Generalization Bounds for Federated Learning: Fast Rates, Unparticipating Clients and Unbounded Losses

彭博睿

中国人民大学统计学院

2024 年 4 月 18 日

汇报大纲

1.1 引入：PAC 理论

定义 (假设空间)

$$\mathcal{H}: \mathcal{X} \to \mathcal{Y}$$

其中 $\mathcal{X}$ 是输入空间，$\mathcal{Y}$ 是输出空间。$\mathcal{H}$ 是假设空间，包含所有可能的假设函数。

定义 (泛化误差)

对于假设 $h \in \mathcal{H}$，真实的函数关系 $y = c(x)$，真实的数据分布 $\mathcal{D}$，h 的 **泛化误差：**

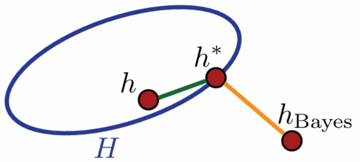
$$R(h) = \mathbb{E}_{x \sim \mathcal{D}}[1_{h(x) \neq c(x)}]$$

1.1 引入：PAC 理论

PAC 理论中重要的概念：过度误差 (excess error) 是假设函数 $h \in \mathcal{H}$ 的误差 $R(h)$ 和最优误差 R^*(也叫贝叶斯误差) 之间的差值，常常被这样分解（本篇论文也用了类似技巧）：

$$R(h) - R^* = (R(h) - \inf_{h \in \mathcal{H}} R(h)) + (\inf_{h \in \mathcal{H}} R(h) - R^*)$$

前者成为估计误差 (estimation error)，后者成为逼近误差 (approximation error)。



理论机器学习采用了 PAC(Probilistically Approximately Correct) 框架，所谓“概率近似正确”。常常得到的结论是：以概率 $1-\delta$，误差小于 ϵ。比如著名的霍夫丁不等式 (Hoeffding's inequality) 就可以用 PAC 的逻辑去理解：

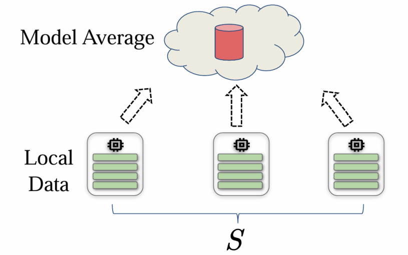
霍夫丁不等式

令 $X_1 \ \ldots \ X_n$ 为独立的随机变量，且 $X_i \in [a,b]$ $i=1 \ \ldots \ n$。这些随机变量的经验均值可表示为：$\bar{X} = \frac{X_1+\cdots+X_n}{n}$
霍夫丁不等式叙述如下：

$$\forall t>0 \quad P(\bar{X}-E[\bar{X}] \geq t) \leq exp(-\frac{2n^2t^2}{\sum_{i=1}^{n}(b_i-a_i)^2})$$

1.2 引入：联邦学习及其研究

联邦学习本质上是一种保证数据隐私的分布式机器学习技术。

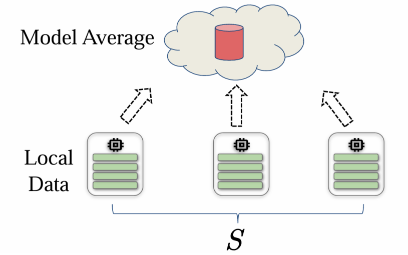


图片来源：`https://www.ai4opt.org/sites/default/files/slides/kale.pdf`

选择模型、本地训练、上传参数、聚合参数、更新模型。

1.2 引入：联邦学习及其研究

联邦学习本质上是一种保证数据隐私的分布式机器学习技术。



图片来源：`https://www.ai4opt.org/sites/default/files/slides/kale.pdf`

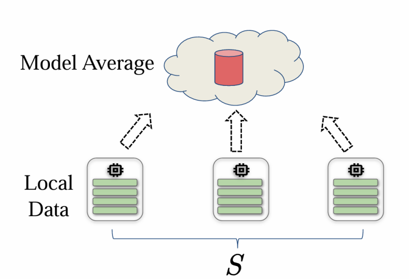
选择模型、本地训练、上传参数、聚合参数、更新模型。

联邦学习遇到的困难

- 客户的真实数据分布收到本地环境影响(Non-IID)，文中用了“Heterogeneous”来表示这种异质性。

1.2 引入：联邦学习及其研究

联邦学习本质上是一种保证数据隐私的分布式机器学习技术。



图片来源：`https://www.ai4opt.org/sites/default/files/slides/kale.pdf`

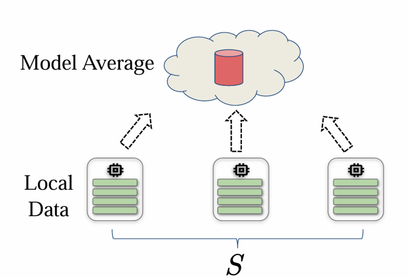
选择模型、本地训练、上传参数、聚合参数、更新模型。

联邦学习遇到的困难

- 客户的真实数据分布收到本地环境影响(Non-IID)，文中用了“Heterogeneous”来表示这种异质性。
- 异质性学习导致泛化误差很难估计

1.2 引入：联邦学习及其研究

联邦学习本质上是一种保证数据隐私的分布式机器学习技术。



图片来源：`https://www.ai4opt.org/sites/default/files/slides/kale.pdf`

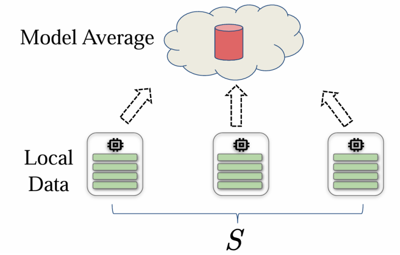
选择模型、本地训练、上传参数、聚合参数、更新模型。

联邦学习遇到的困难

- 客户的真实数据分布收到本地环境影响(Non-IID)，文中用了“Heterogeneous”来表示这种异质性。
- 异质性学习导致泛化误差很难估计
- 许多客户端可能不参与训练(Unparticipating Clients)

1.2 引入：联邦学习及其研究

联邦学习本质上是一种保证数据隐私的分布式机器学习技术。



图片来源：`https://www.ai4opt.org/sites/default/files/slides/kale.pdf`

选择模型、本地训练、上传参数、聚合参数、更新模型。

联邦学习遇到的困难

- 客户的真实数据分布收到本地环境影响(Non-IID)，文中用了“Heterogeneous”来表示这种异质性。
- 异质性学习导致泛化误差很难估计
- 许多客户端可能不参与训练(Unparticipating Clients)
- 未参与客户能否从联邦学习中获益？

1.3 引入：本文相关研究

已有研究成果

- 最优化：训练误差
- 参与客户
- 同质性数据：IID
- 有界损失函数

本文成果

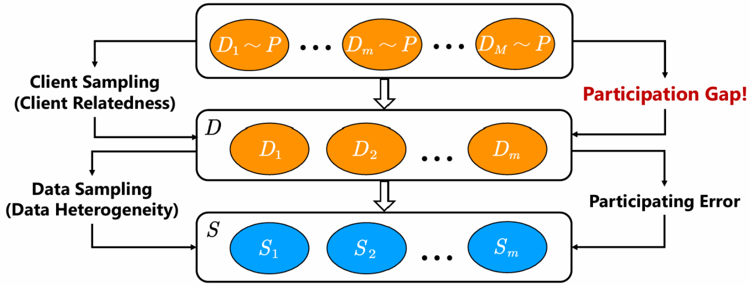
- 泛化误差：测试误差
- 未参与客户
- 异质性数据：Non-IID
- 无界损失函数

2. 研究假设：双层分布框架

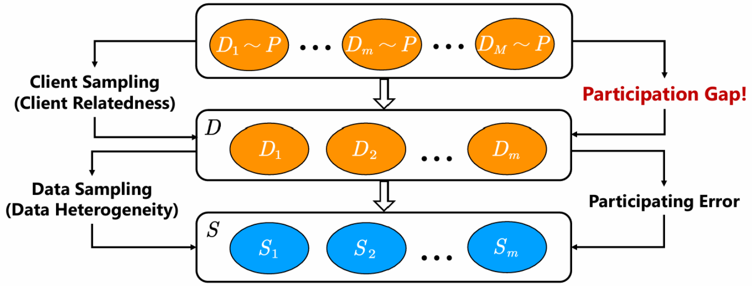
以下是一些符号的定义：

- $\mathcal{X} \subseteq \mathbb{R}^k, \mathcal{Y} \subseteq \mathbb{R}, Z = (X, Y) \in \mathcal{X} \times \mathcal{Y}$：输入输出空间，随机变量 Z
- $\mathcal{D}$：Z 的真实分布；P：分布 $\mathcal{D}$ 的**元分布**。
- 总客户数量 M，参与客户数量 m，而且 $m \ll M$
- D_i 是客户 i 的分布。$\{D_1, \ldots, D_m\}$ 是根据 P 从 $\mathcal{D}$ 独立同分布采样得到的。数据样本 $S_i = \{Z_i^j\}_{j=1}^n$ 是从 D_i 中独立同分布采样得到的。
- $h \in \mathcal{H}$：假设函数，也就是模型
- $l(h, Z_i)$：客户端 i 的损失函数

2. 研究假设：双层分布框架



2. 研究假设：双层分布框架



问题：未参与用户能不能从联邦学习中获益？

2. 研究假设：双层分布框架

定义 (总体风险)

$$\mathcal{L}_P(h) = \mathbb{E}_{D_i \sim P}[\mathbb{E}_{Z \sim D_i}[l(h(X), Y)]]$$

对应的最优总体风险函数是：

$$h^* = \arg\min_{h \in \mathcal{H}} \mathcal{L}_P(h)$$

然而，元分布 P 是未知的、客户的真实分布 D 也是未知的，以上的式子仅用于理论分析。我们要用经验风险来估计总体风险，而不是直接计算。

2. 研究假设：双层分布框架

整篇文章都是围绕该问题展开。为了进一步分析，我们在该框架下定义：

定义 (经验风险)

$$\mathcal{L}_S(h) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{n}\sum_{j=1}^{n} l(h(X_i^j), Y_i^j)$$

其中 (X_i^j, Y_i^j) 表示客户 i 的第 j 个样本，$S = \cup_{i=1}^m S_i$ 对应的最优经验风险函数是：

$$\hat{h} = \arg\min_{h\in\mathcal{H}} \mathcal{L}_S(h)$$

经验风险就是训练误差，可以直接计算。特别地，$\hat{h}$ 就是联邦学习学到的模型。

2. 研究假设：双层分布框架

为了在双层分布框架下分析**泛化误差**，我们定义了半经验风险(semi-empirical risk)。

定义 (半经验风险)

首先定义**半经验分布** $D = \frac{1}{m}\sum_{i=1}^{m} D_i$

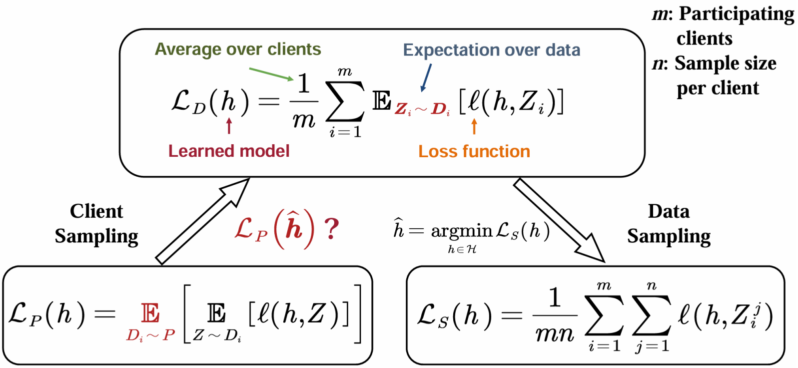
$$\mathcal{L}_D(h) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}_{Z\sim D_i}[l(h(X), Y)]$$

对应的最优半经验风险函数是：

$$\hat{h}^* = \arg\min_{h\in\mathcal{H}} \mathcal{L}_D(h)$$

2. 研究假设：双层分布框架

用一张图来表现它们的关系：



2. 研究假设：双层分布框架

过度误差和半过度误差

- 半过度误差 $\mathcal{L}_D(\hat{h}) - \mathcal{L}_D(\hat{h}^*)$：反映了已经学到的模型 $\hat{h}$ 在半经验分布 D 上的、未出现数据上的表现。
- 过度误差 $\mathcal{L}_P(\hat{h}) - \mathcal{L}_P(h^*)$：反映了已经学到的模型 $\hat{h}$ 在未参与客户上的表现。

2. 研究假设：双层分布框架

在此框架下，我们再借助两个概念，来得到第一个结论。这里是 VC 维（PAC 理论的奠基概念之一）：

定义 (Vapnik-Chervonenkis 维度)

Definition 1 (VC dimension). *Let $(\mathcal{X}, \mathcal{H})$ be a set system that consists of a set and a class H of subsets of X. A set system $(\mathcal{X}, \mathcal{H})$ shatters a set A if each subset of A can be expressed as $A \cap h$ for some h in $\mathcal{H}$. The VC-dimension of $\mathcal{H}$ is the size of the largest set shattered by $\mathcal{H}$.*

Definition 2 (VC subgraph of real valued function). *The subgraph of a function $h(\in \mathcal{H}) : \mathcal{X} \to \mathbb{R}$ is the subset of $\mathcal{X} \times \mathbb{R}$ given by $\{(x,t) : t < h(x)\}$. Then the VC-dimension of the function class $\mathcal{F}$ is defined as the VC-dimension of the set of subgraphs of functions in $\mathcal{H}$.*

VC 维是衡量假设空间的复杂度的一个重要指标。（建议看`https://tangshusen.me/2018/12/09/vc-dimension/`）
def1 依赖数据集，def2 不需要（打散的最大数据集）!
直线 VC 维是 p+1。

2. 研究假设：双层分布框架

Theorem 1 (Generalization error for unparticipating clients). *Let $\mathcal{F}$ be a family of functions related to hypothesis space $\mathcal{H}$: $\mathcal{F} = \{z \mapsto \ell(h,z) : h \in \mathcal{H}\}$. For the* VC *subgraph class $\mathcal{F}$ with* VC *dimension* d*. If the loss function ℓ is bounded by* b*, it follows that with probability at least $1-2\delta$,*

$$\mathcal{L}_P(\widehat{h}) - \mathcal{L}_P(h^*) \le c_1 b\sqrt{\frac{d}{mn}} + b\sqrt{\frac{\ln(1/\delta)}{2mn}} + c_2 b\sqrt{\frac{d}{m}} + b\sqrt{\frac{\ln(1/\delta)}{2m}},$$

where c_1 and c_2 are constants.

2. 研究假设：双层分布框架

Theorem 1 (Generalization error for unparticipating clients). *Let $\mathcal{F}$ be a family of functions related to hypothesis space $\mathcal{H}$: $\mathcal{F} = \{z \mapsto \ell(h, z) : h \in \mathcal{H}\}$. For the* VC *subgraph class $\mathcal{F}$ with* VC *dimension d. If the loss function ℓ is bounded by b, it follows that with probability at least* $1 - 2\delta$,

$$\mathcal{L}_P(\widehat{h}) - \mathcal{L}_P(h^*) \leq c_1 b \sqrt{\frac{d}{mn}} + b\sqrt{\frac{\ln(1/\delta)}{2mn}} + c_2 b\sqrt{\frac{d}{m}} + b\sqrt{\frac{\ln(1/\delta)}{2m}},$$

where c_1 and c_2 are constants.

Remark

- 注意，这里不是一个普通的不等式，而是“概率近似正确”。
- 左边 $\mathcal{L}_P(\hat{h}) - \mathcal{L}_P(h^*)$ 是“过度误差”，也就是模型在 M 个用户中的表现。$\hat{h}$ 是训练模型，是在 m 个参与用户中训练得到的。通俗来讲，定理 1 告诉我们参与训练的 m 越大，过度风险越小。
- 我们可以积极地回答刚开始提出的问题：从平均表现上来说，未参与用户也能从联邦学习中获益。

2. 研究假设：双层分布框架

结论的缺陷

在假设条件下过度风险 $\mathcal{L}_P(\hat{h}) - \mathcal{L}_P(h^*)$ 可以被 $O(\sqrt{\frac{1}{mn}} + \sqrt{\frac{1}{m}})$ 控制。但是当参与客户异质性很高时，$\mathcal{L}_P(h^*)$ 可能会非常大（难以兼顾所有客户）。这导致未参与客户的泛化误差 $\mathcal{L}_P(\hat{h})$ 也会很大。即，**异质性高时，不能指望一个通用的模型能总是表现良好。**

实验结果 (虽然证明有 VC 维，实验用的是神经网络)。

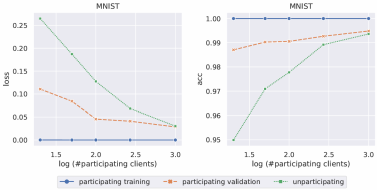


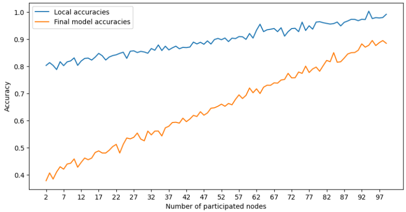
Figure 2: Generalization error versus the number of participating clients.

2. 研究假设：双层分布框架

结论的缺陷

在假设条件下过度风险 $\mathcal{L}_P(\hat{h}) - \mathcal{L}_P(h^*)$ 可以被 $O(\sqrt{\frac{1}{mn}} + \sqrt{\frac{1}{m}})$ 控制。但是当参与客户异质性很高时，$\mathcal{L}_P(h^*)$ 可能会非常大（难以兼顾所有客户）。这导致未参与客户的泛化误差 $\mathcal{L}_P(\hat{h})$ 也会很大。即，**异质性高时，不能指望一个通用的模型能总是表现良好。**

我的复现结果：（没给源码，自己模拟的）



3. 更快的学习率：在双层分布框架下

回忆一下：

3. 更快的学习率：在双层分布框架下

回忆一下：

- $\hat{h}$ 是训练模型，经验风险最小化得到的
- h^* 是最优模型，总体风险最小化得到的

3. 更快的学习率：在双层分布框架下

回忆一下：

- $\hat{h}$ 是训练模型，经验风险最小化得到的
- h^* 是最优模型，总体风险最小化得到的
- 半过度风险 $\mathcal{L}_D(\hat{h}) - \mathcal{L}_D(\hat{h^*})$ 反映了**参与客户**的学习率。
- 过度风险 $\mathcal{L}_P(\hat{h}) - \mathcal{L}_P(h^*)$ 反映了**未参与客户**的学习率。

为了获得更快的学习率，我们对损失函数 l、假设空间 $\mathcal{H}$、半经验分布 D、元分布 P 进行了一些合理的假设。

3. 更快的学习率：在双层分布框架下

假设 1：李普希茨条件

Assumption 1. *Loss function* ℓ *is* L*-Lipschitz in its first argument:* $|\ell(y_1, y) - \ell(y_2, y)| \leq L\,|y_1 - y_2|$.

3. 更快的学习率：在双层分布框架下

假设 1：李普希茨条件

Assumption 1. *Loss function ℓ is L-Lipschitz in its first argument:* $|\ell(y_1, y) - \ell(y_2, y)| \leq L|y_1 - y_2|$.

假设 2：伯恩斯坦条件

Definition 3 (Bernstein condition). *Let μ be a distribution supported on $\mathcal{X} \times \mathcal{Y}$ and let ℓ be a loss function with domain $\mathcal{Y} \times \mathcal{Y}$. The tuple $(\mu, \ell, \mathcal{H}, h^*)$ satisfies the (β, B)-Bernstein condition with parameter $B > 0$ if the following holds for any $h \in \mathcal{H}$:*

$$\mathbb{E}\left(h(X) - h^*(X)\right)^2 \leq B\mathbb{E}\left[\ell(h(X), Y) - \ell(h^*(X), Y)\right]^\beta .$$

Assumption 2. *Theoretical analyses in our two-level distribution framework involve different types of Bernstein conditons:*

(a) *The tuple $(D, \ell, \mathcal{H}, \widehat{h}^*)$ satisfies the Bernstein condition with parameter $B' \geq 1, 0 < \beta' \leq 1$. That is, for any $h \in \mathcal{H}$, $\frac{1}{m}\sum_{i=1}^m \mathbb{E}[h(X_i^1) - \widehat{h}^*(X_i^1)]^2 \leq B'(\mathcal{L}_D(h) - \mathcal{L}_D(\widehat{h}^*))^{\beta'}$.*

(b) *The tuple $(P, \ell, \mathcal{H}, h^*)$ satisfies the Bernstein condition with parameter $B'' \geq 1, 0 < \beta'' \leq 1$. That is, for any $h \in \mathcal{H}$, $\mathbb{E}_{D_i \sim P}[\mathbb{E}_{X \sim D_i}[h(X) - h^*(X)]^2] \leq B''(\mathcal{L}_P(h) - \mathcal{L}_P(h^*))^{\beta''}$.*

3. 更快的学习率：在双层分布框架下

以下是作者的解释：

- 为了得到更快的学习率，大家都会给很多假设，而伯恩斯坦条件常常在理论机器学习中使用。
- 该条件并不苛刻。比如：
  - 任何有界的概率分布函数；
  - 回归问题中严格凸的损失函数；
  - 强凸且李普希茨连续的损失函数。
  - 均方误差且假设函数集合为凸。

3. 更快的学习率：在双层分布框架下

以下是作者的解释：

- 为了得到更快的学习率，大家都会给很多假设，而伯恩斯坦条件常常在理论机器学习中使用。
- 该条件并不苛刻。比如：
  - 任何有界的概率分布函数；
  - 回归问题中严格凸的损失函数；
  - 强凸且李普希茨连续的损失函数。
  - 均方误差且假设函数集合为凸。

以下是我对上面两个假设的理解：

- 根据数学分析的知识，李普希茨条件意味着损失函数 **一致连续**。这个要求比连续性更强，要求连续并且不震荡变化，比如 $y = \sin(x^2)$ 就不满足。我认为该假设很合理，因为常用的损失函数都满足该条件。
- **伯恩斯坦条件**：对半经验分布 D 和元分布 P 进行限制，为了得到更紧的界（后面还会再理解一次）。

3. 更快的学习率：在双层分布框架下

假设 3：一致熵条件 (Uniform Entropy Condition)

Definition 7 (Convering number)**.** *Let $(\mathcal{G}, \rho)$ be a metric space and $\mathcal{F} \subseteq \mathcal{G}$. For any $\epsilon \geq 0$, $\mathcal{F}_\epsilon$ is an ϵ-cover of $\mathcal{F}$ with respect of ρ if for all $f \in F$, we can find $f' \in \mathcal{F}_\epsilon$ such that $\rho(f, f') \leq \epsilon$. The covering number $\mathcal{N}(\epsilon, \mathcal{F}, \rho)$ is defined as the minimum size of an ϵ-cover:*

$$\mathcal{N}(\epsilon, \mathcal{F}, \rho) := \min\{|\mathcal{F}_\epsilon| : \mathcal{F}_\epsilon \text{ is an } \epsilon\text{-cover of } \mathcal{F} \text{ w.r.t } \rho\}.$$

Definition 8 (Uniform entropy number)**.** *The entropy number is defined as the logarithm of the covering number. Let $(\mathcal{G}, \rho)$ be a normed space with $\rho(f, f') = \|f - f'\|$. Let F be an envelope function of $\mathcal{F}$ such that $|f(Z)| \leq F(Z)$, for all Z and f. We further define uniform entropy number of $\mathcal{F}$ as: $\log \mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|_2) = \sup_Q \log \mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|_{L_2(Q)})$, where Q is taken over all probability measures with $0 \leq QF^2 \leq \infty$.*

Definition 9. *A function $\varphi(r) : [0, \infty) \mapsto [0, \infty)$ is sub-root function if it is nondecreasing and $r \mapsto \varphi(r)/\sqrt{r}$ is nonincreasing for $r > 0$.*

3. 更快的学习率：在双层分布框架下

假设 3：一致熵条件 (Uniform Entropy Condition)

Definition 7 (Convering number). *Let $(\mathcal{G}, \rho)$ be a metric space and $\mathcal{F} \subseteq \mathcal{G}$. For any $\epsilon \geq 0$, $\mathcal{F}_\epsilon$ is an ϵ-cover of $\mathcal{F}$ with respect of ρ if for all $f \in F$, we can find $f' \in \mathcal{F}_\epsilon$ such that $\rho(f, f') \leq \epsilon$. The covering number $\mathcal{N}(\epsilon, \mathcal{F}, \rho)$ is defined as the minimum size of an ϵ-cover:*

$$\mathcal{N}(\epsilon, \mathcal{F}, \rho) := \min\{|\mathcal{F}_\epsilon| : \mathcal{F}_\epsilon \text{ is an } \epsilon\text{-cover of } \mathcal{F} \text{ w.r.t } \rho\}.$$

Definition 8 (Uniform entropy number). *The entropy number is defined as the logarithm of the covering number. Let $(\mathcal{G}, \rho)$ be a normed space with $\rho(f, f') = \|f - f'\|$. Let F be an envelope function of $\mathcal{F}$ such that $|f(Z)| \leq F(Z)$, for all Z and f. We further define uniform entropy number of $\mathcal{F}$ as: $\log \mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|_2) = \sup_Q \log \mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|_{L_2(Q)})$, where Q is taken over all probability measures with $0 \leq QF^2 \leq \infty$.*

Definition 9. *A function $\varphi(r) : [0, \infty) \mapsto [0, \infty)$ is sub-root function if it is nondecreasing and $r \mapsto \varphi(r)/\sqrt{r}$ is nonincreasing for $r > 0$.*

Assumption 3 (Uniform entropy number[1]). *Let $\mathcal{H}$ be a family of bounded functions with uniformly entropy number $\log \mathcal{N}(\epsilon, \mathcal{H}, \|\cdot\|_2)$. Assume that there exist positive numbers γ, d and p such that $\log \mathcal{N}(\epsilon, \mathcal{H}, \|\cdot\|_2) \leq d \log^p(\gamma/\epsilon)$ for any $0 < \epsilon \leq \gamma$.*

3. 更快的学习率：在双层分布框架下

作者提到，假设 3 是一个很轻微的假设。下面是一些常见的满足假设 3 的函数：

- 函数集合有界；
- $\mathcal{H}$ 的 VC 维是有限的；
- 如果令 $\epsilon \in (0,1)$，那么所有的欧几里得单位球 $\mathcal{B} \subseteq \mathbb{R}^d$ 都满足。
- 如果 $\mathcal{H}$ 是以 k 为核函数的希尔伯特再生核空间，且 k 的秩为 d，那么满足假设。

3.1 更快的学习率：对于参与客户

- 回忆：半过度风险 $\mathcal{L}_D(\hat{h}) - \mathcal{L}_D(\hat{h}^*)$ 反映了**参与客户**的学习率。

3.1 更快的学习率：对于参与客户

- 回忆：半过度风险 $\mathcal{L}_D(\hat{h}) - \mathcal{L}_D(\hat{h}^*)$ 反映了**参与客户**的学习率。

在之前的假设下，推导出了以下结论：

Theorem 2 (Semi-excess risk for participating clients). *Let $\mathcal{F}$ be a family of functions bounded by b. Under assumptions $\boxed{1}$, $\boxed{3}$ and (a) of Assumption $\boxed{2}$, when $mn \geq cd\log^p(mn)$, it follows that with probability at least $1-\delta$,*

$$\mathcal{L}_D(\widehat{h}) - \mathcal{L}_D(\widehat{h}^*) \leq c_1 \left(\frac{\log^p(mn)}{mn}\right)^{\frac{1}{2-\beta'}} + c_2 \left(\frac{\log(1/\delta)}{mn}\right)^{\frac{1}{2-\beta'}},$$

where c_1 and c_2 are constants depending on γ, p, L, β' and B_1, b, β' respectively.

3.1 更快的学习率：对于参与客户

Theorem 2 (Semi-excess risk for participating clients). *Let $\mathcal{F}$ be a family of functions bounded by b. Under assumptions 1, 3 and (a) of Assumption 2, when $mn \geq cd\log^p(mn)$, it follows that with probability at least $1-\delta$,*

$$\mathcal{L}_D(\widehat{h}) - \mathcal{L}_D(\widehat{h}^*) \leq c_1 \left(\frac{\log^p(mn)}{mn}\right)^{\frac{1}{2-\beta'}} + c_2 \left(\frac{\log(1/\delta)}{mn}\right)^{\frac{1}{2-\beta'}},$$

where c_1 and c_2 are constants depending on γ, p, L, β' and B_1, b, β' respectively.

Remark

- 收敛速率（以概率）在 $O(\frac{1}{\sqrt{mn}})$ 到 $O(\frac{1}{mn})$ 之间，对应 β' 取 0 和 1。（来自伯恩斯坦条件的假设）
- 在伯恩斯坦条件下，当增加参与客户数量 m 和本地数据数量 n 时，半经验风险收敛更快。
- 该结论是 PAC 形式。而之前的研究都是期望形式。

3.2 更快的学习率：对于未参与客户

- 回忆：过度风险 $\mathcal{L}_P(\hat{h}) - \mathcal{L}_P(h^*)$ 反映了**未参与客户**的学习率。

3.2 更快的学习率：对于未参与客户

- 回忆：过度风险 $\mathcal{L}_P(\hat{h}) - \mathcal{L}_P(h^*)$ 反映了**未参与客户**的学习率。

下面的结论是已有研究中最紧的界！

Theorem 3. *Let $\mathcal{F}$ be a family of functions bounded by b. Under assumptions $\boxed{1}$, $\boxed{3}$ and (b) of Assumption $\boxed{2}$, when $m \geq cd\log^p(m)$, for any $\delta > 0$, it follows that with probability at least $1-\delta$,*

$$\mathcal{L}_P(\widehat{h}) - \mathcal{L}_P(h^*) \leq c_0 \left(\mathcal{L}_D(\widehat{h}) - \mathcal{L}_D(\widehat{h}^*)\right) + c_1 \left(\frac{\log^p m}{m}\right)^{\frac{1}{2-\beta''}} + c_2 \left(\frac{\log(1/\delta)}{m}\right)^{\frac{1}{2-\beta''}},$$

where $c_0 = \frac{K}{K-\beta''}$, c_1 and c_2 are constants depending on γ, p, L, β'' and B_2, b, β'' respectively.

3.2 更快的学习率：对于未参与客户

Theorem 3. *Let $\mathcal{F}$ be a family of functions bounded by b. Under assumptions 1, 3 and (b) of Assumption 2, when $m \geq cd\log^p(m)$, for any $\delta > 0$, it follows that with probability at least $1-\delta$,*

$$\mathcal{L}_P(\widehat{h}) - \mathcal{L}_P(h^*) \leq c_0 \left(\mathcal{L}_D(\widehat{h}) - \mathcal{L}_D(\widehat{h}^*)\right) + c_1 \left(\frac{\log^p m}{m}\right)^{\frac{1}{2-\beta''}} + c_2 \left(\frac{\log(1/\delta)}{m}\right)^{\frac{1}{2-\beta''}},$$

where $c_0 = \frac{K}{K-\beta''}$, c_1 and c_2 are constants depending on γ, p, L, β'' and B_2, b, β'' respectively.

Remark

- 右边的 $\mathcal{L}_D(\hat{h}) - \mathcal{L}_D(\hat{h}^*)$ 已经被定理 2 控制住了。
- β', β'' 是伯恩斯坦假设中的常数。当它们取 1 时，超越风险以高概率在 $O(\frac{1}{mn} + \frac{1}{m})$ 之下。

4. 次韦伯分布损失函数：学习率

在这一节，作者给出了双层分布框架下，以次韦伯分布为损失函数的泛化误差的界（学习率）。

定义 (次韦伯分布)

Definition 4 (Sub-Weibull random variables). *A random variable X is said to be sub-Weibull if there is constant $\|X\|_{\psi_\alpha} < \infty$, such that*

$$\mathbb{P}(|X| \geq t) \leq 2\exp(-t^\alpha/\|X\|_{\psi_\alpha}^\alpha), \textit{ for all } t \geq 0.$$

Sub-Gaussian and sub-exponential random variables are two special cases of Sub-Weibull random variables, which correspond to $\alpha = 2$ and $\alpha = 1$, respectively.

4. 次韦伯分布损失函数：学习率

在这一节，作者给出了双层分布框架下，以次韦伯分布为损失函数的泛化误差的界（学习率）。

定义 (次韦伯分布)

Definition 4 (Sub-Weibull random variables). *A random variable X is said to be sub-Weibull if there is constant $\|X\|_{\psi_\alpha} < \infty$, such that*

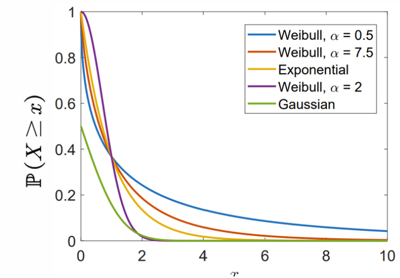
$$\mathbb{P}(|X| \geq t) \leq 2\exp(-t^\alpha / \|X\|_{\psi_\alpha}^\alpha), \text{ for all } t \geq 0.$$

Sub-Gaussian and sub-exponential random variables are two special cases of Sub-Weibull random variables, which correspond to $\alpha = 2$ and $\alpha = 1$, respectively.

次韦伯分布是一种厚尾分布（heavy-tail）。

4. 次韦伯分布损失函数：学习率

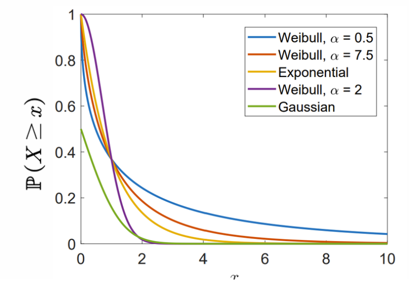
厚尾分布的例子：



图片来源：`https://adamwierman.com/wp-content/uploads/2021/05/book-05-11.pdf`

4. 次韦伯分布损失函数：学习率

厚尾分布的例子：

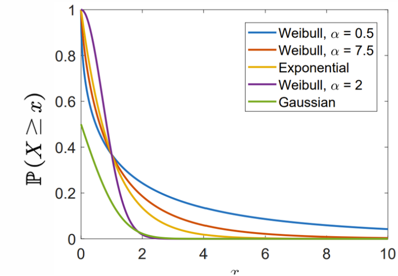


图片来源：https://adamwierman.com/wp-content/uploads/2021/05/book-05-11.pdf

- α 越小，分布的尾部越厚。

4. 次韦伯分布损失函数：学习率

厚尾分布的例子：

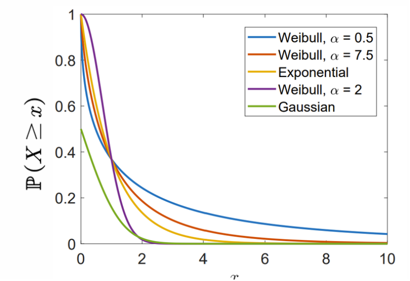


图片来源：`https://adamwierman.com/wp-content/uploads/2021/05/book-05-11.pdf`

- α 越小，分布的尾部越厚。
- 极端事件的概率更高。损失函数在更多数据点上取值更大。（数据不好/模型差/选取 loss 原因）

4. 次韦伯分布损失函数：学习率

厚尾分布的例子：

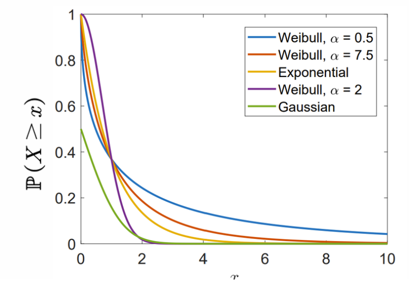


图片来源：`https://adamwierman.com/wp-content/uploads/2021/05/book-05-11.pdf`

- α 越小，分布的尾部越厚。
- 极端事件的概率更高。损失函数在更多数据点上取值更大。（数据不好/模型差/选取 loss 原因）
- 条件更加苛刻（甚至超过了指数分布的尾巴），推导也更困难。

4. 次韦伯分布损失函数：学习率

厚尾分布的例子：



图片来源：`https://adamwierman.com/wp-content/uploads/2021/05/book-05-11.pdf`

- α 越小，分布的尾部越厚。
- 极端事件的概率更高。损失函数在更多数据点上取值更大。（数据不好/模型差/选取 loss 原因）
- 条件更加苛刻（甚至超过了指数分布的尾巴），推导也更困难。
- 和之前伯恩斯坦条件不同！伯恩斯坦条件能推出次高斯的 bound。

4. 次韦伯分布损失函数：学习率

补充：伯恩斯坦条件推出次高斯

https://www.stat.cmu.edu/~arinaldo/Teaching/36709/S19/Scribed_Lectures/Feb5_Aleksandr.pdf

Definition 5.3 (Bernstein condition) *Let X be a random variable with mean μ and variance σ^2. Assume that $\exists b > 0$:*

$$\mathbb{E}|X-\mu|^k \leq \frac{1}{2}k!\sigma^2 b^{k-2}, k = 3, 4, \ldots$$

Then one says that X satisfies Bernstein condition.

Lemma 5.4 *If random variable X satisfies Bernstein condition with parameter b, then:*

$$\mathbb{E}e^{\lambda(X-\mu)} \leq e^{\frac{\lambda^2\sigma^2}{2}\frac{1}{1-b|\lambda|}},\ \forall|\lambda| < \frac{1}{b}$$

Additionally, from the bound on the moment generating function one can obtain the following tail bound (also known as Bernstein inequality):

$$\mathbb{P}(|X-\mu| \geq t) \leq 2\exp\left(-\frac{t^2}{2(\sigma^2+bt)}\right), \forall t > 0$$

Proof: Pick $\lambda : |\lambda| < \frac{1}{b}$ (allowing interchanging summation and taking expectation) and expand the MGF in a Taylor series:

$$\mathbb{E}e^{\lambda(X-\mu)} = 1 + \frac{\lambda^2\sigma^2}{2} + \sum_{k=3}^{\infty}\frac{\mathbb{E}|X-\mu|^k}{k!}\lambda^k \leq 1 + \frac{\lambda^2\sigma^2}{2} + \frac{\lambda^2\sigma^2}{2}\sum_{k=3}^{\infty}(|\lambda|b)^{k-2} =$$

$$= 1 + \frac{\lambda^2\sigma^2}{2}\frac{1}{1-b|\lambda|} \leq e^{\frac{\lambda^2\sigma^2}{2}\frac{1}{1-b|\lambda|}}$$

where we used $1 + x \leq e^x$. To show the final bound, take $\lambda : |\lambda| < \frac{1}{2b}$. Then the bound becomes:

$$e^{\frac{\lambda^2\sigma^2}{2}\frac{1}{1-b|\lambda|}} \leq e^{\lambda^2\sigma^2} = e^{\frac{\lambda^2(2\sigma^2)}{2}}$$

implying that $X \in SE(2\sigma^2, 2b)$. The concentration result then follow by taking $\lambda = \frac{t}{bt+\sigma^2}$.

4. 次韦伯分布损失函数：学习率

对于厚尾分布，许多**集中不等式**（比如之前提的霍夫丁）不能使用，自然需要新的方法。下面是一些定义：

- $\|h\|_{L_2(\mu)}$ 表示 Banach 空间 $L_2(\mathcal{X}, \mu)$。（完备赋范向量空间）
- D 是半经验分布，P 是元分布。
- $\|h\|_{L_2(D)} = (\frac{1}{m}\sum_{i=1}^{m} \mathbb{E}_{X\sim D_i}[h(X)]^2)^{1/2}$
- $\|h\|_{L_2(P)} = (\mathbb{E}_{D_i\sim P}\mathbb{E}_{X\sim D_i}[h(X)]^2)^{1/2}$

4. 次韦伯分布损失函数：学习率

假设：小球条件

Assumption 4 (Small-ball condition)**.** *Let $\mathcal{H} \subset L_2(D)$ be a closed and convex class of functions and $\mathcal{H} - \mathcal{H} := \{h - h' : h, h' \in \mathcal{H}\}$.*

(a) Let $Q_{mn}(\tau) = \inf_{h \in \mathcal{H}-\mathcal{H}} \mathbb{P}(|h(X_i^1)| \geq \tau \|h\|_{L_2(D)})$, where X_i^1 represent the random sample at i-th participating client. There is a $\tau \geq 0$ for which $Q_{mn}(\tau) > 0$.

(b) Let $Q_m(\tau) = \inf_{h \in \mathcal{H}-\mathcal{H}} \mathbb{P}\left(|\mathbb{E}[h(X_i^1)]| \geq \tau \|h\|_{L_2(P)}\right)$, where X_i^1 represent the random sample at i-th participating client. There is a $\tau \geq 0$ for which $Q_m(\tau, P) > 0$.

Remark

- 小球条件是对独立同分布的数据产生过程的假设。

4. 次韦伯分布损失函数：学习率

假设：小球条件

Assumption 4 (Small-ball condition). *Let $\mathcal{H} \subset L_2(D)$ be a closed and convex class of functions and $\mathcal{H} - \mathcal{H} := \{h - h' : h, h' \in \mathcal{H}\}$.*

(a) *Let $Q_{mn}(\tau) = \inf_{h \in \mathcal{H}-\mathcal{H}} \mathbb{P}(|h(X_i^1)| \geq \tau \|h\|_{L_2(D)})$, where X_i^1 represent the random sample at i-th participating client. There is a $\tau \geq 0$ for which $Q_{mn}(\tau) > 0$.*

(b) *Let $Q_m(\tau) = \inf_{h \in \mathcal{H}-\mathcal{H}} \mathbb{P}\left(|\mathbb{E}[h(X_i^1)]| \geq \tau \|h\|_{L_2(P)}\right)$, where X_i^1 represent the random sample at i-th participating client. There is a $\tau \geq 0$ for which $Q_m(\tau, P) > 0$.*

Remark

- 小球条件是对独立同分布的数据产生过程的假设。
- 在高等概率论中，泛化误差常常要用集中不等式。直观上，只有当损失函数的各阶矩很好时，**经验风险**会以高概率**集中**在**总体风险**附近。

4. 次韦伯分布损失函数：学习率

假设：小球条件

Assumption 4 (Small-ball condition). *Let $\mathcal{H} \subset L_2(D)$ be a closed and convex class of functions and $\mathcal{H} - \mathcal{H} := \{h - h' : h, h' \in \mathcal{H}\}$.*

(a) Let $Q_{mn}(\tau) = \inf_{h \in \mathcal{H}-\mathcal{H}} \mathbb{P}(|h(X_i^1)| \geq \tau \|h\|_{L_2(D)})$, where X_i^1 represent the random sample at i-th participating client. There is a $\tau \geq 0$ for which $Q_{mn}(\tau) > 0$.

(b) Let $Q_m(\tau) = \inf_{h \in \mathcal{H}-\mathcal{H}} \mathbb{P}\left(|\mathbb{E}[h(X_i^1)]| \geq \tau \|h\|_{L_2(P)}\right)$, where X_i^1 represent the random sample at i-th participating client. There is a $\tau \geq 0$ for which $Q_m(\tau, P) > 0$.

Remark

- 小球条件是对独立同分布的数据产生过程的假设。
- 在高等概率论中，泛化误差常常要用集中不等式。直观上，只有当损失函数的各阶矩很好时，**经验风险**会以高概率**集中**在**总体风险**附近。
- 但对于厚尾分布不适用！（另一定义：矩母函数发散）

4. 次韦伯分布损失函数：学习率

假设：小球条件

Assumption 4 (Small-ball condition). *Let $\mathcal{H} \subset L_2(D)$ be a closed and convex class of functions and $\mathcal{H} - \mathcal{H} := \{h - h' : h, h' \in \mathcal{H}\}$.*

(a) Let $Q_{mn}(\tau) = \inf_{h \in \mathcal{H}-\mathcal{H}} \mathbb{P}(|h(X_i^1)| \geq \tau \|h\|_{L_2(D)})$, where X_i^1 represent the random sample at i-th participating client. There is a $\tau \geq 0$ for which $Q_{mn}(\tau) > 0$.

(b) Let $Q_m(\tau) = \inf_{h \in \mathcal{H}-\mathcal{H}} \mathbb{P}\left(|\mathbb{E}[h(X_i^1)]| \geq \tau \|h\|_{L_2(P)}\right)$, where X_i^1 represent the random sample at i-th participating client. There is a $\tau \geq 0$ for which $Q_m(\tau, P) > 0$.

Remark

- 于是小球条件被提出。

4. 次韦伯分布损失函数：学习率

假设：小球条件

Assumption 4 (Small-ball condition). *Let $\mathcal{H} \subset L_2(D)$ be a closed and convex class of functions and $\mathcal{H} - \mathcal{H} := \{h - h' : h, h' \in \mathcal{H}\}$.*

(a) Let $Q_{mn}(\tau) = \inf_{h \in \mathcal{H}-\mathcal{H}} \mathbb{P}(|h(X_i^1)| \geq \tau \|h\|_{L_2(D)})$, where X_i^1 represent the random sample at i-th participating client. There is a $\tau \geq 0$ for which $Q_{mn}(\tau) > 0$.

(b) Let $Q_m(\tau) = \inf_{h \in \mathcal{H}-\mathcal{H}} \mathbb{P}\left(|\mathbb{E}[h(X_i^1)]| \geq \tau \|h\|_{L_2(P)}\right)$, where X_i^1 represent the random sample at i-th participating client. There is a $\tau \geq 0$ for which $Q_m(\tau, P) > 0$.

Remark

- 于是小球条件被提出。
- 弱条件 $\|h\|_{L_p(P)} \leq c\|h\|_{L_2(P)}$ 就能推出。

4. 次韦伯分布损失函数：学习率

假设：小球条件

Assumption 4 (Small-ball condition). *Let* $\mathcal{H} \subset L_2(D)$ *be a closed and convex class of functions and* $\mathcal{H} - \mathcal{H} := \{h - h' : h, h' \in \mathcal{H}\}$.

(a) *Let* $Q_{mn}(\tau) = \inf_{h \in \mathcal{H}-\mathcal{H}} \mathbb{P}(|h(X_i^1)| \geq \tau \|h\|_{L_2(D)})$, *where* X_i^1 *represent the random sample at* i*-th participating client. There is a* $\tau \geq 0$ *for which* $Q_{mn}(\tau) > 0$.

(b) *Let* $Q_m(\tau) = \inf_{h \in \mathcal{H}-\mathcal{H}} \mathbb{P}\left(|\mathbb{E}[h(X_i^1)]| \geq \tau \|h\|_{L_2(P)}\right)$, *where* X_i^1 *represent the random sample at* i*-th participating client. There is a* $\tau \geq 0$ *for which* $Q_m(\tau, P) > 0$.

Remark

- 于是小球条件被提出。
- 弱条件 $\|h\|_{L_p(P)} \leq c\|h\|_{L_2(P)}$ 就能推出。
- 假设 (b) 是把 D_i 看成从 P 产生的随机变量

4. 次韦伯分布损失函数：学习率

假设：小球条件

Assumption 4 (Small-ball condition)**.** *Let $\mathcal{H} \subset L_2(D)$ be a closed and convex class of functions and $\mathcal{H} - \mathcal{H} := \{h - h' : h, h' \in \mathcal{H}\}$.*

(a) Let $Q_{mn}(\tau) = \inf_{h \in \mathcal{H}-\mathcal{H}} \mathbb{P}(|h(X_i^1)| \geq \tau \|h\|_{L_2(D)})$, where X_i^1 represent the random sample at i-th participating client. There is a $\tau \geq 0$ for which $Q_{mn}(\tau) > 0$.

(b) Let $Q_m(\tau) = \inf_{h \in \mathcal{H}-\mathcal{H}} \mathbb{P}\left(|\mathbb{E}[h(X_i^1)]| \geq \tau \|h\|_{L_2(P)}\right)$, where X_i^1 represent the random sample at i-th participating client. There is a $\tau \geq 0$ for which $Q_m(\tau, P) > 0$.

Remark

- 假设 (a) 仅需要 $\mathcal{H} \subset L_2(D)$ 以高概率成立。（后面我会解释 $\mathcal{H} - \mathcal{H}$ 的原因）

4. 次韦伯分布损失函数：学习率

假设：小球条件

Assumption 4 (Small-ball condition). *Let $\mathcal{H} \subset L_2(D)$ be a closed and convex class of functions and $\mathcal{H} - \mathcal{H} := \{h - h' : h, h' \in \mathcal{H}\}$.*

(a) Let $Q_{mn}(\tau) = \inf_{h \in \mathcal{H}-\mathcal{H}} \mathbb{P}(|h(X_i^1)| \geq \tau \|h\|_{L_2(D)})$, where X_i^1 represent the random sample at i-th participating client. There is a $\tau \geq 0$ for which $Q_{mn}(\tau) > 0$.

(b) Let $Q_m(\tau) = \inf_{h \in \mathcal{H}-\mathcal{H}} \mathbb{P}\left(|\mathbb{E}[h(X_i^1)]| \geq \tau \|h\|_{L_2(P)}\right)$, where X_i^1 represent the random sample at i-th participating client. There is a $\tau \geq 0$ for which $Q_m(\tau, P) > 0$.

Remark

- 假设 (a) 仅需要 $\mathcal{H} \subset L_2(D)$ 以高概率成立。（后面我会解释 $\mathcal{H} - \mathcal{H}$ 的原因）
- 该条件并不严苛，因为 D 是从 P 中独立同分布抽样得到的。

4. 次韦伯分布损失函数：学习率

假设：小球条件

Assumption 4 (Small-ball condition)**.** *Let $\mathcal{H} \subset L_2(D)$ be a closed and convex class of functions and $\mathcal{H} - \mathcal{H} := \{h - h' : h, h' \in \mathcal{H}\}$.*

(a) Let $Q_{mn}(\tau) = \inf_{h \in \mathcal{H}-\mathcal{H}} \mathbb{P}(|h(X_i^1)| \geq \tau \|h\|_{L_2(D)})$, where X_i^1 represent the random sample at i-th participating client. There is a $\tau \geq 0$ for which $Q_{mn}(\tau) > 0$.

(b) Let $Q_m(\tau) = \inf_{h \in \mathcal{H}-\mathcal{H}} \mathbb{P}\left(|\mathbb{E}[h(X_i^1)]| \geq \tau \|h\|_{L_2(P)}\right)$, where X_i^1 represent the random sample at i-th participating client. There is a $\tau \geq 0$ for which $Q_m(\tau, P) > 0$.

Remark

- 假设 (a) 仅需要 $\mathcal{H} \subset L_2(D)$ 以高概率成立。（后面我会解释 $\mathcal{H} - \mathcal{H}$ 的原因）
- 该条件并不严苛，因为 D 是从 P 中独立同分布抽样得到的。
- 小球条件首次用于异质性数据产生过程。

4.1 小球条件下参与客户的学习率

回忆：$\hat{h}^* = \arg\min_{h\in\mathcal{H}} \mathcal{L}_D(h)$。本节聚焦于度量 $\|h - \hat{h}^*\|^2_{L_2(D)}$，即 h 与 $\hat{h}^*$ 在半经验分布下的 L_2 距离（对应平方损失函数）。

4.1 小球条件下参与客户的学习率

回忆：$\hat{h}^* = \arg\min_{h\in\mathcal{H}} \mathcal{L}_D(h)$。本节聚焦于度量 $\|h-\hat{h}^*\|^2_{L_2(D)}$，即 h 与 $\hat{h}^*$ 在半经验分布下的 L_2 距离（对应平方损失函数）。对于 $\forall h \in \mathcal{H}$，有:

$$\begin{aligned}
\mathcal{L}_S(h) - \mathcal{L}_S(\widehat{h}^*) &= \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\left[(h(X_i^j) - Y_i^j)^2 - (\widehat{h}^*(X_i^j) - Y_i^j)^2\right] && (1)\\
&= \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}(h-\widehat{h}^*)^2(X_i^j) + \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\xi_i^j(h-\widehat{h}^*)(X_i^j), && (2)
\end{aligned}$$

其中 $\xi_i^j = \hat{h}^*(X_i^j) - Y_i^j$。

4.1 小球条件下参与客户的学习率

回忆：$\hat{h}^* = \arg\min_{h\in\mathcal{H}} \mathcal{L}_D(h)$。本节聚焦于度量 $\|h-\hat{h}^*\|^2_{L_2(D)}$，即 h 与 $\hat{h}^*$ 在半经验分布下的 L_2 距离（对应平方损失函数）。
对于 $\forall h \in \mathcal{H}$，有:

$$\mathcal{L}_S(h) - \mathcal{L}_S(\widehat{h}^*) = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\left[(h(X_i^j) - Y_i^j)^2 - (\widehat{h}^*(X_i^j) - Y_i^j)^2\right] \tag{1}$$

$$= \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}(h-\widehat{h}^*)^2(X_i^j) + \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\xi_i^j(h-\widehat{h}^*)(X_i^j), \tag{2}$$

其中 $\xi_i^j = \hat{h}^*(X_i^j) - Y_i^j$。由于 $\hat{h} = \arg\min_{h\in\mathcal{H}} \mathcal{L}_S(h)$，所以 $\mathcal{L}_S(\hat{h}) - \mathcal{L}_S(\hat{h}^*) \le 0$。如果 $\|h-\hat{h}^*\|^2_{L_2(D)}$ 很大，那么上式也有很大的概率大于 0。因为 $\mathcal{L}_S(\hat{h}) - \mathcal{L}_S(\hat{h}^*) \le 0$，所以大概率上 $\|\hat{h}-\hat{h}^*\|^2_{L_2(D)}$ 很小。
前面的假设之所以用 $\mathcal{H}-\mathcal{H}$ 在这里就能看得很清楚了。（是一种泛函）

4.1 小球条件下参与客户的学习率

回忆：$\hat{h}^* = \arg\min_{h\in\mathcal{H}} \mathcal{L}_D(h)$。本节聚焦于度量 $\|h-\hat{h}^*\|^2_{L_2(D)}$，即 h 与 $\hat{h}^*$ 在半经验分布下的 L_2 距离（对应平方损失函数）。
对于 $\forall h \in \mathcal{H}$，有:

$$\mathcal{L}_S(h) - \mathcal{L}_S(\widehat{h}^*) = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\left[(h(X_i^j)-Y_i^j)^2 - (\widehat{h}^*(X_i^j)-Y_i^j)^2\right] \tag{1}$$

$$= \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}(h-\widehat{h}^*)^2(X_i^j) + \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\xi_i^j(h-\widehat{h}^*)(X_i^j), \tag{2}$$

其中 $\xi_i^j = \hat{h}^*(X_i^j) - Y_i^j$。由于 $\hat{h} = \arg\min_{h\in\mathcal{H}} \mathcal{L}_S(h)$，所以 $\mathcal{L}_S(\hat{h}) - \mathcal{L}_S(\hat{h}^*) \le 0$。如果 $\|h-\hat{h}^*\|^2_{L_2(D)}$ 很大，那么上式也有很大的概率大于 0。因为 $\mathcal{L}_S(\hat{h}) - \mathcal{L}_S(\hat{h}^*) \le 0$，所以大概率上 $\|\hat{h}-\hat{h}^*\|^2_{L_2(D)}$ 很小。
前面的假设之所以用 $\mathcal{H}-\mathcal{H}$ 在这里就能看得很清楚了。（是一种泛函）
为了度量第一项，我们引入**拉德马赫复杂度**。

4.1 小球条件下参与客户的学习率

定义 (拉德马赫复杂度)

Definition 5. *We define* $\mathcal{H} - \mathcal{H} = \{h - h' : h, h' \in \mathcal{H}\}$ *and denote by* B_2^m *the* $L_2(D)$ *unit ball entered at* $\widehat{h}^*$*, that is* $B_2^m = \{h \in \mathcal{H} : \|h - \widehat{h}^*\|_{L_2(D)} \leq 1\}$*. For every* $\eta > 0$*, define*

$$\omega_{mn}(\eta) := \inf\left\{s > 0 : \mathbb{E}\left[\sup_{h \in (\mathcal{H}-\mathcal{H}) \cap sB_2^m} \left|\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\sigma_i^j h(X_i^j)\right|\right] \leq \eta s\right\},$$

where σ_i^j *are Rademacher random variables.*

4.1 小球条件下参与客户的学习率

定义 (拉德马赫复杂度)

Definition 5. *We define $\mathcal{H}-\mathcal{H}=\{h-h' : h,h'\in\mathcal{H}\}$ and denote by B_2^m the $L_2(D)$ unit ball entered at $\widehat{h}^*$, that is $B_2^m=\{h\in\mathcal{H} : \|h-\widehat{h}^*\|_{L_2(D)}\le 1\}$. For every $\eta>0$, define*

$$\omega_{mn}(\eta) := \inf\left\{s>0 : \mathbb{E}\left[\sup_{h\in(\mathcal{H}-\mathcal{H})\cap sB_2^m}\left|\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\sigma_i^j h(X_i^j)\right|\right]\le \eta s\right\},$$

where σ_i^j are Rademacher random variables.

标量 $\omega_{mn}(\eta)$ 度量了用户本地函数集合 $\{\mathcal{H}-\mathcal{H}\bigcap sB_2^m\}$。注意，$\omega_{mn}(\eta)$ 仅仅跟假设集 $\mathcal{H}$ 和半经验分布 D 抽出的样本有关。

4.1 小球条件下参与客户的学习率

Theorem 4. *Fix* $\tau > 0$ *for which* $Q_m(2\tau) > 0$ *and set* $\eta < \tau^2 Q_{mn}(2\tau)/32$. *If every random variable* $V_i^j = \xi_i^j h(X_i^j) - \mathbb{E}[\xi_i^j h(X_i^j)]$ *for all* $h \in \mathcal{H} - \widehat{h}^*$ *is Sub-Weibull. For sufficiently large* mn*, with probability at least* $1 - \delta_{mn} - \exp\left(-mnQ_{mn}^2(2\tau)/2\right)$ *one has*

$$\|\widehat{h} - \widehat{h}^*\|_{L_2(D)} \leq 2\max\left\{\omega_{mn}(\tau Q_{mn}(2\tau)/16), (mn)^{-\frac{1}{4}+\iota}\right\},$$

where $0 < \iota < \frac{1}{4}$ *and* $\delta_{mn} = \exp\{-(\frac{c_1\eta^2(mn)^{4\iota}}{\frac{1}{mn}\sum_{i=1}^m\sum_{j=1}^n \left\|V_i^j\right\|_{\psi_\alpha}^2} \wedge \frac{c_2\eta^\alpha(mn)^{\alpha(1/2+2\iota)}}{\max_{(1,1)\leq(i,j)\leq(m,n)}\left\|V_i^j\right\|_{\psi_\alpha}^\alpha})\}$.

4.1 小球条件下参与客户的学习率

Theorem 4. *Fix $\tau > 0$ for which $Q_m(2\tau) > 0$ and set $\eta < \tau^2 Q_{mn}(2\tau)/32$. If every random variable $V_i^j = \xi_i^j h(X_i^j) - \mathbb{E}[\xi_i^j h(X_i^j)]$ for all $h \in \mathcal{H} - \widehat{h}^*$ is Sub-Weibull. For sufficiently large mn, with probability at least $1 - \delta_{mn} - \exp\left(-mnQ_{mn}^2(2\tau)/2\right)$ one has*

$$\|\widehat{h} - \widehat{h}^*\|_{L_2(D)} \leq 2\max\left\{\omega_{mn}(\tau Q_{mn}(2\tau)/16), (mn)^{-\frac{1}{4}+\iota}\right\},$$

where $0 < \iota < \frac{1}{4}$ *and* $\delta_{mn} = \exp\{-(\frac{c_1\eta^2(mn)^{4\iota}}{\frac{1}{mn}\sum_{i=1}^m\sum_{j=1}^n \|V_i^j\|_{\psi_\alpha}^2} \wedge \frac{c_2\eta^\alpha(mn)^{\alpha(1/2+2\iota)}}{\max_{(1,1)\leq(i,j)\leq(m,n)}\|V_i^j\|_{\psi_\alpha}^\alpha})\}$.

Remark

- 异质性联邦学习的、厚尾分布的首个泛化误差结果。

4.1 小球条件下参与客户的学习率

Theorem 4. *Fix $\tau > 0$ for which $Q_m(2\tau) > 0$ and set $\eta < \tau^2 Q_{mn}(2\tau)/32$. If every random variable $V_i^j = \xi_i^j h(X_i^j) - \mathbb{E}[\xi_i^j h(X_i^j)]$ for all $h \in \mathcal{H} - \widehat{h}^*$ is Sub-Weibull. For sufficiently large mn, with probability at least $1 - \delta_{mn} - \exp\left(-mnQ_{mn}^2(2\tau)/2\right)$ one has*

$$\|\widehat{h} - \widehat{h}^*\|_{L_2(D)} \leq 2 \max\left\{\omega_{mn}(\tau Q_{mn}(2\tau)/16), (mn)^{-\frac{1}{4}+\iota}\right\},$$

where $0 < \iota < \frac{1}{4}$ *and* $\delta_{mn} = \exp\{-(\frac{c_1 \eta^2 (mn)^{4\iota}}{\frac{1}{mn}\sum_{i=1}^m \sum_{j=1}^n \|V_i^j\|_{\psi_\alpha}^2} \wedge \frac{c_2 \eta^\alpha (mn)^{\alpha(1/2+2\iota)}}{\max_{(1,1)\leq(i,j)\leq(m,n)} \|V_i^j\|_{\psi_\alpha}^\alpha})\}$.

Remark

- 异质性联邦学习的、厚尾分布的首个泛化误差结果。
- 假设空间的“大小”（复杂度）和噪声的“大小”（厚尾分布）在异质性联邦学习的泛化误差中十分重要。

4.1 小球条件下参与客户的学习率

Theorem 4. *Fix $\tau > 0$ for which $Q_m(2\tau) > 0$ and set $\eta < \tau^2 Q_{mn}(2\tau)/32$. If every random variable $V_i^j = \xi_i^j h(X_i^j) - \mathbb{E}[\xi_i^j h(X_i^j)]$ for all $h \in \mathcal{H} - \widehat{h}^*$ is Sub-Weibull. For sufficiently large mn, with probability at least $1 - \delta_{mn} - \exp\left(-mnQ_{mn}^2(2\tau)/2\right)$ one has*

$$\|\widehat{h} - \widehat{h}^*\|_{L_2(D)} \leq 2\max\left\{\omega_{mn}(\tau Q_{mn}(2\tau)/16), (mn)^{-\frac{1}{4}+\iota}\right\},$$

where $0 < \iota < \frac{1}{4}$ and $\delta_{mn} = \exp\{-(\frac{c_1\eta^2(mn)^{4\iota}}{\frac{1}{mn}\sum_{i=1}^m\sum_{j=1}^n\|V_i^j\|_{\psi_\alpha}^2} \wedge \frac{c_2\eta^\alpha(mn)^{\alpha(1/2+2\iota)}}{\max_{(1,1)\leq(i,j)\leq(m,n)}\|V_i^j\|_{\psi_\alpha}^\alpha})\}$.

Remark

- 异质性联邦学习的、厚尾分布的首个泛化误差结果。
- 假设空间的“大小”（复杂度）和噪声的“大小”（厚尾分布）在异质性联邦学习的泛化误差中十分重要。
- V_i^j 的尾部越厚，δ_{mn} 越大。

4.1 小球条件下参与客户的学习率

Theorem 4. *Fix* $\tau > 0$ *for which* $Q_m(2\tau) > 0$ *and set* $\eta < \tau^2 Q_{mn}(2\tau)/32$. *If every random variable* $V_i^j = \xi_i^j h(X_i^j) - \mathbb{E}[\xi_i^j h(X_i^j)]$ *for all* $h \in \mathcal{H} - \widehat{h}^*$ *is Sub-Weibull. For sufficiently large* mn*, with probability at least* $1 - \delta_{mn} - \exp\left(-mnQ_{mn}^2(2\tau)/2\right)$ *one has*

$$\|\widehat{h} - \widehat{h}^*\|_{L_2(D)} \leq 2\max\left\{\omega_{mn}(\tau Q_{mn}(2\tau)/16), (mn)^{-\frac{1}{4}+\iota}\right\},$$

where $0 < \iota < \frac{1}{4}$ *and* $\delta_{mn} = \exp\{-(\frac{c_1\eta^2(mn)^{4\iota}}{\frac{1}{mn}\sum_{i=1}^m\sum_{j=1}^n \left\|V_i^j\right\|_{\psi_\alpha}^2} \wedge \frac{c_2\eta^\alpha(mn)^{\alpha(1/2+2\iota)}}{\max_{(1,1)\leq(i,j)\leq(m,n)}\left\|V_i^j\right\|_{\psi_\alpha}^\alpha})\}$.

4.1 小球条件下参与客户的学习率

Theorem 4. *Fix* $\tau > 0$ *for which* $Q_m(2\tau) > 0$ *and set* $\eta < \tau^2 Q_{mn}(2\tau)/32$. *If every random variable* $V_i^j = \xi_i^j h(X_i^j) - \mathbb{E}[\xi_i^j h(X_i^j)]$ *for all* $h \in \mathcal{H} - \widehat{h}^*$ *is Sub-Weibull. For sufficiently large* mn*, with probability at least* $1 - \delta_{mn} - \exp\left(-mnQ_{mn}^2(2\tau)/2\right)$ *one has*

$$\|\widehat{h} - \widehat{h}^*\|_{L_2(D)} \leq 2\max\left\{\omega_{mn}(\tau Q_{mn}(2\tau)/16), (mn)^{-\frac{1}{4}+\iota}\right\},$$

where $0 < \iota < \frac{1}{4}$ *and* $\delta_{mn} = \exp\{-(\frac{c_1\eta^2(mn)^{4\iota}}{\frac{1}{mn}\sum_{i=1}^m\sum_{j=1}^n \|V_i^j\|_{\psi_\alpha}^2} \wedge \frac{c_2\eta^\alpha(mn)^{\alpha(1/2+2\iota)}}{\max_{(1,1)\leq(i,j)\leq(m,n)}\|V_i^j\|_{\psi_\alpha}^\alpha})\}$.

Remark

- 当然，δ_{mn} 随 mn 和 ι（与厚尾正相关）的增大而减小。

4.1 小球条件下参与客户的学习率

Theorem 4. *Fix $\tau > 0$ for which $Q_m(2\tau) > 0$ and set $\eta < \tau^2 Q_{mn}(2\tau)/32$. If every random variable $V_i^j = \xi_i^j h(X_i^j) - \mathbb{E}[\xi_i^j h(X_i^j)]$ for all $h \in \mathcal{H} - \widehat{h}^*$ is Sub-Weibull. For sufficiently large mn, with probability at least $1 - \delta_{mn} - \exp\left(-mnQ_{mn}^2(2\tau)/2\right)$ one has*

$$\|\widehat{h} - \widehat{h}^*\|_{L_2(D)} \le 2\max\left\{\omega_{mn}(\tau Q_{mn}(2\tau)/16), (mn)^{-\frac{1}{4}+\iota}\right\},$$

where $0 < \iota < \frac{1}{4}$ and $\delta_{mn} = \exp\{-(\frac{c_1\eta^2(mn)^{4\iota}}{\frac{1}{mn}\sum_{i=1}^m\sum_{j=1}^n \|V_i^j\|_{\psi_\alpha}^2} \wedge \frac{c_2\eta^\alpha(mn)^{\alpha(1/2+2\iota)}}{\max_{(1,1)\le(i,j)\le(m,n)}\|V_i^j\|_{\psi_\alpha}^\alpha})\}$.

Remark

- 当然，δ_{mn} 随 mn 和 ι（与厚尾正相关）的增大而减小。
- 为了让不等式成立的概率更高，δ_{mn} 要小。

4.1 小球条件下参与客户的学习率

Theorem 4. *Fix $\tau > 0$ for which $Q_m(2\tau) > 0$ and set $\eta < \tau^2 Q_{mn}(2\tau)/32$. If every random variable $V_i^j = \xi_i^j h(X_i^j) - \mathbb{E}[\xi_i^j h(X_i^j)]$ for all $h \in \mathcal{H} - \widehat{h}^*$ is Sub-Weibull. For sufficiently large mn, with probability at least $1 - \delta_{mn} - \exp\left(-mnQ_{mn}^2(2\tau)/2\right)$ one has*

$$\|\widehat{h} - \widehat{h}^*\|_{L_2(D)} \leq 2\max\left\{\omega_{mn}(\tau Q_{mn}(2\tau)/16), (mn)^{-\frac{1}{4}+\iota}\right\},$$

where $0 < \iota < \frac{1}{4}$ and $\delta_{mn} = \exp\{-(\frac{c_1\eta^2(mn)^{4\iota}}{\frac{1}{mn}\sum_{i=1}^m\sum_{j=1}^n\|V_i^j\|_{\psi_\alpha}^2} \wedge \frac{c_2\eta^\alpha(mn)^{\alpha(1/2+2\iota)}}{\max_{(1,1)\leq(i,j)\leq(m,n)}\|V_i^j\|_{\psi_\alpha}^\alpha})\}$.

Remark

- 当然，δ_{mn} 随 mn 和 ι（与厚尾正相关）的增大而减小。
- 为了让不等式成立的概率更高，δ_{mn} 要小。
- 即，固定 mn，V_i^j 的尾部越厚，δ_{mn} 越大，$\|h - \hat{h}^*\|_{L_2(D)}^2$ 收敛越慢。

4.1 小球条件下参与客户的学习率

Corollary 1. *Under the same conditions of Theorem* $\boxed{4}$*, for convex function class* $\mathcal{H}$ *and sufficiently large* mn*, with probability at least* $1-\delta_{mn}-\exp\left(-mnQ_{mn}^2(2\tau)/2\right)$ *one has*

$$\mathcal{L}_D(\widehat{h})-\mathcal{L}_D(\widehat{h}^*)\leq(2+\frac{\tau^2}{4}Q_{mn}(2\tau))\max\left(\omega_{mn}^2(\tau Q_{mn}(2\tau)/16),(mn)^{-\frac{1}{2}+\iota}\right),$$

where $0<\iota<\frac{1}{2}$.

Remark

推论 1 给出了半经验风险在次韦伯分布下的界。收敛速率是 $O(\frac{1}{mn^{\frac{1}{2}-\iota}})$，相比于定理 2 的 $O(\frac{1}{\sqrt{mn}})$，收敛速率更慢。

4.1 小球条件下参与客户的学习率

Corollary 1. *Under the same conditions of Theorem $\boxed{4}$, for convex function class $\mathcal{H}$ and sufficiently large mn, with probability at least $1-\delta_{mn}-\exp\left(-mnQ_{mn}^2(2\tau)/2\right)$ one has*

$$\mathcal{L}_D(\widehat{h})-\mathcal{L}_D(\widehat{h}^*)\leq(2+\frac{\tau^2}{4}Q_{mn}(2\tau))\max\left(\omega_{mn}^2(\tau Q_{mn}(2\tau)/16),(mn)^{-\frac{1}{2}+\iota}\right),$$

where $0<\iota<\frac{1}{2}$.

Remark

推论 1 给出了半经验风险在次韦伯分布下的界。收敛速率是 $O(\frac{1}{mn^{\frac{1}{2}-\iota}})$，相比于定理 2 的 $O(\frac{1}{\sqrt{mn}})$，收敛速率更慢。

回顾定理 2：

Theorem 2 (Semi-excess risk for participating clients). *Let $\mathcal{F}$ be a family of functions bounded by b. Under assumptions $\boxed{1}$, $\boxed{3}$ and (a) of Assumption $\boxed{2}$, when $mn\geq cd\log^p(mn)$, it follows that with probability at least $1-\delta$,*

$$\mathcal{L}_D(\widehat{h})-\mathcal{L}_D(\widehat{h}^*)\leq c_1\left(\frac{\log^p(mn)}{mn}\right)^{\frac{1}{2-\beta'}}+c_2\left(\frac{\log(1/\delta)}{mn}\right)^{\frac{1}{2-\beta'}},$$

where c_1 and c_2 are constants depending on γ,p,L,β' and B_1,b,β' respectively.

4.2 小球条件下未参与客户的学习率

为了分析未参与客户的学习率，我们聚焦于 $\|h - h^*\|^2_{L_2(P)}$。这里集合从半经验分布 D 到元分布 P，所以拉德马赫复杂度的定义也要改变。

Definition 6. *We define $\mathcal{H} - \mathcal{H} = \{h - h' : h, h' \in \mathcal{H}\}$ and denote by B_2 the $L_2(P)$ unit ball entered at h^*. For every $\eta > 0$, define*

$$\omega_m(\eta) := \inf \left\{ s > 0 : \mathbb{E} \left[\sup_{h \in (\mathcal{H}-\mathcal{H}) \cap s B_2} \left| \frac{1}{m} \sum_{i=1}^{m} \sigma_i h(X_i) \right| \right] \leq \eta s, \right\}$$

where σ_i are Rademacher random variables.

4.2 小球条件下未参与客户的学习率

为了分析未参与客户的学习率，我们聚焦于 $\|h-h^*\|^2_{L_2(P)}$。这里集合从半经验分布 D 到元分布 P，所以拉德马赫复杂度的定义也要改变。

Definition 6. *We define $\mathcal{H}-\mathcal{H}=\{h-h' : h,h'\in\mathcal{H}\}$ and denote by B_2 the $L_2(P)$ unit ball entered at h^*. For every $\eta>0$, define*

$$\omega_m(\eta) := \inf\left\{ s>0 : \mathbb{E}\left[\sup_{h\in(\mathcal{H}-\mathcal{H})\cap sB_2}\left|\frac{1}{m}\sum_{i=1}^{m}\sigma_i h(X_i)\right|\right] \le \eta s, \right\}$$

where σ_i are Rademacher random variables.

标量 $\omega_m(\eta)$ 度量了 $\{\mathcal{H}-\mathcal{H}\bigcap sB_2\}$ 的复杂度。

4.2 小球条件下未参与客户的学习率

Theorem 5. *Fix $\tau > 0$ for which $Q_m(2\tau) > 0$ and set $\eta < \tau^2 Q_m(2\tau)/32$. If for all $h \in \mathcal{H} - h^*$ the random variable $V_i = \mathbb{E}[\xi_i^1 h(X_i^1)] - \mathbb{E}[\xi_i h(X_i)]$ is Sub-Weibull. For sufficiently large m, with probability at least $1 - \delta_m - \exp\left(-mQ_m^2(2\tau)/2\right)$ one has*

$$\|\widehat{h}^* - h^*\|_{L_2(P)} \leq 2\max\left\{\omega_m(\tau Q_m(2\tau)/16), m^{-\frac{1}{4}+\iota}\right\},$$

where $0 < \iota < \frac{1}{4}$ and $\delta_m = \exp\{-(\frac{c_1\eta^2 m^{4\iota}}{\frac{1}{m}\sum_{i=1}^m \|V_i\|_{\psi_\alpha}^2} \wedge \frac{c_2\eta^\alpha m^{\alpha(1/2+2\iota)}}{\max_{1\leq i\leq m}\|V_i\|_{\psi_\alpha}^\alpha})\}$.

Remark 7. *Theorem 5 provides the first result on the generalization error of unparticipating clients in heterogeneous federated learning with heavy-tailed losses.*

4.2 小球条件下未参与客户的学习率

Theorem 5. *Fix $\tau > 0$ for which $Q_m(2\tau) > 0$ and set $\eta < \tau^2 Q_m(2\tau)/32$. If for all $h \in \mathcal{H} - h^*$ the random variable $V_i = \mathbb{E}[\xi_i^1 h(X_i^1)] - \mathbb{E}[\xi_i h(X_i)]$ is Sub-Weibull. For sufficiently large m, with probability at least $1 - \delta_m - \exp\left(-mQ_m^2(2\tau)/2\right)$ one has*

$$\|\widehat{h}^* - h^*\|_{L_2(P)} \leq 2\max\left\{\omega_m(\tau Q_m(2\tau)/16), m^{-\frac{1}{4}+\iota}\right\},$$

where $0 < \iota < \frac{1}{4}$ and $\delta_m = \exp\{-(\frac{c_1\eta^2 m^{4\iota}}{\frac{1}{m}\sum_{i=1}^m \|V_i\|_{\psi_\alpha}^2} \wedge \frac{c_2\eta^\alpha m^{\alpha(1/2+2\iota)}}{\max_{1\leq i\leq m}\|V_i\|_{\psi_\alpha}^\alpha})\}$.

Remark 7. *Theorem 5 provides the first result on the generalization error of unparticipating clients in heterogeneous federated learning with heavy-tailed losses.*

定理 5 首次给出了异质性联邦学习的、厚尾分布损失函数的、未参与客户的泛化误差界。

4.2 小球条件下未参与客户的学习率

Corollary 2. *Assume for all* $h \in \mathcal{H} - \widehat{h}^*$ *the random variable* $V_i' = \mathbb{E}[h^2(X_i^j)] - \mathbb{E}[h^2(X_i)]$ *is Sub-Weibull and the noise* $h^*(X_i) - Y_i$ *is independent of* X_i*. Under the same conditions of Theorem 5, for* $0 < \eta < 1$ *and sufficiently large* mn*, with probability at least* $1 - \delta' - \exp\left(-mnQ_{mn}^2(2\tau)/2\right) - \exp\left(-mQ_m^2(2\tau)/2\right)$ *one has*

$$\mathcal{L}_P(\widehat{h}) - \mathcal{L}_P(h^*) \leq c_0 \max\left(\omega_{mn}^2\left(\frac{\tau Q_{mn}(2\tau)}{16}\right), (mn)^{-\frac{1}{2}+\iota}\right) + 2\max\left\{\omega_m^2\left(\frac{\tau Q_m(2\tau)}{16}\right), m^{-\frac{1}{2}+\iota}\right\}$$

where $c_0 = \frac{2}{1-\eta}$, $0 < \iota < \frac{1}{2}$ *and* $\delta' = \delta_{mn} + \delta_m + \exp\{-(\frac{c_1\eta^2 m^{4\iota}}{\frac{1}{m}\sum_{i=1}^m \|V_i'\|_{\psi_\alpha}^2} \wedge \frac{c_2\eta^\alpha m^{\alpha(1/2+2\iota)}}{\max_{1\leq i\leq m}\|V_i'\|_{\psi_\alpha}^\alpha})\}$.

4.2 小球条件下未参与客户的学习率

Corollary 2. *Assume for all $h \in \mathcal{H} - \widehat{h}^*$ the random variable $V_i' = \mathbb{E}[h^2(X_i^j)] - \mathbb{E}[h^2(X_i)]$ is Sub-Weibull and the noise $h^*(X_i) - Y_i$ is independent of X_i. Under the same conditions of Theorem 5, for $0 < \eta < 1$ and sufficiently large mn, with probability at least $1 - \delta' - \exp\left(-mnQ_{mn}^2(2\tau)/2\right) - \exp\left(-mQ_m^2(2\tau)/2\right)$ one has*

$$\mathcal{L}_P(\widehat{h}) - \mathcal{L}_P(h^*) \le c_0 \max\left(\omega_{mn}^2\left(\frac{\tau Q_{mn}(2\tau)}{16}\right), (mn)^{-\frac{1}{2}+\iota}\right) + 2\max\left\{\omega_m^2\left(\frac{\tau Q_m(2\tau)}{16}\right), m^{-\frac{1}{2}+\iota}\right\}$$

where $c_0 = \frac{2}{1-\eta}$, $0 < \iota < \frac{1}{2}$ *and* $\delta' = \delta_{mn} + \delta_m + \exp\{-(\frac{c_1\eta^2 m^{4\iota}}{\frac{1}{m}\sum_{i=1}^m \|V_i'\|_{\psi_\alpha}^2} \wedge \frac{c_2\eta^\alpha m^{\alpha(1/2+2\iota)}}{\max_{1\le i\le m}\|V_i'\|_{\psi_\alpha}^\alpha})\}$.

Remark

- 推论 2 给出了总体风险在次韦伯分布下的界，$O(\frac{1}{mn^{\frac{1}{2}-\eta}} + \frac{1}{m^{\frac{1}{2}-\eta}})$。

4.2 小球条件下未参与客户的学习率

Corollary 2. *Assume for all $h \in \mathcal{H} - \widehat{h}^*$ the random variable $V_i' = \mathbb{E}[h^2(X_i^j)] - \mathbb{E}[h^2(X_i)]$ is Sub-Weibull and the noise $h^*(X_i) - Y_i$ is independent of X_i. Under the same conditions of Theorem 5, for $0 < \eta < 1$ and sufficiently large mn, with probability at least $1 - \delta' - \exp\left(-mnQ_{mn}^2(2\tau)/2\right) - \exp\left(-mQ_m^2(2\tau)/2\right)$ one has*

$$\mathcal{L}_P(\widehat{h}) - \mathcal{L}_P(h^*) \leq c_0 \max\left(\omega_{mn}^2\left(\frac{\tau Q_{mn}(2\tau)}{16}\right), (mn)^{-\frac{1}{2}+\iota}\right) + 2\max\left\{\omega_m^2\left(\frac{\tau Q_m(2\tau)}{16}\right), m^{-\frac{1}{2}+\iota}\right\}$$

where $c_0 = \frac{2}{1-\eta}$, $0 < \iota < \frac{1}{2}$ *and* $\delta' = \delta_{mn} + \delta_m + \exp\{-(\frac{c_1\eta^2 m^{4\iota}}{\frac{1}{m}\sum_{i=1}^m \|V_i'\|_{\psi_\alpha}^2} \wedge \frac{c_2\eta^\alpha m^{\alpha(1/2+2\iota)}}{\max_{1\leq i\leq m}\|V_i'\|_{\psi_\alpha}^\alpha})\}$.

Remark

- 推论 2 给出了总体风险在次韦伯分布下的界，$O(\frac{1}{mn^{\frac{1}{2}-\eta}} + \frac{1}{m^{\frac{1}{2}-\eta}})$。
- 相比于定理 3 的 $O(\frac{1}{mn} + \frac{1}{m})$，收敛速率更慢了。

4.2 小球条件下未参与客户的学习率

Corollary 2. *Assume for all $h \in \mathcal{H} - \widehat{h}^*$ the random variable $V_i' = \mathbb{E}[h^2(X_i^j)] - \mathbb{E}[h^2(X_i)]$ is Sub-Weibull and the noise $h^*(X_i) - Y_i$ is independent of X_i. Under the same conditions of Theorem 5, for $0 < \eta < 1$ and sufficiently large mn, with probability at least $1 - \delta' - \exp\left(-mnQ_{mn}^2(2\tau)/2\right) - \exp\left(-mQ_m^2(2\tau)/2\right)$ one has*

$$\mathcal{L}_P(\widehat{h}) - \mathcal{L}_P(h^*) \leq c_0 \max\left(\omega_{mn}^2\left(\frac{\tau Q_{mn}(2\tau)}{16}\right), (mn)^{-\frac{1}{2}+\iota}\right) + 2\max\left\{\omega_m^2\left(\frac{\tau Q_m(2\tau)}{16}\right), m^{-\frac{1}{2}+\iota}\right\}$$

where $c_0 = \frac{2}{1-\eta}$, $0 < \iota < \frac{1}{2}$ *and* $\delta' = \delta_{mn} + \delta_m + \exp\{-(\frac{c_1\eta^2 m^{4\iota}}{\frac{1}{m}\sum_{i=1}^m \|V_i'\|_{\psi_\alpha}^2} \wedge \frac{c_2\eta^\alpha m^{\alpha(1/2+2\iota)}}{\max_{1\leq i\leq m}\|V_i'\|_{\psi_\alpha}^\alpha})\}$.

Remark

- 推论 2 给出了总体风险在次韦伯分布下的界，$O(\frac{1}{mn^{\frac{1}{2}-\eta}} + \frac{1}{m^{\frac{1}{2}-\eta}})$。
- 相比于定理 3 的 $O(\frac{1}{mn} + \frac{1}{m})$，收敛速率更慢了。
- 概率取决于 $\eta, m, n (\eta \in (0, 1/2))$。

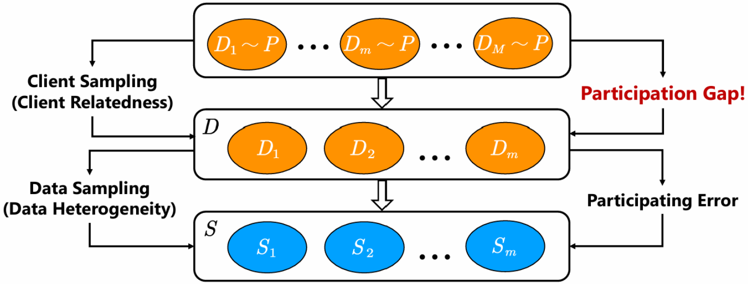
5 相关工作

Table 1: **Generalization Bounds for Heterogeneous Federated Learning**. SC, Pro, and Exp denote Strong convexity, In probability, and In expectation. Sub-expon denotes sub-exponential.

Reference	Loss	Assumption	Part	Unpart	Type
Mohri et al. (2019)	Bounded	Bi-Classifier	$\mathcal{O}(\frac{1}{\sqrt{mn}})$	/	Pro
Chen et al. (2021)	Bounded	Smooth, SC	$\mathcal{O}(\frac{1}{mn})$	/	Exp
Fallah et al. (2021)	Bounded	Smooth, SC	$\mathcal{O}(\frac{1}{mn})$	/	Exp
Our Results	Sub-expon	Lipschitz	$\mathcal{O}(\frac{1}{\sqrt{mn}})$	$\mathcal{O}(\frac{1}{\sqrt{mn}} + \frac{1}{\sqrt{m}})$	Pro
Our Results	Bounded	Bernstein Con	$\mathcal{O}(\frac{1}{mn})$	$\mathcal{O}(\frac{1}{mn} + \frac{1}{m})$	Pro
Our Results	Sub-Weibull	Small-ball	$\mathcal{O}\left((mn)^{\frac{2\iota-1}{2}}\right)$	$\mathcal{O}((mn)^{\frac{2\iota-1}{2}} + m^{\frac{2\iota-1}{2}})$	Pro

6 总结

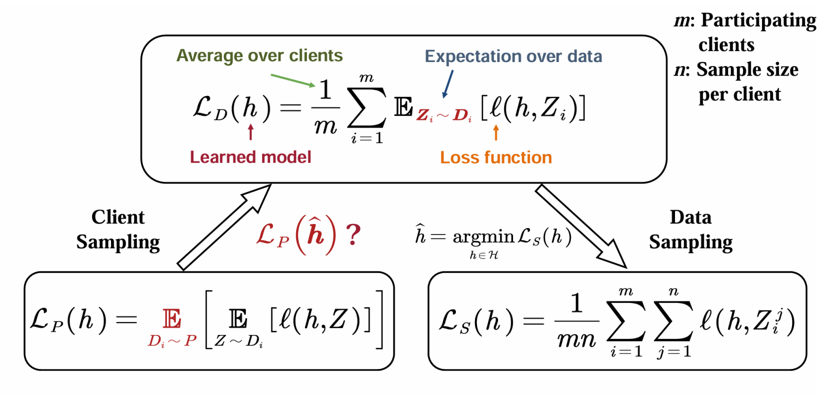
双层分布框架



$D_1 \sim P$
$D_m \sim P$
$D_M \sim P$
Client Sampling
(Client Relatedness)
Participation Gap!
D
D_1
D_2
D_m
Data Sampling
(Data Heterogeneity)
Participating Error
S
S_1
S_2
S_m

6 总结

双层分布框架



Average over clients
Expectation over data
$\mathcal{L}_D(h) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}_{Z_i \sim D_i}[\ell(h, Z_i)]$
Learned model
Loss function
m: Participating clients
n: Sample size per client
Client Sampling
$\mathcal{L}_P(\hat{h})$?
$\hat{h} = \operatorname{argmin}_{h \in \mathcal{H}} \mathcal{L}_S(h)$
Data Sampling
$\mathcal{L}_P(h) = \mathbb{E}_{D_i \sim P}\left[\mathbb{E}_{Z \sim D_i}[\ell(h, Z)]\right]$
$\mathcal{L}_S(h) = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} \ell(h, Z_i^j)$

泛化误差界——更快的速率

- **Learning Rates for unparticipating Client ——Bounded Losses**

$$\hat{h} = \operatorname*{argmin}_{h\in\mathcal{H}} \mathcal{L}_S(h)$$

$$\mathcal{L}_P(\hat{h}) - \min_{h\in\mathcal{H}} \mathcal{L}_P(h) \leq \mathcal{O}\left(\sqrt{\frac{1}{mn}} + \sqrt{\frac{1}{m}}\right)$$

- **Bernstein Condition:** $\mathbf{E}[\ell(h,Z) - \ell(h^\star,Z)]^2 \leq B\,\mathbf{E}[\ell(h,Z) - \ell(h^\star,Z)]$

- **Learning Rates for unparticipating Client——Bernstein Condition**

$$\mathcal{L}_P(\hat{h}) - \min_{h\in\mathcal{H}} \mathcal{L}_P(h) \leq \mathcal{O}\left(\frac{1}{mn} + \frac{1}{m}\right)$$

Unparticipating clients would benefit from the model trained by participating clients!

6 总结

泛化误差界——厚尾分布

- **Heavy-tail Distribution**

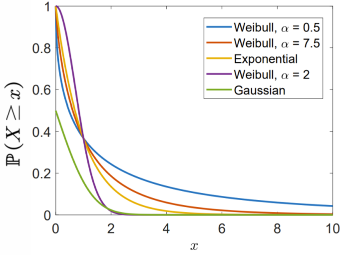


Figure: Illustration of heavy tails[2]

- **Learning Rates for unparticipating Client——Heavy-tail Data**

$$\eta \in (0, 1/2)$$

$$\mathcal{L}_P(\hat{h}) - \min_{h \in \mathcal{H}} \mathcal{L}_P(h) \leq \mathcal{O}\left(\frac{1}{mn^{\frac{1}{2}-\eta}} + \frac{1}{m^{\frac{1}{2}-\eta}} \right)$$

The probability depends on η, m, n

7 附录

- 原文链接：
  `https://openreview.net/pdf?id=-EHqoysUYLx`
- 主要参考了**The Elements of Statistical Learning**以及**Foundations of Machine Learning(Second Edition)**（点击访问）
- 本展示是 LaTeX 制作的 beamer，tex 源码已上传到**github**
- 实验源码同上链接。

谢谢！