A Proof of Proposition 1

Proof of Proposition 1. First of all, we recall the definition of the two risks from (3) and (4):

$$R^{\mathrm{cal}}(h; f) = \mathbb{E}\left[\big(f(X) - \mathbb{E}\left[Y \mid f(X)\right]\big)^2\right]$$
$$R^{\mathrm{sha}}(h; f) = \mathbb{E}\left[\big(\mathbb{E}\left[Y \mid h \circ f(X)\right] - \mathbb{E}\left[Y \mid f(X)\right]\big)^2\right].$$

To keep our notation concise, we use $Z = f(X)$ as a shorthand notation, and also let $Y_Z := \mathbb{E}[Y \mid Z]$ and $Y_{h(Z)} = \mathbb{E}[Y \mid h(Z)]$ throughout this proof. We can decompose the recalibration risk from Definition 5:

$$\begin{aligned}
R(h) &= \mathbb{E}[(h(Z) - Y_Z)^2] \\
&= \mathbb{E}[(h(Z) - Y_{h(Z)} + Y_{h(Z)} - Y_Z)^2] \\
&= \mathbb{E}[(h(Z) - Y_{h(Z)})^2] + \mathbb{E}[(Y_{h(Z)} - Y_Z)^2] + 2\mathbb{E}[(h(Z) - Y_{h(Z)})(Y_{h(Z)} - Y_Z)] \\
&= \mathbb{E}[(h(Z) - Y_{h(Z)})^2] + \mathbb{E}[(Y_{h(Z)} - Y_Z)^2] + 2\mathbb{E}[\mathbb{E}[(h(Z) - Y_{h(Z)})(Y_{h(Z)} - Y_Z) \mid h(Z)]] \\
&= \mathbb{E}[(h(Z) - Y_{h(Z)})^2] + \mathbb{E}[(Y_{h(Z)} - Y_Z)^2] + 2\mathbb{E}[(h(Z) - Y_{h(Z)})(Y_{h(Z)} - \mathbb{E}[Y_Z \mid h(Z)])] \\
&= \mathbb{E}[(h(Z) - Y_{h(Z)})^2] + \mathbb{E}[(Y_{h(Z)} - Y_Z)^2] + 2\mathbb{E}[(h(Z) - Y_{h(Z)})(Y_{h(Z)} - Y_{h(Z)})] \\
&= \underbrace{\mathbb{E}[(h(Z) - Y_{h(Z)})^2]}_{R^{\mathrm{cal}}(h)} + \underbrace{\mathbb{E}[(Y_{h(Z)} - Y_Z)^2]}_{R^{\mathrm{sha}}(h)}.
\end{aligned}$$

□

B Proof of Theorem 1

In this section, we present a proof of Theorem 1. Let $(Y, Z) \in \mathcal{Y} \times \mathcal{Z}$ be random variables that admits a joint distribution $P_{Y,Z}$, which we assume to be fixed throughout this section. Let $S = \{(y_i, z_i) \in \mathcal{Y} \times \mathcal{Z} : i \in [n]\}$ and let $\mathcal{B} = \{I_1, I_2, \ldots, I_B\}$ be the uniform-mass binning scheme (cf. Definition 6) of size B induced by (z_i's in) S. Note that if S is a random sample from $P_{Y,Z}$, then the binning scheme $\mathcal{B}$ induced by S is also a random variable following a derived distribution. To facilitate our analysis, we introduce the notion of well-balanced binning.

Definition 8 (Well-balanced binning; [29])**.** *Let $B \in \mathbb{N}$, let Z be a random variable that takes value in $[0, 1]$, and let $\alpha \in \mathbb{R}$ such that $\alpha \geq 1$. A binning scheme $\mathcal{B}$ of size B is* α-well-balanced with respect to Z *if*

$$\frac{1}{\alpha B} \leq P[Z \in I_b] \leq \frac{\alpha}{B}, \qquad \forall b \in [B].$$

In addition, we define two (parameterized families of) Boolean-valued functions Φ_{balance} and Φ_{approx} as follows: for any binning scheme $\mathcal{B}$,

$$\forall \alpha \in \mathbb{R}, \quad \Phi_{\mathrm{balance}}(\mathcal{B}; \alpha) := \mathbb{1}\left\{\frac{1}{\alpha|\mathcal{B}|} \leq P\left[Z \in I\right] \leq \frac{\alpha}{|\mathcal{B}|}, \ \forall I \in \mathcal{B}\right\}, \tag{21}$$

$$\forall \varepsilon \in \mathbb{R}, \quad \Phi_{\mathrm{approx}}(\mathcal{B}; \varepsilon) := \mathbb{1}\left\{\max_{I \in \mathcal{B}} |\hat{\mu}_I - \mu_I| \leq \varepsilon\right\}, \tag{22}$$

where $\mathbb{1}(A) = 1$ if and only if the predicate A is true, and for each interval $I \in \mathcal{B}$,

$$\hat{\mu}_I = \frac{\sum_{i=1}^n y_i \cdot \mathbb{1}_I(z_i)}{\sum_{i=1}^n \cdot \mathbb{1}_I(z_i)} \qquad \text{and} \qquad \mu_I = \mathbb{E}_{(Y,Z) \sim P_{Y,Z}}\left[Y \cdot \mathbb{1}_I(Z)\right]. \tag{23}$$

Note that if $\Phi_{\mathrm{balance}}(\mathcal{B}; \alpha) = 1$ for $\alpha \geq 1$, then $\mathcal{B}$ is α-well-balanced with respect to Z (cf. Definition 8). Also, if $\Phi_{\mathrm{approx}}(\mathcal{B}; \varepsilon) = 1$ for $\varepsilon \geq 0$, then the conditional empirical mean of Y in each bin $I \in \mathcal{B}$ approximates the conditional expectation with error at most ε, uniformly for all bins.

The rest of this section is organized as follows. In Section B.1, we ensure that for an appropriate choice of $\alpha, \varepsilon \in \mathbb{R}$, it holds with high probability (with respect to the randomness in $\mathcal{B}$) that $\Phi_{\mathrm{balance}}(\mathcal{B}; \alpha) = \Phi_{\mathrm{approx}}(\mathcal{B}; \varepsilon) = 1$. In Section B.2, we establish upper bounds on the reliability risk R^{cal} and the sharpness risk R^{sha} under the premise that $\Phi_{\mathrm{balance}}(\mathcal{B}; \alpha) = \Phi_{\mathrm{approx}}(\mathcal{B}; \varepsilon) = 1$. Finally, in Section B.3, we conclude the proof of Theorem 1 by combining these results together.

B.1 High-probability certification of the conditions

Well-balanced binning scheme. First of all, we observe that the uniform-bass binning scheme $\mathcal{B}$ induced by an IID random sample from $P_{Y,Z}$ is 2-well-balanced with high probability, if the sample size is sufficiently large. Here we paraphrase a result from [29] in our language.

Lemma 3 ([29, Lemma 4.3])**.** *Let $S = \{Z_i : i \in [n]\}$ be an IID sample drawn from P_Z and let $\mathcal{B}$ be the uniform-mass binning scheme of size B induced by S. There exists a universal constant $c' > 0$ such that for any $\delta \in (0,1)$, if $n \geq c' \cdot B \log(B/\delta)$, then $\Phi_{\mathrm{balance}}(\mathcal{B}, 2) = 1$ with probability at least $1 - \delta$.*

Lemma 3 states that

$$n \geq c' \cdot B \log\left(\frac{B}{\delta}\right) \qquad \Longrightarrow \qquad P\left[\mathcal{B} \text{ is 2-well-balanced with respect to } P_Z\right] \geq 1 - \delta.$$

While the value of the universal constant c was not specified in the original reference [29], we remark that one may set, for example, $c' = 2420$, which can be verified by following their proof with c' kept explicit.

The proof of Lemma 3 in [29] relies on a discretization argument that considers a fine-grained cover of $\mathcal{Z} = [0,1]$ consisting of disjoint intervals—namely, $\left\{I'_j : j \in [10B]\right\}$ such that $P[Z \in I'_j] = \frac{1}{10B}$ for all $j \in [10B]$—and then approximates each I_b by a subset of the cover. As the authors of [29] remarked, this argument provides a tighter sample complexity upper bound than naïvely applying Chernoff bounds or a standard VC dimension argument, which would yield an upper bound of order $O\left(B^2 \log\left(\frac{B}{\delta}\right)\right)$. We omit the proof of Lemma 3 and refer interested readers to the referenced paper [29] for more details.

Uniform concentration of bin-wise means. Next, we argue that for the uniform-mass binning scheme $\mathcal{B}$ induced by an IID sample, the conditional empirical means of each bin concentrates to the population conditional expectation, uniformly for all bins in $\mathcal{B}$. Here we restate a result from [20].

Lemma 4 ([20, Corollary 1])**.** *Let P_Z be an absolutely continuous probability measure on $\mathcal{Z} = [0,1]$, and $S = \{Z_i : i \in [n]\}$ be an IID sample drawn from P_Z. Let $B \in \mathbb{N}$ such that $B \leq \frac{n}{2}$ and $\mathcal{B}$ be the uniform-mass binning scheme of size B induced by S. Then for any $\delta \in (0,1)$,*

$$P\left[\Phi_{\mathrm{approx}}(\mathcal{B}; \varepsilon_\delta) = 1\right] \geq 1 - \delta \qquad \textit{where} \qquad \varepsilon_\delta = \sqrt{\frac{1}{2(\lfloor n/B \rfloor - 1)} \log\left(\frac{2B}{\delta}\right)} + \frac{1}{\lfloor n/B \rfloor}. \tag{24}$$

Lemma 4 states that under the mild regularity condition of P_Z being absolutely continuous, the uniform-mass binning accurately approximates all bin-wise conditional means as long as there are at least two samples per bin in the sense that

$$n \geq 2B \qquad \Longrightarrow \qquad P\left[\sup_{b \in [B]} |\hat{\mu}_b - \mu_b| \leq \sqrt{\frac{1}{2(\lfloor n/B \rfloor - 1)} \log\left(\frac{2B}{\delta}\right)} + \frac{1}{\lfloor n/B \rfloor}\right] \geq 1 - \delta.$$

B.2 Conditional upper bounds on reliability risk and sharpness risk

In this section, we establish upper bounds on the reliability risk R^{cal} and the sharpness risk R^{sha} for $\hat{h}$ under the premise that $\Phi_{\mathrm{balance}}(\mathcal{B}; \alpha) = 1$ and $\Phi_{\mathrm{approx}}(\mathcal{B}; \varepsilon) = 1$ for appropriate parameters $\alpha, \varepsilon \in \mathbb{R}$.

Preparation. To avoid clutter in the lemma statements to follow, here we recall our problem setting and set several notation that will be used throughout this section. Recall that $P = P_{X,Y}$ is a joint distribution on $\mathcal{X} \times \mathcal{Y}$ and let $f : \mathcal{X} \to \mathcal{Z}$ is a measurable function. In addition, we let $\tilde{S} = \{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y} : i \in [n]\}$ be an IID sample drawn from P, and let $S = \left\{(z, y) \in \mathcal{Z} \times \mathcal{Y} : (x, y) \in \tilde{S} \text{ and } z = f(x)\right\}$. Let $\mathcal{B}$ be the uniform-mass binning scheme induced by (z's in) S, and let $\hat{h} = \hat{\mathcal{B}} : \mathcal{Z} \to \mathcal{Z}$ be the recalibration function derived from $\mathcal{B}$ as we described in Section 4.1; see (9). The dependence among $P, f, \tilde{S}, S, \mathcal{B}$, and $\hat{h}$ are summarized by a diagram in Figure 4.

Furthermore, we define the index function for a binning scheme to facilitate our analysis.

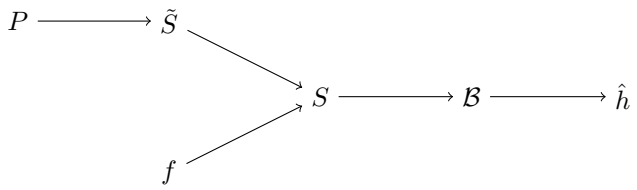


Figure 4: Stochastic dependence among $P, f, \tilde{S}, S, \mathcal{B}$, and $\hat{h}$.

Definition 9. *Let $\mathcal{B}$ be a binning scheme. The* index function *for $\mathcal{B}$ is the function $\beta : \mathcal{Z} \to [|\mathcal{B}|]$ such that*

$$\beta(z) = \sum_{I \in \mathcal{B}} \mathbb{1}_{(0, \sup I]}(z). \tag{25}$$

Remark 6. *Note that β is a measurable function and defines an index function that identifies which bin of $\mathcal{B}$ the argument $z \in [0,1]$ belongs to. Specifically, suppose that $\mathcal{B} = \{I_1, \ldots, I_B\}$ for some $B \in \mathbb{N}$ and there exists $u_0, u_1, \ldots, u_B \in [0,1]$ such that (i) $0 = u_0 < u_1 < \cdots < u_B = 1$ and (ii) $I_b = (u_{b-1}, u_b]$ for all $b \in [B] \setminus \{1\}$ and $I_1 = [u_0, u_1]$. Then $\beta(z) = b$ if and only if $z \in I_b$.*

B.2.1 Calibration risk upper bound

We observe that if a binning scheme $\mathcal{B}$ produces empirical means $\hat{\mu}_I$ that approximate the true means μ_I with error at most ε, then the calibration risk is upper bounded by ε^2.

Lemma 5 (Calibration risk bound)**.** *For any $\varepsilon \geq 0$, if $\Phi_{\mathrm{approx}}(\mathcal{B}; \varepsilon) = 1$, then*

$$R^{\mathrm{cal}}(\hat{h}; f, P) \leq \varepsilon^2.$$

Proof of Lemma 5. To begin with, we recall the definition of the calibration risk (Definition 3), and let $Z = f(X)$. Then we may write

$$\begin{aligned}
R^{\mathrm{cal}}\big(\hat{h}; f, P\big) &= \mathbb{E}\left[\Big(\hat{h}(Z) - \mathbb{E}[Y \mid \hat{h}(Z)]\Big)^2\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\Big(\hat{h}(Z) - \mathbb{E}[Y \mid \hat{h}(Z)]\Big)^2 \,\middle|\, \beta(Z)\right]\right] && \because \text{the law of total expectation} \\
&= \mathbb{E}\left[\big(\hat{\mu}_{I_{\beta(Z)}} - \mu_{I_{\beta(Z)}}\big)^2\right] && \text{cf. (23)} \\
&\leq \max_{I \in \mathcal{B}} \big(\hat{\mu}_I - \mu_I\big)^2.
\end{aligned}$$

Note that if $\Phi_{\mathrm{approx}}(\mathcal{B}; \varepsilon) = 1$, then $\max_{I \in \mathcal{B}} \big(\hat{\mu}_I - \mu_I\big)^2 \leq \varepsilon^2$. □

We remark that the proof of Lemma 5 is a simple application of applying Hölder's inequality. Also, we note that a similar argument was considered in [20, Proposition 1] to establish the inequalities between the L^p-counterparts of the calibration risk, which they call the ℓ_p-expected calibration error (ECE). In this work, we focus on the case $p = 2$.

B.2.2 Sharpness risk upper bound

Next, we present an upper bound for the sharpness risk that diminishes as the binning scheme $\mathcal{B}$ becomes more balanced.

Lemma 6 (Sharpness risk bound)**.** *Suppose that the optimal post-hoc recalibration function $h^*_{f,P}$, cf. (7), is monotonically non-decreasing. Let $\alpha \in \mathbb{R}$ such that $\alpha \geq 1$. If $\Phi_{\mathrm{balance}}(\mathcal{B}, \alpha) = 1$, then*

$$R^{\mathrm{sha}}(\hat{h}; f, P) \leq \frac{\alpha}{|\mathcal{B}|}.$$

Proof of Lemma 6. Letting $Z = f(X)$, we can write the sharpness risk of $\hat{h}$ over f with repsect to P as

$$R^{\mathrm{sha}}(\hat{h}; f, P) := \mathbb{E}\left[\big(\mathbb{E}\big[Y \mid \hat{h}(Z)\big] - \mathbb{E}\left[Y \mid Z\right]\big)^2\right].$$

We recall the definition of the index function β for $\mathcal{B}$ (Definition 9) and observe that

$$
\begin{aligned}
&\mathbb{E}\left[\left(\mathbb{E}\left[Y \mid \hat{h}(Z)\right] - \mathbb{E}\left[Y \mid Z\right]\right)^2\right] \\
&\quad\le \mathbb{E}\left[\left|\mathbb{E}\left[Y \mid \hat{h}(Z)\right] - \mathbb{E}\left[Y \mid Z\right]\right|\right] && \because \left|\mathbb{E}\left[Y \mid \hat{h}(Z)\right] - \mathbb{E}\left[Y \mid Z\right]\right| \le 1 \\
&\quad= \sum_{I\in\mathcal{B}} \mathbb{E}\left[\left|\mathbb{E}[Y \mid \hat{h}(Z)] - \mathbb{E}\left[Y \mid Z\right]\right| \cdot \mathbb{1}_I(Z)\right] \\
&\quad= \sum_{I\in\mathcal{B}} \mathbb{E}\left[\mathbb{E}\left[\left|\mathbb{E}[Y \mid \hat{h}(Z)] - \mathbb{E}\left[Y \mid Z\right]\right| \cdot \mathbb{1}_I(Z) \,\middle|\, \beta(Z)\right]\right] && \because \text{the law of total expectation} \\
&\quad= \sum_{I\in\mathcal{B}} P[Z\in I] \cdot \mathbb{E}\left[\mathbb{E}\left[\left|\mathbb{E}[Y \mid \hat{h}(Z)] - \mathbb{E}\left[Y \mid Z\right]\right| \,\middle|\, Z\in I\right]\right] && \because \text{Remark 6} \\
&\quad\le \sum_{I\in\mathcal{B}} P[Z\in I] \cdot \left(\sup_{z\in I} h^*_{f,P}(z) - \inf_{z\in I} h^*_{f,P}(z)\right) && \because \text{by definition of } h^*_{f,P}\text{; cf. (7)} \\
&\quad\le \sum_{I\in\mathcal{B}} \frac{\alpha}{|\mathcal{B}|} \cdot \left(\sup_{z\in I} h^*_{f,P}(z) - \inf_{z\in I} h^*_{f,P}(z)\right) && \because \Phi_{\text{balance}}(\mathcal{B},\alpha) = 1 \\
&\quad\le \frac{\alpha}{|\mathcal{B}|}.
\end{aligned}
$$

The inequality in the last line follows from the facts that (i) $I \in \mathcal{B}$ are mutually exclusive and (ii) $h^*_{f,P}(z) \in [0,1]$ and $h^*_{f,P}$ is monotone non-decreasing. □

Our proof of Lemma 6 relies on similar techniques that are used in [29, Lemmas D.5 and D.6]. However, we note that we obtain an improved constant — 1 as opposed to 2 in [29, Lemma D.6] — with a more refined analysis.

An improved rate with additional assumptions. It is possible to improve the rate of the sharpness risk upper bound from $O(|\mathcal{B}|^{-1})$ to $O(|\mathcal{B}|^{-2})$ with an additional regularity assumption on $h^*_{f,P}$.

Recall that we assumed in `(A3)` that there exists $K > 0$ such that if $z_1 \le z_2$, then $h^*_{f,P}(z_2) - h^*_{f,P}(z_1) \le K \cdot \big(F_Z(z_2) - F_Z(z_1)\big)$, that is, $h^*_{f,P}$ is K-smooth with respect to F_Z. This posits that the conditional probability $P[Y = 1|Z = z]$ of the target variable Y given a forecast variable Z cannot vary too much in regions where the density of Z is low, or where the forecast is rarely issued. This is a reasonable assumption because if $P[Y = 1|Z]$ changes too rapidly with respect to Z, then it suggests that we need additional information about Y beyond what Z can provide in order to improve the quality of forecasts. We remark that `(A3)` is indeed a fairly mild assumption to impose on, however, is not a trivial one.

Remark 7 (Mildness of `(A3)`)**.** *Suppose that $Z = f(X)$ has a density p_Z that is uniformly lower bounded by ϵ on the support of Z. If $h^*_{f,P}$ is L-Lipschitz, then $h^*_{f,P}$ is (L/ϵ)-smooth with respect to F_Z. This also provides a sufficient condition to verify `(A3)` in practice.*

Remark 8 (Non-triviality of `(A3)`)**.** *Notice that even if F_Z is absolutely continuous and $h^*_{f,P}$ is continuous, the smoothness constant K could become large if the prediction Z is heavily miscalibrated. For instance, in Figure 6, $h^*_{f,P}(z)$ is changing fast in the interval $[0.5, 0.75]$ where $p_Z(z)$ is small, which results in a larger value of K that can even diverge if $p_Z(z) \to 0$.*

Here we define the notion of ψ-smoothness to formalize Assumption `(A3)`, and then present an improved upper bound for the sharpness risk.

Definition 10 (ψ-smoothness)**.** *Let $K \in \mathbb{R}_+$ and $\psi : [0,1] \to [0,1]$ be a monotone non-decreasing function. A function $\phi : [0,1] \to [0,1]$ is K-smooth with respect to ψ if for any $z_1, z_2 \in [0,1]$ such that $z_1 \le z_2$,*

$$\big|\phi(z_2) - \phi(z_1)\big| \le K \cdot \big(\psi(z_2) - \psi(z_1)\big). \tag{26}$$

Lemma 7 (Improved sharpness risk bound)**.** *Suppose that the function $h^*_{f,P}(z)$ defined in (7) is monotonically non-decreasing and K-smooth with respect to F_Z for some $K \ge 0$, where F_Z is the cumulative distribution function of $Z = f(X)$. If $\Phi_{\text{balance}}(\mathcal{B},\alpha) = 1$, then*

$$R^{\text{sha}} \le \frac{K^2\alpha^3}{B^2}.$$

Proof of Lemma 7. Let $Z = f(X)$ and $B = |\mathcal{B}|$. For each $b \in [B]$, we let $z_{b,\max} := \sup I_b$ and $z_{b,\min} := \inf I_b$. Then we have

$$
\begin{aligned}
&R^{\mathrm{sha}}(\hat{h}; f, P) \\
&\quad = \mathbb{E}\left[\left(\mathbb{E}\big[Y \mid \hat{h}(Z)\big] - \mathbb{E}\left[Y \mid Z\right]\right)^2\right] \\
&\quad = \mathbb{E}\left[\mathbb{E}\left[\left(\mathbb{E}\big[Y \mid \hat{h}(Z)\big] - \mathbb{E}\left[Y \mid Z\right]\right)^2 \,\middle|\, \beta(Z)\right]\right] \\
&\quad = \sum_{b=1}^{B} P[Z \in I_b] \cdot \mathbb{E}\left[\left(\mathbb{E}\big[Y \mid \hat{h}(Z)\big] - \mathbb{E}\left[Y \mid Z\right]\right)^2 \,\middle|\, \beta(Z) = b\right] \\
&\quad \le \sum_{b=1}^{B} P[Z \in I_b] \cdot \left(h^*_{f,P}(z_{b,\max}) - h^*_{f,P}(z_{b,\min})\right)^2 && \because h^*_{f,P} \text{ is non-decreasing} \\
&\quad \le \sum_{b=1}^{B} P[Z \in I_b] \cdot \left(K \cdot \big(F_Z(z_{b,\max}) - F_Z(z_{b,\min})\big)\right)^2 && \because h^*_{f,P} \text{ is K-smooth w.r.t. } F_Z \\
&\quad = \sum_{b=1}^{B} K^2 \cdot P[Z \in I_b]^3 \\
&\quad \le K^2 \sum_{b=1}^{B} \left(\frac{\alpha}{B}\right)^3 && \because \Phi_{\mathrm{balance}}(\mathcal{B}, \alpha) = 1 \\
&\quad = \frac{K^2 \alpha^3}{B^2}.
\end{aligned}
$$

□

Remark 9 (Tightness of the rate $O(B^{-2})$)**.** *The asymptotic rate $R^{\mathrm{sha}} = O(B^{-2})$ is tight and cannot be further improved without additional assumptions. For instance, let's consider a uniform-mass binning of size B on $Z \sim$ Uniform$[0,1]$. In the population limit, each bin has width $1/B$ and within-bin variance $1/(12B^2)$. Thus, the sharpness risk, obtained by taking expectation of the conditional variance (per each bin), is $1/(12B^2)$, attaining the rate B^{-2}.*

B.3 Completing the proof of Theorem 1

Proof of Theorem 1. For given $\delta \in (0,1)$, let $\delta_1 = \delta_2 = \delta/2$. Then we observe that

$$
\begin{aligned}
&n \ge c' \cdot |\mathcal{B}| \log\left(\frac{|\mathcal{B}|}{\delta_1}\right) && \Longrightarrow && P\left[\Phi_{\mathrm{balance}}(\mathcal{B}, 2) = 1\right] \ge 1 - \delta_1 && \text{by Lemma 3} \\
&n \ge 2|\mathcal{B}| && \Longrightarrow && P\left[\Phi_{\mathrm{approx}}(\mathcal{B}, \varepsilon_{\delta_2}) = 1\right] \ge 1 - \delta_2 && \text{by Lemma 4}
\end{aligned}
$$

where $c' > 0$ is the universal constant that appears in Lemma 3 and

$$
\varepsilon_{\delta_2} = \sqrt{\frac{1}{2(\lfloor n/|\mathcal{B}| \rfloor - 1)} \log\left(\frac{2|\mathcal{B}|}{\delta_2}\right)} + \frac{1}{\lfloor n/|\mathcal{B}| \rfloor}.
$$

Observe that $\delta_1 = \frac{\delta}{2} < \frac{1}{2}$ and $|\mathcal{B}| \ge 1$, and thus, $\log\left(\frac{|\mathcal{B}|}{\delta_1}\right) \ge \log 2$. Letting $c := \max\{c', \frac{2}{\log 2}\}$ and applying the union bound, we have

$$
n \ge c \cdot |\mathcal{B}| \log\left(\frac{2|\mathcal{B}|}{\delta}\right) \qquad \Longrightarrow \qquad P\big[\Phi_{\mathrm{balance}}(\mathcal{B}, 2) = 1 \text{ and } \Phi_{\mathrm{approx}}(\mathcal{B}, \varepsilon_{\delta/2}) = 1\big] \ge 1 - \delta.
$$

Next, we observe that if $\Phi_{\mathrm{balance}}(\mathcal{B}, 2) = 1$ and $\Phi_{\mathrm{approx}}(\mathcal{B}, \varepsilon_{\delta_2}) = 1$, then

$$
\begin{aligned}
&R^{\mathrm{cal}}(\hat{h}; f; P) \le (\varepsilon_{\delta_2})^2 && \text{by Lemma 5,} \\
&R^{\mathrm{sha}}(\hat{h}; f, P) \le \frac{2}{|\mathcal{B}|}, && \text{by Lemma 6.}
\end{aligned}
$$

Additionally, if the Assumption (A3) also holds, then we obtain a stronger upper bound on $R^{\mathrm{sha}}(\hat{h}; f, P)$ by Lemma 7:

$$R^{\mathrm{sha}}(\hat{h}; f, P) \leq \frac{8K^2}{|\mathcal{B}|^2}.$$

□

C Proof of Theorem 2

This section contains a proof of Theorem 2. Prior to the proof, in Section C.1, we provide several lemmas that will be useful in our proof. Thereafter, we present a proof of Theorem 2 in its entirety in Section C.2.

C.1 Useful lemmas

C.1.1 Concentration of $\hat{w}_k$ to w_k^*

First of all, we recall the binomial Chernoff bound, which is a classical result about the concentration of measures that can be found in standard textbooks on probability theory.

Lemma 8 (Binomial Chernoff bound)**.** *Let X_i be IID Bernoulli random variables with parameters $p \in (0,1)$, and let $S_n := \frac{1}{n}\sum_{i=1}^n X_i$. Then for any $\delta \in \mathbb{R}$ such that $0 < \varepsilon < 1$,*

$$P\left[S_n \geq (1+\varepsilon)p\right] \leq \exp\left(-\frac{\varepsilon^2 p}{3}n\right),$$
$$P\left[S_n \leq (1-\varepsilon)p\right] \leq \exp\left(-\frac{\varepsilon^2 p}{2}n\right).$$

It follows from Lemma 8 that for any $\varepsilon, \delta \in (0,1)$,

$$n \geq \frac{3}{\varepsilon^2 p}\log\left(\frac{2}{\delta}\right) \qquad \Longrightarrow \qquad P\left(\frac{|S_n - p|}{p} > \varepsilon\right) \leq \delta. \tag{27}$$

Let P, Q be two distributions on $\mathcal{Y} = \{0,1\}$, and let $\mathcal{D}_P \sim P$, $\mathcal{D}_Q \sim Q$ denote IID samples of size n_P, n_Q, respectively. Recall from (13) and (16) that for each $k \in \{0,1\}$, we define

$$w_k^* = \frac{P_Q[Y=k]}{P_P[Y=k]}, \qquad \text{and} \qquad \hat{w}_k = \frac{P_{\mathcal{D}_Q}[Y=k]}{P_{\mathcal{D}_P}[Y=k]}.$$

Then, we let

$$\rho_0 := \frac{\hat{w}_0}{w_0^*} \qquad \text{and} \qquad \rho_1 := \frac{\hat{w}_1}{w_1^*}. \tag{28}$$

Now we define another parameterized family of Boolean-valued functions $\Phi_{\mathrm{ratio}}(\mathcal{D}_P, \mathcal{D}_Q; \beta)$ as follows. Given $\mathcal{D}_P \sim P$, $\mathcal{D}_Q \sim Q$, and $\beta \in \mathbb{R}$ such that $1 < \beta \leq 2$,

$$\Phi_{\mathrm{ratio}}(\mathcal{D}_P, \mathcal{D}_Q; \beta) := \mathbb{1}\left\{\frac{1}{\beta} \leq \rho_k \leq \beta, \ \ \forall k \in \{0,1\}\right\}. \tag{29}$$

Corollary 9. *Let P, Q be two distributions on $\mathcal{Y} = \{0,1\}$, and let $\mathcal{D}_P \sim P$, $\mathcal{D}_Q \sim Q$ denote IID samples of size n_P, n_Q, respectively. For each $k \in \{0,1\}$, let $p_k := P_P[Y=k]$ and $q_k := P_Q[Y=k]$. Likewise, we let $\hat{p}_k = \frac{1}{n_P}\sum_{y_i \in \mathcal{D}_P} \mathbb{1}\{y_i = k\}$ and $\hat{q}_k = \frac{1}{n_Q}\sum_{y_i \in \mathcal{D}_Q} \mathbb{1}\{y_i = k\}$. For any $\delta \in (0,1)$ and any $\beta \in (1,2]$, if*

$$n_P \geq \frac{27}{(\beta-1)^2 \min\{p_0, p_1\}}\log\left(\frac{8}{\delta}\right) \quad \textit{and} \quad n_Q \geq \frac{27}{(\beta-1)^2 \min\{q_0, q_1\}}\log\left(\frac{8}{\delta}\right),$$

then

$$P\left(\Phi_{\mathrm{ratio}}(\mathcal{D}_P, \mathcal{D}_Q; \beta) = 1\right) \geq 1 - \delta.$$

Proof of Corollary 9. Let $\varepsilon = \frac{\beta-1}{3}$. Since $\frac{1+x}{1-x} \le 1+3x$ for all $x \in [0, 1/3]$, we have $\frac{1}{\beta} \le \frac{1-\varepsilon}{1+\varepsilon} < \frac{1+\varepsilon}{1-\varepsilon} \le \beta$. Then it follows from (27) that for each $k \in \{0,1\}$,

$$n_P \ge \frac{3}{\varepsilon^2 p_k} \log\left(\frac{8}{\delta}\right) \qquad \Longrightarrow \qquad P\left(\frac{|\hat{p}_k - p_k|}{p_k} > \varepsilon\right) \le \frac{\delta}{4},$$
$$n_Q \ge \frac{3}{\varepsilon^2 q_k} \log\left(\frac{8}{\delta}\right) \qquad \Longrightarrow \qquad P\left(\frac{|\hat{q}_k - q_k|}{q_k} > \varepsilon\right) \le \frac{\delta}{4}.$$

Applying the union bound, we obtain the following implication:

$$\begin{aligned} n_P \ge \frac{3}{\varepsilon^2 \min\{p_0, p_1\}} \log\left(\frac{8}{\delta}\right) \text{ and } n_Q \ge \frac{3}{\varepsilon^2 \min\{q_0, q_1\}} \log\left(\frac{8}{\delta}\right) \\ \Longrightarrow \qquad P\left(\max_{k\in\{0,1\}} \frac{|\hat{p}_k - p_k|}{p_k} > \varepsilon \text{ or } \max_{k\in\{0,1\}} \frac{|\hat{q}_k - q_k|}{q_k} > \varepsilon\right) \le \delta \\ \Longrightarrow \qquad P\left(\max_{k\in\{0,1\}} \rho_k > \frac{1+\varepsilon}{1-\varepsilon} \text{ or } \min_{k\in\{0,1\}} \rho_k < \frac{1-\varepsilon}{1+\varepsilon}\right) \le \delta \\ \Longrightarrow \qquad P\left(\max_{k\in\{0,1\}} \rho_k > \beta \text{ or } \min_{k\in\{0,1\}} \rho_k < \frac{1}{\beta}\right) \le \delta. \end{aligned}$$

□

C.1.2 Regularity of the Shift Correction Function

Lemma 10. *Let $w = (w_0, w_1) \in \mathbb{R}^2$ such that $w_0, w_1 > 0$ and $w_0 + w_1 = 1$. The function $g_w : [0,1] \to [0,1]$ such that $g_w(z) = \frac{w_1 z}{w_1 z + w_0(1-z)}$ is L-Lipschitz where $L = \max\left\{\frac{w_1}{w_0}, \frac{w_0}{w_1}\right\}$.*

Proof of Lemma 10. First of all, consider the first-order derivative of g_w:

$$\frac{d}{dz} g_w(z) = \frac{w_1 \cdot \big[w_1 z + w_0(1-z)\big] - w_1 z \cdot (w_1 - w_0)}{\big[w_1 z + w_0(1-z)\big]^2} = \frac{w_1 w_0}{\big[w_1 z + w_0(1-z)\big]^2}.$$

We observe that g_w is monotone increasing as $\frac{d}{dz} g_w(z) > 0$ for all $z \in [0,1]$. Next, we consider the second-order derivative of g_w:

$$\frac{d^2}{dz^2} g_w(z) = \frac{2 w_0 w_1 \cdot (w_0 - w_1)}{\big[w_1 z + w_0(1-z)\big]^3} \begin{cases} > 0, & \forall z \in [0,1] \quad \text{if } w_0 > w_1, \\ = 0, & \forall z \in [0,1] \quad \text{if } w_0 = w_1, \\ < 0, & \forall z \in [0,1] \quad \text{if } w_0 < w_1. \end{cases}$$

Therefore,

$$\sup_{z\in[0,1]} \frac{d}{dz} g_w(z) = \begin{cases} \frac{d}{dz} g_w(z)\Big|_{z=1} = \frac{w_0}{w_1} & \text{if } w_0 > w_1, \\ \frac{d}{dz} g_w(z)\Big|_{z=0} = \frac{w_1}{w_0} & \text{if } w_0 \le w_1. \end{cases}$$

□

Lemma 11. *Let P, Q be joint distributions of $(X, Y) \in \mathcal{X} \times \{0,1\}$, and let $w_k = \frac{P[Y=k]}{Q[Y=k]}$ for $k \in \{0,1\}$. If P, Q satisfy the label shift assumption (Definition 7), i.e., if Assumptions (B1) and (B2) hold, then for any measurable function $f : \mathcal{X} \to \mathbb{R}$, the following two-sided inequality holds:*

$$\min_{k\in\{0,1\}} w_k \le \frac{\mathbb{E}_Q[f(X)]}{\mathbb{E}_P[f(X)]} \le \max_{k\in\{0,1\}} w_k. \tag{30}$$

Proof of Lemma 11. First of all, we observe that

$$\begin{aligned} \mathbb{E}_Q\left[f(X)\right] &= \mathbb{E}_Q\left[\mathbb{E}_Q\big[f(X) \mid Y\big]\right] && \text{by the law of total expectation} \\ &= \sum_{k=0}^{1} P_Q[Y=k] \cdot \mathbb{E}_Q\big[f(X) \mid Y\big] \\ &= \sum_{k=0}^{1} \big(w_k \cdot P_P[Y=k]\big) \cdot \mathbb{E}_P\big[f(X) \mid Y\big]. && \text{by definition of } w_k \text{ \& the label shift assumption} \end{aligned}$$

Thus, it follows that $\min_k w_k \cdot \mathbb{E}_P[f(X)] \le \mathbb{E}_Q[f(X)] \le \max_k w_k \cdot \mathbb{E}_P[f(X)]$. □

C.2 Completing the proof of Theorem 2

Proof of Theorem 2. This proof is presented in four steps. In Step 1, we establish a simple upper bound for the risk $R_Q(\hat{h}_Q; f)$ that consists of two error terms: the first term quantifies the error introduced by the estimated label shift correction, $\hat{g}$, while the second term quantifies the error due to the estimated recalibration function, $\hat{h}_P$. In Steps 2 and 3, we derive separate upper bounds for these two error terms. Finally, in Step 4, we combine the results from Steps 1-3 to obtain a comprehensive upper bound for R_Q, which concludes the proof.

Step 1. Decomposition of R_Q. Recalling the definition of the risk R_Q, cf. (5), we obtain the following inequality:

$$\begin{aligned} R_Q(\hat{h}_Q; f) &= \mathbb{E}_Q\left[\Big(\hat{h}_Q \circ f(X) - \mathbb{E}_Q[Y|f(X)]\Big)^2\right] \\ &= \mathbb{E}_Q\left[\Big(\hat{g} \circ \hat{h}_P \circ f(X) - g^* \circ \hat{h}_P \circ f(X) + g^* \circ \hat{h}_P \circ f(X) - \mathbb{E}_Q[Y|f(X)]\Big)^2\right] \\ &\overset{(a)}{\leq} 2 \cdot \Bigg\{ \underbrace{\mathbb{E}_Q\left[\Big(\hat{g} \circ \hat{h}_P \circ f(X) - g^* \circ \hat{h}_P \circ f(X)\Big)^2\right]}_{=:T_1} \qquad (31) \\ &\quad + \underbrace{\mathbb{E}_Q\left[\Big(g^* \circ \hat{h}_P \circ f(X) - \mathbb{E}_Q[Y|f(X)]\Big)^2\right]}_{=:T_2} \Bigg\}, \qquad (32) \end{aligned}$$

where (a) follows from the simple inequality $(a+b)^2 \leq 2(a^2+b^2)$ for all $a, b \in \mathbb{R}$.

In Step 2 and Step 3 of this proof, we establish separate upper bounds for the two terms, T_1, T_2.

Step 2. An upper bound for T_1. Recall from (13) and (16) that

$$g^*(z) = \frac{w_1^* z}{w_1^* z + w_0^*(1-z)} \qquad \text{where} \qquad w_k^* = \frac{Q[Y=k]}{P[Y=k]}, \ \ \forall k \in \{0,1\},$$

$$\hat{g}(z) = \frac{\hat{w}_1 z}{\hat{w}_1 z + \hat{w}_0(1-z)} \qquad \text{where} \qquad \hat{w}_k = \frac{\hat{Q}[Y=k]}{\hat{P}[Y=k]}, \ \ \forall k \in \{0,1\}.$$

Let

$$\rho_0 := \frac{\hat{w}_0}{w_0^*} \qquad \text{and} \qquad \rho_1 := \frac{\hat{w}_1}{w_1^*}. \qquad (33)$$

Then we observe that for any $z \in (0,1)$,

$$\begin{aligned} |\hat{g}(z) - g^*(z)| &= \left| \frac{\hat{w}_1 z}{\hat{w}_1 z + \hat{w}_0(1-z)} - \frac{w_1^* z}{w_1^* z + w_0^*(1-z)} \right| \\ &= \left| \frac{\big(\hat{w}_1 w_0^* - w_1^* \hat{w}_0\big) \cdot z(1-z)}{\big[\hat{w}_1 z + \hat{w}_0(1-z)\big] \cdot \big[w_1^* z + w_0^*(1-z)\big]} \right| \\ &\leq \left| \frac{\big(\hat{w}_1 w_0^* - w_1^* \hat{w}_0\big) \cdot z(1-z)}{\big(\hat{w}_1 w_0^* + w_1^* \hat{w}_0\big) \cdot z(1-z)} \right| \\ &= \left| \frac{\hat{w}_1 w_0^* - w_1^* \hat{w}_0}{\hat{w}_1 w_0^* + w_1^* \hat{w}_0} \right| \\ &= \frac{\big|\rho_0 - \rho_1\big|}{\rho_0 + \rho_1}. \end{aligned}$$

Moreover, $\hat{g}(0) = g^*(0) = 0$ and $\hat{g}(1) = g^*(1) = 1$. Letting $Z_{\hat{h}} := \hat{h}_P \circ f(X)$, we obtain

$$T_1 = \mathbb{E}_Q\left[\Big(\hat{g}(Z_{\hat{h}}) - g^*(Z_{\hat{h}})\Big)^2\right] \leq \left(\frac{\rho_0 - \rho_1}{\rho_0 + \rho_1}\right)^2. \qquad (34)$$

It remains to establish probabilistic tail bounds for ρ_0, ρ_1, which we will accomplish in Step 4 of this proof.

Step 3. An upper bound for T_2. We observe that

$$\begin{aligned}
T_2 &= \mathbb{E}_Q\left[\left(g^* \circ \hat{h}_P \circ f(X) - \mathbb{E}_Q[Y \mid f(X)]\right)^2\right] \\
&= \mathbb{E}_Q\left[\left(g^* \circ \hat{h}_P \circ f(X) - g^*\big(\mathbb{E}_P[Y \mid f(X)]\big)\right)^2\right] && \because \text{Label shift assumption, cf. (14)} \\
&\le \left(\frac{w^*_{\max}}{w^*_{\min}}\right)^2 \cdot \mathbb{E}_Q\left[\left(\hat{h}_P \circ f(X) - \mathbb{E}_P[Y \mid f(X)]\right)^2\right] && \because g^* \text{ is } \frac{w^*_{\max}}{w^*_{\min}}\text{-Lipschitz, cf. Lemma 10} \\
&\le \left(\frac{w^*_{\max}}{w^*_{\min}}\right)^2 \cdot w^*_{\max} \cdot \mathbb{E}_P\left[\left(\hat{h}_P \circ f(X) - \mathbb{E}_P[Y \mid f(X)]\right)^2\right] && \because \text{by Lemma 11} \\
&= \frac{{w^*_{\max}}^3}{{w^*_{\min}}^2} \cdot R_P\big(\hat{h}_P; f\big).
\end{aligned}$$

Step 4. Concluding the proof. For given $\delta \in (0,1)$, let[2] $\delta_1 = \delta_2 = \delta/4$ and $\delta_3 = \delta/2$. We observe that

$$\begin{aligned}
&n_P \ge c' \cdot |\mathcal{B}| \log\left(\frac{|\mathcal{B}|}{\delta_1}\right) && \Longrightarrow && P\left[\Phi_{\text{balance}}(\mathcal{B}, 2) = 1\right] \ge 1 - \delta_1 && \text{by Lemma 3} \\
&n_P \ge 2|\mathcal{B}| && \Longrightarrow && P\left[\Phi_{\text{approx}}(\mathcal{B}, \varepsilon_{\delta_2}) = 1\right] \ge 1 - \delta_2 && \text{by Lemma 4}
\end{aligned}$$

where $c' > 0$ is the universal constant that appears in Lemma 3 and

$$\varepsilon_{\delta_2} = \sqrt{\frac{1}{2(\lfloor n/|\mathcal{B}|\rfloor - 1)} \log\left(\frac{2|\mathcal{B}|}{\delta_2}\right)} + \frac{1}{\lfloor n/|\mathcal{B}|\rfloor}.$$

Furthermore, assuming

$$n_P \ge \frac{27}{\min\{p_0, p_1\}} \log\left(\frac{8}{\delta_3}\right) \qquad \text{and} \qquad n_Q \ge \frac{27}{\min\{q_0, q_1\}} \log\left(\frac{8}{\delta_3}\right),$$

we may define β_{δ_3} as a function of n_P, n_Q and δ_3 such that

$$\beta_{\delta_3} = \beta_{\delta_3}(n_P, n_Q) := 1 + \sqrt{\max\left\{\frac{1}{n_P \cdot \min\{p_0, p_1\}},\ \frac{1}{n_Q \cdot \min\{q_0, q_1\}}\right\} \cdot 27 \log\left(\frac{8}{\delta_3}\right)}. \quad (35)$$

Then it follows from Corollary 9 that

$$P\left(\Phi_{\text{ratio}}(\mathcal{D}_P, \mathcal{D}_Q; \beta_0) = 1\right) \ge 1 - \delta_3.$$

Observe that $\delta_1 = \frac{\delta}{4} < \frac{1}{4}$ and $|\mathcal{B}| \ge 4$, and thus, $\log\left(\frac{|\mathcal{B}|}{\delta_1}\right) \ge \log 16 \ge 2$. Let $c = c'$. Since $c' \ge 1$ and $\log\left(\frac{|\mathcal{B}|}{\delta_1}\right) \ge \log\left(\frac{16}{\delta}\right) = \log\left(\frac{8}{\delta_3}\right)$, we notice that

$$\begin{aligned}
&n_P \ge \max\left\{c,\ \frac{27}{\min\{p_0, p_1\}}\right\} \cdot |\mathcal{B}| \log\left(\frac{4|\mathcal{B}|}{\delta}\right) \\
&\qquad \Longrightarrow \qquad n_P \ge \max\left\{c' \cdot |\mathcal{B}| \log\left(\frac{|\mathcal{B}|}{\delta_1}\right),\ 2|\mathcal{B}|,\ \frac{27}{\min\{p_0, p_1\}} \log\left(\frac{8}{\delta_3}\right)\right\}.
\end{aligned}$$

In summary, we obtain that for any given $\delta \in (0,1)$,

$$\begin{aligned}
&n_P \ge \max\left\{c,\ \frac{27}{\min\{p_0, p_1\}}\right\} \cdot |\mathcal{B}| \log\left(\frac{4|\mathcal{B}|}{\delta}\right) \quad \text{and} \quad n_Q \ge \frac{27}{\min\{q_0, q_1\}} \log\left(\frac{16}{\delta}\right) \\
&\qquad \Longrightarrow \qquad P\left[\Phi_{\text{balance}}(\mathcal{B}, 2) = 1\ \&\ \Phi_{\text{approx}}(\mathcal{B}, \varepsilon_{\delta/4}) = 1\ \&\ \Phi_{\text{ratio}}(\mathcal{D}_P, \mathcal{D}_Q; \beta_{\delta/2}) = 1\right] \ge 1 - \delta.
\end{aligned} \quad (36)$$

[2]We remark that our decomposition of δ into $\delta_1, \delta_2, \delta_3$ is arbitrary, and is intended to simplify the subsequent analysis.

Conditioned on the event $\Phi_{\text{balance}}(\mathcal{B}, 2) = 1$ & $\Phi_{\text{approx}}(\mathcal{B}, \varepsilon_{\delta/4}) = 1$ & $\Phi_{\text{ratio}}(\mathcal{D}_P, \mathcal{D}_Q; \beta_{\delta/2}) = 1$,

$$T_1 \leq \left(\frac{\left|\rho_0 - \rho_1\right|}{\rho_0 + \rho_1}\right)^2 \leq \left(\frac{\beta_{\delta/2} - \frac{1}{\beta_{\delta/2}}}{\beta_{\delta/2} + \frac{1}{\beta_{\delta/2}}}\right)^2 \leq \left(\beta_{\delta/2} - 1\right)^2, \quad \because \text{(34); also, see (29)}$$

$$T_2 \leq \frac{{w^*_{\max}}^3}{{w^*_{\min}}^2} \cdot R_P\big(\hat{h}_P; f\big)$$

$$\leq \frac{{w^*_{\max}}^3}{{w^*_{\min}}^2} \cdot \left(\varepsilon^2_{\delta/4} + \frac{2}{|\mathcal{B}|}\right). \qquad \because \text{proof of Theorem 1; Lemmas 5 \& 6}$$

Note that if Assumption (A3) holds, then we additionally have

$$T_2 \leq \frac{{w^*_{\max}}^3}{{w^*_{\min}}^2} \cdot \left(\varepsilon^2_{\delta/4} + \frac{8K^2}{|\mathcal{B}|^2}\right).$$

Inserting these upper bounds for T_1 and T_2 into (31), (32) and recalling the expression for β in (35), we complete the proof. □

D Proof sketch of the argument in Remark 5

Recall our composite recalibration function,

$$\hat{h}_Q = \hat{g} \circ \hat{h}_P, \tag{37}$$

does not use the features in $\mathcal{D}_Q$. Specifically, $\hat{g}$ is parameterized by $\hat{w} = (\hat{w}_0, \hat{w}_1)$, which can be estimated using only labels in $\mathcal{D}_P$ and $\mathcal{D}_Q$, cf. (16). According to Theorem 2, $R_Q(\hat{h}_Q) = O(n_Q^{-1})$ with high probability for sufficiently large n_P.

Now suppose we are given an unlabeled target sample with unknown label shift. We can estimate w using the target features via a maximum likelihood label shift estimation approach [14], yielding $\hat{w}^{\text{ML}} = (\hat{w}_0^{\text{ML}}, \hat{w}_1^{\text{ML}})$. and the calibrated classifier $\hat{h} \circ f$. This results in a different composite recalibration function than Equation (37),

$$\hat{h}_Q^{\text{ML}} = \hat{g}^{\text{ML}} \circ \hat{h}_P, \tag{38}$$

where $\hat{g}^{\text{ML}} : [0, 1] \to [0, 1]$ is defined as $\hat{g}^{\text{ML}}(z) = \hat{w}_1^{\text{ML}} z / (\hat{w}_1^{\text{ML}} z + \hat{w}_0^{\text{ML}}(1 - z))$. We claim in Remark 5 that, for sufficiently large n_P, the composite recalibration function in Equation (38) achieves $R_Q(\hat{h}_Q^{\text{ML}}) = O(n_Q^{-1})$ with high probability, enjoying the same convergence rate as $\hat{h}_Q$ (Equation 37). Here we give a proof sketch.

Proof sketch. Suppose we use the maximum likelihood approach in [14] to estimate w. We want to show $R_Q(\hat{h}_Q^{\text{ML}}) = O(n_Q^{-1})$ with high probability for sufficiently large n_P. Recall $R_Q(\hat{h}_Q) \leq 2(T_1 + T_2)$ according to Equation (31) and (32), and label shift estimation error only affects T_1, so it is sufficient to show $T_1 = O(n_Q^{-1})$ with high probability.

For sufficiently large n_P, $\hat{h}_P \circ f$ is sufficient calibrated, so $\|\hat{w} - w\|_2^2 = O(n_Q^{-1})$ by Theorem 3 in [14]. Since

$$\|\hat{w} - w\|_2^2 = \sum_{k \in \{0,1\}} (\rho_k - 1)^2 w_k^2 \geq {w^*_{\min}}^2 \sum_{k \in \{0,1\}} (\rho_k - 1)^2 \geq {w^*_{\min}}^2 \max_{k \in \{0,1\}} (\rho_k - 1)^2, \tag{39}$$

we have $\rho_k \in [1-\alpha, 1+\alpha]$ for $k \in \{0, 1\}$, where $\alpha = \frac{\|\hat{w} - w\|_2}{w^*_{\min}} > 0$. For sufficiently small $\|\hat{w} - w\|_2^2$, we can control $\alpha < 0.5$, which bounds T_1 in (34):

$$T_1 \leq \left(\frac{\rho_0 - \rho_1}{\rho_0 + \rho_1}\right)^2 \leq \frac{2\alpha}{2 - 2\alpha} \leq 2\alpha = 2\frac{\|\hat{w} - w\|_2}{w^*_{\min}} = O(n_Q^{-1}). \tag{40}$$

The rest of the proof are the same with Appendix C.2. □

E Details on the experiments

In Section E.1 and E.3, we consider a family of joint distributions $\mathcal{D}(\pi)$ of X and Y, where $Y \sim \text{Bernoulli}(\pi)$, $X \mid Y = 0 \sim N(-2, 1)$, and $X \mid Y = 1 \sim N(2, 1)$. Suppose we are given $f(x) = \sigma(x) := 1/(1 + e^{-x})$, for $x \in \mathbb{R}$, as a probabilistic classifier. The optimal recalibration function can be derived as

$$h^*_{f,P}(z) = P[Y = 1 \mid f(X) = z] = \sigma(4\sigma^{-1}(z)). \tag{41}$$

In Section E.2, we consider a parametric family of recalibration functions called beta calibration [28]: $\mathcal{H}_{\text{beta}} = \{h_{\text{beta}}(\cdot; a, b, c) : a \geq 0, b \geq 0, c \in \mathbb{R}\}$, where $h_{\text{beta}}(\cdot; a, b, c) : [0, 1] \to [0, 1]$ is defined as

$$h_{\text{beta}}(z; a, b, c) = \frac{1}{1 + 1/\left(e^c \frac{z^a}{(1-z)^b}\right)}. \tag{42}$$

In addition, consider a subfamily $\mathcal{H}_{\text{logit-normal}} \subset \mathcal{H}_{\text{beta}}$ defined as $\mathcal{H}_{\text{logit-normal}} = \{h_{\text{logit-normal}}(\cdot; a, c) := h_{\text{beta}}(\cdot; a, a, c) = \sigma(a\sigma^{-1}(\cdot) + c) : a \geq 0, c \in \mathbb{R}\}$[3]. Apparently, the optimal recalibration function in Equation (41), $h^*_{f,P} \in \mathcal{H}_{\text{logit-normal}}$.

E.1 Verifying results for UMB

First, we recalibrate f on data distributed as $\mathcal{D}(0.5)$ using UMB.

Verifying the risk convergence in Theorem 1 We vary $n \in [10^2, 10^7]$ and $B \in [6, 10^3]$ in the log scale. For each combination of (n, B), we use UMB to recalibrate f on data generated from $\mathcal{D}(0.5)$, and compute quadrature estimates of population $R^{\text{cal}}(\hat{h})$, $R^{\text{sha}}(\hat{h})$, and $R(\hat{h})$, as well as their high probability bounds based on Theorem 1. The constant K in Assumption `(A3)` is selected by numerical maximization as

$$K = \max_{0 \leq z_1 < z_2 \leq 1} \frac{h^*(z_2) - h^*(z_1)}{P[Z \in [z_1, z_2]]}.$$

Figure 1 shows the bounds follow the same trends as their associated population quantities, providing valid upper bounds in all cases.

Verifying the optimal choice of the number of bins. We find empirically optimal $B^{*\text{experiment}}$ that achieves the minimal risk for each choice of n. We compute the theoretically optimal choice of the number of bins, $B^{*\text{theory}}$, by minimizing the finite-sample upper bounds. Figure 2 shows $B^{*\text{experiment}}$ follows the same trend with $B^{*\text{theory}}$, both scales in $O(n^{1/3})$.

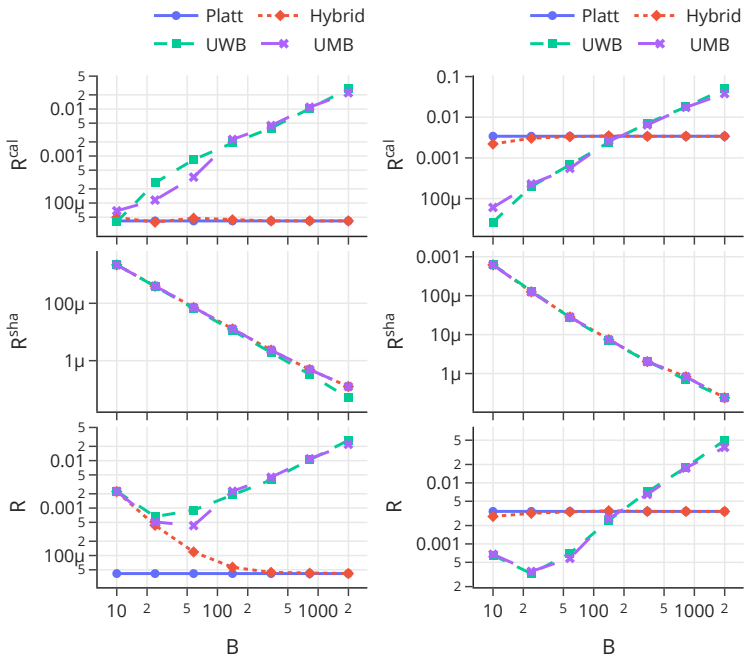
E.2 Comparing recalibration methods

To highlight the benefits and drawbacks of UMB's nonparametric nature, we compare UMB with (semi-)parametric recalibration methods in scenarios where the parametric assumption is correct and where it is misspecified. We compare the method under study, uniform-mass binning (UMB), with 3 other recalibration methods: uniform-width binning (UWB) [18], Platt scaling [40][4], and a hybrid parametric-binning method [29]. Note that Platt scaling and the hybrid method adopt the parametric assumption $h^* \in \mathcal{H}_{\text{logit-normal}}$.

For the first setting, we construct optimal recalibration function $h^* \in \mathcal{H}_{\text{logit-normal}}$ so that the parametric assumption of Platt scaling and the hybrid method holds. In particular, we consider the distribution $Z \in \text{Uniform}[0, 1]$ and $Y \mid Z \sim \text{Bernoulli}(h_{\text{logit-normal}}(Z; a, c))$ with $a = 4$ and $c = 0$. For the second setting, we construct $h^* \in \mathcal{H}_{\text{beta}}$ but $h^* \notin \mathcal{H}_{\text{logit-normal}}$ so that the parametric assumption fails. In particular, we consider the distribution $Z \sim \text{Uniform}[0, 1]$ and $Y \mid Z \sim \text{Bernoulli}(h_{\text{beta}}(z; a, b, c))$ with $a = 0.1$, $b = 4$, and $c = 0$. For each setting, we fix calibration sample size to be $n = 5000$.

[3]We say $Z \sim \text{Logit-Normal}(\mu, \tau^2)$ if $\sigma^{-1}(Z) \sim N(\mu, \tau^2)$ [2]. Similar to beta calibration [28], we adopt the name "logit-normal calibration" after a simple example: if $Y = \text{Bernoulli}(0.5)$, $Z \mid Y = i \sim \text{Logit-Normal}(\mu_i, \tau_i^2)$ for $i \in \{0, 1\}$, then the optimal recalibration function $\mathbb{E}[Y \mid Z = z] = h_{\text{logit-normal}}(z; a, c)$ for some a, c depending on μ_i's and τ_i's.

[4]The original Platt scaling operates on outputs of real-valued SVM outputs [40]. For probabilistic classifiers, we follow [38, 29, 22] and implement Platt scaling by first transforming probabilities onto the real line via the logit transform σ^{-1}.



(a) Correct parametric assumption (b) Misspecified parametric assumption

Figure 5: Risks vs. number of bins B.

Risks as functions of the number of bins B We traverse the number of bins $B \in [10, 2000]$ in the log scale and compare how each method behaves as B changes. When the parametric assumption is correct, the hybrid method achives significantly lower R^{cal} and overall R than UMB and UWB for sufficiently large number of bins (Figure 5a), an advantage highlighted in [29]. In contrast, when the parametric assumption fails, the binning methods UMB and UWB has better performance with the optimal number of bins (Figure 5b). This is because Platt scaling and hybrid methods are intrinsicly biased when $h^* \notin \mathcal{H}_{\text{logit-normal}}$, as noted in Section 4.2.

Quantitative results of risks under optimal B For each setting, we fix B that achieves low recalibration risk for UWB and UMB in Figure 5. Specifically, we choose $B = 2\lfloor n^{1/3} \rfloor = 34$ for the correct parametric assumption setting, and $B = \lfloor n^{1/3} \rfloor = 17$ for the misspecified parametric assumption setting. Then, for each setting, we compare the 90% quantiles of risks of each recalibration method fitted on 100 random replicates of calibration datasets of size $n = 5000$.

Table 1 quantititively verifies that Platt and the Hybrid method achieves lower R^{cal} and overall R if the parametric assumption is correct, and UWB and UMB achieves lower R^{cal} and overall R when the parametric assumption fails.

Visualization of calibration curves We fix the calibration dataset and visualize the calibration curves for all methods under the two settings. Figure 3 shows that the binning methods (UWB and UMB) closely track the optimal recalibration function h^* in both settings. In contrast, the hybrid approach follows the Platt scaling estimates, leading to an inherent bias from h^* when the parametric assumption is invalid (Figure 3b).

E.3 Comparing recalibration schemes under label shift

We consider the label shift with source distribution $\mathcal{D}(0.5)$ and target distribution $\mathcal{D}(\pi_Q)$, where π_Q varies in $\{0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5\}$. The results where $\pi_Q > 0.5$ can be inferred by symmetry and hence not experimented. We vary n_P in $\{10, 10^3, 10^5, 10^7\}$ and n^Q in $\{10, 10^3, 10^5\}$.

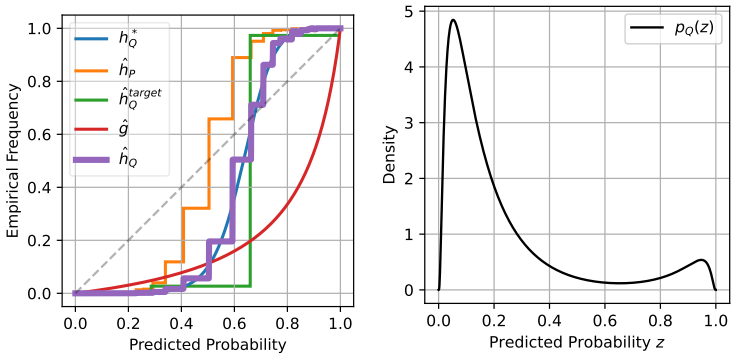


Figure 6: *Left*: calibration curves of for COMPOSITE $\hat{h}_Q$, SOURCE $\hat{h}_P$, TARGET $\hat{h}_Q^{\text{target}}$, and LABEL-SHIFT $\hat{g}$. *Right*: the marginal density of $Z = f(X)$ under Q.

Aside from our proposed recalibration function $\hat{h}_Q = \hat{g} \circ \hat{h}_P$ (17), referred to as COMPOSITE, we consider three other calibration approaches as baselines: (1) SOURCE, denoted as $\hat{h}_P$, which is only calibrated on the source data, (2) LABEL-SHIFT, denoted as $\hat{g}$, which performs label shift correction without calibration, and (3) TARGET, denoted as $\hat{g}_Q^{\text{target}}$, which is only calibrated on the target data. The number of bins B are chosen to be $n_P^{1/3}$ for COMPOSITE and SOURCE, and $n_Q^{1/3}$ for TARGET.

Table 2 shows the risks for different approaches with $\pi_Q = 0.1$, $n_P = 10^3$, and $n_Q = 10^2$. In terms of R^{cal}, COMPOSITE performs the best, as it is calibrated to the target distribution by taking advantage of the abundant source data. In terms of R^{sha}, LABEL-SHIFT achieves $R^{\text{sha}} = 0$ due to the strictly increasing $\hat{g}$, but it suffers from high R^{cal}. COMPOSITE and SOURCE achieve smaller R^{sha} than TARGET, as a result of using more bins on a larger sample. Considering the combined impact of calibration and sharpness, our approach COMPOSITE attains the lowest overall recalibration risk R as well as MSE.

Figure 6 shows the optimal recalibration function h^* and the recalibration functions for the four approaches. It can be seen that COMPOSITE best estimates h^* with the highest resolution.