A. Proofs

A.1. Proof of Proposition 1

Define the random variables $u(X) = |\mathbb{E}[Y \mid h(X)] - h(X)|^p$ and $v(X) = 1$. Then, by Hölder's inequality for $r = q/p$ and $s = (1 - 1/r)^{-1}$,

$$\begin{aligned}
(\ell_p\text{-ECE}(h))^p &= \mathbb{E}[u(X)] \\
&= \mathbb{E}[|u(X)v(X)|] \\
&\leqslant \mathbb{E}[|u(X)|^r]^{1/r}\,\mathbb{E}[|v(X)|^s]^{1/s} \\
&= \mathbb{E}[|u(X)|^r]^{1/r} \\
&= \mathbb{E}[|\mathbb{E}[Y \mid h(X)] - h(X)|^q]^{p/q} \\
&= (\ell_q\text{-ECE}(h))^p,
\end{aligned}$$

which proves (5). If h satisfies (3), then $u(X) \leqslant \varepsilon^p$ a.s. Thus $\ell_p\text{-ECE}(h) = \mathbb{E}[u(X)]^{1/p} \leqslant \varepsilon$. □

A.2. Proof of Lemma 1

Let F denote the cdf corresponding to f. The structure of the proof is as follows:

- We first compute the conditional density of the order statistics $S_{(l+1)}, S_{(l+2)}, \ldots, S_{(u-1)}$, given $S_{(l)}$ and $S_{(u)}$, in terms of f and F (the expression for this is (15)). The basic building block for this computation is a result on the conditional density of order statistics given a single order statistic (equation (12)).
- Next, we compute the conditional density of the order statistics of the independent random variables $\{S_i'\}_{i\in[u-l-1]}$, given $S_{(l)}$, $S_{(u)}$, and $S_{(l)} < S_i' < S_{(u)}$ for all $i \in [u-l-1]$ (the expression for this is (16)).
- We verify that (15) and (16) are identical, which shows that the conditional density of the order statistics matches. Finally, we conclude that the unordered random variables must themselves have the same conditional density. This completes the argument.

Let $0 \leqslant s_1 < \ldots < s_{l-1} < a < s_{l+1} < \ldots < s_n \leqslant 1$. The conditional density of all the order statistics given $S_{(l)}$

$$f(S_{(1)} = s_1, S_{(2)} = s_2, \ldots, S_{(l-1)} = s_{l-1}, S_{(l+1)} = s_{l+1}, \ldots, S_{(n)} = s_n \mid S_{(l)} = a)$$

is given by

$$\left((l-1)!\,\Pi_{i=1}^{l-1}\frac{f(s_i)}{F(a)}\right)\cdot\left((n-l)!\,\Pi_{i=l}^{n}\frac{f(s_i)}{1-F(a)}\right).$$

For one derivation, see Ahsanullah et al. (2013, Chapter 5, equation (5.2)). This implies that the order statistics larger than $S_{(l)}$ are independent of the order statistics smaller than $S_{(l)}$ given $S_{(l)}$, and

$$f(S_{(l+1)} = s_{l+1}, \ldots, S_{(n)} = s_n) \mid S_{(l)} = a) = \left((n-l)!\,\Pi_{i=l+1}^{n}\frac{f(s_i)}{1-F(a)}\right). \tag{12}$$

Suppose we draw $n-l$ independent samples $T_1, T_2, \ldots, T_{n-l}$ from the distribution whose density is given by

$$g(s) = \begin{cases} \frac{f(s)}{1-F(a)} & \text{if } s \in [a, 1], \\ 0 & \text{otherwise.} \end{cases}$$

(This is the conditional density of S given $S > S_{(l)} = a$ where S is an independent random variable distributed as Q_S.) Consider the order statistics $T_{(1)}, T_{(2)}, \ldots, T_{(n-l)}$ of these $n-l$ samples. It is a standard result — for example, see Arnold et al. (2008, Chapter 2, equation (2.2.3)) — that the density of the order statistics is

$$g(T_{(1)} = s_{l+1}, T_{(2)} = s_{l+2}, \ldots, T_{(n-l)} = s_n) = (n-l)!\,\Pi_{i=1}^{n-l} g(s_{l+1}),$$

which is identical to (12). Thus we can see the following fact:

$$\begin{gathered}\text{the density of the order statistics larger than } S_{(l)}, \text{ given } S_{(l)} = a, \\ \text{is the same as the density of the order statistics } T_{(1)}, T_{(2)}, \ldots, T_{(n-l)}.\end{gathered} \tag{13}$$

Now consider the distribution of the order statistics $T_{(1)}, T_{(2)}, \ldots, T_{(u-l-1)}$ given $T_{(u-l)}$. Let $0 < s_{l+1} < \ldots < s_{u-1} < b \leqslant 1$. Using the same series of steps that led to equation (12), we have

$$\begin{aligned} &g(T_{(1)} = s_{l+1}, T_{(2)} = s_{l+2}, \ldots, T_{(u-l-1)} = s_{u-1} \mid T_{(u-l)} = b) \\ &\qquad = (u-l-1)!\, \Pi_{i=1}^{u-l-1} \frac{g(s_{l+i})}{G(b)}, \end{aligned} \tag{14}$$

where G is the cdf of g:

$$G(s) = \begin{cases} \frac{F(s)-F(a)}{1-F(a)} & \text{if } s \in [a, 1]\ , \\ 0 & \text{if } s \in (-\infty, a)\ , \\ 1 & \text{if } s \in (1, \infty)\ . \end{cases}$$

Due to fact (13), the density of $(T_{(1)}, \ldots, T_{(u-l-1)})$ given $T_{(u-l)} = b$ is the same as the density of $(S_{(l+1)}, \ldots, S_{(u-1)})$ given $S_{(u)} = b$ and $S_{(l)} = a$. Thus,

$$f(S_{(l+1)} = s_{l+1}, \ldots, S_{(u-1)} = s_{u-1} \mid S_{(l)} = a, S_{(u)} = b) = (u-l-1)!\, \Pi_{i=1}^{u-l-1} \frac{g(s_{l+i})}{G(b)}.$$

Writing g and G in terms of f and F, we get

$$f(S_{(l+1)} = s_{l+1}, \ldots, S_{(u-1)} = s_{u-1} \mid S_{(l)} = a, S_{(u)} = b) = (u-l-1)!\, \Pi_{i=1}^{u-l-1} \frac{f(s_{l+i})}{F(b) - F(a)}. \tag{15}$$

Now consider the independent random variables $\{Z_i\}_{i=1}^{u-l-1}$, where the density of each Z_i is the same as the conditional density of S'_i, given $S_{(l)} = a < S'_i < b = S_{(u)}$.

Thus the density h of each Z_i is given by

$$h(s) = \begin{cases} \frac{f(s)}{F(b)-F(a)} & \text{if } s \in [a, b]\ , \\ 0 & \text{otherwise.} \end{cases}$$

The density of the order statistics $Z_{(1)}, \ldots, Z_{(u-l-1)}$ is given by

$$h(Z_{(1)} = s_{l+1}, \ldots, Z_{(u-l-1)} = s_{u-1}) = (u-l-1)!\, \Pi_{i=1}^{u-l-1} h(s_{l+i}), \tag{16}$$

which exactly matches the right hand side of (15). Thus,

$$\begin{aligned} &f(S_{(l+1)} = s_{l+1}, \ldots, S_{(u-1)} = s_{u-1} \mid S_{(l)} = a, S_{(u)} = b) \\ &= h(Z_{(1)} = s_{l+1}, \ldots, Z_{(u-l-1)} = s_{u-1}) \\ &= f(S'_{(1)} = s_{l+1}, \ldots, S'_{(u-l-1)} = s_{u-1} \mid S_{(l)} = a, S_{(u)} = b, \text{for every } i \in [u-l-1], S_{(l)} < S'_i < S_{(u)}). \end{aligned}$$

Since the conditional densities of the order statistics match, the conditional densities of the unordered random variables must also match. This gives us the claimed result.

□

A.3. Proof of Lemma 2

The sequence of order statistics $S_{(1)}, S_{(2)}, \ldots, S_{(n)}$ form a Markov chain (Arnold et al., 2008, Theorem 2.4.3). Thus

$$\left(S_{(k_{i-1}+1)}, \ldots, S_{(k_i - 1)} \perp\!\!\!\perp S_{(k_0)}, \ldots, S_{(k_{i-2})}, S_{(k_{i+1})}, \ldots, S_{(k_B)}\right) \mid S_{(k_{i-1})}, S_{(k_i)}.$$

Consequently, for the unordered set of random variables $S_{\{k_{i-1}+1\}}, \dots, S_{\{k_i-1\}}$, we have:

$$\left(S_{\{k_{i-1}+1\}}, \dots, S_{\{k_i-1\}} \perp\!\!\!\perp S_{(k_0)}, \dots, S_{(k_{i-2})}, S_{(k_{i+1})}, \dots, S_{(k_B)}\right) \mid S_{(k_{i-1})}, S_{(k_i)}.$$

Thus,

$$f(S_{\{k_{i-1}+1\}}, \dots, S_{\{k_i-1\}} \mid S_{(k_0)}, \dots, S_{(k_B)}) = f(S_{\{k_{i-1}+1\}}, \dots, S_{\{k_i-1\}} \mid S_{(k_{i-1})}, S_{(k_i)}).$$

Using Lemma 1, the result follows. □

A.4. Proof of Theorem 3

For $b \in \{0, 1, \dots, B\}$, define $k_b = \lceil b(n + 1/B) \rceil$. Let $S_{(0)} := 0$ and $S_{(n+1)} := 1$ be fixed hypothetical 'order-statistics'. The rest of this proof is conditional on the observed set $\mathcal{S} := (S_{(k_1)}, S_{(k_2)}, \dots, S_{(k_{B-1})})$. (Marginalizing over $\mathcal{S}$ gives the theorem result as stated.) Let $\mathcal{B} : \mathcal{X} \to [B]$ be the binning function: for all x, $\mathcal{B}(x) = b \iff S_{(k_{b-1})} \leqslant g(x) < S_{(k_b)}$. Note that given $\mathcal{S}$, the binning function $\mathcal{B}$ is deterministic. In particular, this means that for every $b \in [B]$, $\mathbb{E}\left[Y \mid \mathcal{B}(X) = b\right]$ is a fixed number that is not random on the calibration data or (X, Y).

Let us fix some $b \in [B]$ and denote $l = k_{b-1}, u = k_b$. By Lemma 2, the scores $S_{\{l+1\}}, S_{\{l+2\}}, \dots, S_{\{u-1\}}$ are independent and identically distributed given $\mathcal{S}$, and the conditional distribution of each of them equals that of $g(X)$ given $\mathcal{B}(X) = b$. Thus $Y_{\{l+1\}}, Y_{\{l+2\}}, \dots, Y_{\{u-1\}}$ are independent and identically distributed given $\mathcal{S}$, and the conditional distribution of each of them is Bernoulli($\mathbb{E}\left[Y \mid \mathcal{B}(X) = b\right]$). Thus for any $t \in (0, 1)$, by Hoeffding's inequality, with probability at least $1 - t$,

$$\left|\mathbb{E}\left[Y \mid \mathcal{B}(X) = b\right] - \widehat{\Pi}_b\right| \leqslant \sqrt{\frac{\log(2/t)}{2\lfloor u - l - 1 \rfloor}} \leqslant \sqrt{\frac{\log(2/t)}{2(\lfloor n/B \rfloor - 1)}}. \tag{17}$$

The second inequality holds since for any b,

$$\begin{aligned} u - l &= k_b - k_{b-1} \\ &= \lfloor (b+1)(n+1)/B \rfloor - \lfloor b(n+1)/B \rfloor \\ &= \lfloor U + (n+1)/B \rfloor - \lfloor U \rfloor, \text{ where } U = b(n+1)/B, \\ &\geqslant \lfloor (n+1)/B \rfloor \geqslant \lfloor n/B \rfloor. \end{aligned}$$

Next, we set $t = \alpha/B$ in (17), and take a union bound over all $b \in B$. Thus, with probability at least $1 - \alpha$, the event

$$E : \qquad \text{for every } b \in [B],\ \left|\mathbb{E}\left[Y \mid \mathcal{B}(X) = b\right] - \widehat{\Pi}_b\right| \leqslant \varepsilon$$

occurs. To prove the final calibration guarantee, we need to change the conditioning from $\mathcal{B}(X)$ to $h(X)$. Specifically, we have to be careful about the possibility of multiple bins having the same $\widehat{\Pi}$ values, in which case, conditioning on $\mathcal{B}(X)$ and conditioning on $h(X)$ is not the same. Given that E occurs (which happens with probability at least $1 - \alpha$),

$$\begin{aligned} &\left|\mathbb{E}\left[Y \mid h(X)\right] - h(X)\right| \\ &= \left|\mathbb{E}\left[\mathbb{E}\left[Y \mid \mathcal{B}(X), h(X)\right] \mid h(X)\right] - h(X)\right| && \text{(applying tower rule)} \\ &= \left|\mathbb{E}\left[\mathbb{E}\left[Y \mid \mathcal{B}(X)\right] \mid h(X)\right] - h(X)\right| && (\mathbb{E}\left[Y \mid \mathcal{B}(X), h(X)\right] = \mathbb{E}\left[Y \mid \mathcal{B}(X)\right]) \\ &= \left|\mathbb{E}\left[\mathbb{E}\left[Y \mid \mathcal{B}(X)\right] - h(X) \mid h(X)\right]\right| \\ &= \left|\mathbb{E}\left[\mathbb{E}\left[Y \mid \mathcal{B}(X)\right] - \widehat{\Pi}_{\mathcal{B}(X)} \mid h(X)\right]\right| && \text{(by definition of } h\text{)} \\ &\leqslant \mathbb{E}\left[\left|\mathbb{E}\left[Y \mid \mathcal{B}(X)\right] - \widehat{\Pi}_{\mathcal{B}(X)}\right| \mid h(X)\right] && \text{(Jensen's inequality)} \\ &\leqslant \varepsilon && \text{(since } E \text{ occurs).} \end{aligned}$$

This completes the proof of the conditional calibration guarantee. The ECE bound follows by Proposition 1. □

A.5. Proof of Corollary 1

Conditioned on $\mathcal{S}$ (defined in the proof of Theorem 3), for some $b \in [B]$, $l = k_{b-1}$ and $u = k_b$, we showed in the proof of Theorem 3 that with probability at least $1 - \alpha/B$,

$$\left|\mathbb{E}\left[Y \mid \mathcal{B}(X) = b\right] - \text{Mean}(Y_{(l+1)}, Y_{(l+2)}, \ldots, Y_{(u-1)})\right| \leqslant \sqrt{\frac{\log(2B/\alpha)}{2(\lfloor n/B \rfloor - 1)}}.$$

Thus for $b \in [B-1]$,

$$\begin{aligned}\left|\mathbb{E}\left[Y \mid \mathcal{B}(X) = b\right] - \widehat{\Pi}_b\right| &\leqslant \left|\mathbb{E}\left[Y \mid \mathcal{B}(X) = b\right] - \text{Mean}(Y_{(l+1)}, Y_{(l+2)}, \ldots, Y_{(u-1)})\right| \\ &\qquad + \left|\text{Mean}(Y_{(l+1)}, Y_{(l+2)}, \ldots, Y_{(u-1)}) - \text{Mean}(Y_{(l+1)}, Y_{(l+2)}, \ldots, Y_{(u)})\right| \\ &\leqslant \sqrt{\frac{\log(2B/\alpha)}{2(\lfloor n/B \rfloor - 1)}} + \frac{1}{\lfloor n/B \rfloor} \qquad \text{(by fact (9))} \\ &\leqslant \varepsilon.\end{aligned}$$

The rest of the argument can be completed exactly as in the proof of Theorem (3) after equation (17). □

A.6. Proof of Theorem 4

Let $\{\widehat{\Pi}_b'\}_{b \in [B]}$ denote the the pre-randomization values of $\widehat{\Pi}_b$ as computed in line 13 of Algorithm 2. Due to the randomization in line (15), no two $\widehat{\Pi}_b$ values are the same. Formally, consider any two indices $1 \leqslant a \neq b \leqslant B$. Then, $\widehat{\Pi}_a = \widehat{\Pi}_b$ if and only if $\delta(V_a - V_b) = \widehat{\Pi}_a' - \widehat{\Pi}_b'$, which happens with probability zero. Thus for any $1 \leqslant a \neq b \leqslant B$, $\widehat{\Pi}_a \neq \widehat{\Pi}_b$ (with probability one).

The rest of the proof is conditional on $\mathcal{S}$, as defined in the proof of Theorem 3. (Marginalizing over $\mathcal{S}$ gives the theorem result as stated.) As noted in that proof, conditioning on $\mathcal{S}$ makes the binning function $\mathcal{B}$ deterministic, which simplifies the proof significantly.

First, we prove a per bin concentration bound for $\widehat{\Pi}_b$ of the form of (17). The δ randomization changes this bound as follows. For any $b \in [B], t \in (0,1)$, with probability at least $1 - t$,

$$\begin{aligned}\left|\mathbb{E}\left[Y \mid \mathcal{B}(X) = b\right] - \widehat{\Pi}_b\right| &\leqslant \left|\mathbb{E}\left[Y \mid \mathcal{B}(X) = b\right] - \widehat{\Pi}_b'\right| + \left|\widehat{\Pi}_b - \widehat{\Pi}_b'\right| \\ &\leqslant \sqrt{\frac{\log(2/t)}{2(\lfloor n/B \rfloor - 1)}} + \left|(1+\delta)^{-1}(\widehat{\Pi}_b' + \delta) - \widehat{\Pi}_b'\right| \qquad \text{(Hoeffding's inequality (17))} \\ &\leqslant \sqrt{\frac{\log(2/t)}{2(\lfloor n/B \rfloor - 1)}} + \delta. \end{aligned} \tag{18}$$

Given this concentration bound for every bin, the (ε_2, α)-conditional calibration bound can be shown following the arguments in the proof of Theorem 3 after inequality (17). We now show the marginal calibration guarantee. Note that since no two $\widehat{\Pi}_b$ values are the same, $\mathcal{B}(X)$ is known given $\widehat{\Pi}_{\mathcal{B}(X)}$, and so $\mathbb{E}\left[Y \mid h(X)\right] = \mathbb{E}\left[Y \mid \mathcal{B}(X)\right]$. Thus,

$$\begin{aligned}&P(\left|\mathbb{E}\left[Y \mid h(X)\right] - h(X)\right| \leqslant \varepsilon_1) \\ &= \sum_{b=1}^{B} P(\left|\mathbb{E}\left[Y \mid h(X)\right] - h(X)\right| \leqslant \varepsilon_1 \mid \mathcal{B}(X) = b)\ P(\mathcal{B}(X) = b) \qquad \text{(law of total probability)} \\ &= \sum_{b=1}^{B} P(\left|\mathbb{E}\left[Y \mid \mathcal{B}(X)\right] - h(X)\right| \leqslant \varepsilon_1 \mid \mathcal{B}(X) = b)\ P(\mathcal{B}(X) = b) \qquad (\mathbb{E}\left[Y \mid h(X)\right] = \mathbb{E}\left[Y \mid \mathcal{B}(X)\right]) \\ &= \sum_{b=1}^{B} P(\left|\mathbb{E}\left[Y \mid \mathcal{B}(X)\right] - \widehat{\Pi}_{\mathcal{B}(X)}\right| \leqslant \varepsilon_1 \mid \mathcal{B}(X) = b)\ P(\mathcal{B}(X) = b) \qquad \text{(by definition of } h)\end{aligned}$$

$$\begin{aligned}
&\geqslant \sum_{b=1}^{B}(1-\alpha)\, P(\mathcal{B}(X)=b) && (t=\alpha \text{ in } (18))\\
&= 1-\alpha.
\end{aligned}$$

This proves (ε_1, α)-marginal calibration.

For the ECE bound, note that for every bin $b \in [B]$, $\widehat{\Pi}'_b$ is the average of at least $\lfloor n/B \rfloor - 1$ Bernoulli random variables with bias $\mathbb{E}\left[Y \mid \mathcal{B}(X)=b\right]$. We know the exact form of the variance of averages of Bernoulli random variables with a given bias, giving the following:

$$\text{Var}(\widehat{\Pi}'_b) \leqslant \frac{\mathbb{E}\left[Y \mid \mathcal{B}(X)=b\right](1-\mathbb{E}\left[Y \mid \mathcal{B}(X)=b\right])}{\lfloor n/B \rfloor - 1} \leqslant \frac{1}{4(\lfloor n/B \rfloor - 1)}. \tag{19}$$

We now rewrite the expectation of the square of the ℓ_2-ECE in terms of $\text{Var}(\widehat{\Pi}'_b)$. Recall that all expectations and probabilities in the entire proof are conditional on $\mathcal{S}$, so that $\mathcal{B}$ is known; the same is true for all expectations in the forthcoming panel of equations. To aid readability, when we apply the tower law, we are explicit about the remaining randomness in $\mathcal{D}_n$.

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_n}\left[(\ell_2\text{-ECE}(h))^2\right] &= \mathbb{E}_{\mathcal{D}_n}\left[\mathbb{E}_{(X,Y)}\left[(\mathbb{E}\left[Y \mid h(X)\right] - h(X))^2 \mid \mathcal{D}_n\right]\right]\\
&= \mathbb{E}_{\mathcal{D}_n}\left[\sum_{b=1}^{B}(\mathbb{E}\left[Y \mid \mathcal{B}(X)=b\right] - \widehat{\Pi}_b)^2 P(\mathcal{B}(X)=b)\right]\\
&= \sum_{b=1}^{B}\mathbb{E}_{\mathcal{D}_n}\left[(\mathbb{E}\left[Y \mid \mathcal{B}(X)=b\right] - \widehat{\Pi}_b)^2 P(\mathcal{B}(X)=b)\right]\\
&= \sum_{b=1}^{B}\mathbb{E}_{\mathcal{D}_n}\left[(\mathbb{E}\left[Y \mid \mathcal{B}(X)=b\right] - \widehat{\Pi}_b)^2\right] P(\mathcal{B}(X)=b).
\end{aligned}$$

The first equality is by the tower rule. The second equality uses the same simplifications as the panel of equations used to prove the marginal calibration guarantee (law of total probability, using $\mathbb{E}\left[Y \mid h(X)\right] = \mathbb{E}\left[Y \mid \mathcal{B}(X)\right]$, and the definition of h). The third equality uses linearity of expectation. The fourth equality follows since $\mathcal{B}$ is deterministic given $\mathcal{S}$. Now note that

$$\mathbb{E}_{\mathcal{D}_n}(\mathbb{E}\left[Y \mid \mathcal{B}(X)=b\right] - \widehat{\Pi}_b)^2 = \mathbb{E}_{\mathcal{D}_n}(\mathbb{E}\left[Y \mid \mathcal{B}(X)=b\right] - \widehat{\Pi}'_b + \widehat{\Pi}'_b - \widehat{\Pi}_b)^2 \leqslant \text{Var}(\widehat{\Pi}'_b) + \delta^2,$$

since $\mathbb{E}\left[Y \mid \mathcal{B}(X)=b\right] = \mathbb{E}_{\mathcal{D}_n}(\widehat{\Pi}'_b)$ and $\left|\widehat{\Pi}'_b - \widehat{\Pi}_b\right| \leqslant \delta$ deterministically. Thus by bound (19),

$$\mathbb{E}_{\mathcal{D}_n}\left[(\ell_2\text{-ECE}(h))^2\right] \leqslant \sum_{b=1}^{B}\left(\frac{1}{4(\lfloor n/B \rfloor - 1)} + \delta^2\right) P(\mathcal{B}(X)=b) = \frac{1}{4(\lfloor n/B \rfloor - 1)} + \delta^2 \leqslant \frac{B}{2n} + \delta^2.$$

The last inequality holds since $n \geqslant 2B$ implies that $\lfloor n/B \rfloor - 1 \geqslant n/2B$. Jensen's inequality now gives the final result:

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_n}\left[\ell_2\text{-ECE}(h)\right] &\leqslant \sqrt{\mathbb{E}_{\mathcal{D}_n}\left[(\ell_2\text{-ECE}(h))^2\right]} && \text{(Jensen's inequality)}\\
&\leqslant \sqrt{\frac{B}{2n} + \delta^2} \leqslant \sqrt{\frac{B}{2n}} + \delta.
\end{aligned}$$

The bound on $\mathbb{E}_{\mathcal{D}_n}\left[\ell_p\text{-ECE}(h)\right]$ for $p \in [1, 2)$ follows by Proposition 1. □

B. Assessing the Theoretical Guarantee of UMS

We compute the number of calibration points n required to guarantee $(\varepsilon, \alpha) = (0.1, 0.1)$-marginal calibration with $B = 10$ bins using UMS, based on Theorem 5 of Gupta et al. (2020). Following their notation, if the minimum number of calibration

Algorithm 2 Randomized UMD

```
1:  **Input:** Scoring function $g : \mathcal{X} \to [0, 1]$, #bins $B$, calibration data $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$,
    randomization parameter $\delta > 0$ (arbitrarily small)
2:  **Output:** Approximately calibrated function $h$
3:  $(U_1, U_2, \ldots, U_n) \sim$ Unif$[0, 1]^n$
4:  $(S_1, S_2, \ldots, S_n) \leftarrow (1 + \delta)^{-1}(g(X_1) + \delta U_1, g(X_2) + \delta U_2, \ldots, g(X_n) + \delta U_n)$
5:  $(S_{(1)}, S_{(2)}, \ldots, S_{(n)}) \leftarrow$ order-stats$(S_1, S_2, \ldots, S_n)$
6:  $(Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}) \leftarrow (Y_1, Y_2, \ldots, Y_n)$ ordered as per the ordering of $(S_{(1)}, S_{(2)}, \ldots, S_{(n)})$
7:  $\Delta \leftarrow (n + 1)/B$
8:  $\widehat{\Pi} \leftarrow$ empty array of size $B$
9:  $A \leftarrow$ 0-indexed array$([0, \lceil\Delta\rceil, \lceil 2\Delta\rceil, \ldots, n + 1])$
10: **for** $b \leftarrow 1$ **to** $B$ **do**
11:     $l \leftarrow A_{b-1}$
12:     $u \leftarrow A_b$
13:     $\widehat{\Pi}_b \leftarrow$ Mean$(Y_{(l+1)}, Y_{(l+2)}, \ldots, Y_{(u-1)})$
14:     $V_b \sim$ Unif$[0, 1]$
15:     $\widehat{\Pi}_b \leftarrow (1 + \delta)^{-1}(\widehat{\Pi}_b + \delta V_b)$
16: **end for**
17: $(S_{(0)}, S_{(n+1)}) \leftarrow (0, 1)$
18: $h(\cdot) \leftarrow \sum_{b=1}^{B} \mathbb{1}\left\{S_{(A_{b-1})} \leqslant (1 + \delta)^{-1}(g(\cdot) + \delta U) < S_{(A_b)}\right\} \widehat{\Pi}_b$, for $U \sim$ Unif$[0, 1]$
```

points in a bin is denoted as $N_{b^\star}$, then the Hoeffding-based bound on ε, with probablity of failure δ, is $\sqrt{\log(2B/\delta)/2N_{b^\star}}$. (The original bound is based on empirical-Berstein which is often tighter in practice, but Hoeffding is tighter in the worst case.) Let us set $\delta = \alpha/2 = 0.05$ since the remaining failure budget $\alpha/2$ is for the bin estimation to ensure that $N_{b^\star}$ is lower bounded. Thus, the requirement $\sqrt{\log(2 \cdot 10/0.05)/2N_{b^\star}} \leqslant \varepsilon = 0.1$ translates roughly to $N_{b^\star} \geqslant 300$.

To ensure $N_{b^\star} \geqslant 300$, we define the bins to each have roughly $1/B$ fraction of the calibration points in the first split of the data. Lemma 4.3 (Kumar et al., 2019) shows that w.p. $\geqslant 1 - \delta$, the true mass of the estimated bins is at least $1/2B$, as long as the first split of the data has at least $cB \log(10B/\delta)$ points, for a universal constant c. The original proof is for a $c \geqslant 2000$, but let us suppose that with a tighter analysis it can be improved to (say) $c = 100$. Then for $\delta = \alpha/4 = 0.025$, the first split of the data must have at least $100 \cdot 10 \cdot \log(100/0.025) \geqslant 8000$ calibration points. Finally, we use Theorem 5 (Gupta et al., 2020) to bound $N_{b^\star}$. If n' is the cardinality of the second split (denoted as $\left|\mathcal{D}^2_{\text{cal}}\right|$ in the original result), then they show that for $\delta = 0.025$, $N_{b^\star} \geqslant n'/2B - \sqrt{n'/\log(2B/\delta)/2} \approx n'/20 - 1.8\sqrt{n'}$. Since we require $N_{b^\star} \geqslant 300$, we must have approximately $n' \geqslant 9500$. Overall, the theoretical guarantee for UMS requires $n \geqslant 17500$ points to guarantee $(0.1, 0.1)$-marginal calibration with 10 bins.

C. Randomized UMD

We now describe the randomized version of UMD (Algorithm 2) that is nearly identical to the non-randomized version in practice, but for which we are able to show better theoretical properties. In this sense, we view randomized UMD as a theoretical tool rather than a novel algorithm (nevertheless, all experimental results in this paper use randomized UMD). Algorithm 2 takes as input a randomization parameter $\delta > 0$ which can be arbitrarily small, such as 10^{-20}. The specific lines that induce randomization, in comparison to Algorithm 1, are lines 3, 4, 14, 15 and 18. This δ perturbation leads to a better theoretical result than the non-randomized version — in comparison to Theorem 3, Theorem 4 does not require absolute continuity of $g(X)$ and provides an improved marginal calibration guarantee.

C.1. Absolute Continuity of $g(X)$

In Theorem 3, we assumed that $g(X)$ is absolutely continuous with respect to the Lebesgue measure, or equivalently, it has a pdf. This may not always be the case. For example, X may contain atoms, or g may have discrete outputs in $[0, 1]$. If $g(X)$ does not have a pdf, a simple randomization trick can be used to ensure that the results hold in full generality (we performed this randomization in our experiments as well).

First, we append the features X with Unif$[0,1]$ random variables U so that $(X,U) \sim P_X \times$ Unif$[0,1]$. Next, for an arbitrarily small value $\delta > 0$, such as 10^{-20}, we define $\widetilde{g} : \mathcal{X} \times [0,1] \to [0,1]$ as $\widetilde{g}(x,u) = (1+\delta)^{-1}(g(x) + \delta u)$. Thus for every x, $\widetilde{g}(x,\cdot)$ is arbitrarily close to $g(x)$, and we do not lose the informativeness of g. However, now $\widetilde{g}(X,U)$ is guaranteed to be absolutely continuous with respect to the Lebesgue measure. The precise implementation details are as follows: (a) to train, draw $(U_i)_{i\in[n]} \sim$ Unif$[0,1]^n$ and call Algorithm 1 with $\widetilde{g}, \{((X_i,U_i),Y_i)\}_{i\in[n]}$; (b) to test, draw a new Unif$[0,1]$ random variable for each test point. Algorithm 2 packages this randomization into the pseudocode; see lines 3, 4 and 18.

The above process is a technical way of describing the following intuitive methodology: "break ties among the scores arbitrarily but consistently". Lemmas 1 and 2 fail if two data points have $S_i = S_j$ and one of them is the order statistics we are conditioning on. However, if we fix an arbitrary secondary order through which ties can be broken even if $S_i = S_j$ or $S = S_i$, the lemmas can be made to go through. The noise term δU in $\widetilde{g}$ implicitly provides a strict secondary order.

C.2. Improved Marginal Calibration Guarantee

The marginal calibration guarantee of Theorem 4 hinges on the bin biases $\widehat{\Pi}_b$ being unique. Lines 14 and 15 in Algorithm 2 ensure that this is satisfied almost surely by adding an infinitesimal random perturbation to each $\widehat{\Pi}_b$. This is identical to the technique described in Section C.1. Due to the perturbation, the ε required to satisfy calibration as per equation (11) has an additional δ term. However the δ can be chosen to be arbitrarily small, and this term is inconsequential.

We make an informal remark that may be relevant to practitioners. In practice, we expect that the bin biases computed using Algorithm 1 are unique with high probability without the need for randomization. As long as the bin biases are unique, the marginal calibration and ECE guarantees of Theorem 4 apply to Algorithm 1 as well. Thus, the $\widehat{\Pi}$-randomization can be skipped if 'simplicity' or 'interpretability' is desired. Note that the $g(X)$ randomization (Section C.1) is still crucial since we envision many practical scenarios where $g(X)$ is not absolutely continuous. In summary, randomized UMD uses a small random perturbation to ensure that (a) the score values and (b) the bin bias estimates, are unique. The particular randomization strategy we proposed is not special; any other strategy that achieves the aforementioned goals is sufficient (for example, using a (truncated) Gaussian random variable instead of uniform).

D. Additional Experiments

We present additional experiments to supplement those presented in the main paper.

In Section 4, we compared UMD to other binning methods on the CREDIT dataset, for $n = 3$K and $n = 7$K. Here, we present plots for $n = 1$K and $n = 5$K (for easier comparison, we also show the plots for $n = 3$K and $n = 7$K). The marginal validity plots are in Figure 4, and the conditional validity plots are in Figure 5. Apart from additional evidence for the same observations made in Section 4, we also see some interesting behavior in the low sample case ($n = 1$K). First, the Theorem 4 curve does not explain performance as well as the other plots. We tried the Clopper Pearson exact confidence interval (Clopper and Pearson, 1934) instead of Hoeffding and obtained nearly identical results (plots not presented). It would be interesting to explore if a tighter guarantee can be shown for small sample sizes. Second, for $n = 1$K, scaling-binning performs better than UMD in both the marginal and conditional validity plots, and is competitive with isotonic regression in the marginal validity plot. This behavior occurs since in the small sample regime, while all other binning methods attempt to re-estimate the biases of the bins using very little data, scaling-binning relies on the statistical efficiency of the learnt g which was trained on 15K training points. A similar phenomenon was observed by Niculescu-Mizil and Caruana (2005) when comparing Platt scaling and isotonic regression: Platt scaling performs better at small sample sizes since it relies more on the underlying efficiency of g, compared to isotonic regression.

While the experiments considered so far use 10K points for training logistic regression, 5K points for Platt scaling, and between 0.5-10K points for binning, a practically common setting is where most points are used for training the base model, and a small fraction of points are used for recalibration. On recommendation of one of the ICML reviewers, we ran experiments with 14K points for training logistic regression, 1K for Platt scaling, and 1K for binning. The marginal and conditional validity plots for this experiment are displayed in Figure 6. We observe that these plots are very similar to the marginal and conditional validity plots in Figures 4 and 5 for $n = 1$K, and the same conclusions described in the previous paragraph can be drawn.

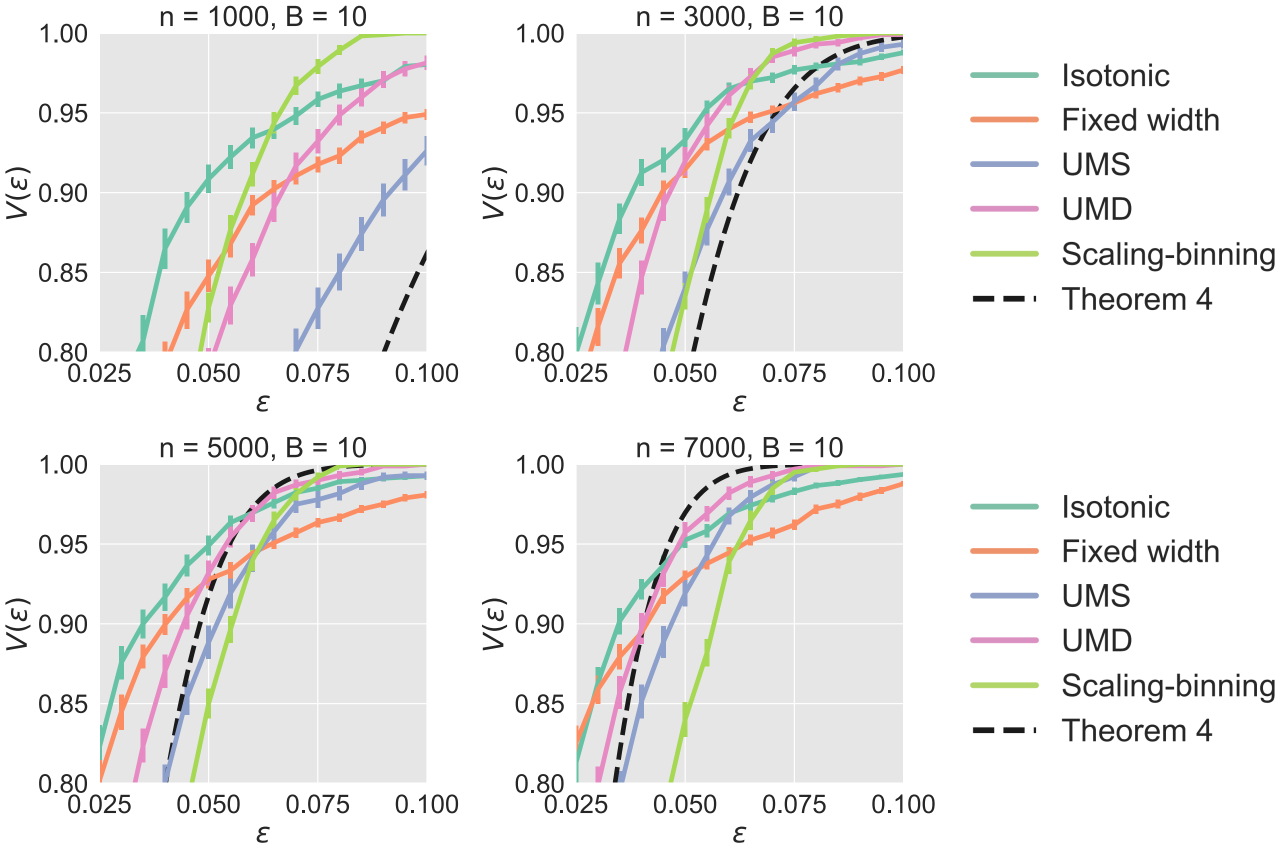


Figure 4. Marginal validity plots comparing UMD to other binning methods. The performance of UMD improves at higher values of n and ε, and the performance of UMD is closely explained by its theoretical guarantee. Isotonic regression and fixed-width binning perform well at small values of ε.

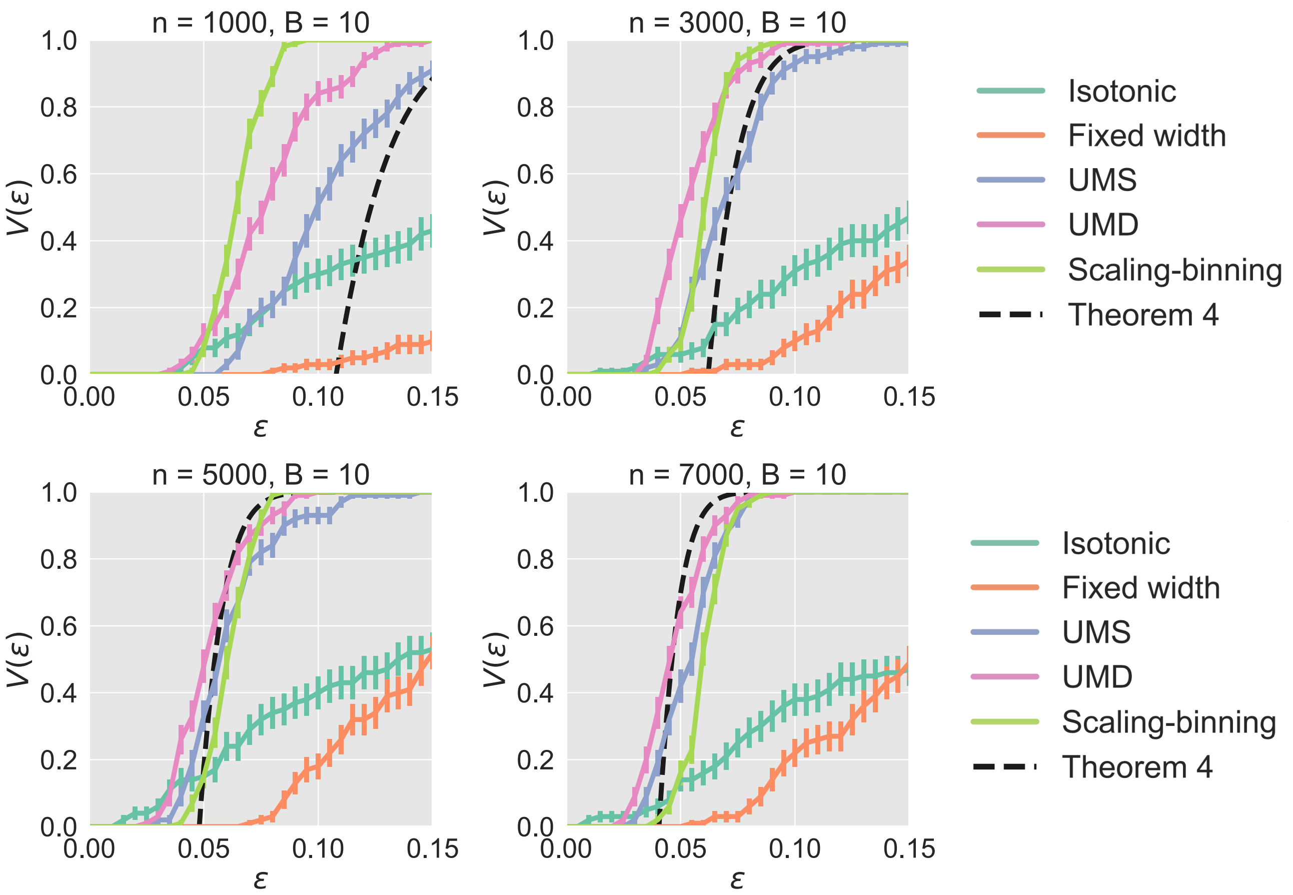
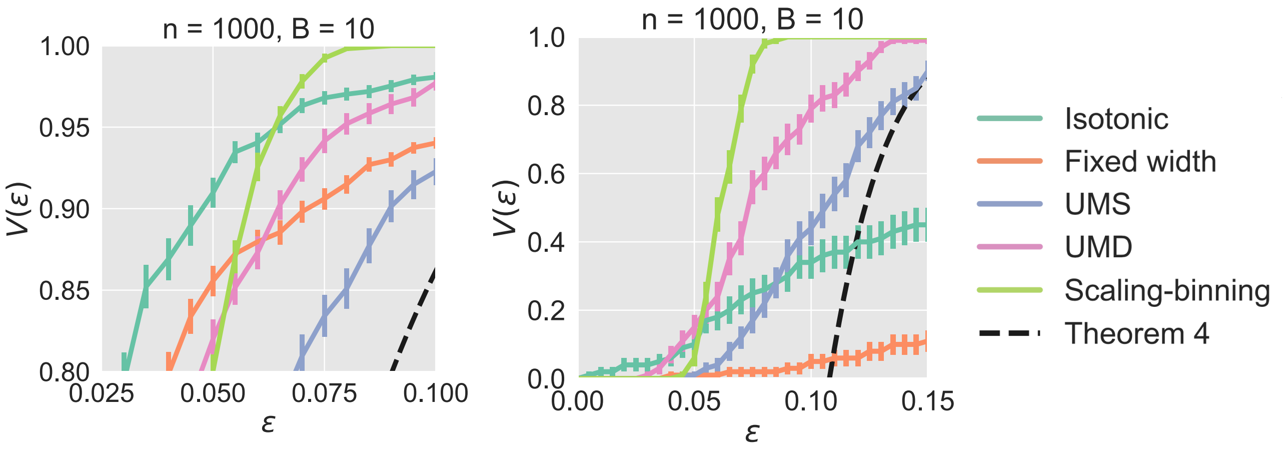


Figure 5. Conditional validity plots comparing UMD to other binning methods. UMD and scaling-binning are the best methods for conditional calibration at nearly all values of n, ε. Scaling-binning performs slightly better for small n whereas UMD performs slightly better for large n. The performance of UMD is closely explained by its theoretical guarantee.



(a) Marginal validity plot.

(b) Conditional validity plot.

Figure 6. Validity plots comparing UMD to other binning methods with fewer points used for recalibration. Namely, 14K points are used for training logistic regression, 1K for Platt scaling, and 1K for binning. Overall, scaling-binning performs quite well, since it relies on the underlying efficiency of logistic regression more than the other methods.