arXiv:2302.07869v1 [cs.LG] 15 Feb 2023

通过强自适应在线学习改进在线保形预测

Aadyot Bhatnagar* Huan Wang* Caiming Xiong* Yu Bai*

2023年2月16日

摘要

我们研究在线设置中的预测集不确定性量化问题，其中数据分布可能随时间任意变化。近期研究提出了 *online conformal prediction*技术，利用在线学习领域的遗憾最小化算法，学习具有近似有效覆盖率且遗憾较小的预测集。然而，标准遗憾最小化可能不足以应对变化环境；在这类环境中，人们不仅希望在整个时间范围内获得性能保证，也希望在所有（子）时间区间内获得保证。我们提出了新的在线保形预测方法，以最小化 *strongly adaptive regret*，该指标衡量所有固定长度区间上的最坏情况遗憾。我们证明，这些方法能够同时对所有区间长度实现近乎最优的强自适应遗憾，并达到近似有效覆盖率。实验表明，在时间序列预测和分布偏移下的图像分类等现实任务中，与现有方法相比，我们的方法始终能够获得更好的覆盖率和更小的预测集。

1 引言

现代机器学习模型在许多场景中都能做出高度准确的预测。在高风险决策任务中，通过量化真实标签可能偏离模型预测的程度来估计模型的不确定性，同样至关重要。不确定性量化的一种常见方法是学习预测集，将每个输入与一组候选标签关联起来，例如用于回归的预测区间和用于分类的标签集。学习预测集最重要的要求是实现有效覆盖率，即以至少 $1-\alpha$ （例如 90%）的概率覆盖真实标签。

保形预测（Vovk 等人，2005年）是一个强大的框架，可将任何基础预测器（例如预训练模型）转换为具有有效覆盖保证的预测集（Angelopoulos 和 Bates，2021）。这些保证对数据分布几乎不作任何假设，唯一例外是可交换性（独立同分布数据是其充分条件）。然而，在许多现实场景中，可交换性并不成立，例如时间序列数据（Chernozhukov 等人，2018）或数据损坏（Hendrycks 等人，2018），此时数据可能会出现分布偏移。针对这类分布偏移，人们提出了多种应对方法，例如重新加权（Tibshirani 等人，2019；Barber 等人，2022）或分布鲁棒优化（Cauchois 等人，2022）。

近期一系列研究针对数据按序列顺序到达的情形，开发了在线保形预测方法（Gibbs and Cand`es, 2021, 2022；Zaffran et al., 2022；Feldman et al., 2022）。在每一步中，这些方法都会输出一个由单一半径参数确定的预测集，该参数控制集合的大小。收到真实标签后，它们会通过遗憾最小化技术（例如在线梯度下降（OGD）（Zinkevich, 2003）），针对半径上的某种分位数损失自适应地调整该参数。研究表明，无论数据分布如何，这些方法都能使经验覆盖频率接近 $1-\alpha$（Gibbs and Cand`es, 2021）。除了覆盖率之外，重要的是，这些方法还能针对分位数损失实现次线性遗憾（Gibbs and Cand`es, 2022）。此类遗憾保证确保

*Salesforce AI Research。邮箱：{abhatnagar, huan.wang, cxiong, yu.bai}@salesforce.com。

预测集的大小合理，并排除通过交替预测空集和全集来实现有效覆盖率的“平凡”算法（参见第 2 节的讨论）。

尽管遗憾最小化技术能够实现覆盖率与遗憾保证，但在更动态的环境中，它们可能表现不佳。在这类环境下，我们不仅希望算法在整个时间范围内（如遗憾所体现的那样）表现出色，也希望它在每个子区间内都表现出色。例如，如果数据分布发生了几次突变，我们不仅希望算法在整个预测步长内表现出色，也希望它在每两次连续突变之间的每个连续区间内都表现出色。Gibbs 和 Cand`es (2022年) 部分解决了这一问题，他们提出了完全自适应保形推断（FACI）算法。这是一种元算法，它聚合了多个专家（基学习器），这些基学习器是采用不同学习率的 OGD 实例。然而，他们的算法可能并不适合实现这类基于区间的保证，因为每个专家仍在整个预测步长内运行，并未真正局部化。这也体现在 FACI 虽然能在固定长度为 $\widetilde{\mathcal{O}}(\sqrt{k}$ 的区间内实现近乎最优的 k遗憾，却无法同时对所有长度 $k \in [T]$ 实现这一点。

本文利用强自适应遗憾最小化这一在线学习技术，同时在所有子区间上取得优异表现（Daniely 等人，2015；Jun 等人，2017年），设计了改进的在线保形预测算法。我们提出的算法——强自适应在线保形预测（SAOCP）——是一种新的元算法，用于管理多个专家，关键区别在于现在每位专家仅在自己的活动区间内运行。我们的贡献总结如下：

- 我们提出了 SAOCP，这是一种新的在线保形预测算法。SAOCP 是一种元算法，它维护多个专家，每位专家都有自己的活动区间，并以强自适应遗憾最小化技术为基础（第3节）。我们将专家实例化为无尺度 OGD（SF-OGD），这是 OGD 的任意时刻变体，我们也将其作为一种独立算法进行研究。

- 我们证明，SAOCP 在所有长度为 $\widetilde{\mathcal{O}}$（$\sqrt{k}$）的区间上同时实现近乎最优的强自适应遗憾，并且 SAOCP 和 SF-OGD 都能实现近似有效覆盖率（第4节）。

- 我们通过实验表明，在两个真实世界任务——时间序列预测和分布偏移下的图像分类（第5节）中，与现有方法相比，SAOCP 和 SF-OGD 在局部化窗口中实现了更高的覆盖率，并生成了更小的预测集。[1]

1.1 相关工作

保形预测 保形预测的最初思想（利用可交换数据）见于 Vovk 等（1999，2005）以及 Shafer 和 Vovk（2008）的早期研究。此后，通过保形预测学习预测集已成为回归不确定性量化的一种主要方法（Papadopoulos，2008；Vovk，2012；Lei 和 Wasserman，2014；Vovk 等，2018；Romano 等，2019；Gupta 等，2019；Barber 等人，2021，2022）和分类（Lei 等，2013；Romano 等，2020；Cauchois 等，2020，2022；Angelopoulos 等，2021b）中的一种主要方法；此外，还应用于一般风险控制（Bates 等，2021；Angelopoulos 等，2021a，2022a）、生物成像（Angelopoulos 等，2022b）和蛋白质设计（Fannjiang 等，2022）等领域。

近期研究还提出，除了覆盖率之外，也要优化预测集的 效率（例如宽度或基数）（Pearce 等人，2018；Park 等人，2020年；Yang 和 Kuchibhotla，2021；Stutz 等人，2022年；Angelopoulos 等人，2021a，b；Bai 等人，2022）。我们考虑的遗憾可以视为在线设置下效率的一种（替代）度量。

分布偏移下的保形预测 对于更具挑战性的情形，即数据可能出现分布偏移（因此不再满足可交换性），已有多种方法被提出，以实现近似有效的覆盖率，例如重新加权（利用关于数据依赖结构的先验知识）（Tibshirani 等人，2019年；Podkopaev 和 Ramdas，2021；Cand`es 等人，2021；Barber 等人，2022年）、分布鲁棒优化（Cauchois 等人，2020年）或双重稳健校准（Yang 等人，2022年）。

[1]我们实验的代码可在 https://github.com/salesforce/online_conforma l 找到.

我们的工作为在线保形预测这一研究方向增添了新的成果（Gibbs 和 Cand`es，2021，2022年；Zaffran 等人，2022年；Feldman 等人，2022年），该方向利用在线学习文献中的遗憾最小化技术（Zinkevich，2003；Hazan，2022年），根据近期观测自适应地调整预测集的大小。与我们的工作密切相关的是Gibbs 和 Cand`es的FACI 算法（2022年），它是一种利用多个专家应对变化环境的元算法。我们的元算法 SAOCP 与其在风格上有所不同：我们的专家只在各自的活动区间内运行，并且在强自适应遗憾方面提供了更好的保证。

另一类相关研究探讨了时间序列数据的保形预测。 Chernozhukov 等人 (2018年)；Xu 和 Xie (2021年)；Sousa 等人 (2022年) 通过随机化和集成，为某种意义上遍历的时间序列生成有效的预测集。其他一些研究则将标准保形预测直接应用于时间序列，要么不提供理论保证，要么只要求较弱的可交换性概念（Dashevskiy 和 Luo，2008年；Wisniewski 等人，2020年； Kath 和 Ziel，2021； Stankeviciute 等人，2021； Sun 和 Yu，2022）。

强自适应在线学习 我们的算法借鉴了在线学习领域的技术，特别是强自适应遗憾最小化（Daniely 等人，2015； Jun 等人，2017年； Zhang 等人，2018）以及无尺度算法，以实现其他类型的自适应遗憾保证（例如，随时遗憾保证）（Orabona 和 P´等，2018年）。

2 预备知识

我们考虑标准的学习问题：在这些问题中，我们观测到样本 $(x,\ y) \in \mathcal{X} \times \mathcal{Y}$，并希望根据输入 x预测标签 y。预测集 $C : \mathcal{X} \to 2^{\mathcal{Y}}$ 是一个集合值函数，它将任意输入 x 映射到一个由预测标签 $C(x)$ 构成的集合 $\subset \mathcal{Y}$。两个常见的例子是用于回归的预测区间，其中 $\mathcal{Y} = \mathbb{R}$且 $C(x)$ 是一个区间；以及用于（m类）分类的标签预测集，其中 $\mathcal{Y} = [m]$ 且 $C(x)$ 是 $[m]$ 的子集。预测集是量化黑箱模型点预测 $\widehat{y} = f(x)$ 所伴随不确定性的常用方法。

我们研究在线 设置下学习预测集的问题。在这种设置中，数据 $(X_1,\ Y_1) \ldots, (X_T,\ Y_T)$ 依次到达。在每个时间步 $t \in [T]$，我们根据当前输入 X_t 和过去的观测 $\{(X_i,\ Y_i)\}_{i \le t-1}$输出一个预测集 $C_t = C_t(X_t)$，然后才观测真实标签 Y_t。预测集的主要目标是实现有效覆盖率： $\mathbb{P}[Y_t \in C_t(X_t)] = 1 - \alpha$，其中 $1 - \alpha \in (0,\ 1)$ 是用户预先确定的目标覆盖率。 α 的标准取值包括 $\{0.1, 0.05\}$，分别对应{90%，95%} 的目标覆盖率。

在线保形预测现在我们回顾在线保形预测这一思想，它由Gibbs 和 Cand`es (2021，2022年)。这种在线设置下学习预测集的框架，即使在分布偏移的情况下，也能实现覆盖保证。

在每个时刻 $t \in [T]$，在线保形预测假设我们有一族预测集 $\mathcal{C}_t = \{C_t(x, s)\}_{x \in \mathcal{X}, s \in \mathbb{R}}$，由半径参数 $s \in \mathbb{R}$ 指定；我们需要预测 $\widehat{s}_t \in \mathbb{R}$ 并输出预测集 $C_t = C_t(X_t, \widehat{s}_t) \subset \mathcal{Y}$。集合族 $(\mathcal{C}_t)_{t \in [T]}$ 通常通过基础预测器 f_t 来定义（例如， $f_t \equiv f$ 可以是一个固定的预训练模型）。回归中的一个标准例子是，我们有一个基础预测器 $f_t : \mathcal{X} \to \mathbb{R}$，并且可以选择 $C_t(X_t, s) := [f_t(X_t) - s, f_t(X_t) + s]$ 作为围绕 $f_t(X_t)$ 的预测区间，此时半径 s 是区间宽度的一半。一般来说，我们允许任何满足的集合 $\mathcal{C}_t$，这些集合是嵌套集合（Gupta 等，2019），即对于所有 $x \in \mathcal{X}$ 和 $s \le s'$， $C_t(x, s) \subseteq C_t(x, s')$，因此更大的半径总是会产生更大的集合。

在线保形预测采用在线学习技术，根据过去的观测学习 $\widehat{s}_t$。定义“真实半径”为 $S_t := \inf\{s \in \mathbb{R} : Y_t \in C_t(X_t, s)\}$ (即满足 s 覆盖 C_t 的最小半径 Y_t)，我们考虑 $1 - \alpha$-分位数损失 (又称 弹球损失 (Koenker 和 Bassett Jr，1978))，用于衡量 S_t 与任意预测半径 $\widehat{s}$之间的差异：

$$\ell^{(t)}(\widehat{s}) = \ell_{1-\alpha}(S_t, \widehat{s}) := \max\{(1-\alpha)(S_t - \widehat{s}), \alpha(\widehat{s} - S_t)\}. \tag{1}$$

在本文余下部分中，我们假设所有真实半径都有界：对所有 $t \in [T]$， $S_t \in [0, D]$ 几乎必然成立。

观测到 X_t、预测半径 $\widehat{s}_t$并观测到标签 Y_t (因此 S_t) 后，可以计算梯度[2] $\nabla \ell^{(t)}(\widehat{s}_t)$，其形式如下，非常简洁：

$$\nabla \ell^{(t)}(\widehat{s}_t) = \alpha - \mathbb{1}[\widehat{s}_t < S_t] = \alpha - \underbrace{\mathbb{1}[Y_t \notin \widehat{C}_t]}_{:=\mathrm{err}_t}, \tag{2}$$

其中 err_t 是时刻 t (如果 $t = 1$ 未能覆盖 C_t，则 err_t) 的未覆盖指示变量。 Gibbs 和 Cand`es(2021) 执行一次在线梯度下降 (OGD) 步骤，得到 $\widehat{s}_{t+1}$:

$$\widehat{s}_{t+1} = \widehat{s}_t - \eta \nabla \ell^{(t)}(\widehat{s}_t) = \widehat{s}_t + \eta(\mathrm{err}_t - \alpha), \tag{3}$$

其中 $\eta > 0$ 是学习率，算法从某个 $\widehat{s}_1 \in \mathbb{R}$初始化。更新（3）会在未能覆盖 C_t 时增大预测半径 Y_t ($\mathrm{err}_t = 1$)，否则减小半径。这种方法通过近期观测来调整半径，直观上是合理的。

自适应保形推断 (ACI) Gibbs 和 Cand`es (2021) 提出的 ACI 算法将更新（3）与特定的 $\mathcal{C}_t = \mathcal{C}_t^{\mathsf{ACI}}$选择结合使用，其中

$$\widehat{C}_t^{\mathsf{ACI}}(X_t, \widehat{s}_t) = \left\{ y : \widetilde{S}_t(X_t, y) \le Q_{\widehat{s}_t}\left(\{\widetilde{S}_\tau\}_{\tau=1}^{t-1} \right) \right\}, \tag{4}$$

其中 $\widetilde{S}_t : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ 是任意函数（称为评分函数）， $S_\tau := S_\tau(X_\tau, y_\tau)$ 表示第 τ 个观测的分数，而 $Q_s(\cdot)$ 表示集合的 $s \in (0, 1)$ 经验分位数（参见（32））。换言之，ACI 的置信集包含所有满足其评分 $S_t(X_t, y)$ 低于过往分数的 $\widehat{s}_t$ 分位数的 y，并使用（3）来学习这个分位数水平。本文提出的框架将 ACI 算法推广到允许任意选择 $\mathcal{C}_t$ （即嵌套集合），其中包括但不限于（4）。为便于讨论，除非另有明确说明，本文也将这一广义版本的（3）称为 ACI。

在实证研究中，我们在第 5 节表明，FACI（Gibbs 和 Cand`es，2022年）（ACI 的扩展）在采用我们更一般的参数化直接预测 S_t 时表现 $C(X_t, \widehat{s}_t) = \{y : \widetilde{S}_t(X_t, y) \le \widehat{s}_t\}$。

覆盖率与遗憾保证 Gibbs 和 Cand`es （2021）表明，ACI[3] 能够实现近似有效的（经验）覆盖，即经验未覆盖频率接近目标水平α:

$$\mathrm{CovErr}(T) := \left| \tfrac{1}{T} \textstyle\sum_{t=1}^{T} \mathrm{err}_t - \alpha \right| \le \tfrac{D+\eta}{\eta T}. \tag{5}$$

除覆盖率外，根据标准的在线学习分析，ACI 对分位数损失还具有遗憾 界 $\{\ell^{(t)}\}_{t\in[T]}$: 我们有 $\mathrm{Reg}(T) \le \mathcal{O}(D^2/\eta + \eta T) \le \mathcal{O}(D\sqrt{T})$（其中 η 取最优值），其中

$$\mathrm{Reg}(T) := \textstyle\sum_{t=1}^{T} \ell^{(t)}(\widehat{s}_t) - \inf_{s^\star \in \mathbb{R}} \sum_{t=1}^{T} \ell^{(t)}(s^\star). \tag{6}$$

将遗憾作为覆盖率之外的另一项性能度量，其优势在于它能排除某些以“平凡”的方式实现良好覆盖、但在实践中并无用处的算法——例如，C_t 可能只是在空集和全集之间交替 $\{\alpha, 1-\alpha\}$ 比例的时间；这种做法在任意数据分布上都满足覆盖率界（5），但即使在某些简单的数据分布上也会遭受线性遗憾（参见附录 A.2）。

我们指出，遗憾与覆盖误差之间存在联系，即 $\mathrm{CovErr}(T) = |\frac{1}{T}\sum_{t=1}^{T} \nabla \ell^{(t)}(\widehat{s}_t)|$，也就是说，覆盖误差等于损失的平均梯度（导数）。然而，在没有进一步分布假设的情况下，遗憾界与覆盖率界并不能以黑箱方式相互推出（例如参见 Gibbs 和 Cand`es (2021，附录 A），因此必须针对每种算法分别加以证明。

[2]更准确地说， $\nabla \ell^{(t)}(\widehat{s}_t)$ 是一个次梯度。 [3]他们的结果是基于对 $\mathcal{C}_t$ 在 (4) 中的特定选择得出的，但可以直接推广到任何 $\mathcal{C}_t$ 构成嵌套集合的集合。

算法 1： 强自适应在线保形预测（SAOCP），改编自 Jun 等人(2017年)。

输入： 目标覆盖率 $1-\alpha \in (0,1)$; 最大可能半径 $D > 0$

对于 $t = 1, \ldots, T$，**执行**

// Obtain prediction interval by aggregating active experts

初始化新的专家 $\mathcal{A}_t = \texttt{SF-OGD}(\alpha \leftarrow \alpha;\ \eta \leftarrow D/\sqrt{3};\ \widehat{s}_1 \leftarrow \widehat{s}_{t-1})$（算法 2），并设置权重 $w_{t,t} = 0$

计算活跃集 $\text{Active}(t) = \{i \in [T] :\ t - L(i) < i \le t\}$，其中 $L(i)$ 定义见（8）

计算先验概率 $\pi_i \propto i^{-2}(1 + \lfloor \log_2 i \rfloor)^{-1}\mathbb{1}[i \in \text{Active}(t)]$

为所有 $i \in [t]$计算未归一化概率 $\widehat{p}_i = \pi_i [w_{t,i}]_+$

将 $p = \widehat{p}/\|\widehat{p}\|_1 \in \Delta^t$ 归一化（如果 $\|\widehat{p}\|_1 > 0$，否则 $p = \pi$）

计算预测半径 $\widehat{s}_t = \sum_{i \in \text{激活}\ (t)} p_i \widehat{s}_{i,\,t}$（针对 $t \ge 2$），以及 $\widehat{s}_t = 0$ 针对 $t = 1$

观察输入 $X_t \in \mathcal{X}$ 并 **返回**预测集 $\widehat{C}_t(X_t,\ \widehat{s}_t)$

// Use meta loss and per-expert losses to update experts

观察真实标签 $Y_t \in \mathcal{Y}$，计算真实半径 $S_t = \inf\{s \in \mathbb{R} :\ Y_t \in \widehat{C}_t(X_t,\ s)\}$ 和损失函数 $\ell^{(t)}(\cdot) = \ell_{1-\alpha}(S_t, \cdot)$

对于 $i \in \text{Active}(t)$ **执行**

更新专家 $\mathcal{A}_i$，使用（X_t，Y_t）并获得下一个预测半径 $\widehat{s}_{i,\,t+1}$

计算 $g_{i,\,t} = \begin{cases} -\frac{1}{D}\left(\ell^{(t)}(\widehat{s}_t) - \ell^{(t)}(\widehat{s}_{i,t})\right) & w_{i,\,t} > 0 \\ -\frac{1}{D}\left[\ell^{(t)}(\widehat{s}_t) - \ell^{(t)}(\widehat{s}_{i,t})\right]_+ & w_{i,\,t} \le 0 \end{cases}$

更新专家权重 $w_{i,\,t+1} = \frac{1}{t-i+1}\left(\sum_{j=i}^{t} g_{i,j}\right)\left(1 + \sum_{j=i}^{t} w_{i,j} g_{i,j}\right)$

3 强自适应在线保形预测

我们的方法源于这样一个观察：从某种意义上说，遗憾最小化存在局限，因为遗憾衡量的是整个预测步长 $[T]$上的表现；当算法遇到变化环境时，这可能不足以反映其表现。例如，如果 $S_t = 1$ 在 $1 \le t \le T/2$ 期间，而 $S_t = 100$ 在 $T/2 < t \le T$期间，那么在所有大小为 $T/2$ 的（子）区间 上都实现较小的遗憾，比在 $[T]$ 上实现较小的遗憾是强得多的保证。因此，我们寻求同时适用于所有区间的局部化 保证，以避免最坏情况，例如在特定区间内出现显著的覆盖不足或较大的半径。

强自适应遗憾 (SARegret)（Daniely 等人，2015；Zhang 等人，2018）已在在线学习文献中被提出，作为遗憾的一种推广，用以刻画在线学习算法在所有区间上的整体表现。具体而言，对于任意 $k \in [T]$，任意算法在长度为 k 时的 SARegret 定义为

$$\text{SAReg}(T, k) := \max_{[\tau, \tau+k-1] \subseteq [T]} \left(\sum_{t=\tau}^{\tau+k-1} \ell^{(t)}(\widehat{s}_t) - \inf_{s^\star} \sum_{t=\tau}^{\tau+k-1} \ell^{(t)}(s^\star) \right) \tag{7}$$

SAReg(T，k) 衡量所有长度为 k 的区间上的最大遗憾。在 $k = T$ 时，它退化为通常的遗憾，但对于更小的 k，其值还可能更小。

算法：SAOCP 我们利用最小化强自适应遗憾的技术来进行在线保形预测。我们的主要算法——强自适应在线保形预测（SAOCP，见算法 1）——将 Jun 等人 (2017年的工作推广到在线保形预测设定。从整体上看，SAOCP 是一种元算法，用于管理多个专家；每个专家本身都是一个任意的在线学习算法，负责处理自己的活动区间，且该区间具有有限的生命周期。在每个 $t \in [T]$，算法 1 都会实例化一个新的专家 $\mathcal{A}_t$，其活动区间为 $[t, t + L(t) - 1]$，其中 $L(t)$ 是其生命周期：

$$L(t) := g \cdot \max_{n \in \mathbb{Z}} \{2^n : t \equiv 0 \bmod 2^n\}, \tag{8}$$

且 $g \in \mathbb{Z}_{\ge 1}$ 是每位专家生命周期的乘数。显然，在选择（$g\lfloor \log_2 t \rfloor$）时，任意时刻最多有 t 个专家处于活跃状态，这使算法 8 下的算法1 总运行时间为 $\mathcal{O}(T \log T)$，适用于任意

算法 2： 无尺度在线梯度下降（SF-OGD），改编自Orabona 和 P´等人(2018)。

输入： $\alpha \in (0,\ 1)$，学习率 $\eta > 0$，初始化 $\widehat{s}_1 \in \mathbb{R}$**1 对于** $t \geq 1$ **执行2** 观察输入 $X_t \in \mathcal{X}$**3 返回** 预测集 $C_t(X_t,\ \widehat{s}_t)$**4** 观察真实标签 $Y_t \in \mathcal{Y}$，并计算真实半径 $S_t = \inf\{s \in \mathbb{R} : Y_t \in \widehat{C}_t(X_t,\ s)\}$。**5** 计算损失 $\ell^{(t)}(\cdot) = \ell_{1-\alpha}(S_t,\ \cdot)$**6** 更新预测半径 $\widehat{s}_{t+1} = \widehat{s}_t - \eta \nabla\ell^{(t)}(\widehat{s}_t)\Big/\sqrt{\sum_{i=1}^{t} \|\nabla\ell^{(i)}(\widehat{s}_i)\|_2^2}$

(10)

$g = \Theta(1)$。然后，在任意时刻 t，预测半径 $\widehat{s}_t$ 由活跃专家的预测聚合得到（第7行）：

$$\widehat{s}_t = \sum_{i \in \text{Active}(t)} p_{i,t}\widehat{s}_{i,t},$$

其中，权重 $\{p_{i,t}\}_{i\in[t]}$（第4-6行）依赖于第 $\{w_{i,t}\}_{i\in[t]}$ 行中通过硬币投注方案计算得出的（Orabona 和 P´等人，2016；Jun 等人，2017年）第12-13行的结果。

专家的选择原则上，SAOCP 允许选择任何一种专家，只要它是在自身活动区间上具有良好遗憾最小化能力、并满足任意时刻遗憾保证的算法。我们选择无尺度 OGD（SF-OGD；算法2）（Orabona 和 P´等人，2018）作为专家。它是 OGD 的一种变体，会根据过去梯度范数的累积值衰减其有效学习率（参见（10））。

对于分位数损失（1）（在整个预测步长 $[T]$ 上以学习率 $\eta = \Theta(D)$ 执行）；$\eta = D/\sqrt{3}$ 为最优），SF-OGD 享有任意时刻遗憾保证（命题A.2）

$$\text{Reg}(t) \leq \mathcal{O}(D\sqrt{t}) \quad \text{for all } t \in [T], \tag{9}$$

这可直接通过应用 Orabona 和 P´al （2018，定理 2）得出。将 SF-OGD 代入算法 1 第2 行，即可得到完整的 SAOCP 算法。

将 SF-OGD 作为独立算法 作为一种强遗憾最小化算法，SF-OGD 也可以独立运行（在 $[T]$上），作为在线保形预测算法（见算法 2）。我们通过实证发现，在许多场景中，SF-OGD 本身已能取得出色的表现（第 5 节）。

4 理论

4.1 强自适应遗憾

我们首先证明 SAOCP 的 SARegret 保证。由于我们将 SF-OGD 用作专家来实例化 SAOCP，证明可直接将 SF-OGD 的遗憾界（9）代入 SAOCP 的 SARegret 保证（Jun 等人，2017年），详见附录 B.1。

命题 4.1 (SAOCP 的 SARegret 遗憾界)。 算法 *1* 同时达到以下 *SARegret* 遗憾界，适用于所有长度 $k \in [T]$:

$$\text{SAReg}(T, k) \leq 15D\sqrt{k(\log T + 1)} \leq \widetilde{\mathcal{O}}(D\sqrt{k}). \tag{11}$$

在 $\widetilde{\mathcal{O}}(D\sqrt{k})$速率下，SAOCP 对一般在线凸优化问题所达到的性能接近最优，这是因为任何长度为 k (Orabona，2019年，定理 5.1）对应的固定区间上，都存在标准遗憾下界 $\Omega(D\sqrt{k}$。

与 FACI 的比较 SAOCP 在 (11) 中给出的 SARegret 保证，相比 FACI（完全自适应保形推断）算法（Gibbs 和 Cand`es，2022年）有了显著改进；FACI 是 ACI 的扩展。具体而言，(11) 对所有长度 k均同时成立。相比之下，在我们的设定中，FACI 达到 SAReg(T, k) $\le \widetilde{\mathcal{O}}(D^2/\eta + \eta k)$（参见其定理 3.2），其中 $\eta > 0$ 是其元算法学习率。这可以通过优化 η，对单个 k 得到相同的速率 $\widetilde{\mathcal{O}}(D\sqrt{k})$，但无法同时适用于多个 k 值。

此外，就算法风格而言，尽管 SAOCP 和 FACI 都是维护多个专家（基础算法）的元算法，但两者的一个主要区别在于，FACI 中的所有专家学习率各不相同，且都在 $[T]$ 上保持活跃；而 SAOCP 中的专家则在其活动区间上有所不同（参见（8））。

动态遗憾 动态遗憾衡量在线学习算法相对于任意比较器序列的表现，是对遗憾的一种推广，用于刻画算法在变化环境中的表现（Zinkevich，2003；Besbes 等人，2015）。基于从动态遗憾到强自适应遗憾的归约（Zhang 等人，2018），我们证明，SAOCP 相对于 $[T]$ 中任意区间上的最优比较器，也能实现近乎最优的动态遗憾，其收敛速率取决于真实半径 $\{S_t\}_{t\ge 1}$ 的某种路径长度；详见命题 B.1 及其后的讨论。

4.2 覆盖率

回顾式（5）中定义的经验覆盖误差：

$$\mathrm{CovErr}(T) = \left|\tfrac{1}{T}\textstyle\sum_{t=1}^{T} \mathrm{err}_t - \alpha\right|.$$

在不作任何分布假设的情况下，我们证明，对于任意t，SF-OGD 达到 CovErr($\le \mathcal{O}(t^{-1/4}\log t)$。因此，随着$t \in [T]$，其经验覆盖率会收敛到目标值 $1-\alpha$，$T \to \infty$与 ACI 类似（但收敛速度略慢）。证明（附录 B.3）基于分组论证，~~以及 SF-OGD~~ 的有效学习率 $\eta/\sqrt{\sum_{\tau=1}^{t} \left\|\nabla \ell^{(\tau)}(\widehat{s}_\tau)\right\|_2^2}$ 随 t缓慢变化这一事实。

定理 4.2（SF-OGD 的覆盖率界）。 算法 *2* 采用任意学习率 $\eta = \Theta(D)$ 和任意初始化 $\widehat{s}_1 \in [0, D]$，可达到 $\mathrm{CovErr}(T) \le \mathcal{O}(\alpha^{-2}T^{-1/4}\log T)$，对任意 $T \ge 1$。

现在，我们基于与定理 4.2 中类似的分组论证，为 SAOCP 提供一个无分布假设的覆盖率界。

定理 4.3（SAOCP 覆盖率界；定理 B.3 的非正式版本）。 对于任意 $T \ge 1$，将算法 *1* 的第 *7* 行替换为采样一个专家 $i \sim p_t$ 并进行预测 $\widehat{s}_t := \widehat{s}_{t,\,i}$ 的随机化变体可以达到

$$\mathrm{CovErr}(T) \le \mathcal{O}\big(\inf_\beta (T^{1/2-\beta} + T^{\beta-1}S_\beta(T))\big). \tag{12}$$

定理 4.3 考虑了 SAOCP 的一个随机化变体，其覆盖率界取决于一个量 $S_\beta(T)$（完整定义见定理 B.3），该量衡量专家权重的平滑性以及每个单独专家的累积梯度范数。出于技术原因，这两者都是必要的；它们也出现在其他专家式元算法（例如 FACI）的覆盖率界中（Gibbs 和 Cand`es，2022年）。例如，如果 ~存在 $\beta \in (1/2, 1)$，使得对于某个 $\gamma < 1-\beta$有 $S_\beta(T) \le \mathcal{O}(T^\gamma)$，则（12）意味着覆盖率界 CovErr($T$) $\le \widetilde{\mathcal{O}}($ $T^{-\min\{1/2-\beta,\beta-1+\gamma\}}$) $= o_T(1)$。

我们指出，定理 4.3 对于专家权重 $\{p_t\}_{t\in[T]}$（第 5-6 行）和活动区间的其他选择也普遍成立，而不局限于算法 1 中指定的情形。特别地，SF-OGD 是仅有一个活跃专家处于活动状态的特殊情形 $[T]$。在这种情况下，我们可以得到定理 $\widetilde{\mathcal{O}}$（$\alpha^{-2}T^{-1/4}$）的界 4.2 （详细讨论见附录B.4.1 ）。

分布假设下的额外覆盖率保证在对 S_1、……、S_T的分布作出一些温和的正则性假设后，我们在定理C.3 中证明，SAOCP 在每个时间子区间上都能实现近似有效覆盖率。它在任意区间 $I = [\tau, \tau+k-1] \subseteq [1, T]$ 上的覆盖误差为

$\widetilde{\mathcal{O}}\left(k^{-1/(2q)}+\left(\mathrm{Var}_I/k\right)^{1/q}\right)$，其中 $q\geq 2$ 是一个用于量化分布正则性的量，而 Var_I 是衡量 S_t 在 $t\in I$ 上的真实分位数之间变化程度的一种度量（参见（29））。特别地，对于满足 $\mathrm{Var}_I=o(k)$ I 的任意区间，我们都能实现近似有效覆盖率。

5 实验

我们在两个具有代表性的现实世界在线不确定性量化任务上，对 SF-OGD（算法 2）和 SAOCP（算法 1）进行实证检验：时间序列预测（第 5.1 节）以及分布偏移^情况下的图像分类（第 5.2 节）。预测集 $\{C_t(x,\ s)\}_{x,\ s}$ 的选取将在各项实验中分别说明。在两项实验中，我们都与以下方法进行比较：

1. SCP：针对在线设置调整的标准分割保形预测（Vovk 等人，2005年），仅根据过去的半径预测其 $(1-\alpha)$ 分位数。在我们的设置中，SCP 不具备有效覆盖率保证，因为数据通常可能不可交换；2. NExCP：非可交换SCP（Barber 等人，2022年），这是 SCP 的一种变体，通过重新加权处理不可交换数据。我们遵循其建议，采用指数加权方案，对较近的观测赋予更高权重；3. FACI（Gibbs 和 Cand`es，2022年）采用其特定的“分位数参数化”（4）和与我们所选 ~ ~S_t 相对应的评分函数 C_t；^4. FACI-S：FACI 的广义版本，直接用于预测我们所选 $\widehat{s}_t$对应的半径 C_t 。

所有方法的更多细节请参见附录 E。本节中，我们将目标覆盖率设为标准的 $1-\alpha=90\%$。

5.1 时间序列预测

设置 我们考虑具有实值观测 $\{y_t\}_{t\geq 1}\in$~ ~$\mathbb{R}$ 的多时域时间序列预测问题，其中基础预测器 f 使用历史数据 $X_t:=y_{1:t}$ 来预测未来 H 步，即 $f(X_t)=\{\widehat{f}^{(h)}(X_t)\}_{h\in[H]}=\{\widehat{y}_{t+h}^{(h)}\}_{h\in[H]}$，其中 $\widehat{y}_{t+h}^{(h)}$是对 y_{t+h} 的预测。使用 $\widehat{f}(X_t)$，我们生成固定宽度的预测区间

$$\widehat{C}_t^{(h)}(X_t,\widehat{s}_t^{(h)}):=\left[\widehat{y}_{t+h}^{(h)}-\widehat{s}_t^{(h)},\widehat{y}_{t+h}^{(h)}+\widehat{s}_t^{(h)}\right], \tag{13}$$

其中，对每个 $h\in[H]$，均由在线保形预测算法的一个独立副本来预测 $\widehat{s}_t^{(h)}$（因此有 H 个此类算法并行运行）。我们使用标准的滚动窗口评估循环构建在线设置，其中每个批次 包括预测全部 H 个区间 $\{\widehat{C}_t^{(h)}\}_{h\in[H]}$、观测全部 H 个真实值 $\{y_{t+h}\}_{h\in[H]}$，然后通过设置 $t\to t+H$ 进入下一个批次。对于每个 $h\in[H]$，我们仅使用一个区间 $\widehat{C}_t^{(h)}$（X_t，$\widehat{s}_t^{(h)}$）进行评估。评估完成后，我们使用所有配对 $\{(y_{t+k},\widehat{y}_{t+k}^{(h)})\}_{k\in[H]}$ 来更新 $\widehat{s}_t^{(h)}\to\widehat{s}_{t+H}^{(h)}$。

基础预测器我们考虑三种不同类型的基础预测器（模型），并使用它们在 Merlion v2.0.0 中的实现（Bhatnagar 等人，2021）：

1. LGBM：一种使用梯度提升树进行预测的模型 $\widehat{y}_{t+h}^{(h)}=\widehat{f}^{(h)}$（$y_{t-L+1}$、……、$y_t$）。该方法在许多时间序列基准上都取得了出色的性能（Elsayed 等，2021；Bhatnagar 等，2021）。2. ARIMA(10，$d^\star$，10)：针对时间序列的经典自回归积分滑动平均随机过程模型，其中差分阶数 $d^\star$ 通过 KPSS 平稳性检验确定（Kwiatkowski 等，1992年）。3. Prophet（Taylor 和 Letham，2017年）：一种流行的贝叶斯模型，直接将值 y 表示为^时间的函数，即 $\widehat{y}_t=f(t)$。

方法	LGBM（平均绝对误差 = 0.06）				自回归积分滑动平均模型（平均绝对误差 = 0.18）				Prophet（平均绝对误差 = 0.12）			
	覆盖率	宽度	局部条件误差$_k$	SAReg$_k$	覆盖率	宽度	局部条件误差$_k$	SAReg$_k$	覆盖率	宽度	局部条件误差$_k$	SAReg$_k$
SCP	0.844	0.127	0.252	0.017	0.871	0.245	0.237	0.039	0.783	0.178	0.355	0.019
NExCP	0.875	0.134	0.197	0.013	0.871	0.245	0.227	0.040	0.856	0.187	0.231	0.010
FACI	0.866	**0.113**	0.180	**0.009**	0.866	0.232	0.214	0.034	0.867	0.175	0.184	0.006
SF-OGD	0.889	0.138	0.154	0.011	0.877	0.250	0.195	0.037	0.888	0.186	**0.138**	0.007
FACI-S	0.883	0.128	0.163	0.010	0.872	0.238	0.201	0.035	0.885	0.180	0.144	0.006
SAOCP	0.882	0.121	**0.143**	**0.009**	0.864	**0.221**	**0.190**	**0.033**	0.872	**0.173**	0.143	**0.005**

表 1：M4 小时（414 个时间序列）上在目标覆盖率 $1-\alpha = 0.9$ 和区间大小 $k = 20$下的结果。当方法的全局覆盖率位于 (0.85，0.95)（绿色）范围内时，最佳结果以**粗体**显示，次佳结果加下划线。对于所有基础预测器，SAOCP 在宽度、局部覆盖误差和强自适应遗憾方面均取得了最佳或次佳结果。

方法	LGBM（平均绝对误差 = 0.14）				自回归积分滑动平均模型（平均绝对误差 = 0.06）				Prophet（平均绝对误差 = 0.32）			
	覆盖率	宽度	局部条件误差$_k$	SAReg$_k$	覆盖率	宽度	局部条件误差$_k$	SAReg$_k$	覆盖率	宽度	局部条件误差$_k$	SAReg$_k$
SCP	0.769	0.184	0.466	0.031	0.896	0.122	0.290	0.018	0.599	0.349	0.614	0.051
NExCP	0.818	0.183	0.420	0.015	0.891	0.116	0.296	0.012	0.715	0.356	0.559	0.019
FACI	0.846	0.169	0.308	0.008	0.886	0.101	0.259	**0.008**	0.767	0.344	0.397	0.014
SF-OGD	0.873	0.173	0.246	0.011	0.892	0.106	0.245	0.011	0.862	0.354	0.220	0.008
FACI-S	0.875	0.169	0.240	0.010	0.891	0.103	0.243	0.010	0.866	0.352	0.210	0.007
SAOCP	0.869	**0.162**	**0.213**	**0.007**	0.875	**0.093**	**0.238**	**0.008**	0.867	**0.349**	**0.172**	**0.005**

表 2：M4 Daily（4227 条时间序列）在目标覆盖率 $1-\alpha = 0.9$ 和区间大小 $k = 20$下的结果。最佳结果以**粗体**表示，只要该方法的全局覆盖率处于 (0.85,0.95)（绿色）范围内，次佳结果则加下划线。对于所有基础预测器，SAOCP 在宽度、局部覆盖误差和强自适应遗憾方面均表现最佳。对于 LGBM 和 Prophet，只有直接预测 $\widehat{s}_{t+1}$ 而非将其作为 $S_1, \ldots, S_t$ 的分位数来预测的方法，才能达到 (0.85,0.95) 范围内的全局覆盖率。

数据集 我们在四个数据集上进行评估，这些数据集共包含 5111 条时间序列：M4 竞赛的小时频率（414 条时间序列）、日频率（4227 条时间序列）和周频率（359 条时间序列）子集。M4 是一个涵盖多个领域时间序列的数据集，包括工业、人口统计、环境、金融和交通运输（Makridakis 等人，2018）；另一个是 NN5，这是一个包含 111 条银行日数据时间序列的数据集（Ben Taieb 等人，2012）。我们将每条时间序列归一化，使其取值落在 $[0, 1]$。

对于小时、日度和周度数据，我们分别使用预测步长 H 24、30 和 26。每条长度为 L 的时间序列被划分为长度为 $L - 120$ 的训练集，其中 80% 用于训练基础预测器，20% 用于初始化不确定性量化方法；另设长度为 120 的测试集，用于测试不确定性量化方法。

评估指标 对于每项实验，我们对所有时间序列的以下统计量取平均值：全局覆盖率、中位宽度、最坏情况局部覆盖误差

$$\mathrm{LCE}_k := \max_{[\tau, \tau+k-1] \subseteq [1,T]} \left| \alpha - \tfrac{1}{k} \textstyle\sum_{t=\tau}^{\tau+k-1} \mathrm{err}_t \right|, \tag{14}$$

以及强自适应遗憾 SAReg(T, k) (7)，简称为 SAReg$_k$。所有情况下，我们均使用长度为 $k = 20$的区间。我们还报告了各基础预测器的平均绝对误差（MAE）均值。

结果 我们在表 1 和 2 中报告 M4 小时和 M4 Daily 的结果，并在表 3 和 4(附录 D) 中报告 M4 Weekly 和 NN5 Daily 的结果。

SAOCP 始终能在 (0.85,0.95) 内达到全局覆盖率，并且在全部 3 个 M4 数据集上，对于所有基础预测器，其区间宽度、局部覆盖误差和强自适应遗憾均排名第一或第二。FACI-S 的局部条件误差$_k$ 和 SAReg$_k$ 通常优于 FACI，这表明直接预测 $\widehat{s}_{t+1}$ 而不是将其作为 $S_1, \ldots, S_t$ 的分位数进行预测具有优势。FACI-S 和 SF-OGD 的相对表现有所不同，不过 FACI-S 通常略胜一筹。然而，SAOCP 的局部条件误差$_k$ 和 SAReg$_k$ 始终优于 FACI-S 和 SF-OGD。

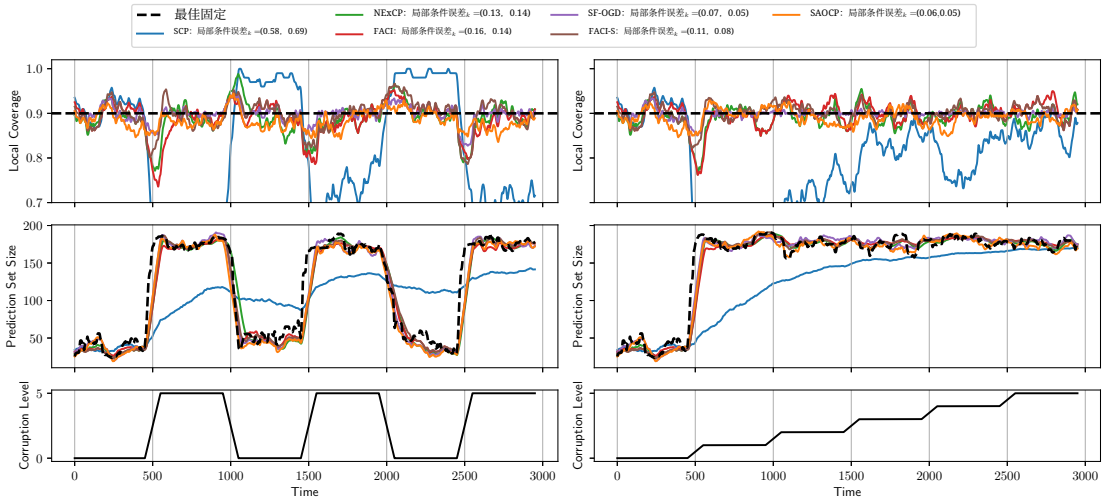

图 1：当分布每 500 步在 TinyImageNet 和 TinyImageNet-C 之间偏移时，各种不确定性量化方法实现的局部覆盖率（上排）和预测集大小（第二排）。我们绘制窗口大小为 100 的移动平均值。左：损坏等级在 0 和 5 之间突变偏移。右：损坏等级从 0 逐渐变化到 5。SAOCP 和 SF-OGD 的局部覆盖率最接近 0.9 的目标值，尤其是在变点处。虽然它们的预测集大小相近，但 SAOCP 的局部条件误差$_k$ 低于 SF-OGD。

在多种情况下，SCP/NExCP/FACI 均未能在 (0.85,0.95) 内达到全局覆盖率（表 2 和 3）。在所有这些情况下，基础预测器的平均绝对误差至少为 0.14，这表明当底层基础预测器不准确时，像 SF-OGD/SAOCP 那样直接预测 $\hat{s}_{t+1}$ 具有优势。

关于集成模型的补充实验在附录F中，我们使用 EnbPI（Xu 和 Xie，2021）训练自助集成，并将 EnbPI 的结果与对该集成产生的残差应用 NExCP、FACI、SF-OGD 和 SAOCP 得到的结果进行比较。结果总体上与正文中的结果一致。

5.2 分布偏移下的图像分类

数据集与设置 我们评估保形预测方法在底层分布以系统性方式发生偏移时维持覆盖率的能力。我们使用在 ImageNet 上预训练并通过 PyTorch 实现的 ResNet-50 分类器（Heetal.，2016）。此处，$x \in \mathcal{X}$ 是一幅图像，而$y \in \mathcal{Y} = [m]$ 是其类别。为了构造相对于训练分布的结构化分布偏移，我们使用 TinyImageNet-C 和 ImageNet-C（Hendrycks 和 Dietterich，2019年），它们分别是 TinyImageNet（$m = 200$ 个类别）（Le 和 Yang，2015）和 ImageNet（$m = 1000$ 个类别）（Deng et al.，2009）测试集的受损版本，旨在评估模型鲁棒性。这些受损数据集对原始测试集中的每幅图像施加 15 种视觉损坏，且每种损坏均有 5 个不同的严重程度级别。

我们考虑两种情形：**突变偏移**，其中损坏等级在 0（基础测试集）和 5 之间交替变化；以及**渐进偏移**，其中损坏等级按 $\{0, 1, \ldots, 5\}$ 的顺序递增。在切换到下一个等级之前，我们会为每个损坏等级随机抽取 500 个数据点。

预测集 我们遵循 Angelopoulos 等人 (2021b) 构造预测集。设 $\widehat{f}\colon \mathbb{R}^d \to \Delta^m$是一个在 m-单纯形上输出概率分布的分类器。在每个 t，我们采样 $U_t \sim \mathrm{Unif}[0,1]$并令

$$S_t(x,y) = \lambda\sqrt{[k_y - k_{reg}]_+} + U_t\widehat{f}_y(x) + \sum_{i=1}^{k_y-1}\widehat{f}_{\pi(i)}(x)$$
$$\widehat{C}_t(X_t) = \{y : S_t(X_t,y) \le \widehat{s}_t\}, \tag{15}$$

^其中，π 是按降序排列 $f(x)$ 的排列，$\pi(k_y) = y$，且 λ 和 k_{reg} 是用于减小预测集大小的正则化参数。对于 TinyImageNet，我们使用 $\lambda = 0.01$ 和$k_{reg} = 20$。对于 ImageNet，我们使用 $\lambda = 0.01$ 和 $k_{reg} = 10$。

评估指标 在评估不确定性量化方法时，我们使用长度为 $k = 100$ 的区间，绘制每种方法的局部覆盖率和预测集大小（PSS），

$$\text{LocalCov}(t) = \frac{1}{100}\sum_{i=t}^{t+99} \mathbb{1}[Y_i \in \widehat{C}_i(X_i)]$$
$$\text{LocalPSS}(t) = \frac{1}{100}\sum_{i=t}^{t+99} \left|\widehat{C}_i(X_i)\right|.$$

我们将局部覆盖率与 $1-\alpha$这一目标值进行比较，同时将局部预测集大小与 $1-\alpha$ 预言机集合大小的经验分位数 PSS $^\star_t = |\{y$ 进行比较：$S_t(X_t, y) \leq S_t(X_t, Y_t)\}|$。这些目标值是每个窗口中的“最佳固定”值。我们还报告了最坏情况局部覆盖误差 LCE_{100}（14）。

结果 我们在图 1 中对 TinyImageNet 和 TinyImageNet-C 上的不确定性量化方法进行评估，并在图 2（附录 G）中对 Ima- geNet 和 ImageNet-C 上的方法进行评估。在突发和渐进的分布偏移下，SAOCP 和 SF-OGD 的局部覆盖率始终最接近 0.9 这一目标值。当分布发生突发偏移时，这种差异更加明显。当分布偏移更加渐进时，在第一个变点处，NExCP、FACI 和 FACI-S 的覆盖率比 SAOCP 和 SF-OGD 更差；该变点处最佳集合大小发生的变化最大。

除SCP之外，所有方法预测集的大小都相近，不过 FACI、FACI-S 和 NExCP 的预测集大小对最佳固定大小变化的适应更慢（例如 $t \in [500, 700]$ 图 2 中的渐变偏移）。在 TinyImageNet 上，SAOCP 获得的局部覆盖率略高于 SF-OGD，且二者的预测集大小相近（图 1）。在 ImageNet 上，SAOCP 和 SF-OGD 达到的局部覆盖率相近，但 SAOCP 往往能够以更小的预测集达到该覆盖率（图 2）。

6 结论

本文针对任意分布偏移下的在线保形预测，提出了新的算法。我们提出的算法实现了近似有效覆盖率，并取得了优于现有工作的强自适应遗憾。在真实世界实验中，我们提出的算法在局部窗口内获得了更接近目标值的覆盖率，并且生成的预测集比现有方法更小。我们的工作为未来研究开启了许多问题，例如如何获得更强的覆盖保证，或刻画在具有分布偏移的各种设置下学习得到的半径的最优性。

参考文献

Anastasios N Angelopoulos、Stephen Bates、Emmanuel J Cand`es、Michael I Jordan 和 Lihua Lei。先学习后检验：校准预测算法以实现风险控制。*arXiv* 预印本 *arXiv:2110.01052*，2021a。

Anastasios N Angelopoulos、Stephen Bates、Adam Fisch、Lihua Lei 和 TalSchuster。保形风险控制。*arXiv* 预印本 *arXiv:2208.02814*，2022a。

Anastasios N Angelopoulos、Amit Pal Kohli、Stephen Bates、Michael Jordan、Jitendra Malik、Thayer Al- shaabi、Srigokul Upadhyayula 和 Yaniv Romano。图像到图像回归中的无分布假设不确定性量化及其在成像中的应用。载于国际机器学习大会，第 717–730 页。PMLR，2022b。

Anastasios Nikolas Angelopoulos 和 Stephen Bates。保形预测与分布无关不确定性量化简介，2021。网址 https://arxiv.org/abs/2107.07511。

Anastasios Nikolas Angelopoulos、Stephen Bates、Michael Jordan 和 Jitendra Malik。使用保形预测为图像分类器构建不确定性集合。载于 国际学习表征会议，2021b。网址https://openreview.net/forum?id=eNdiU_DbM9。

Yu Bai、Song Mei、Huan Wang、Yingbo Zhou 和 Caiming Xiong。基于一般函数类的高效且可微保形预测。载于《国际学习表征会议》，2022年。网址 https://openreview.net/forum?id=Ht85_jyihxp。Rina Foygel Barber、Emmanuel J. Cand`es、Aaditya Ramdas 和 Ryan J. Tibshirani。使用刀切法进行预测推断。《统计年刊》，49(1)：486–507，2021。doi：10.1214/20-AOS1965。网址 https://doi.org/10.1214/20-AOS1965。Rina Foygel Barber、Emmanuel J. Cand`es、Aaditya Ramdas 和 Ryan J. Tibshirani。超越可交换性的保形预测，2022年。网址 https://arxiv.org/abs/2202.13415。Stephen Bates、Anastasios Angelopoulos、Lihua Lei、Jitendra Malik 和 Michael Jordan。无分布假设、风险控制的预测集。《美国计算机学会会刊》（JACM），68(6)：1–34，2021年。Souhaib Ben Taieb、Gianluca Bontempi、Amir F. Atiya 和 Antti Sorjamaa。基于NN5预测竞赛的多步时间序列预测策略综述与比较。《应用专家系统》，39(8)：7067–7083，2012。ISSN 0957-4174。doi：https://doi.org/10.1016/j.eswa. 2012.01.039。网址 https://www.sciencedirect.com/science/article/pii/S0957417412000528。Omar Besbes、Yonatan Gur 和 Assaf Zeevi。非平稳随机优化。《运筹学》，63(5)：1227–1244，2015年。Aadyot Bhatnagar、Paul Kassianik、Chenghao Liu、Tian Lan、Wenzhuo Yang、Rowan Cassius、Doyen Sa- hoo、Devansh Arpit、Sri Subramanian、Gerald Woo、Amrita Saha、Arun Kumar Jagota、Gokulakrishnan Gopalakrishnan、Manpreet Singh、K C Krithika、Sukumar Maddineni、Daeki Cho、Bo Zong、Yingbo Zhou、Caiming Xiong、Silvio Savarese、Steven Hoi 和 Huan Wang。Merlion：时间序列机器学习库。2021年。网址 https://arxiv.org/abs/2109.09265。Emmanuel J. Cand`es、Lihua Lei 和 Zhimei Ren。保形生存分析，2021年。网址 https: //arxiv.org/abs/2103.09763。Maxime Cauchois、Suyash Gupta、Alnur Ali 和 John C. Duchi。稳健验证：即使分布发生偏移，也能做出可信预测，2020年。网址 https://arxiv.org/abs/2008.04267。Maxime Cauchois、Suyash Gupta 和 John C. Duchi。知其所知：多类别和多标签预测中的有效且经过验证的置信集。《机器学习研究杂志》，22(1)，2022年7月。ISSN 1532-4435。Victor Chernozhukov、Kaspar W¨uthrich 和 Zhu Yinchu。针对相关数据的预测性机器学习的精确稳健保形推断方法。载于 S´ebastien Bubeck、Vianney Perchet 和 Philippe Rigollet 编，《第31届学习理论会议论文集》，第75卷《机器学习研究论文集》，第732–749页。PMLR，2018年7月6–9日。网址 https://proceedings.mlr. press/v75/chernozhukov18a.html。Amit Daniely、Alon Gonen 和 Shai Shalev-Shwartz。强自适应在线学习。载于 Francis Bach 和 David Blei 编，《第32届国际机器学习会议论文集》，第37卷《机器学习研究论文集》，第1405–1411页，法国里尔，2015年7月7–9日。PMLR。网址 https://proceedings.mlr.press/v37/daniely15.html。Mikhail Dashevskiy 和 Zhiyuan Luo。带置信度的网络流量需求预测。载于 IEEE GLOBE- COM 2008—2008年IEEE全球电信会议，第1–5页，2008年。doi：10.1109/GLOCOM. 2008.ECP.284。Jia Deng、Wei Dong、Richard Socher、Li-Jia Li、Kai Li 和 Li Fei-Fei。ImageNet：大型层次化图像数据库。载于2009年IEEE计算机视觉与模式识别会议，第248–255页。IEEE，2009年。Shereen Elsayed、Daniela Thyssens、Ahmed Rashed、Hadi Samer Jomaa 和 Lars Schmidt-Thieme。时间序列预测真的需要深度学习模型吗？，2021年。网址 https://arxiv.org/abs/2101. 02118。

Clara Fannjiang、Stephen Bates、Anastasios Angelopoulos、Jennifer Listgarten 和 Michael I Jordan。针对设计问题的保形预测。arXiv 预印本 arXiv:2202.03613，2022年。Shai Feldman、Liran Ringel、Stephen Bates 和 Yaniv Romano。在线学习模型的风险控制，2022年。网址：https://arxiv.org/abs/2205.09095。Isaac Gibbs 和 Emmanuel Cand`es。分布偏移下的自适应保形推断。载于 A. Beygelzimer、Y. Dauphin、P. Liang 和 J. Wortman Vaughan 编，《神经信息处理系统进展》，2021。网址：https://openreview.net/forum?id=6vaActvpcp3。Isaac Gibbs 和 Emmanuel Cand`es。任意分布偏移下的在线预测保形推断，2022年。网址：https://arxiv.org/abs/2208.08401。Chirag Gupta、Arun K Kuchibhotla 和 Aaditya K Ramdas。嵌套保形预测与分位数袋外集成方法。arXiv 预印本 arXiv:1910.10562，2019年。Elad Hazan。《在线凸优化导论》。麻省理工学院出版社，剑桥，美国马萨诸塞州，2022年。ISBN 9780262046985。何恺明、张祥雨、任少卿和孙剑。用于图像识别的深度残差学习。载于《2016年IEEE计算机视觉与模式识别会议》（CVPR），第770–778页，2016年。doi：10.1109/CVPR.2016.90。Dan Hendrycks 和 Thomas Dietterich。对神经网络在常见损坏和扰动下的鲁棒性进行基准测试。载于《国际学习表征会议》，2019年。网址：https: //openreview.net/forum?id=HJz6tiCqYm。Dan Hendrycks、Mantas Mazeika、Duncan Wilson 和 Kevin Gimpel。利用可信数据训练深度网络，以应对严重噪声污染的标签。《神经信息处理系统进展》，31，2018年。Kwang-Sung Jun、Francesco Orabona、Stephen Wright 和 Rebecca Willett。利用硬币投注改进强自适应在线学习。载于 Aarti Singh 和 Jerry Zhu 编，《第20届国际人工智能与统计会议论文集》，《机器学习研究论文集》第54卷，第943–951页。PMLR，2017年4月20–22日。网址：https://proceedings.mlr.press/v54/jun17a.html。Christopher Kath 和 Florian Ziel。保形预测区间估计及其在日前和日内电力市场中的应用。《国际预测学杂志》，37(2)：777–799，2021年。ISSN 0169-2070。doi：https://doi.org/10.1016/j.ijforecast.2020.09.006。网址：https://www.sciencedirect.com/science/article/pii/S0169207020301473。Roger Koenker 和 Gilbert Bassett Jr.。回归分位数。《计量经济学》：计量经济学学会杂志，第33–50页，1978年。Denis Kwiatkowski、Peter C.B. Phillips、Peter Schmidt 和 Yongcheol Shin。检验平稳性原假设与单位根备择假设：我们有多大把握认为经济时间序列存在单位根?《计量经济学杂志》，54(1)：159–178，1992年。ISSN 0304-4076。doi：https: //doi.org/10.1016/0304-4076(92)90104-Y。网址：https://www.sciencedirect.com/science/article/pii/030440769290104Y。Ya Le 和 Xuan S. Yang。Tiny ImageNet 视觉识别挑战赛。2015。Jing Lei 和 Larry Wasserman。非参数回归的无分布假设预测带。《英国皇家统计学会杂志：B辑（统计方法学）》，76(1)：71–96，2014年。doi：https://doi.org/ 10.1111/rssb.12021。网址：https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12021。Jing Lei、James Robins 和 Larry Wasserman。无分布假设预测集。《美国统计学会杂志》，108(501)：278–287，2013年。doi：10.1080/01621459.2012.751873。网址：https://doi. org/10.1080/01621459.2012.751873。PMID：25237208。

Spyros Makridakis、Evangelos Spiliotis，以及 Vassilios Assimakopoulos。M4 竞赛：结果，fi发现、结论与未来方向。《国际预测学杂志》，34(4):802–808，2018。

Francesco Orabona。在线学习现代导论。*arXiv* 预印本 *arXiv:1912.13213*，2019年。

Francesco Orabona 和 D´avid P´al。硬币投注与无参数在线学习。神经信息处理系统进展，29，2016。

Francesco Orabona 和 D´avid P´al。无尺度在线学习。《理论计算机科学》，716:50–69，2018。ISSN 0304-3975。doi: https://doi.org/10.1016/j.tcs.2017.11.021。URL https://www.sciencedirect.com/science/article/pii/S0304397517308514。ALT 2015 特刊。

Harris Papadopoulos。归纳保形预测：理论及其在神经网络中的应用。由 Paula Fritzsche 编。《人工智能工具》，第 18 章。IntechOpen，里耶卡，2008年。doi：10.5772/6078。URL https://doi.org/10.5772/6078。

Sangdon Park、Osbert Bastani、Nikolai Matni 和 Insup Lee。通过校准预测为深度神经网络构建PAC置信集。载于 国际学习表征会议，2020年。URL https://openreview.net/forum?id=BJxVI04YvB。

Adam Paszke、Sam Gross、Francisco Massa、Adam Lerer、James Bradbury、Gregory Chanan、Trevor Killeen、Zeming Lin、Natalia Gimelshein、Luca Antiga、Alban Desmaison、Andreas Kopf、Edward Yang、Zachary DeVito、Martin Raison、Alykhan Tejani、Sasank Chilamkurthy、Benoit Steiner、Lu Fang、Junjie Bai 和 Soumith Chintala。PyTorch：一种命令式风格的高性能深度学习库。载于神经信息处理系统进展 *32*，第 8024–8035 页。Curran Associates, Inc.，2019年。

Tim Pearce、Alexandra Brintrup、Mohamed Zaki 和 Andy Neely。深度学习的高质量预测区间：一种无分布假设的集成方法。载于 Jennifer Dy 和 Andreas Krause 编，第*35*届国际机器学习大会论文集，第80卷，机器学习研究论文集，第4075–4084页。PMLR，2018年7月10日至15日。网址：https://proceedings.mlr.press/v80/pearce18a.html。

Aleksandr Podkopaev 和 Aaditya Ramdas。标签偏移条件下分类的无分布假设不确定性量化。载于 Cassio de Campos 和 Marloes H. Maathuis 编，第三十七届人工智能不确定性会议论文集，第161卷，机器学习研究论文集，第844–853页。PMLR，2021年7月27日至30日。网址：https://proceedings.mlr.press/v161/podkopaev21a.html。

Yaniv Romano、Evan Patterson 和 Emmanuel Candes。《共形化分位数回归》。载于 H. Wallach、H. Larochelle、A. Beygelzimer、F. d'Alch´e-Buc、E. Fox 和 R. Garnett 编，神经信息处理系统进展，第 32 卷。Curran Associates, Inc.，2019年。网址：https://proceedings.neurips.cc/paper/2019/file/5103c3584b063c431bd1268e9b5e76fb-Paper.pdf。

Yaniv Romano、Matteo Sesia 和 Emmanuel J. Cand`es。《具有有效且自适应的覆盖率的分类》。载于神经信息处理系统进展第 *34* 届国际会议论文集，NIPS'20，美国纽约州雷德胡克，2020年。Curran Associates Inc. ISBN 9781713829546。

Glenn Shafer 和 Vladimir Vovk。《保形预测教程》。机器学习研究杂志，9(3)，2008年。

Martim Sousa、Ana Maria Tom´e 和 Jos´e Moreira。多步超前自适应共形异方差时间序列预测的通用框架，2022年。网址 https://arxiv.org/abs/2207.14219。Kamile Stankeviciute、Ahmed M. Alaa 和 Mihaela van der Schaar。共形时间序列预测。载于 M. Ranzato、A. Beygelzimer、Y. Dauphin、P.S. Liang 和 J. Wortman Vaughan 编，神经信息处理系统进展，第34卷，第6216–6228页。Curran Associates, Inc.，2021年。网址 https://proceedings.neurips.cc/paper/2021/file/312f1ba2a72318edaaa995a67835fad5-Paper.pdf。

Ingo Steinwart 和 Andreas Christmann。借助弹球损失估计条件分位数。伯努利，17(1):211–225，2011。doi：10.3150/10-BEJ267。网址 https://doi.org/10.3150/10-BEJ267。

David Stutz、Krishnamurthy Dj Dvijotham、Ali Taylan Cemgil 和 Arnaud Doucet。学习最优保形分类器。载于《国际学习表征会议》，2022年。网址 https:// openreview.net/forum?id=t8O-4LKFVx。Sophia Sun 和 Rose Yu。用于多步时间序列预测的 Copula 保形预测，2022年。网址 https://arxiv.org/abs/2212.03281。Sean J. Taylor 和 Benjamin Letham。大规模预测。PeerJ预印本，5(e3190v2)，2017年9月。doi：10.7287/peerj.preprints.3190v2。Ryan J Tibshirani、Rina Foygel Barber、Emmanuel Candes 和 Aaditya Ramdas。协变量偏移下的保形预测。载于 H. Wallach、H. Larochelle、A. Beygelzimer、F. d'Alch´e-Buc、E. Fox 和 R. Garnett 编，《神经信息处理系统进展》，第32卷。Curran Associates, Inc.，2019年。网址 https://proceedings.neurips.cc/paper/2019/file/8fb21ee7a2207526da55a679f0332de2-Paper.pdf。Vladimir Vovk。归纳式保形预测器的条件有效性。载于 Steven C. H. Hoi 和 Wray Buntine 编，《亚洲机器学习会议论文集》，《机器学习研究论文集》第25卷，第475–490页，新加坡管理大学，新加坡，2012年11月4日至6日。PMLR。网址 https://proceedings.mlr.press/v25/vovk12.html。Vladimir Vovk、Alex Gammerman 和 Glenn Shafer。《随机世界中的算法学习》。Springer-Verlag，柏林，海德堡，2005年。ISBN 0387001522。Vladimir Vovk、Ilia Nouretdinov、Valery Manokhin 和 Alexander Gammerman。交叉保形预测分布。载于 Alex Gammerman、Vladimir Vovk、Zhiyuan Luo、Evgueni Smirnov 和 Ralf Peeters 编，《第七届保形与概率预测及应用研讨会论文集》，《机器学习研究论文集》第91卷，第37–51页。PMLR，2018年6月11日至13日。网址 https://proceedings.mlr.press/v91/vovk18a.html。Volodya Vovk、Alexander Gammerman 和 Craig Saunders。算法随机性的机器学习应用。载于《第十六届国际机器学习会议论文集》，ICML '99，第444–453页，美国加利福尼亚州旧金山，1999年。Morgan Kaufmann出版社。ISBN 1558606122。Wojciech Wisniewski、David Lindsay 和 Sian Lindsay。保形预测区间估计在做市商净头寸中的应用。载于 Alexander Gammerman、Vladimir Vovk、Zhiyuan Luo、Evgueni Smirnov 和 Giovanni Cherubin 编，《第九届保形与概率预测及应用研讨会论文集》，《机器学习研究论文集》第128卷，第285–301页。PMLR，2020年9月9日至11日。网址 https://proceedings.mlr.press/v128/wisniewski20a.html。Chen Xu 和 Yao Xie。动态时间序列的保形预测区间。载于 Marina Meila 和 Tong Zhang 编，《第38届国际机器学习会议论文集》，《机器学习研究论文集》第139卷，第11559–11569页。PMLR，2021年7月18日至24日。网址 https: // proceedings.mlr.press/v139/xu21h.html。Yachong Yang 和 Arun Kumar Kuchibhotla。有限样本高效保形预测，2021年。网址 https://arxiv.org/abs/2104.13871。Yachong Yang、Arun Kumar Kuchibhotla 和 Eric Tchetgen Tchetgen。协变量偏移下预测集的双重稳健校准，2022年。网址 https://arxiv.org/abs/2203.01761。Margaux Zaffran、Olivier Feron、Yannig Goude、Julie Josse 和 Aymeric Dieuleveut。时间序列的自适应保形预测。载于 Kamalika Chaudhuri、Stefanie Jegelka、Le Song、Csaba Szepesvari、Gang Niu 和 Sivan Sabato 编，《第39届国际机器学习会议论文集》，《机器学习研究论文集》第162卷，第25834–25866页。PMLR，2022年7月17日至23日。网址 https://proceedings.mlr.press/v162/zaffran22a.html。

Lijun Zhang、Tianbao Yang、rong jin 和 Zhi-Hua Zhou。《强自适应方法的动态遗憾》。载于 Jennifer Dy 和 Andreas Krause 编，第*35*届机器国际会议论文集

学习，第80卷，机器学习研究论文集，第5882–5891页。PMLR，2018年7月10日至15日。网址：https://proceedings.mlr.press/v80/zhang18o.html。

Martin Zinkevich。在线凸规划与广义无穷小梯度上升。载于第二十届国际机器学习大会论文集，ICML'03，第 928–935 页。AAAI Press，2003 年。ISBN 1577351894。

A 在线保序预测算法的基本性质

A.1 SF-OGD 的性质

我们考虑 SF-OGD 算法（算法 2）。我们首先证明，SF-OGD 的迭代值均被限制在一个范围内，该范围略大于真实半径的范围。证明与 Gibbs 和 Cand`es(2021，引理 4.1）类似。

引理 A.1（SF-OGD 的有界迭代值）。 假设真实半径有界： $S_t \in [0, D]$ 对所有 $t \in [T]$。那么，算法 2 采用任意初始化 $\widehat{s}_1 \in [-\eta, D+\eta]$ 和学习率 $\eta > 0$ 均具有有界迭代值：

$$\widehat{s}_t \in [-\eta, D+\eta] \quad \textit{for all } t \in [T].$$

证明。 根据（2），回顾可知

$$\nabla \ell^{(t)}(\widehat{s}_t) = \alpha - \mathbb{1}\left[\widehat{s}_t < S_t\right] = \alpha - \mathrm{err}_t \in \{-(1-\alpha), \alpha\} \subset [-1, 1]. \tag{16}$$

对于所有 $t \in [T]$。因此，算法 2 满足：对于任意 $t \geq 1$，有

$$|\widehat{s}_{t+1} - \widehat{s}_t| = \eta \left| \frac{\alpha - \mathrm{err}_t}{\sqrt{\sum_{\tau=1}^{t} (\alpha - \mathrm{err}_\tau)^2}} \right| \leq \eta. \tag{17}$$

我们用反证法证明该引理。假设存在某个 t 使得 $\widehat{s}_t \notin [-\eta, D+\eta]$。令 $t \geq 2$为满足条件的最小时间索引（稍微滥用一下记号）。假设 $\widehat{s}_t > D + \eta$，那么根据（17）我们必须有 $\widehat{s}_{t-1} > D$，但 $\widehat{s}_{t-1} \leq D + \eta$。注意到 $\widehat{s}_{t-1} > D \geq S_{t-1}$ 根据我们的前提条件，因此第（$t-1$个预测集必须覆盖，故会出错$_{t-1} = 0$。因此，根据算法更新（10），我们有

$$\widehat{s}_t = \widehat{s}_{t-1} - \eta \frac{\alpha - \mathrm{err}_{t-1}}{\sqrt{\sum_{\tau=1}^{t-1} (\alpha - \mathrm{err}_\tau)^2}} < \widehat{s}_{t-1} \leq D + \eta,$$

这与我们的假设 $\widehat{s}_t > D + \eta$ 矛盾。对于另一种情形 $\widehat{s}_t < -\eta$，也可以推导出类似的矛盾。由此得证。 □

以下遗憾界可直接将无尺度 OGD 的通用遗憾界（Orabona 和 P´al，2018，定理 2）应用于分位数损失（1）得到。

命题 A.2（SF-OGD 的任意时刻遗憾界）。 假设真实半径有界： $S_t \in [0, D]$ 对所有 $t \in [T]$。_则算法 2 以任意初始化 $\widehat{s}_1 \in [0, D]$ 和学习率 $\eta = D/\sqrt{3}$ 对任意 $t \in [T]$均可达到以下遗憾界：

$$\mathrm{Reg}(t) \leq (\sqrt{3}+1) D \sqrt{\sum_{\tau=1}^{t} \left\| \nabla \ell^{(\tau)}(\widehat{s}_\tau) \right\|_2^2} \leq \mathcal{O}(D\sqrt{t}).$$

证明。 第二个不等式可直接由（16）得到。

为证明第一个不等式（遗憾界），我们注意到，算法 2 是 Orabona 和 P´al (2018，第4节）中的无尺度镜像下降算法的一个特例，其凸损失为 $\ell^{(t)}(\cdot) = \ell_{1-\alpha}(S_t, \cdot)$，正则化项为 $R(s) := s^2/(2\eta)$（采用他们的记号），它相对于 $\lambda = 1/\eta$范数是 ℓ_2 强凸的

在 $\mathbb{R}$上。此外，根据引理 A.1，我们有 $\widehat{s}_t \in [-\eta, D+\eta]$ 对所有 $t \in [T]$均成立。因此，应用 Orabona 和 P´al (2018，定理 2）可得，对任意 $t \in [T]$,

$$\begin{aligned}
&\sum_{\tau=1}^{t} \ell^{(\tau)}(\widehat{s}_\tau) - \inf_{s^\star \in [0,D]} \sum_{\tau=1}^{t} \ell^{(\tau)}(s^\star) \le \left(\frac{1}{\lambda} + \sup_{\tau \ge 1} B_R(s^\star, \widehat{s}_\tau)\right) \cdot \sqrt{\sum_{\tau=1}^{t} \left\|\nabla \ell^{(\tau)}(\widehat{s}_\tau)\right\|_2^2} \\
&= \left(\eta + \sup_{s^\star \in [0,D], s' \in [-\eta, D+\eta]} \frac{1}{2\eta}(s^\star - s')^2\right) \cdot \sqrt{\sum_{\tau=1}^{t} \left\|\nabla \ell^{(\tau)}(\widehat{s}_\tau)\right\|_2^2} \\
&= \left(\eta + \frac{(D+\eta)^2}{2\eta}\right) \cdot \sqrt{\sum_{\tau=1}^{t} \left\|\nabla \ell^{(\tau)}(\widehat{s}_\tau)\right\|_2^2},
\end{aligned}$$

其中， $B_R(\cdot, \cdot)$ 表示与 R 相关联的布雷格曼散度。选择 $\eta = D/\sqrt{3}$ 后，首项系数为 $(\sqrt{3}+1)D$。注意到

$$\mathrm{Reg}(t) = \sum_{\tau=1}^{t} \ell^{(\tau)}(\widehat{s}_\tau) - \inf_{s^\star \in \mathbb{R}} \sum_{\tau=1}^{t} \ell^{(\tau)}(s^\star) = \sum_{\tau=1}^{t} \ell^{(\tau)}(\widehat{s}_\tau) - \inf_{s^\star \in [0,D]} \sum_{\tau=1}^{t} \ell^{(\tau)}(s^\star)$$

by 我们的假设 $S_t \in [0, D]$ 以及分位数损失的基本性质 $\left\{\ell^{(\tau)}(\cdot)\right\}_{\tau \ge 1}$. □

A.2 “平凡”算法及其覆盖率保证示例

我们考虑以下不使用任何数据的“平凡”在线保形预测算法：在 $(1-\alpha)$ 的时间步中，有 D 的比例直接预测最大半径，然后在 α 的时间步中预测最小半径 0:

$$\begin{cases} \widehat{s}_t := D \quad \text{for} \quad t \in \{1, \ldots, \lfloor (1-\alpha)T \rfloor\} =: T_{\text{full}}, \\ \widehat{s}_t := 0 \quad \text{for} \quad t \in \{\lfloor (1-\alpha)T \rfloor + 1, \ldots, T\} =: T_{\text{empty}}. \end{cases} \tag{18}$$

根据我们的假设 $S_t \in [0, D]$ 几乎处处成立，以及 $C_t(X_t, \cdot)$ 的嵌套集结构，我们有 $\mathbf{err}_t = \mathbb{1}[Y_t \in C_t] = \mathbb{1}[\widehat{s}_t \ge S_t] = 0$对所有 $t \in T_{\text{full}}$均成立，同样地，$\mathbf{err}_t = 1$ 对所有 $t \in T_{\text{empty}}$均成立。因此，算法 (18) 直接满足

$$\mathrm{CovErr}(T) = \left|\frac{1}{T}\sum_{t=1}^{T} \mathrm{err}_t - \alpha\right| = \left|\frac{|T_{\text{empty}}|}{T} - \alpha\right| = \left|\frac{T - \lfloor (1-\alpha)T \rfloor}{T} - \alpha\right| \le \frac{1}{T}, \tag{19}$$

也就是说，该算法实现了约为 $(1-\alpha)$ 的经验覆盖率，误差为 $O(1/T)$。此外，很容易看出，只需稍微修改 T_{full}, T_{empty} 的定义（使两个索引集交替出现），就能使上述覆盖率界以任意时刻均成立的方式保持有效（对于 $t \in [T]$）。

然而，很容易构造出这样的数据分布，使得朴素算法（18）在分位数损失 $\ell^{(t)}$ （在式（1）中定义）上遭受线性遗憾，而且这些数据分布可以选择得相当简单。例如，假设所有数据点都具有相同的真实半径 $D/2$，也就是说，

$$S_t \equiv D/2 \quad \text{for all } t \in [T].$$

于是，对于 $s^\star = D/2$，我们有 $\ell^{(t)}(s^\star) = \ell_{1-\alpha}(D/2, D/2) = 0$ 对所有 $t \in [T]$均成立，其总损失为 $\sum_{t=1}^{T} \ell^{(t)}(s^\star) = 0$ （这是可能达到的最小值，因为 $\ell^{(t)}(\cdot) \ge 0$）。另一方面，算法 (18) 的损失为

$$\ell^{(t)}(\widehat{s}_t) = \ell_{1-\alpha}(S_t, \widehat{s}_t) = \begin{cases} \ell_{1-\alpha}(D/2, D) = \alpha(D/2) \quad \text{for} \quad t \in T_{\text{full}}, \\ \ell_{1-\alpha}(D/2, 0) = (1-\alpha)(D/2) \quad \text{for} \quad t \in T_{\text{empty}}. \end{cases}$$

因此，我们有

$$\mathrm{Reg}(T) = \sum_{t=1}^{T} \ell^{(t)}(\widehat{s}_t) - \inf_{s^\star} \ell^{(t)}(s_\star) = \sum_{t=1}^{T} \ell^{(t)}(\widehat{s}_t) = \alpha D/2 \cdot |T_{\text{full}}| + (1-\alpha)D/2 \cdot |T_{\text{empty}}|$$
$$= \alpha D/2 \cdot \lfloor (1-\alpha)T \rfloor + (1-\alpha)D/2 \cdot (T - \lfloor (1-\alpha)T \rfloor) \geq \alpha(1-\alpha)DT = \Omega(T),$$

即算法 (18) 遭受线性遗憾。这表明，次线性遗憾是排除类似 (18) 这样的平凡算法的合理标准。

B 第 4 节的证明

B.1 命题 4.1 的证明

将 SF-OGD 的遗憾界（命题 A.2）代入 Jun 等人(2017年，定理 2）即可得到证明。定义 $u(t) :=$ 最大$_n\{2^n : t \equiv 0 \text{ 模 } 2^n\}$。任取 $k \in [T]$ 和 $\tau \in [T-k+1]$。他们的证明首先将区间 $[\tau, \tau+k-1]$ 拆分为连续的子区间 $\bar{J}^{(1)}$、.....、$\bar{J}^{(n)}$，其中 $\bar{J}^{(i)} = [\tau_i$,最大$\{\tau+k, \tau_i + u(\tau_i)\} - 1]$ 是专家 $\mathcal{A}_{\tau_i}$的活动区间的一个前缀。

对于任意固定的 $s^\star \in \mathbb{R}$，我们有

$$\begin{aligned}
\mathrm{Regret}_\tau^k(s^\star) &:= \sum_{t=\tau}^{\tau+k-1} \ell^{(t)}(\widehat{s}_t) - \sum_{t=\tau}^{\tau+k-1} \ell^{(t)}(s^\star) \\
&= \sum_{i=1}^{n} \sum_{t \in \bar{J}^{(i)}} \left(\ell^{(t)}(\widehat{s}_t) - \ell^{(t)}(\widehat{s}_{t,\tau_i}) \right) + \sum_{i=1}^{n} \sum_{t \in \bar{J}^{(i)}} \left(\ell^{(t)}(\widehat{s}_{t,\tau_i}) - \ell^{(t)}(s^\star) \right) \\
&\leq D \sum_{i=1}^{n} \underbrace{\sqrt{|\bar{J}^{(i)}|(7\log T + 5)}}_{\text{Jun et al. (2017, Lemma 2)}} + D \sum_{i=1}^{n} \underbrace{\sqrt{|\bar{J}^{(i)}|(1+\sqrt{3})}}_{\text{Proposition A.2}} \\
&\leq D\left(\sqrt{7\log T + 5} + \sqrt{1+\sqrt{3}} \right) \underbrace{\sum_{j=0}^{\infty} \sqrt{k 2^{-j}}}_{\text{Jun et al. (2017, Lemma 3)}} \\
&= \frac{D\sqrt{2}}{\sqrt{2}-1} \sqrt{k} \left(\sqrt{7\log T + 5} + \sqrt{1+\sqrt{3}} \right) \leq 15D\sqrt{k(\log T + 1)}
\end{aligned}$$

对所有 $s^\star \in \mathbb{R}$ 和所有区间 $[\tau, \tau+k-1] \subset [T]$ 取上确界，可得 SAReg(T, k) 的所需界。

□

B.2 SAOCP 的动态遗憾

命题 B.1 (SAOCP 的动态遗憾界)。 算法 *1* 可达到如下动态遗憾界：对于任意长度为 $[\tau, \tau+k-1] \subset [T]$ 的区间 $k \in [T]$，有

$$\sum_{t=\tau}^{\tau+k-1} \ell^{(t)}(\widehat{s}_t) - \min_{s^\star_{\tau:\tau+k-1}} \sum_{t=\tau}^{\tau+k-1} \ell^{(t)}(s_t^\star) = \sum_{t=\tau}^{\tau+k-1} \left[\ell^{(t)}(\widehat{s}_t) - \ell^{(t)}(S_t) \right] \leq \widetilde{\mathcal{O}}\left(D\left[V_{[\tau,\tau+k-1]}^{1/3} k^{2/3} + \sqrt{k} \right] \right), \quad (20)$$

其中

$$V_{[\tau,\tau+k-1]} := \sum_{t=\tau+1}^{\tau+k-1} |S_t - S_{t-1}|$$

是 $[\tau, \tau+k-1]$ 内真实半径的 路径长度。

我们指出，动态遗憾[4] $\widetilde{\mathcal{O}}(V_{[\tau,\tau+k-1]}^{1/3}k^{2/3}+\sqrt{k})$ 是一般在线凸优化问题中极小极大最优的动态遗憾（Besbes 等人，2015）。

与 FACI 的动态遗憾比较 将 (20) 除以 k，可得 SAOCP 在 $[\tau,\tau+k-1]$ 上的如下平均动态遗憾界：

$$\widetilde{\mathcal{O}}\Big(D\Big[(V_{[\tau,\tau+k-1]}/k)^{1/3}+1/\sqrt{k}\Big]\Big),$$

同时对所有长度 k 和 $\tau\in[T-k+1]$均成立。

相比之下，采用学习率 η 的 FACI 算法（经调整以适用于我们的设定）可达到平均动态遗憾界（Gibbs 和 Cand`es，2022年，定理 3.2）

$$\widetilde{\mathcal{O}}\Big(D\Big[(V_{[\tau,\tau+k-1]}/k)^{1/2}+\eta/D+D/(\eta k)\Big]\Big).$$

当路径长度 $V_{[\tau,\tau+k-1]}=o(k)$ 时，FACI 对平均路径长度（$V_{[\tau,\tau+k-1]}/k=o(1)$ 的依赖更优，但由于无法同时为所有 k 选择最优的 η，其对 k本身的依赖更差；这与对比 SARegret 界（第 4.1 节）类似。

命题 B.1 的证明 我们将 Zhang 等人 的动态遗憾界 (2018，推论 5) 应用于区间 $[\tau,\tau+k-1]$ 上的 SAOCP 算法，并注意到根据引理 $\widehat{s}_t\in[-\eta,D+\eta]\subset[-D,2D]$A.1，我们的迭代点 以及算法 $\eta=D/\sqrt{3}$ 1 中的选择 。因此我们得到

$$\sum_{t=\tau}^{\tau+k-1}\ell^{(t)}(\widehat{s}_t)-\min_{s^\star_{\tau:\tau+k-1}}\sum_{t=\tau}^{\tau+k-1}\ell^{(t)}(s_t^\star)\le\widetilde{\mathcal{O}}\Big(D\Big[\widetilde{V}_{[\tau,\tau+k-1]}^{1/3}k^{2/3}+\sqrt{k}\Big]\Big),$$

其中

$$\widetilde{V}_{[\tau,\tau+k-1]}=\sum_{t=\tau+1}^{\tau+k-1}\sup_{s'\in[0,D]}\left|\ell^{(t)}(s')-\ell^{(t-1)}(s')\right|\overset{(i)}{\le}\sum_{t=\tau+1}^{\tau+k-1}|S_t-S_{t-1}|=V_{[\tau,\tau+k-1]},\tag{21}$$

其中 (i) 是由以下事实得出的：$\left|\ell^{(t)}(s')-\ell^{(t-1)}(s')\right|=|\ell_{1-\alpha}(s',\ S_t)-\ell_{1-\alpha}(s',\ S_{t-1})|\le|S_t-S_{t-1}|$ 根据分位数损失关于第二个参数的 1- 利普希茨连续性（1）。这就证明了所需结果。 □

B.3 定理4.2的证明

首先注意到，根据（2）和（10），算法2 可简化为以下更新式

$$\widehat{s}_{t+1}=\widehat{s}_t+\eta\frac{\mathrm{err}_t-\alpha}{\sqrt{\sum_{s=1}^t(\mathrm{err}_s-\alpha)^2}}=\widehat{s}_1+\eta\sum_{s=1}^t\frac{\mathrm{err}_s-\alpha}{\sqrt{\sum_{i=1}^s(\mathrm{err}_i-\alpha)^2}}.\tag{22}$$

注意到，对于所有 $\widehat{s}_{t+1}\in[-\eta,D+\eta]$，都有 $t\ge0$（引理A.1），这意味着 t

$$\left|\sum_{t=t_0+1}^{t_f}\frac{\mathrm{err}_t-\alpha}{\sqrt{\sum_{s=1}^t(\mathrm{err}_s-\alpha)^2}}\right|=\frac{1}{\eta}\left|\widehat{s}_{t_f+1}-\widehat{s}_{t_0+1}\right|\le\frac{D+2\eta}{\eta}\quad\text{for any }0\le t_0<t_f.$$

注意到 $|\mathrm{err}_t-\alpha|\in[\alpha,1]$ 对于所有 t。因此，我们可以在下面用B.2 引理，并取 $a_t=\mathrm{err}_t-\alpha$ 和$M=(D+2\eta)/\eta$，从而得到对于任意 $T\ge1$,

$$\left|\frac{1}{T}\sum_{t=1}^T\mathrm{err}_t-\alpha\right|\le2\left(\frac{D+3\eta}{\eta}+\alpha^{-2}\log T\right)T^{-1/4}\le\mathcal{O}(\alpha^{-2}T^{-1/4}\log T),$$

其中，后一个界对任意 $\eta=\Theta(D)$ 均成立。这就证明了定理 4.2。 □

[4]更准确地说，i ~中间结果，其中 $V_{[\tau,\tau+k-1]}$ 替换为标准的损失总变差 $V_{[\tau,\tau+k-1]}$ 在（21） .

引理 B.2。 假设序列 $\{a_t\}_{t\in[T]} \in \mathbb{R}$ 满足 $\alpha \le |a_t| \le 1$，其中某个 $\alpha > 0$，并且

$$\left| \sum_{t=t_0+1}^{t_f} \frac{a_t}{\sqrt{\sum_{s=1}^{t} a_s^2}} \right| \le M \quad \textit{for any} \;\; 0 \le t_0 < t_f \le T.$$

于是我们有

$$\left| \frac{1}{T} \sum_{t=1}^{T} a_t \right| \le 2\big(M + 1 + \alpha^{-2} \log T\big) T^{-1/4}.$$

证明。 证明基于分组论证。定义整数

$$L = \lceil T^{\beta} \rceil, \quad K = \lceil T/L \rceil \le T^{1-\beta} + 1,$$

wh其中 $\beta \in (0, 1)$ 是一个待选参数。对于任意 $k \in [K]$，定义第 k 组为 e

$$G_k = \{t_{k-1} + 1, \ldots, t_k\} := \{(k-1)L + 1, \ldots, \min\{kL, T\}\}, \tag{23}$$

使得对于所有 $k \in [K-1]$ 和 $|G_K| \le L$，我们都有 $\bigcup_{k=1}^{K} G_k = [T]$, $|G_k| = L$。

接下来，对于任意固定的 $k \ge 2$，定义求和式

$$S_k := \sum_{t\in G_k} \frac{a_t}{\sqrt{\sum_{s=1}^{t} a_s^2}}, \quad \widetilde{S}_k := \sum_{t\in G_k} \frac{a_t}{\sqrt{\sum_{s=1}^{t_{k-1}} a_s^2}}.$$

根据我们的前提条件，对于所有 $k \in [K]$，我们都有 $|S_k| \le M$。此外，我们有

$$\begin{aligned}
\left| S_k - \widetilde{S}_k \right| &\le \sum_{t\in G_k} |a_t| \cdot \left(\frac{1}{\sqrt{\sum_{s=1}^{t_{k-1}} a_s^2}} - \frac{1}{\sqrt{\sum_{s=1}^{t} a_s^2}} \right) \le |G_k| \cdot \left(\frac{1}{\sqrt{\sum_{s=1}^{t_{k-1}} a_s^2}} - \frac{1}{\sqrt{\sum_{s=1}^{t_k} a_s^2}} \right) \\
&\overset{(i)}{\le} L \cdot \frac{\sum_{s=t_{k-1}+1}^{t_k} a_s^2}{2\Big(\sum_{s=1}^{t_{k-1}} a_s^2\Big)^{3/2}} \overset{(ii)}{\le} L \cdot \frac{L}{2(\alpha^2(k-1)L) \cdot \Big(\sum_{s=1}^{t_{k-1}} a_s^2\Big)^{1/2}} = \frac{L}{2\alpha^2(k-1) \cdot \sqrt{\sum_{s=1}^{t_{k-1}} a_s^2}},
\end{aligned}$$

其中，(i) 对 $x, y \ge 0$ 使用不等式 $\frac{1}{\sqrt{x}} - \frac{1}{\sqrt{x+y}} \le \frac{y}{2x^{3/2}}$，而 (ii) 使用界 $\sum_{s=t_{k-1}+1}^{t_k} a_s^2 \le (t_k - t_{k-1}) \le L$ 和 $\sum_{s=1}^{t_{k-1}} a_s^2 \ge \alpha^2 t_{k-1} = \alpha^2(k-1)L$。根据三角不等式，可得

$$\left| \widetilde{S}_k \right| \le |S_k| + \left| \widetilde{S}_k - S_k \right| \le M + \frac{L}{2\alpha^2(k-1) \cdot \sqrt{\sum_{s=1}^{t_{k-1}} a_s^2}},$$

因此，对于任意满足 $k \ge 2$ 的

$$\begin{aligned}
\left| \sum_{t\in G_k} a_t \right| &= \underbrace{\left| \sum_{t\in G_k} \frac{a_t}{\sqrt{\sum_{s=1}^{t_{k-1}} a_s^2}} \right|}_{|\widetilde{S}_k|} \cdot \sqrt{\sum_{s=1}^{t_{k-1}} a_s^2} \le \left(M + \frac{L}{2\alpha^2(k-1) \cdot \sqrt{\sum_{s=1}^{t_{k-1}} a_s^2}} \right) \cdot \sqrt{\sum_{s=1}^{t_{k-1}} a_s^2} \\
&\le M \sqrt{\sum_{s=1}^{t_{k-1}} a_s^2} + \frac{L}{2\alpha^2(k-1)} \le M\sqrt{(k-1)L} + \frac{L}{2\alpha^2(k-1)}.
\end{aligned}$$

对于 $k = 1$，显然有 $\left|\sum_{t\in G_1} a_t\right| \le |G_1| \le L$。对 $k \in [K]$ 的界求和可得

$$\left| \sum_{t=1}^{T} a_t \right| \le L + \sum_{k=1}^{K} \left| \sum_{t\in G_k} a_t \right| \le L + M\sqrt{L} \cdot \sum_{k=2}^{K} \sqrt{k-1} + \frac{L}{2\alpha^2} \sum_{k=2}^{K} \frac{1}{k-1}$$

$$
\begin{aligned}
&\le L + \frac{2}{3}M\sqrt{L}K^{3/2} + \frac{L}{2\alpha^2}\log_2 K \\
&\le \lceil T^\beta \rceil + \frac{2}{3}M\sqrt{\lceil T^\beta \rceil}\cdot T^{3(1-\beta)/2} + \frac{1}{2\alpha^2}\lceil T^\beta \rceil \log_2(T^{1-\beta}) \\
&\le 2T^\beta + 2MT^{3/2-\beta} + \frac{2}{\alpha^2}T^\beta \log T.
\end{aligned}
$$

取 $\beta = 3/4$，可得

$$
\left|\sum_{t=t_0+1}^{t_f} a_t\right| \le 2\big(M + 1 + \log T/\alpha^2\big)T^{3/4}.
$$

两边同时除以 T，即可得到所需结果。 □

B.4 SAOCP 覆盖率

定理 B.3 (SAOCP 覆盖率界)。 考虑算法 *1* 的一个随机化版本，其中将第*7*行改为抽样选择一个专家 $i \sim p_{t,\cdot} \in \Delta([t])$ 并输出半径 $\widehat{s}_{t,\,i}$。考虑相应的期望误覆盖误差

$$
\widetilde{\mathrm{err}}_t := \sum_{i=1}^{t} p_{t,i}\underbrace{\mathbb{1}\left[\widehat{s}_{t,i} < S_t\right]}_{:=\mathrm{err}_{t,i}}. \tag{24}
$$

那么，对于任意 $T \ge 1$，我们有

$$
\left|\frac{1}{T}\sum_{t=1}^{T}\widetilde{\mathrm{err}}_t - \alpha\right| \le \mathcal{O}\left(\inf_{\beta\in(1/2,1)}\left\{T^{1/2-\beta} + T^{\beta-1}\times\underbrace{\left(1 + \sum_{j=2}^{\lceil T^{1-\beta}\rceil}\max_{t\in G_j}\sum_{i=1}^{t}\left|p_{t,i} - p_{t_{j-1},i}\frac{G^i_{i:t_{j-1}}}{G^i_{i:t}}\right|\right)}_{S_\beta(\{p_t\}_{t\in[T]},\{\sum_{\tau=i}^{t}\|\nabla\ell^{(\tau)}(\widehat{s}_{\tau,i})\|_2^2\}_{i\le t})=:S_\beta(T)}\right\}\right)
$$

(理解为 $p_{t_{j-1},\,i} := 0$ 对于任意 $i > t_{j-1}$)，其中对于每个 $\beta \in (1/2,\ 1)$，$\{G_j\}_{j=1}^{\lceil T^{1-\beta}\rceil}$ 满足 $|G_j| \le \lceil T^\beta \rceil$，
$G_j = \{t_{j-1}+1, \ldots,$ $\{t_j, T\}\}$ $[T]$ 23 min 是均匀分组，如 ()中所示，并且

$$
G^i_{i:t} := \sqrt{\sum_{\tau=i}^{t}\left\|\nabla\ell^{(\tau)}(\widehat{s}_{\tau,i})\right\|_2^2} = \sqrt{\sum_{\tau=i}^{t}(\mathrm{err}_{\tau,i} - \alpha)^2}
$$

是专家 $\mathcal{A}_i$ 对于任意 $t > i$ 所收到的累积平方梯度（将专家理解为即使变为非活跃状态，也会一直运行到时间 T）。

证明。 任取 $i \in [T]$。由于算法 1 将每个专家 $\mathcal{A}_i$ 选为 SF-OGD（算法 2），根据（10）可知，对所有 $t \ge i$,

$$
\widehat{s}_{t+1,i} - \widehat{s}_{t,i} = \frac{\eta}{G^i_{i:t}}\cdot(\mathrm{err}_{t,i} - \alpha). \tag{25}
$$

现在固定任意一个 $\beta \in (1/2,\ \mathbf{1})$。对于任意群 $2 \le j \le \lceil T^{1-\beta}\rceil$ 和 $t \in G_j$，将上述结果代入定义（24）可得

$$
\begin{aligned}
&\widetilde{\mathrm{err}}_t - \alpha = \sum_{i=1}^{t} p_{t,i}(\mathrm{err}_{t,i} - \alpha) = \frac{1}{\eta}\sum_{i=1}^{t} p_{t,i}G^i_{i:t}(\widehat{s}_{t+1,i} - \widehat{s}_{t,i}) \\
&= \frac{1}{\eta}\sum_{i=1}^{t_{j-1}} p_{t_{j-1},i}G^i_{i:t_{j-1}}(\widehat{s}_{t+1,i} - \widehat{s}_{t,i}) + \frac{1}{\eta}\sum_{i=1}^{t}\left(p_{t,i}G^i_{i:t} - p_{t_{j-1},i}G^i_{i:t_{j-1}}\right)(\widehat{s}_{t+1,i} - \widehat{s}_{t,i})
\end{aligned}
$$

$$= \frac{1}{\eta}\sum_{i=1}^{t_{j-1}} p_{t_{j-1},i} G^i_{i:t_{j-1}}(\widehat{s}_{t+1,i} - \widehat{s}_{t,i}) + \frac{1}{\eta}\sum_{i=1}^{t}\left(p_{t,i} - p_{t_{j-1},i}\frac{G^i_{i:t_{j-1}}}{G^i_{i:t}}\right)\cdot G^i_{i:t}(\widehat{s}_{t+1,i} - \widehat{s}_{t,i}).$$

对 $t \in G_j$ 求和，并注意到第一项求和中的系数 $p_{t_{j-1},\,i} G^i_{i:t_{j-1}}$ 不依赖于t，可得

$$\begin{aligned}
&\left|\sum_{t\in G_j}(\widetilde{\mathrm{err}}_t - \alpha)\right| \\
&\le \left|\frac{1}{\eta}\sum_{i=1}^{t_{j-1}} p_{t_{j-1},i} G^i_{i:t_{j-1}}(\widehat{s}_{t_j+1,i} - \widehat{s}_{t_{j-1}+1,i})\right| + |G_j|\cdot\max_{t\in G_j}\left|\frac{1}{\eta}\sum_{i=1}^{t}\left(p_{t,i} - p_{t_{j-1},i}\frac{G^i_{i:t_{j-1}}}{G^i_{i:t}}\right)\cdot G^i_{i:t}(\widehat{s}_{t+1,i} - \widehat{s}_{t,i})\right| \\
&\le \frac{1}{\eta}\max_{i\in[t_{j-1}]} G^i_{i:t_{j-1}}\left|\widehat{s}_{t_j+1,i} - \widehat{s}_{t_{j-1}+1,i}\right| + |G_j|\cdot\max_{t\in G_j}\sum_{i=1}^{t}\left|\frac{1}{\eta}\left(p_{t,i} - p_{t_{j-1},i}\frac{G^i_{i:t_{j-1}}}{G^i_{i:t}}\right)\cdot G^i_{i:t}(\widehat{s}_{t+1,i} - \widehat{s}_{t,i})\right| \\
&\overset{(i)}{\le} \frac{D+2\eta}{\eta}\sqrt{T} + |G_j|\cdot\max_{t\in G_j}\sum_{i=1}^{t}\left|p_{t,i} - p_{t_{j-1},i}\frac{G^i_{i:t_{j-1}}}{G^i_{i:t}}\right| \\
&\overset{(ii)}{\le} C\sqrt{T} + |G_j|\cdot\max_{t\in G_j}\sum_{i=1}^{t}\left|p_{t,i} - p_{t_{j-1},i}\frac{G^i_{i:t_{j-1}}}{G^i_{i:t}}\right|
\end{aligned}$$

上文中，(i) 根据 $G^i_{i:t_{j-1}}$ 的定义使用了 ~~$G^i_{i:t_{j-1}} \le \sqrt{t_{j-1}-i+1} \le \sqrt{T}$~~，利用每个专家都在 $[-\eta, D+\eta]$ 内初始化这一事实并应用引理 A.1 得到界 $\left|\widehat{s}_{t_j+1,i} - \widehat{s}_{t_{j-1}+1,i}\right| \le (D+2\eta)$，以及由 (25) 得到界 $\left|G^i_{i:t}(\widehat{s}_{t+1,i} - \widehat{s}_{t,i})\right| \le \eta$；(ii) 利用了算法 1 中的 $\eta = D/\sqrt{3}$ 这一事实，因此 $(D+2\eta)/\eta = 2+\sqrt{3} =: C$ 是一个绝对常数。另外，请注意，对于组 $j=1$，我们可以直接得到

$$\left|\sum_{t\in G_1}(\widetilde{\mathrm{err}}_t - \alpha)\right| \le |G_1|.$$

将以上所有界对 $j \in \left[\left\lceil T^{1-\beta}\right\rceil\right]$ 求和可得

$$\begin{aligned}
\left|\sum_{t=1}^{T}(\widetilde{\mathrm{err}}_t - \alpha)\right| &\le \sum_{j=1}^{\lceil T^{1-\beta}\rceil}\left|\sum_{t\in G_j}(\widetilde{\mathrm{err}}_t - \alpha)\right| \\
&\le \mathcal{O}\left(T^{3/2-\beta} + |G_1| + \sum_{j=2}^{\lceil T^{1-\beta}\rceil}|G_j|\times\max_{t\in G_j}\sum_{i=1}^{t}\left|p_{t,i} - p_{t_{j-1},i}\frac{G^i_{i:t_{j-1}}}{G^i_{i:t}}\right|\right) \\
&\le \mathcal{O}\left(T^{3/2-\beta} + T^{\beta}\left(1 + \sum_{j=2}^{\lceil T^{1-\beta}\rceil}\max_{t\in G_j}\sum_{i=1}^{t}\left|p_{t,i} - p_{t_{j-1},i}\frac{G^i_{i:t_{j-1}}}{G^i_{i:t}}\right|\right)\right)
\end{aligned}$$

将两边同时除以 T，即可得到该固定 β 对应的所需界。再对 $\beta \in (1/2,\ 1)$ 取上确界，即可得到所需结果。

□

B.4.1 讨论：&SF-OGD 作为特例

首先注意，定理 B.3 的证明并不依赖于专家权重 $\{p_{t,i}\}_{i<t}$ 或活动区间的具体结构。因此，除了算法 B.3 中指定的方案外，定理 1的结果普遍适用于任何其他针对具有任意活动区间的专家的聚合方案。

特别地，通过设置 $p_{t,\,1} = 1$ 和 $p_{t,\,i} = 0$ 对于 $i \ge 2$，并定义第一位专家 $\mathcal{A}_1$ 在以下范围内处于活跃状态$[T]$，算法 1 （无论是否采用随机化，因为只有一位活跃专家）都可恢复

算法 2。在这种情况下，我们证明对于任意 $S_\beta(T) \le \widetilde{\mathcal{O}}(\alpha^{-2})$，均有 $\beta \in (1/2, 1)$，因此定理 B.3（及其在定理 4.3中的非正式表述）确实将定理 4.2 作为特殊情形包含在内，只需选择 $\beta = 3/4$，正如第 4.2节所述。

我们有

$$S_\beta(T) = 1 + \sum_{j=2}^{\lceil T^{1-\beta}\rceil} \max_{t\in G_j} \sum_{i=1}^{t} \left| p_{t,i} - p_{t_{j-1},i} \frac{G^i_{i:t_{j-1}}}{G^i_{i:t}} \right| \overset{(i)}{=} 1 + \sum_{j=2}^{\lceil T^{1-\beta}\rceil} \max_{t\in G_j} \left| 1 - \frac{G^1_{1:t_{j-1}}}{G^1_{1:t}} \right|, \tag{26}$$

其中，(i) 使用了以下事实：对于 $i \ge 2$，$p_{t,1} = 1$ 且 $p_{t,i} = 0$。对于任意 $t \in G_j$，我们有

$$\left| 1 - \frac{G^1_{1:t_{j-1}}}{G^1_{1:t}} \right| = 1 - \frac{\sqrt{\sum_{s=1}^{t_{j-1}} (\mathrm{err}_s - \alpha)^2}}{\sqrt{\sum_{s=1}^{t} (\mathrm{err}_s - \alpha)^2}} \overset{(i)}{\le} \frac{\sum_{s=t_{j-1}+1}^{t} (\mathrm{err}_s - \alpha)^2}{2\sum_{s=1}^{t_{j-1}} (\mathrm{err}_s - \alpha)^2} \overset{(ii)}{\le} \frac{t - t_{j-1}}{2\alpha^2 t_{j-1}} \overset{(iii)}{\le} \frac{\lceil T^\beta \rceil}{2\alpha^2 \cdot (j-1) \lceil T^\beta \rceil} = \frac{1}{2\alpha^2 (j-1)},$$

其中，(i) 由不等式 $1 - \frac{\sqrt{x}}{\sqrt{x+y}} = \frac{\sqrt{x+y} - \sqrt{x}}{\sqrt{x+y}} \le \frac{\sqrt{x+y} - \sqrt{x}}{\sqrt{x}} = \sqrt{1 + \frac{y}{x}} - 1 \le \frac{y}{2x}$ 对任意 x，$y \ge 0$成立而得；(ii) 由界 $|\mathrm{err}_s - \alpha| \in [\alpha, 1]$ 对任意 s成立而得；(iii) 由分组的定义（23）而得。将上述界代入（26），我们得到

$$S_\beta(T) \le 1 + \sum_{j=2}^{\lceil T^{1-\beta}\rceil} \frac{1}{2\alpha^2 (j-1)} \le \mathcal{O}\big(\alpha^{-2} \log T\big) = \widetilde{\mathcal{O}}(\alpha^{-2}),$$

从而证明该结论。

C SAOCP 的分布感知覆盖率保证

在本节中，我们证明，在对真实半径作出温和的密度下界假设的条件下，对于长度为 k 的每个区间，SAOCP（算法 1）的覆盖误差的一个概率变体由 $\widetilde{\mathcal{O}}(k^{-1/(2q)}) + \widetilde{\mathcal{O}}((\mathrm{Var}_k/k)^{1/q})$ 所界，其中 $q \ge 2$ 是密度下界假设的一个参数，而 Var_k 衡量真实半径的 $1-\alpha$ 条件分位数中的某种方差（在长度为 k 的区间上）。该证明基于 SAOCP 的强自适应遗憾保证（就分位数损失而言；命题 4.1），并利用一种 自校准不等式类型的论证（Steinwart 和 Christmann，2011）通过超额分位数损失来界定参数估计误差。

设定 我们考虑第 2 节所述的在线保形预测设定。对于任意 $t \ge 1$，令$\mathcal{F}_t := \sigma(\{X_i, \widehat{s}_i, S_i\}_{i\in[t-1]}, X_t)$ 为由所有观测数据生成的 σ-代数 $\{(X_i, \widehat{s}_i, S_i)\}_{i\le t-1}$ 以及 X_t。注意，根据在线保形预测设定的定义，预测半径 $\widehat{s}_t$ 只能依赖于 $\mathcal{F}_t$ 中的信息，以及（可能存在的）外部随机性。因此，我们有 $S_t \perp\!\!\!\perp \widehat{s}_t \mid \mathcal{F}_t$，即$S_t$ 与 $\widehat{s}_t$ 在给定 $\mathcal{F}_t$的条件下相互独立。

现在我们陈述关于真实半径分布的假设。

假设 C.1 （密度上界）。对于所有 $t \in [T]$，存在常数 $L > 0$，使得 $S_t \mid \mathcal{F}_t$ 是一个取值范围在 $[0, D]$ 内且具有密度 $f_t : [0, D] \to \mathbb{R}_{\ge 0}$、满足$f_t(s) \le L/D$ 的连续随机变量，对所有 $s \in [0, D]$ 均成立。

假设 C.2 （密度下界）。对于所有 $t \in [T]$，$S_t \mid \mathcal{F}_t$ 是一个取值有界于 $[0, D]$ 且具有密度 $f_t : [0, D] \to \mathbb{R}_{\ge 0}$ 的连续随机变量。以概率 1，存在常数$b > 0, q \ge 2, \Delta_t > 0$ 使得

$$f_t(s) \ge \frac{2b}{D} \left| \frac{2(s - s_t^\star)}{D} \right|^{q-2} \tag{27}$$

对于所有 $s \in [s_t^\star - \Delta_t, s_t^\star + \Delta_t]$，其中

$$s_t^\star := Q_{1-\alpha}(S_t \mid \mathcal{F}_t) \tag{28}$$

是 S_t 的 $1-\alpha$ 条件分位数。

假设 C.2 的示例如下：当 $q = 2$ 时，这对应于条件密度 $f_t(\cdot)$ 在 $s_t^\star$附近局部存在常数下界；例如，若每个 $f_t(\cdot$ 本身在 $[0, D]$ 上都有常数下界（这是 Gibbs 和 Cand`es (2022年) 所作的假设）。较大的 q 会使密度下界（27）更容易满足，因此对应着更宽松的假设。我们还注意到， $s_t^\star$ 本身是一个在 $\mathcal{F}_t$ 上可测的随机变量。

对于任意区间 $I = [\tau, \tau + k - 1] \subseteq [T]$，我们的覆盖率结果取决于 $s_\tau^\star$、……、 $s_{\tau+k-1}^\star$ 之间的某种方差。具体而言，定义区间分位数变差

$$\mathrm{Var}_I := \sum_{t=\tau}^{\tau+k-1} \mathbb{E}\left[\left(\frac{s_t^\star}{D} - \frac{1}{kD}\sum_{i=\tau}^{\tau+k-1} \mathbb{E}[s_i^\star \mid \mathcal{F}_\tau]\right)^2\right]. \tag{29}$$

那么，SAOCP 的预测值 $\widehat{s}_\tau$、……、 $\widehat{s}_{\tau+k-1}$ 与真实条件分位数 $s_\tau^\star$、……、 $s_{\tau+k-1}^\star$ 之间的期望绝对差为 $\widetilde{\mathcal{O}}($ $k^{-1/(2q)}) + \widetilde{\mathcal{O}}(k^{-1/q}$方差$_I^{1/q}$)。由于累积分布函数的利普希茨连续性（根据假设 C.1），覆盖误差也具有类似形式。因此，区间分位数变差越小，SAOCP 的覆盖率越高；只要 $\mathrm{Var}_I \le o(|I|)$，它就能实现近似有效覆盖率。更正式地说，我们有：

定理 C.3。 设假设 *C.1* 与 *C.2* 成立。固定任意区间 $I = [\tau, \tau + k - 1] \subseteq [T]$。则，令$\Delta_I = \min\{\Delta_t : \ t \in I\}$，算法 *1* 实现如下分位数估计误差

$$\frac{1}{|I|}\sum_{t\in I} \mathbb{E}\left[\left|\frac{\widehat{s}_t - s_t^\star}{D}\right|\right] \le \mathcal{O}\left(\frac{1}{b^{1/q}}\frac{D}{\Delta_I}\left(\frac{\log T}{|I|}\right)^{1/(2q)}\right) + \mathcal{O}\left(\frac{L^{1/q}}{b^{1/q}}\frac{D}{\Delta_I}\left(\frac{\mathrm{Var}_I}{|I|}\right)^{1/q}\right)$$

以及区间漏覆盖误差

$$\frac{1}{|I|}\sum_{t\in I}\left|\mathbb{P}[Y_t \in \widehat{C}_t(X_t)] - (1-\alpha)\right| \le \mathcal{O}\left(\frac{L}{b^{1/q}}\frac{D}{\Delta_I}\left(\frac{\log T}{|I|}\right)^{1/(2q)}\right) + \mathcal{O}\left(\frac{L^{1+1/q}}{b^{1/q}}\frac{D}{\Delta_I}\left(\frac{\mathrm{Var}_I}{|I|}\right)^{1/q}\right).$$

在定理 C.3中， q 是一个用于量化以下难度的参数：将 $S_t \mid \mathcal{F}_t$ 的分布在远离其 $1-\alpha$ 条件分位数 $s_t^\star$的区域进行下界估计。如果 q更高，则与 $s_t^\star$的接近程度与分位数损失上的期望遗憾（1）的相关性更低。与此同时，随着 D/Δ_I 增大而 α 减小，该项会变得更大。该项的引入与定理 α^{-2} 4.2中 的引入相呼应，这表明更极端的分位数更难学习。

定理 C.3 的证明 关键在于技术性引理 C.4，该引理利用序列 $\widehat{s}_\tau$，……， $\widehat{s}_{\tau+k-1}$ 的期望动态遗憾，对该序列与 $1-\alpha$，……， S_τ的真实 $S_{\tau+k-1}$条件分位数之间的期望距离进行上界估计。我们将期望动态遗憾分解为算法 1的区间遗憾（可由命题 4.1给出上界）以及一项可使用 Var_I 进行上界估计的项。所需的覆盖率界由这些 S_t的累积分布函数的利普希茨连续性得出（这由密度上界假设 C.1蕴含）。在本证明中，我们使用 $\mathbb{E}_{\mathcal{F}_t}[X]$ 作为条件期望 $\mathbb{E}[X \mid \mathcal{F}_t]$的简称。

引理 C.4 （用动态遗憾界定分位数估计误差）。 任取区间 $I = [\tau, \tau+k-1] \subseteq [T]$，并令 $\Delta_I = \min\{\Delta_t : \ t \in I\}$。在假设 *C.2*下，我们有

$$\sum_{t\in I}\mathbb{E}\left[\left|\frac{\widehat{s}_t - s_t^\star}{D}\right|\right]^q \le \frac{2q(q-1)}{bD}\left(\frac{D}{2\Delta_I}\right)^q \sum_{t\in I}\mathbb{E}[\ell_{1-\alpha}(S_t, \widehat{s}_t) - \ell_{1-\alpha}(S_t, s_t^\star)].$$

为证明定理 C.3，我们采用与 Zhang 等人 (2018类似的技术，并将期望动态遗憾分解为

$$\mathbb{E}\left[\sum_{t\in I}\ell_{1-\alpha}(S_t, \widehat{s}_t) - \ell_{1-\alpha}(S_t, s_t^\star)\right] = \underbrace{\mathbb{E}\left[\sum_{t\in I}\ell_{1-\alpha}(S_t, \widehat{s}_t) - \inf_s \sum_{t\in I}\ell_{1-\alpha}(S_t, s)\right]}_{A} + \underbrace{\mathbb{E}\left[\inf_s \sum_{t\in I}\ell_{1-\alpha}(S_t, s) - \ell_{1-\alpha}(S_t, s_t^\star)\right]}_{B}.$$

我们首先注意到，项 A 只是 I上的期望区间遗憾。由于命题 4.1 以概率 1 界定了强自适应遗憾，因此我们可以界定 $A \le 15D\sqrt{|I|}(\log T+1$。现在，我们分析项 B，并注意到

$$\mathbb{E}\left[\inf_s \sum_{t\in I} \ell_{1-\alpha}(S_t,s)-\ell_{1-\alpha}(S_t,s_t^\star)\right] \le \inf_s \mathbb{E}\left[\sum_{t\in I}\mathbb{E}_{\mathcal{F}_t}[\ell_{1-\alpha}(S_t,s)-\ell_{1-\alpha}(S_t,s_t^\star))]\right]$$

根据 Jensen 不等式和条件期望的塔式性质。现在，对于任意 $t\in I$ 和 $s\in[0,D]$,

$$\begin{aligned}
&\mathbb{E}_{\mathcal{F}_t}[\ell_{1-\alpha}(S_t,s)-\ell_{1-\alpha}(S_t,s_t^\star)]\\
&=\mathbb{E}_{\mathcal{F}_t}[(1-\alpha)(S_t-s)\mathbb{1}[S_t>s]+\alpha(s-S_t)\mathbb{1}[S_t\le s]-(1-\alpha)(S_t-s_t^\star)\mathbb{1}[S_t>s_t^\star]-\alpha(s_t^\star-S_t)\mathbb{1}[S_t\le s_t^\star]]\\
&=\begin{cases}\mathbb{E}_{\mathcal{F}_t}[(S_t-s)\mathbb{1}[S_t>s]-\alpha(S_t-s)-\alpha(s_t^\star-S_t)] & S_t\le s_t^\star\\ \mathbb{E}_{\mathcal{F}_t}[(s-S_t)\mathbb{1}[S_t\le s]-(1-\alpha)(s-S_t)-(1-\alpha)(s_t^\star-S_t)] & S_t>s_t^\star\end{cases}\\
&=\begin{cases}\mathbb{E}_{\mathcal{F}_t}[(S_t-s)\mathbb{1}[s<S_t\le s_t^\star]+\alpha(s-s_t^\star)] & S_t\le s_t^\star\\ \mathbb{E}_{\mathcal{F}_t}[(s-S_t)\mathbb{1}[s_t^\star\le S_t\le s]-(1-\alpha)(s-s_t^\star)] & S_t>s_t^\star\end{cases}\\
&\overset{(i)}{=}\mathbb{E}_{\mathcal{F}_t}[(S_t-s)\mathbb{1}[s\le S_t\le s_t^\star]+(s-S_t)\mathbb{1}[s_t^\star\le S_t\le s]]+\alpha(s-s_t^\star)\mathbb{P}_{\mathcal{F}_t}[S_t\le s_t^\star]-(1-\alpha)(s-s_t^\star)\mathbb{P}_{\mathcal{F}_t}[S_t\le s_t^\star]\\
&\overset{(ii)}{=}\mathbb{E}_{\mathcal{F}_t}[(S_t-s)\mathbb{1}[s\le S_t\le s_t^\star]+(s-S_t)\mathbb{1}[s_t^\star\le S_t\le s]]\\
&=\left|\int_s^{s_t^\star}(x-s)f_t(x)\mathrm{d}x\right| \overset{(iii)}{\le} \frac{L(s_t^\star-s)^2}{2D}
\end{aligned}$$

上文中，(i) 利用了以下事实： $s_t^\star$ 是 $\mathcal{F}_t$-可测的，(ii) 利用了以下事实： $\mathbb{P}_{\mathcal{F}_t}[S_t\le s_t^\star]=1-\alpha$ 根据定义，以及 (iii) 利用了密度上界 $f_t(x)\le \frac{L}{D}$ (假设C.1)。因此，对$s\in\mathbb{R}$ 取下确界，并根据 Var 的定义$_I$ (29)，可得 $B\le \frac{LD}{2}\mathrm{Var}_I$，并且

$$\mathbb{E}\left[\sum_{t\in I}\ell_{1-\alpha}(S_t,\widehat{s}_t)-\ell_{1-\alpha}(S_t,s_t^\star)\right]\le 15D\sqrt{|I|(\log T+1)}+\frac{LD}{2}\mathrm{Var}_I.$$

我们将这一结果与幂均值不等式、引理C.4以及以下事实结合起来：$(q(q-1))^{1/q}=\mathcal{O}(1)$ 且 $(x+y)^{1/q}\le x^{1/q}+y^{1/q}$，以证明定理C.3的第一部分，

$$\frac{1}{|I|}\sum_{t\in I}\mathbb{E}\left[\left|\frac{\widehat{s}_t-s_t^\star}{D}\right|\right]\le\left(\frac{1}{|I|}\sum_{t\in I}\mathbb{E}\left[\left|\frac{\widehat{s}_t-s_t^\star}{D}\right|\right]^q\right)^{1/q}\le\mathcal{O}\left(\frac{D}{b^{1/q}\Delta_I}\left(\frac{\log T}{|I|}\right)^{1/(2q)}\right)+\mathcal{O}\left(\frac{D}{\Delta_I}\left(\frac{L\mathrm{Var}_I}{b|I|}\right)^{1/q}\right). \tag{30}$$

为证明所需的覆盖率界，我们注意到

$$\frac{1}{|I|}\sum_{t\in I}\left|\mathbb{P}[Y_t\in\widehat{C}_t(X_t)]-(1-\alpha)\right|=\frac{1}{|I|}\sum_{t\in I}|\mathbb{E}[\mathbb{E}_{\mathcal{F}_t}[\mathbb{1}[S_t\le\widehat{s}_t]-\mathbb{1}[S_t\le s_t^\star]]]|\le\frac{1}{|I|}\sum_{t\in I}\mathbb{E}[|F_t(\widehat{s}_t)-F_t(s_t^\star)|],$$

其中，最后一个不等式使用了 Jensen 不等式以及以下事实： $\mathbb{E}_{\mathcal{F}_t}[\mathbb{1}[S_t\le s]]=F_t(s)$ 是 $S_t\mid\mathcal{F}_t$的累积分布函数。结合式（30）与假设C.1即可得到该结果；该假设表明累积分布函数 F_t 是L/D-利普希茨的。 □

引理 C.4 的证明 我们首先考虑以下一般情形，其中 S 是一个有界于 $[0,D]$ 且密度为 f的连续随机变量。定义 $s^\star=Q_{1-\alpha}(S)$。令 g为归一化随机变量 $X=\frac{2S-D}{D}\in[-1,1]$的密度，并令 $x^\star=Q_{1-\alpha}(X)$。如 Steinwart 和 Christmann (2011，例 2.3) 中所述，假设存在常数 $b>0$、$q\ge 2$、$\Delta>0$ 使得

$$g(x)\ge b|x-x^\star|^{q-2}\iff f(s)\ge\frac{2b}{D}\left|\frac{2(s-s^\star)}{D}\right|^{q-2} \tag{31}$$

方法	LGBM（平均绝对误差 = 0.19）				自回归积分滑动平均模型（平均绝对误差 = 0.09）				Prophet（平均绝对误差 = 0.41）			
	覆盖率	宽度	局部条件误差$_k$	SAReg$_k$	覆盖率	宽度	局部条件误差$_k$	SAReg$_k$	覆盖率	宽度	局部条件误差$_k$	SAReg$_k$
SCP	0.706	0.288	0.443	0.040	0.911	0.203	0.257	0.017	0.555	0.453	0.571	0.058
NExCP	0.764	0.268	0.411	0.014	0.904	0.185	0.254	0.009	0.681	0.462	0.508	0.017
FACI	0.822	0.254	0.282	0.008	0.899	0.169	0.205	**0.007**	0.776	0.458	0.338	0.007
SF-OGD	0.872	0.262	0.208	0.011	0.901	0.170	0.191	0.010	0.871	0.475	0.209	0.009
FACI-S	0.863	**0.240**	0.207	0.010	0.891	**0.152**	0.197	0.010	0.856	**0.459**	0.197	**0.007**
SAOCP	0.872	0.248	**0.170**	**0.008**	0.886	0.155	**0.178**	**0.007**	0.868	0.473	**0.158**	**0.007**

表 3：M4 Weekly（359 个时间序列）上的结果，目标覆盖率为 $1-\alpha=0.9$，区间大小为 $k=20$。最佳结果以**粗体**显示，次佳结果加下划线，前提是该方法的全局覆盖率位于 (0.85,0.95)（绿色）内。对于所有基础预测器，SAOCP 均实现了最佳的局部覆盖误差、最佳的强自适应遗憾和第二佳的宽度。对于 LGBM 和 Prophet，唯一能使全局覆盖率位于 (0.85,0.95) 内的方法，是直接预测 $\widehat{s}_{t+1}$，而不是将其作为 $S_1,\dots,S_t$ 的分位数。

对所有 $s\in[s^\star-\Delta,s^\star+\Delta]$。令 $\beta=\frac{b}{q-1}$ 且 $\gamma=\beta(\frac{2\Delta}{D})^{q-1}$。根据 Steinwart 和 Christmann (2011，定理 2.7），

$$\begin{aligned}|x-x^\star| &\le 2^{1-1/q}q^{1/q}\gamma^{-1/q}(\mathbb{E}[\ell_{1-\alpha}(X,x)-\ell_{1-\alpha}(X,x^\star)])^{1/q}\\ &= 2\left(\frac{q}{2\gamma}\mathbb{E}[\ell_{1-\alpha}(X,x)-\ell_{1-\alpha}(X,x^\star)]\right)^{1/q}\\ &= 2\left(\frac{q(q-1)}{b}\left(\frac{D}{2\Delta}\right)^{q-1}\mathbb{E}[\ell_{1-\alpha}(X,x)-\ell_{1-\alpha}(X,x^\star)]\right)^{1/q}\end{aligned}$$

由于 $|x-x^\star|=\frac{2}{D}|s-s^\star|$, $\ell_{1-\alpha}(X,\ x)=\frac{2}{D}\ell_{1-\alpha}(S,\ s)$，我们可以得到

$$\left|\frac{s-s^\star}{D}\right|^q\le\frac{2q(q-1)}{bD}\left(\frac{D}{2\Delta}\right)^q\mathbb{E}[\ell_{1-\alpha}(S,s)-\ell_{1-\alpha}(S,s^\star)].$$

引理 C.4 现在只需固定任意一个 $t\in I$，定义 $\mathcal{G}_t=\sigma(\mathcal{F}_t,\ \widehat{s}_t)$，并注意到 $S_t\mid\mathcal{G}_t\overset{\text{dist}}{=}S_t\mid\mathcal{F}_t$（因为$\mathcal{G}_t$ 只涉及在线预测算法关于 $\mathcal{F}_t$ 的可能额外“外部”随机性），因此 $s_t^\star=Q_{1-\alpha}(S_t\mid\mathcal{F}_t)=Q_{1-\alpha}(S_t\mid\mathcal{G}_t)$。由于 $\widehat{s}_t$ 和 $s_t^\star$ 是 $\mathcal{G}_t$-可测的，我们可以界定

$$\mathbb{E}_{\mathcal{G}_t}\left[\left|\frac{\widehat{s}_t-s_t^\star}{D}\right|^q\right]\le\frac{2q(q-1)}{bD}\left(\frac{D}{2\Delta_I}\right)^q\mathbb{E}_{\mathcal{G}_t}[\ell_{1-\alpha}(S_t,\widehat{s}_t)-\ell_{1-\alpha}(S_t,s_t^\star)].$$

对两边取无条件期望，并注意到根据 Jensen 不等式，$\mathbb{E}[|s_t-s_t^\star|]^q\le\mathbb{E}[|s_t-s_t^\star|^q]$，再对所有 $t\in I$ 求和，即可得到该结果。□

D 更多时间序列实验

此处，我们在表3和表4 中分别报告了时间序列实验（如第 5.1节所述）在 M4 Weekly 和 NN5 Daily（两个较小的数据集）上的结果。M4 Weekly（表 3）上的结果与 M4 Daily（表 2）上的结果十分相似。所有方法在 NN5 上的表现都不错。考虑到即使是拆分保形预测也能达到很低的最坏情况局部覆盖误差，NN5 上的残差可能服从近似可交换分布（Barber 等人，2022年，定理 2a、3）。

方法	LGBM（平均绝对误差 = 0.08）				自回归积分滑动平均模型（平均绝对误差 = 0.07）				Prophet（平均绝对误差 = 0.11）			
	覆盖率	宽度	局部条件误差$_k$	SAReg$_k$	覆盖率	宽度	局部条件误差$_k$	SAReg$_k$	覆盖率	宽度	局部条件误差$_k$	SAReg$_k$
SCP	0.932	0.205	**0.123**	0.012	0.938	0.199	0.126	0.012	0.912	0.221	0.154	0.010
NExCP	0.922	0.187	0.133	0.011	0.922	0.175	0.136	0.012	0.908	0.208	0.146	0.010
FACI	0.910	**0.179**	0.130	**0.010**	0.906	**0.162**	0.132	**0.011**	0.900	**0.200**	0.131	**0.009**
SF-OGD	0.904	0.190	**0.123**	0.011	0.901	0.176	0.130	0.012	0.898	0.216	0.128	0.011
FACI-S	0.909	**0.179**	0.127	**0.010**	0.910	0.166	**0.123**	**0.011**	0.904	0.203	**0.125**	**0.009**
SAOCP	0.892	**0.179**	0.125	0.012	0.895	0.166	0.127	0.012	0.885	0.207	0.128	0.011

表4：NN5 Daily（111 条时间序列）的结果，目标覆盖率为 $1-\alpha = 0.9$，区间大小为 $k = 20$。只要方法的全局覆盖率处于 (0.85，0.95)（绿色），最佳结果以**粗体**显示，次佳结果加下划线。所有方法的表现都同样出色。

E 更多实验细节

我们在此提供所有方法的具体实现细节。我们用 $Q_{1-\alpha}(\cdot)$ 表示一组标量的（经验）第 $1-\alpha$ 分位数，其定义如下：

$$Q_{1-\alpha}\left(\{S_\tau\}_{\tau=1}^{t}\right) := \inf\left\{s \in \mathbb{R} : \frac{1}{t}\sum_{\tau=1}^{t} \mathbb{1}\left[S_\tau \le s\right] \ge 1-\alpha\right\}. \tag{32}$$

1. SCP：分割保形预测（Vovk 等人，2005年）预测 $\widehat{s}_{t+1} = Q_{1-\alpha}(\frac{1}{t}\sum_{\tau=1}^{t}\delta_{S_t})$。

2. NExCP：非交换保形预测通过使用加权分位数函数$\widehat{s}_{t+1} = Q_{1-\alpha}(\frac{1}{t}\sum_{\tau=1}^{t} w_\tau \delta_{S_t})$，扩展了 SCP。Barber 等人 (2022年) 建议使用几何衰减权重，使 NExCP 能适应分布偏移的情况，因此我们使用 $w_t = (1-3\alpha/4)^{1-t} w_1$。

3. FACI：完全自适应保形推断有 4 个超参数：各个专家的学习率 $\gamma_1, \ldots, \gamma_N$；目标区间长度 k；元算法学习率 η；以及平滑参数 σ。我们设定 $k = 100$，并遵循Gibbs 和 Cand`es (2022年) 来设定 $N = 8$，$\sigma = \frac{1}{2k}$，$\gamma = \{0.001, 0.002, 0.004, 0.008, 0.016, 0.032, 0.064, 0.128\}$，以及

$$\eta_t = \sqrt{\frac{\log(Nk)+2}{\sum_{\tau=t-k}^{t-1}\mathbb{E}[\ell_\alpha(\beta_t,\alpha_t)^2]}},$$

其中期望是对 α_t 求取的。我们还针对时间序列实验尝试了 $k = 20$（以匹配我们的评估指标），但结果更差。

4. SF-OGD：无尺度在线梯度下降。唯一的超参数是最大半径 D。对于时间序列实验（第 5.1 节，附录 F），我们将每个预测步长 $D/\sqrt{3}$ 的 h 设置为训练数据校准划分中观测到的最大 h 步残差。对于 m-分类图像分类实验（第 5.2 节，附录 G），我们设置 $D = 1 + \lambda\sqrt{m - k_{reg}}$，~~其中 λ~~ 和 k_{reg} 是式（15）中的宽度正则化参数。

5. FACI-S：将 FACI 应用于 S_t，而不是 α_t。超参数相同，但用于计算 η_t 的损失为 $\ell_{1-\alpha}(S_t, \widehat{s}_t)$，且学习率乘以 D。我们以与 SF-OGD 相同的方式设置 D。

6. SAOCP：有 2 个超参数：最大半径 D 和生命周期乘数 g，见（8）。我们按照与 SF-OGD 相同的方式设置 D。对于时间序列实验，我们设置 $g = 8$；对于图像分类实验，我们设置 $g = 32$。

F 时间序列实验与集成模型

在本节中，我们复现第5.1 节中的实验，使用通过 EnbPI 方法（Xu 和 Xie，2021训练的集成模型。具体来说，我们在（$\widehat{f}^{(b)}, X_t, Y_T$）$_{t\in I_b}$上训练基学习器，其中 I_b 是采样得到的

	LGBM（平均绝对误差 = 0.05）				自回归积分滑动平均模型（平均绝对误差 = 0.14）				Prophet（平均绝对误差 = 0.08）			
方法	覆盖率	宽度	局部条件误差$_k$	SAReg$_k$	覆盖率	宽度	局部条件误差$_k$	SAReg$_k$	覆盖率	宽度	局部条件误差$_k$	SAReg$_k$
EnbPI	0.803	0.088	0.331	0.017	0.916	0.220	0.186	0.032	0.834	0.141	0.299	0.018
EnbNEx	0.864	0.097	0.218	0.010	0.907	0.198	0.193	0.027	0.892	0.151	0.195	0.010
EnbFACI	0.856	**0.081**	0.200	0.007	0.900	**0.181**	0.161	**0.021**	0.884	**0.130**	0.164	0.007
EnbSF-OGD	0.871	0.098	0.173	0.008	0.906	0.201	0.165	0.027	0.898	0.145	0.144	0.008
EnbSAOCP	0.884	0.091	**0.134**	**0.005**	0.893	0.192	**0.150**	0.053	0.888	**0.130**	**0.130**	**0.005**

表 5：M4 小时数据集上的集成结果，目标覆盖率为 $1-\alpha=0.9$，区间大小为 $k=20$。只要方法的全局覆盖率处于 (0.85,0.95)（绿色），最佳结果以**粗体**显示，次佳结果以下划线标出。对于所有模型，EnbSAOCP 的局部覆盖误差和强自适应遗憾均为最佳，但 ARIMA 除外；在 ARIMA 上，其强自适应遗憾略高。

	LGBM（平均绝对误差 = 0.11）				自回归积分滑动平均模型（平均绝对误差 = 0.10）				Prophet（平均绝对误差 = 0.18）			
方法	覆盖率	宽度	局部条件误差$_k$	SAReg$_k$	覆盖率	宽度	局部条件误差$_k$	SAReg$_k$	覆盖率	宽度	局部条件误差$_k$	SAReg$_k$
EnbPI	0.512	0.093	0.716	0.077	0.894	0.121	0.289	0.056	0.791	0.288	0.374	0.062
EnbNEx	0.749	0.142	0.520	0.020	0.894	0.116	0.293	0.035	0.847	0.343	0.367	0.029
EnbFACI	0.776	0.130	0.392	0.015	0.887	0.100	0.248	**0.015**	0.853	**0.232**	0.230	0.025
EnbSF-OGD	0.798	0.143	0.396	0.021	0.898	0.106	0.241	0.047	0.900	0.292	0.195	0.023
EnbSAOCP	0.875	**0.138**	**0.203**	**0.007**	0.908	**0.096**	**0.187**	0.034	0.917	0.233	**0.139**	**0.011**

表 6：在 M4 Daily 上使用集成，目标覆盖率为 $1-\alpha=0.9$，区间大小为 $k=20$。最佳结果以**粗体**标示，次佳结果加下划线，前提是方法的全局覆盖率位于 (0.85,0.95)（绿色）。对于所有模型，EnbSAOCP 的宽度、局部覆盖误差和强自适应遗憾均排名第一或第二。它也是唯一能为 LGBM 实现有效覆盖率的方法。

从 $[T]$ 中随机抽取。然后，我们通过聚合所有未在 (X_t,Y_t) 上训练的模型，得到残差 $S_t^y=|y-\phi(\widehat{f}^{(b)}(X_t):I_b\not\ni t)|$。最后，新观测 (X_{T+1},Y_{T+1}) 的残差为 $|Y_{T+1}-\phi(\widehat{f}_{(b)}(X_{T+1}):b\in[B])|$。我们在集成中使用 $B=5$ 个模型。

EnbPI 将半径预测为 $\widehat{s}_{t+1}$，即 $1-\alpha$ 此前观测到的残差的经验分位数，与拆分保形预测中的做法相同。不过，这些预测集也可以通过分数的任意函数获得，包括 NExCP、FACI、SF-OGD 或 SAOCP。除 EnbPI 外，我们分别将这些混合方法称为 EnbNEx、EnbFACI、EnbSF-OGD 和 EnbSAOCP。我们使用附录 E 中所述的相同超参数。

这种方法使其他方法与 EnbPI 处于更加公平的比较基础上，因为 EnbPI 不再需要训练/校准划分，并将底层模型从单个学习器改为更准确的集成。我们认为 EnbPI 的贡献与我们的工作是正交的，本节表明，他们的方法可以成功地与我们的方法结合。

结果与第 5.1 节的结果相似。在所有 M4 数据集上，EnbSAOCP 通常都能取得最佳或次佳的区间宽度、最坏情况局部覆盖误差和强自适应遗憾（表 5、6、7）。在 NN5（表 8）上，所有方法取得的强自适应遗憾都相近。不过，EnbSF-OGD 和 EnbSAOCP 分别取得了最佳和次佳的最坏情况局部覆盖误差，但代价是区间略宽。在所有情况下，EnbSAOCP 的区间都比 EnbSF-OGD 更窄。

G TinyImageNet/TinyImageNet-C 上的图像分类

我们使用 TinyImageNet（5.2）基础数据集及其损坏版本 TinyImageNet-C（Le 和 Yang，2019年）上的 ResNet-50 分类器，复现第 Hendrycks 和 Dietterich，2019年）节中的实验。我们使用随机梯度下降训练模型，学习率为 0.1（每 7 个 epoch 衰减 10 倍）、动量为 0.9、批量大小为 256；如果验证准确率连续 10 个 epoch 没有提升，则提前停止。模型最终取得了 52.8% 的测试准确率。图 1 展示了结果。

在执行未　　不确定性量化方面，我们使用带有宽度正则化的保形分数（15）　　n

	LGBM（平均绝对误差 = 0.12）				自回归积分滑动平均模型（平均绝对误差 = 0.07）				Prophet（平均绝对误差 = 0.16）			
方法	覆盖率	宽度	局部条件误差$_k$	SAReg$_k$	覆盖率	宽度	局部条件误差$_k$	SAReg$_k$	覆盖率	宽度	局部条件误差$_k$	SAReg$_k$
EnbPI	0.540	0.118	0.596	0.079	0.893	0.174	0.266	0.021	0.785	0.288	0.369	0.048
EnbNEx	0.720	0.168	0.483	0.015	0.910	0.161	0.252	0.012	0.875	0.375	0.305	0.027
EnbFACI	0.784	0.159	0.336	0.010	0.905	0.142	0.193	0.009	0.881	0.235	0.201	0.017
EnbSF-OGD	0.795	0.172	0.329	0.016	0.908	0.152	0.184	0.012	0.915	0.304	0.164	0.023
EnbSAOCP	0.874	**0.165**	**0.173**	**0.005**	0.909	**0.131**	**0.151**	**0.008**	0.907	**0.232**	**0.133**	**0.012**

表7：在 M4 Weekly 上的集成结果，目标覆盖率为 $1-\alpha = 0.9$，区间大小为 $k = 20$。在方法的全局覆盖率位于 (0.85,0.95)（绿色）时，最佳结果以**粗体**标示，次佳结果以下划线标示。EnbSAOCP 在所有模型上均取得了最佳宽度、局部覆盖误差和强自适应遗憾。它也是唯一在 LGBM 上实现有效覆盖率的方法。

	LGBM（平均绝对误差 = 0.09）				自回归积分滑动平均模型（平均绝对误差 = 0.07）				Prophet（平均绝对误差 = 0.07）			
方法	覆盖率	宽度	局部条件误差$_k$	SAReg$_k$	覆盖率	宽度	局部条件误差$_k$	SAReg$_k$	覆盖率	宽度	局部条件误差$_k$	SAReg$_k$
EnbPI	0.859	**0.164**	0.225	0.012	0.910	0.163	0.159	0.010	0.904	0.163	0.160	**0.010**
EnbNEx	0.882	0.177	0.177	0.010	0.920	0.174	0.138	0.012	0.923	0.178	0.126	0.012
EnbFACI	0.883	0.177	0.166	**0.009**	0.905	**0.156**	0.131	**0.009**	0.907	**0.156**	0.123	**0.010**
EnbSF-OGD	0.895	0.194	0.132	0.010	0.912	0.176	**0.120**	0.012	0.912	0.178	0.120	0.013
EnbSAOCP	0.886	0.188	**0.131**	0.012	0.906	0.168	0.122	0.012	0.905	0.170	**0.118**	0.013

表 8：NN5 Daily 上集成方法的结果，目标覆盖率为 $1-\alpha = 0.9$，区间大小为 $k = 20$。在方法的全局覆盖率处于 (0.85，0.95)（绿色）时，最佳结果以**粗体**显示，次佳结果带下划线。EnbSAOCP 和 EnbSF-OGD 分别以略宽的区间为代价，取得了最好的和第二好的最坏情况局部覆盖误差误差。

参数 $\lambda = 0.01$ 和 $k_{reg} = 20$。如第5.2节所述，SAOCP 和 SF-OGD 的局部覆盖率仍最接近目标值 0.9。分布偏移突然发生时，差异最为明显，这表明与其他方法相比，它们能够更快地适应这些分布偏移。虽然所有方法的预测集大小相近，但与 SAOCP 和 SF-OGD 相比，NExCP 和 FACI 适应最佳固定预测集大小的速度较慢。SAOCP 的覆盖率也优于 SF-OGD。

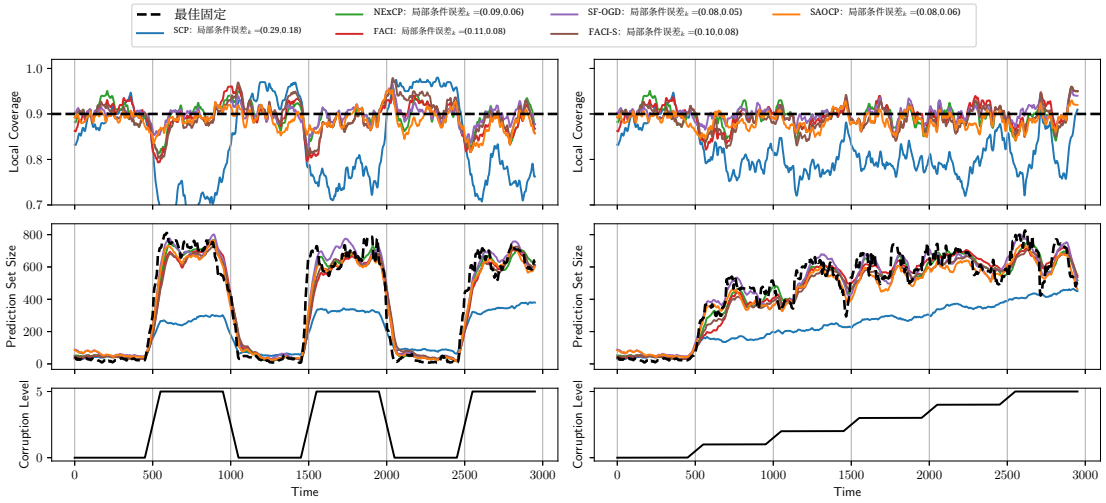

图2：当分布每隔 500 步在 ImageNet 和 ImageNet-C 之间发生偏移时，各种不确定性量化方法所达到的局部覆盖率（上排）和预测集大小（第二排）。我们绘制了窗口大小为 $k = 100$ 的移动平均。左：损坏等级在 0 和 5 之间发生突变偏移。右：损坏等级从 0 逐渐变化到 5。SAOCP 和 SF-OGD 的局部覆盖率最接近 0.9 这一目标值，尤其是在变点处。尽管这两种方法达到的局部覆盖率相近，但 SAOCP 返回的预测集比 SF-OGD 更小。