尽管在该水平下，400次拟合中没有一次同时满足加权均方误差的两个界。在设置29中，2%预算使覆盖率达到95.5%，区间比校正后的Wald区间宽36.1%，但 B_q 仅在53.5%的拟合中界定了真实绝对条件偏差。抽样宽度有助于弥合这一差距，这表明形式覆盖保证既需要合理的偏差余量，也需要有效的中心化推断。

对学习得到的方向所残留误差的敏感性。表 19 补充了表7中的覆盖率结果。共同的Riesz尺度使同一次抽样中容许误差的差异可归因于结果残差尺度。在设置29中，2%预算在线性结果拟合下产生的平均容许误差为 0.0231，在提升拟合下则为 0.0103。这些容许误差分别在 31.0% 和 40.5% 的拟合中覆盖了森林方向估计量的绝对条件偏差。

表19 19*ACIC* 秩一修正的容许量、调整后的宽度和绝对条件偏差

		$q = 1\%$			$q = 2\%$		
初始拟合 / 方向	抽样均值	B_q	宽度	偏差覆盖率	均值 B_q	宽度	偏差覆盖率
设置*27*							
线性回归、逻辑回归／随机森林	200	0.0113	0.133	48.0	0.0227	0.156	72.5
提升，提升树 / 森林	200	0.0037	0.087	43.0	0.0074	0.095	74.0
线性回归、逻辑回归／PFN	20	0.0113	0.111	90.0	0.0227	0.134	95.0
PFN、逻辑回归／森林	20	0.0037	0.091	10.0	0.0075	0.099	40.0
PFN, 提升树 / 森林	20	0.0037	0.089	30.0	0.0075	0.097	65.0
设置*29*							
线性回归、逻辑回归／森林	200	0.0115	0.141	14.5	0.0231	0.164	31.0
提升，提升树 / 森林	200	0.0051	0.092	17.5	0.0103	0.103	40.5
线性、逻辑回归 / PFN	20	0.0110	0.117	55.0	0.0220	0.139	85.0
PFN、逻辑回归 / 森林	20	0.0036	0.090	20.0	0.0073	0.097	25.0
PFN, 提升树 / 森林	20	0.0036	0.087	30.0	0.0073	0.094	35.0

容许量对每次抽样使用一个去偏 Riesz 尺度，以及每个结果拟合的残差尺度。容许量和调整后的宽度以边际结果标准差表示。偏差覆盖率是满足 $|b_K| \le B_q$ 的抽样百分比，其中 $|b_K| \le B_q$ 使用已知的条件偏差计算。相同抽样和预算下目标量的覆盖率见表7。