Both Jackknife and bootstrap are generic methods that can be used to reduce the bias of statistical estimators. However, the traditional theory proves incapable of answering whether the bootstrap or jackknife can help reduce the bias of the empirical entropy so as to achieve the minimax rates of entropy estimation. We provide the answers in this lecture using advanced theory of approximation.

To simplify the presentation we focus on the binomial model \( n \hat{p}_n \sim B(n,p) \) throughout this lecture. However, we emphasize that the methodology applies to any statistical models of interest.

## 1 Jackknife intuition

Let \( e_1,n(p) = f(p) - \mathbb{E}_p f(\hat{p}_n), \ f \in C([0,1]) \) be the bias term. Construct the jackknife bias corrected estimator as

\[
\hat{f}_2 = nf(\hat{p}_n) - (n-1)f(\hat{p}_{n-1})
\]

We now heuristically claim that this approach can reduce bias. Suppose that

\[
e_1,n(p) = \frac{a(p)}{n} + \frac{b(p)}{n^2} + O\left(\frac{1}{n^3}\right),\tag{1}
\]

where \( a(p), b(p) \) are unknown functions of \( p \) which do no depend on \( n \). We also have

\[
e_1,n-1(p) = \frac{a(p)}{n-1} + \frac{b(p)}{(n-1)^2} + O\left(\frac{1}{(n-1)^3}\right).\tag{2}
\]

Hence, the overall bias of \( \hat{f}_2 \) is:

\[
f(p) - \mathbb{E}_p \hat{f}_2 = ne_1,n(p) - (n-1)e_1,n-1(p)
\]

\[
= \frac{b(p)}{n} - \frac{b(p)}{n-1} + O\left(\frac{1}{n^2}\right) \tag{3}
\]

\[
= -\frac{b(p)}{n(n-1)} + O\left(\frac{1}{n^2}\right), \tag{4}
\]

which shows that the bias has been reduced to order \( \frac{1}{n^2} \) instead of order \( \frac{1}{n} \).

**Note:** although the result seems solid, it does not capture the dependence in \( p \). For example, if \( f(p) = p \log \frac{1}{p} \), standard Taylor expansion shows that the function \( b(p) \) and the final \( O\left(\frac{1}{n^2}\right) \) term contain functions of \( p \) that explode to infinity as \( p \to 0 \). Traditional statistical theory fails to analyze the uniform bias reduction properties of the jackknife for functions such as \( f(p) = p \log \left(\frac{1}{p}\right) \). This task can be done via the advanced theory of approximation.

## 2 Approximation-theoretic analysis of Jackknife

**Recap:**

- \( r \)-th order symmetric difference: \( \Delta^r_h f(x) = \sum_{k=0}^{r} (-1)^k \binom{r}{k} f(x + (r-k)h) \)
- \( \omega^r_{\varphi}(f,t) = \sup_{0 \leq h \leq r} \|\Delta^r_{\varphi} f(x)\|, \varphi(x) = \sqrt{x(1-x)} \)
• \( f(p) = p \log \frac{1}{p} \Rightarrow \omega_r^\varphi(f, t) \asymp t^2 \) for fixed \( r \geq 2 \)

Let \( n_1 < n_2, \ldots < n_r = kn_1 = n \), where \( k \) is fixed and \( n_i \in \mathbb{N} \) and choose \( c_1, \ldots, c_r \) such that \( \sum_{i=1}^{r} c_i = 1 \) and
\[
\text{for } \rho \in \mathbb{Z}, 1 \leq \rho \leq r - 1 : \sum_{i=1}^{r} \frac{c_i}{n_i^\rho} = 0
\]

Then the estimator would be:
\[
\hat{f}_r = \sum_{i=1}^{r} c_i \hat{f}(\hat{p}_{n_i})
\]

Note that this framework incorporates the one described in section 1 with \( n_1 = n - 1, n_2 = n \), with \( c_1 = -(n-1), c_2 = n \).

**Lemma 1.** The solutions of \( c_i, 1 \leq i \leq r \) are given by
\[
c_i = \prod_{j \neq i} \frac{n_i}{n_i - n_j}, 1 \leq i \leq r.
\]

**Theorem 2.** Suppose that \( \sum_{i=1}^{r} |c_i| \leq C \), where \( C \) is a constant independent of \( n \). Fix \( r > 0 \). Then:

1. \( \| f - \mathbb{E} \hat{f}_r \|_{\infty} \leq C(\omega_r^\varphi(f, \frac{1}{\sqrt{r}}) + n^{-r} ||f||) \)

2. for fixed \( \alpha \in (0, 2r) \):
\[
\| f - \mathbb{E} \hat{f}_r \| = O(n^{-\alpha/2}) \Leftrightarrow \omega_r^\varphi(f, t) = O(t^\alpha)
\]

3. if \( f(p) = p \log \frac{1}{p} \), then \( \| f - \mathbb{E} \hat{f}_2 \| \asymp \frac{1}{n} \)

There are several interesting interpretations of this results:

1. Iterating the jackknife one more time is equivalent to raising the modulus order by 2

2. Raising the modulus order does not reduce the error for functions such as \( f(p) = p \log \frac{1}{p} \), hence doing jackknife does not improve the worst case bias of empirical entropy in terms of order

3. The traditional jackknife, which has \( n_1 = n - 1, n_2 = n \), does not satisfy the conditions of this theorem.

We now show that the reason why the traditional jackknife is not included in the previous theorem is that it does not satisfy the nice properties enlisted in it.

**Example 3.** Denote the \( r \)-th order jackknife with \( n_r = r, n_j - n_{j-1} = 1 \) as \( \hat{f}_r \). Then there exists a fixed function \( f \in C((0, 1]) \) and \( ||f|| \leq 1 \), such that \( \| \mathbb{E} \hat{f}_r - f \| \geq n^{r-1} \). Also, there exists \( f \in C([0; 1]) \) such that \( \text{Var}(\hat{f}_2) \geq \frac{e^2}{n} \).

This result shows that the traditional jackknife may have very bad bias and variance properties if the function is not very smooth. However, this does not imply that the traditional jackknife does not work for functions like \( f(p) = p \log \frac{1}{p} \). Since the general theorem does not apply to this situation, one needs to carefully compute the bias of the jackknife estimator for entropy. It proves to be a highly challenging task. We have the following conjecture:

**Conjecture 4.** Consider \( f(p) = p \log \frac{1}{p} \). We conjecture that \( \| f - \mathbb{E} \hat{f}_r \| \asymp \frac{1}{n} \) for any fixed \( r \geq 2 \). It is proved to be true when \( r = 2 \).
3 Bootstrap

The bootstrap relies on two principles: the plug-in principle and the Monte–Carlo principle.

1. Plug-in principle: if \( \hat{P} \) is a good estimate of \( P \) then \( F(\hat{P}) \) is a good estimate of \( F(P) \).

2. Monte–Carlo principle: in order to compute the plug-in estimator \( F(\hat{\theta}) \) of \( F(\theta) = \int g(X)dP_\theta \), where \( X \sim P_\theta \), we sample \( B \) random objects \( X_1, X_2, \ldots, X_B \) \( \overset{i.i.d.}{\sim} P_\hat{\theta} \), and then use

\[
\frac{1}{B} \sum_{i=1}^{B} g(X_i)
\]

(7)

to approximately compute \( F(\hat{\theta}) \).

Now we investigate the bias correction problem. Since \( e_1(p) = f(p) - E_p \hat{f}(\hat{p}_n) \), the plug-in principle suggests that \( e_1(\hat{p}_n) \approx e_1(p) \), which further suggests the bias corrected estimator

\[
\hat{f}_2 = f(\hat{p}_n) + e_1(\hat{p}_n).
\]

(8)

Note that one in general needs to use the Monte–Carlo principle to approximately compute \( e_1(\hat{p}_n) \). It is also clear that one can do this correction infinitely many times. Indeed, we have

\[
e_2(p) = e_1(p) - E e_1(\hat{p}_n)
\]

\[
\hat{f}_r = f(\hat{p}_n) + \sum_{i=1}^{r-1} e_i(\hat{p}_n)
\]

It begs the question: how can we understand the bias correction properties of this bootstrap approach?

1. Question 1: Does \( \hat{f}_r \) converge as \( r \to \infty \)? The answer is YES!

2. Question 2: What does it converge to?

\[\text{Lemma 5. } \lim_{r \to \infty} \|e_r(p) - (f(p) - L_n[f](p))\| = 0 \text{ for any } f \in C([0;1]), \text{ where } L_n[f](p) \text{ is the unique Lagrange interpolating polynomial of } f \text{ at } n+1 \text{ points } \{0, 1/n, 2/n, \ldots, 1\}.\]

3. Question 3: does \( L_n[f](p) \) have good approximation properties?

\[\text{Example 6 (Bernstein). } \text{For } f(p) = |p - 1/2|, \lim_{r \to \infty} |f(p) - L_n[f](p)| = \infty \text{ for any } p \text{ except } p = 0, p = 1/2.\]

These results show that iterating the bootstrap bias correction too many times may not be a wise idea. What about doing it only a few times? We now show that it essentially has the similar performance as doing the jackknife exactly the same number of times.

\[\text{Theorem 7. } \text{Fix } r > 0. \text{ Then} \]

1. \( \|f(p) - E_p \hat{f}_r\|_\infty \leq C(\omega^{2r}(f, \frac{1}{\sqrt{n}}) + n^{-r}\|f\|) \)

2. for fixed \( \alpha \in (0, 2r] \):

\( \|f(p) - E_p \hat{f}_r\| = O(n^{-\alpha/2}) \Leftrightarrow \omega^{2r}(f, t) = O(t^\alpha) \)

3. if \( f(p) = p \log \frac{1}{p} \), then \( \|f - E \hat{f}_r\| \asymp \frac{1}{n} \)