A Additional derivations

A.1 Derivation of the MSE decomposition

Definition A.1 (Mean Squared Error (MSE))**.** The mean squared error of an estimator is

$$\mathrm{MSE}(f) := \mathbb{E}[(f(x)-y)^2]. \tag{11}$$

Proposition A.2. $\mathrm{MSE}(f) \geq \mathrm{CE}_2(f)^2$

Proof.

$$\begin{aligned}
\mathrm{MSE}(f) :=& \mathbb{E}[(f(x)-y))^2] = \mathbb{E}[((f(x)-\mathbb{E}[y \mid f(x)]) + (\mathbb{E}[y \mid f(x)]-y))^2] && (12)\\
=& \underbrace{\mathbb{E}[(f(x)-\mathbb{E}[y \mid f(x)])^2]}_{=CE_2^2} + \mathbb{E}[(\mathbb{E}[y \mid f(x)]-y)^2] && (13)\\
&+ 2\mathbb{E}[(f(x)-\mathbb{E}[y \mid f(x)])(\mathbb{E}[y \mid f(x)]-y)]
\end{aligned}$$

which implies

$$\begin{aligned}
\mathrm{MSE}(f) - \mathrm{CE}_2(f)^2 =& \mathbb{E}[(\mathbb{E}[y \mid f(x)]-y)^2] && (14)\\
&+ 2\mathbb{E}[(f(x)-\mathbb{E}[y \mid f(x)])(\mathbb{E}[y \mid f(x)]-y)]\\
=& \mathbb{E}[(\mathbb{E}[y \mid f(x)]-y)^2] + 2\mathbb{E}[(f(x)\mathbb{E}[y \mid f(x)]] && (15)\\
&- 2\mathbb{E}[f(x)y] - 2\mathbb{E}[\mathbb{E}[y \mid f(x)]^2] + 2\mathbb{E}[\mathbb{E}[y \mid f(x)]y]]\\
=& \mathbb{E}[\mathbb{E}[y \mid f(x)]^2] + \mathbb{E}[y^2] - 2\mathbb{E}[\mathbb{E}[y \mid f(x)]y] && (16)\\
&+ 2\mathbb{E}[(f(x)\mathbb{E}[y \mid f(x)]] - 2\mathbb{E}[f(x)y]\\
&- 2\mathbb{E}[\mathbb{E}[y \mid f(x)]^2] + 2\mathbb{E}[\mathbb{E}[y \mid f(x)]y]]\\
=& \mathbb{E}[y^2] + 2\mathbb{E}[(f(x)\mathbb{E}[y \mid f(x)]] - 2\mathbb{E}[f(x)y] && (17)\\
&- \mathbb{E}[\mathbb{E}[y \mid f(x)]^2]\\
=& \mathbb{E}[(2f(x)-y-\mathbb{E}[y \mid f(x)])(\mathbb{E}[y \mid f(x)])-y] && (18)\\
=& \mathbb{E}[(f(x)-y)(\mathbb{E}[y \mid f(x)]-y)] && (19)\\
&+ \mathbb{E}[(f(x)-\mathbb{E}[y \mid f(x)])(\mathbb{E}[y \mid f(x)]-y)].
\end{aligned}$$

By the law of total expectation, we will write the above as

$$\begin{aligned}
\mathrm{MSE}(f) - \mathrm{CE}_2(f)^2 = \mathbb{E}[\mathbb{E}[&(f(x)-y)(\mathbb{E}[y \mid f(x)]-y) && (20)\\
&+ (f(x)-\mathbb{E}[y \mid f(x)])(\mathbb{E}[y \mid f(x)]-y) \mid f(x)]].
\end{aligned}$$

Focusing on the inner conditional expectation, we have that

$$\begin{aligned}
&\mathbb{E}[(f(x)-y)(\mathbb{E}[y \mid f(x)]-y) + (f(x)-\mathbb{E}[y \mid f(x)])(\mathbb{E}[y \mid f(x)]-y) \mid f(x)]\\
&\quad= \mathbb{E}[y \mid f(x)](f(x)-1)(\mathbb{E}[y \mid f(x)]-1) + (1-\mathbb{E}[y \mid f(x)])f(x)\mathbb{E}[y \mid f(x)]\\
&\qquad+ \mathbb{E}[y \mid f(x)](f(x)-\mathbb{E}[y \mid f(x)])(\mathbb{E}[y \mid f(x)]-1)\\
&\qquad+ (1-\mathbb{E}[y \mid f(x)])(f(x)-\mathbb{E}[y \mid f(x)])\mathbb{E}[y \mid f(x)] && (21)\\
&\quad= (1-\mathbb{E}[y \mid f(x)])\mathbb{E}[y \mid f(x)] \geq 0 \quad \forall f(x) && (22)
\end{aligned}$$

and therefore

$$\mathrm{MSE}(f) - \mathrm{CE}_2(f)^2 = \mathbb{E}[(1-\mathbb{E}[y \mid f(x)])\mathbb{E}[y \mid f(x)]] \geq 0. \tag{23}$$

□

The expectation in Equation (23) is over variances of Bernoulli random variables with probabilities $\mathbb{E}[y \mid f(x)]$.

A.2 Derivation of Equation (3)

By considering $y \in \{0, 1\}$, we have the following:

$$\begin{aligned}
\mathbb{E}[y \mid f(x)] &= \sum_{y_k \in \mathcal{Y}} y_k \, p_{y|f(x)}(y_k) = \frac{\sum_{y_k \in \mathcal{Y}} y_k \, p_{f(x),y}(f(x), y_k)}{p_{f(x)}(f(x))} && (24) \\
&= \frac{p_{f(x),y}(f(x), y_k = 1)}{p_{f(x)}(f(x))} = \frac{p_{f(x)|y}(f(x)|y_k = 1) p_y(y_k = 1)}{p_{f(x)}(f(x))} && (25) \\
&\approx \frac{\frac{1}{\sum_{i=1}^n y_i} \sum_{i=1}^n k(f(x); f(x_i)) y_i \frac{\sum_{i=1}^n y_i}{n}}{\frac{1}{n} \sum_{i=1}^n k(f(x); f(x_i))} && (26) \\
&\approx \frac{\sum_{i=1}^n k(f(x); f(x_i)) y_i}{\sum_{i=1}^n k(f(x); f(x_i))} =: \widehat{\mathbb{E}[y \mid f(x)]} && (27)
\end{aligned}$$

A.3 Derivation of Equation (6)

We consider the optimization problem for some $\lambda > 0$:

$$f = \arg\min_{f \in \mathcal{F}} \Big(\mathrm{MSE}(f) + \lambda \, \mathrm{CE}_2(f)^2 \Big). \tag{28}$$

Using Equation (23) we rewrite:

$$\begin{aligned}
\mathrm{MSE}(f) + \lambda \, \mathrm{CE}_2(f)^2 &= (1 + \lambda) \, \mathrm{MSE}(f) - \lambda \Big(\mathrm{MSE}(f) - \mathrm{CE}_2(f)^2 \Big) \\
&= (1 + \lambda) \, \mathrm{MSE}(f) - \lambda \mathbb{E}\Big[\Big(1 - \mathbb{E}[y \mid f(x)] \Big) \mathbb{E}[y \mid f(x)] \Big]. && (29)
\end{aligned}$$

Rescaling Equation (29) by a factor of $(1 + \lambda)^{-1}$ and a variable substitution $\gamma = \frac{\lambda}{1+\lambda} \in [0, 1)$, we have that:

$$\begin{aligned}
f &= \arg\min_{f \in \mathcal{F}} \Big(\mathrm{MSE}(f) + \lambda \, \mathrm{CE}_2(f)^2 \Big) \\
&= \arg\min_{f \in \mathcal{F}} \Big(\mathrm{MSE}(f) - \gamma \mathbb{E}\Big[\Big(1 - \mathbb{E}[y \mid f(x)] \Big) \mathbb{E}[y \mid f(x)] \Big] \Big) \\
&= \arg\min_{f \in \mathcal{F}} \Big(\mathrm{MSE}(f) + \gamma \mathbb{E}\Big[\mathbb{E}[y \mid f(x)]^2 \Big] \Big). && (30)
\end{aligned}$$

B Bias of ratio of U-statistics

The unbiased estimator for the square of a mean μ_X^2 is given by:

$$\widehat{\mu_X^2} = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n X_i X_j = \frac{1}{n(n-1)} \left(\left(\sum_{i=1}^n X_i \right)^2 - \sum_{i=1}^n X_i^2 \right). \tag{31}$$

This is a second order U-statistics with kernel $h(x_1, x_2) = x_1 x_2$. The bias of the ratio of two of these estimators converges as $\mathcal{O}\left(\frac{1}{n}\right)$, as the following lemma proves.

Lemma B.1. *Let θ_1 and θ_2 be two estimable parameters and let U_1 and U_2 be the two corresponding U-statistics of order m_1 and m_2, respectively, based on a sample of n i.i.d. RVs. The bias of the ratio U_1/U_2 of these two U-statistics will converge as $\mathcal{O}\left(\frac{1}{n}\right)$.*

Proof. Let $R = \theta_1/\theta_2$ be the ratio of two estimable parameters and $r = U_1/U_2$ the ratio of the corresponding U-statistics. Note, that U_i is an unbiased estimator of θ_i, $\mathbb{E}[U_i] = \theta_i$, $i = 1, 2$, however, the ratio is usually biased. To investigate the bias of that ratio we rewrite

$$r = R \left(1 + \frac{U_1 - \theta_1}{\theta_1} \right) \left(1 + \frac{U_2 - \theta_2}{\theta_2} \right)^{-1}. \tag{32}$$

If $\left|\frac{U_2-\theta_2}{\theta_2}\right| < 1$, we can expand $\left(1+\frac{U_2-\theta_2}{\theta_2}\right)^{-1}$ in a geometric series:

$$r = R\left(1+\frac{(U_1-\theta_1)}{\theta_1}\right)\left(1-\frac{(U_2-\theta_2)}{\theta_2}+\frac{(U_2-\theta_2)^2}{\theta_2^2}-\frac{(U_2-\theta_2)^3}{\theta_2^3}+\frac{(U_2-\theta_2)^4}{\theta_2^4}-...\right) \tag{33}$$

$$\begin{aligned} &= R\left(1+\frac{(U_1-\theta_1)}{\theta_1}-\frac{(U_2-\theta_2)}{\theta_2}-\frac{(U_2-\theta_2)(U_1-\theta_1)}{\theta_2\theta_1}\right. \\ &\quad +\frac{(U_2-\theta_2)^2}{\theta_2^2}+\frac{(U_2-\theta_2)^2(U_1-\theta_1)}{\theta_2^2\theta_1}-\frac{(U_2-\theta_2)^3}{\theta_2^3}-\frac{(U_2-\theta_2)^3(U_1-\theta_1)}{\theta_2^3\theta_1} \\ &\quad \left.+\frac{(U_2-\theta_2)^4}{\theta_2^4}+\frac{(U_2-\theta_2)^4(U_1-\theta_1)}{\theta_2^4\theta_1}-...\right). \end{aligned} \tag{34}$$

If $\zeta_1 > 0$, a U-statistic U of order m obtained from a sample of n observations converges in distribution [Shao, 2003]:

$$\sqrt{n}\,(U-\mathbb{E}[U]) \xrightarrow{\mathrm{d}} N(0, m^2\zeta_1). \tag{35}$$

Keeping the terms up to $\Theta\left(\frac{1}{n}\right)$:

$$r = R\left(1+\frac{(U_1-\theta_1)}{\theta_1}-\frac{(U_2-\theta_2)}{\theta_2}-\frac{(U_2-\theta_2)(U_1-\theta_1)}{\theta_2\theta_1}+\frac{(U_2-\theta_2)^2}{\theta_2^2}+o\left(\frac{1}{n}\right)\right) \tag{36}$$

To examine the bias, we take the expectation value of this expression:

$$\mathbb{E}[r] = R\left(1+\frac{\mathbb{E}\big[(U_1-\theta_1)\big]}{\theta_1}-\frac{\mathbb{E}\big[(U_2-\theta_2)\big]}{\theta_2}-\frac{\mathbb{E}\big[(U_2-\theta_2)(U_1-\theta_1)\big]}{\theta_2\theta_1}+\frac{\mathbb{E}\big[(U_2-\theta_2)^2\big]}{\theta_2^2}+o\left(\frac{1}{n}\right)\right) \tag{37}$$

We now make use of the following expressions:

$$\mathbb{E}\big[(U_1-\theta_1)\big] = \mathbb{E}\big[(U_2-\theta_2)\big] = 0 \tag{38}$$

$$\mathbb{E}\big[(U_2-\theta_2)(U_1-\theta_1)\big] = \mathrm{Cov}(U_2, U_1) \tag{39}$$

$$\mathbb{E}\big[(U_2-\theta_2)^2\big] = \mathrm{Var}(U_2) \tag{40}$$

$$\tag{41}$$

Using these expressions the expectation of r becomes:

$$\mathbb{E}[r] = R\left(1-\frac{\mathrm{Cov}(U_2,U_1)}{\theta_2\theta_1}+\frac{\mathrm{Var}(U_2)}{\theta_2^2}+o\left(\frac{1}{n}\right)\right) \tag{42}$$

Using Equation (35), the linearity of covariance and with $\mathrm{Var}(aX) = a^2\,\mathrm{Var}(X)$ we obtain:

$$\mathrm{Cov}(U_2,U_1), \mathrm{Var}(U_2) \in \mathcal{O}\left(\frac{1}{n}\right) \implies \mathbb{E}[r] = R\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right). \tag{43}$$

□

C De-biasing of ratios of straight averages

Let X and Y be random variables and let μ_X and μ_Y be the means of their distributions, respectively. Consider the problem of finding an unbiased estimator for the ratio of means:

$$R = \frac{\mu_Y}{\mu_X}. \tag{44}$$

A first approach to estimate this ratio R is to compute the ratio of the sample means: Let $(X_1, Y_1), ..., (X_n, Y_n)$ be pairs of i.i.d. random variables that are jointly distributed:

$$r = \hat{R} = \frac{\hat{\mu_Y}}{\hat{\mu_X}} = \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i}{\frac{1}{n}\sum_{i=1}^{n} X_i} = \frac{\bar{Y}}{\bar{X}}. \tag{45}$$

This, however, is a biased estimator, which can be seen as follows (we follow [Tin, 1965, Ogliore et al., 2011] here):

$$r = \frac{\bar{Y}}{\bar{X}} = \frac{\mu_Y}{\mu_X}\left(\frac{\bar{Y}}{\mu_Y}\right)\left(\frac{\bar{X}}{\mu_X}\right)^{-1} = R\left(1 + \frac{\bar{Y} - \mu_Y}{\mu_Y}\right)\left(1 + \frac{\bar{X} - \mu_X}{\mu_X}\right)^{-1}. \tag{46}$$

This has now the form of a converging geometric series. Thus, if

$$\left|\frac{\bar{X} - \mu_X}{\mu_X}\right| < 1, \tag{47}$$

we can expand $\left(1 + \frac{\bar{X} - \mu_X}{\mu_X}\right)^{-1}$ in a geometric series, which is defined as:

$$\sum_{k=0}^{\infty} a\, b^k = a + ab + ab^2 + ... = \frac{a}{1 - b}. \tag{48}$$

In our case we can identify $a = R\left(1 + \frac{\bar{Y} - \mu_Y}{\mu_Y}\right)$ and $b = -\frac{\bar{X} - \mu_X}{\mu_X}$.
Thus, using the geometric series expansion, we can write:

$$r = R\left(1 + \frac{\bar{Y} - \mu_Y}{\mu_Y}\right)\left(1 - \frac{(\bar{X} - \mu_X)}{\mu_X} + \frac{(\bar{X} - \mu_X)^2}{\mu_X^2} - \frac{(\bar{X} - \mu_X)^3}{\mu_X^3} + \frac{(\bar{X} - \mu_X)^4}{\mu_X^4} - ...\right) \tag{49}$$

$$\begin{aligned} = R\Bigg(1 &+ \frac{(\bar{Y} - \mu_Y)}{\mu_Y} - \frac{(\bar{X} - \mu_X)}{\mu_X} - \frac{(\bar{X} - \mu_X)(\bar{Y} - \mu_Y)}{\mu_Y \mu_X} + \frac{(\bar{X} - \mu_X)^2}{\mu_X^2} \\ &+ \frac{(\bar{X} - \mu_X)^2(\bar{Y} - \mu_Y)}{\mu_X^2 \mu_Y} - \frac{(\bar{X} - \mu_X)^3}{\mu_X^3} - \frac{(\bar{X} - \mu_X)^3(\bar{Y} - \mu_Y)}{\mu_X^3 \mu_Y} + \frac{(\bar{X} - \mu_X)^4}{\mu_X^4} + ...\Bigg) \end{aligned} \tag{50}$$

Neglecting higher order terms Since $\bar{X}$ and $\bar{Y}$ are U-statistics, we make use of the asymptotic behaviour of U-statistics. If $\zeta_1 > 0$, a U-statistics U_n of order m obtained from a sample of n observations behaves as $n \to \infty$ like ([Shao, 2003]):

$$\sqrt{n}\,(U_n - \mathbb{E}[U_n]) \xrightarrow{\mathrm{d}} N(0, m^2\zeta_1). \tag{51}$$

As we seek an estimator that is unbiased up until order n^{-2} and since $\mathbb{E}[\bar{X}] = \mu_X$, we can neglect all terms of order 5 or higher since for $n \to \infty$:

$$(\bar{X} - \mu_X)^5 \in \mathcal{O}(n^{-2.5}) \tag{52}$$

$$(\bar{X} - \mu_X)^4(\bar{Y} - \mu_Y) \in \mathcal{O}(n^{-2.5}) \tag{53}$$

Therefore, we obtain:

$$\begin{aligned} r \approx R\Bigg(1 &+ \frac{(\bar{Y} - \mu_Y)}{\mu_Y} - \frac{(\bar{X} - \mu_X)}{\mu_X} - \frac{(\bar{X} - \mu_X)(\bar{Y} - \mu_Y)}{\mu_Y \mu_X} + \frac{(\bar{X} - \mu_X)^2}{\mu_X^2} \\ &+ \frac{(\bar{X} - \mu_X)^2(\bar{Y} - \mu_Y)}{\mu_X^2 \mu_Y} - \frac{(\bar{X} - \mu_X)^3}{\mu_X^3} - \frac{(\bar{X} - \mu_X)^3(\bar{Y} - \mu_Y)}{\mu_X^3 \mu_Y} + \frac{(\bar{X} - \mu_X)^4}{\mu_X^4}\Bigg) \end{aligned} \tag{54}$$

Identities to compute the terms of the series expansion of r

$$\mathbb{E}[\bar{X}-\mu_X] = \mathbb{E}[\bar{Y}-\mu_Y] = 0 \tag{55}$$

$$\mathbb{E}[(\bar{X}-\mu_X)^2] = \mathrm{Var}(\bar{X}) = \frac{1}{n}\,\mathrm{Var(X)} \tag{56}$$

$$\mathbb{E}\Big[\Big(\bar{X}-\mu_X\Big)\Big(\bar{Y}-\mu_Y\Big)\Big] = \mathrm{Cov}(\bar{X},\bar{Y}) = \frac{1}{n}\,\mathrm{Cov}(X,Y) \tag{57}$$

$$\mathbb{E}\Big[\Big(\bar{X}-\mu_X\Big)^2\Big(\bar{Y}-\mu_Y\Big)\Big] = \mathrm{Cov}(\bar{X}^2,\bar{Y}) - 2\mu_X\,\mathrm{Cov}(\bar{X},\bar{Y}) \tag{58}$$

$$= \frac{1}{n^2}\Big(\mathrm{Cov}(X^2,Y) - 2\mu_X\,\mathrm{Cov}(X,Y)\Big) \tag{59}$$

$$\mathbb{E}\Big[\Big(\bar{X}-\mu_X\Big)^3\Big] = \mathrm{Cov}(\bar{X}^2,\bar{X}) - 2\mu_X\,\mathrm{Var}(\bar{X}) \tag{60}$$

$$= \frac{1}{n^2}\,\mathrm{Cov}(X^2,X) - \frac{2}{n^2}\mu_X\,\mathrm{Var}(X) \tag{61}$$

$$\mathbb{E}\Big[\Big(\bar{X}-\mu_X\Big)^3\Big(\bar{Y}-\mu_Y\Big)\Big] = \mathrm{Cov}(\bar{X}^3,\bar{Y}) - 3\mu_X\,\mathrm{Cov}(\bar{X}^2,\bar{Y}) + 3\mu_X^2\,\mathrm{Cov}(\bar{X},\bar{Y}) \tag{62}$$

$$= \frac{3}{n^2}\,\mathrm{Var(X)}\,\mathrm{Cov}(X,Y) + \mathcal{O}(n^{-3}) \tag{63}$$

$$\mathbb{E}\Big[\Big(\bar{X}-\mu_X\Big)^4\Big] = \mathrm{Cov}(\bar{X}^3,\bar{X}) - 3\mu_X\,\mathrm{Cov}(\bar{X}^2,\bar{X}) + 3\mu_X^2\,\mathrm{Var}(\bar{X}) \tag{64}$$

$$= \frac{3}{n^2}\,\mathrm{Var}(X)^2 + \mathcal{O}(n^{-3}) \tag{65}$$

Bias Using these expressions we can compute the expectation value of $r = \hat{R}$:

$$\mathbb{E}[r] \approx R\Bigg(1 + \frac{1}{n}\Bigg(\frac{\mathrm{Var(X)}}{\mu_X^2} - \frac{\mathrm{Cov}(X,Y)}{\mu_X\mu_Y}\Bigg) + \frac{1}{n^2}\Bigg(\frac{(\mathrm{Cov}(X^2,Y) - 2\mu_X\,\mathrm{Cov}(X,Y))}{\mu_X^2\mu_Y} - \frac{(\mathrm{Cov}(X^2,X) - 2\mu_X\,\mathrm{Var}(X))}{\mu_X^3} - \frac{3\,\mathrm{Var}(X)\,\mathrm{Cov}(X,Y)}{\mu_X^3\mu_Y} + \frac{3\,\mathrm{Var}(X)^2}{\mu_X^4}\Bigg)\Bigg) \tag{66}$$

The bias or $r = \widehat{R}$ is defined as:

$$\mathrm{Bias}(r) = \mathbb{E}[r] - R \tag{67}$$

$$= R\Bigg(\frac{1}{n}\Bigg(\frac{\mathrm{Var}(X)}{\mu_X^2} - \frac{\mathrm{Cov}(X,Y)}{\mu_X\mu_Y}\Bigg) + \frac{1}{n^2}\Bigg(\frac{(\mathrm{Cov}(X^2,Y) - 2\mu_X\,\mathrm{Cov}(X,Y))}{\mu_X^2\mu_Y} \tag{68}$$

$$- \frac{(\mathrm{Cov}(X^2,X) - 2\mu_X\,\mathrm{Var}(X))}{\mu_X^3} - \frac{3\,\mathrm{Var}(X)\,\mathrm{Cov}(X,Y)}{\mu_X^3\mu_Y} + \frac{3\,\mathrm{Var}(X)^2}{\mu_X^4}\Bigg)\Bigg) \tag{69}$$

Therefore an unbiased version of r is:

$$r_{\text{unbiased}} = r - R\Bigg(\frac{1}{n}\Bigg(\frac{\mathrm{Var}(X)}{\mu_X^2} - \frac{\mathrm{Cov}(X,Y)}{\mu_X\mu_Y}\Bigg) + \frac{1}{n^2}\Bigg(\frac{(\mathrm{Cov}(X^2,Y) - 2\mu_X\,\mathrm{Cov}(X,Y))}{\mu_X^2\mu_Y} \tag{70}$$

$$- \frac{(\mathrm{Cov}(X^2,X) - 2\mu_X\,\mathrm{Var}(X))}{\mu_X^3} - \frac{3\,\mathrm{Var}(X)\,\mathrm{Cov}(X,Y)}{\mu_X^3\mu_Y} + \frac{3\,\mathrm{Var}(X)^2}{\mu_X^4}\Bigg)\Bigg) \tag{71}$$

A corrected version of the estimator $r = \hat{R}$ is consequently given by:

$$r_{corr} := r\Bigg(1 - \frac{1}{n}\Bigg(\frac{\widehat{\mathrm{Var}(X)}}{\widehat{\mu_X^2}} - \frac{\widehat{\mathrm{Cov}(X,Y)}}{\widehat{\mu_X\mu_Y}}\Bigg) - \frac{1}{n^2}\Bigg(\frac{(\widehat{\mathrm{Cov}(X^2,Y)} - 2\widehat{\mu_X}\widehat{\mathrm{Cov}(X,Y)})}{\widehat{\mu_X^2}\widehat{\mu_Y}} - \frac{(\widehat{\mathrm{Cov}(X^2,X)} - 2\widehat{\mu_X}\widehat{\mathrm{Var}(X)})}{\widehat{\mu_X^3}} - \frac{3\widehat{\mathrm{Var}(X)}\widehat{\mathrm{Cov}(X,Y)}}{\widehat{\mu_X^3}\widehat{\mu_Y}} + \frac{3\widehat{\mathrm{Var}(X)}^2}{\widehat{\mu_X^4}}\Bigg)\Bigg) \tag{72}$$

In the above equation we again encounter rations of estimators which again might be biased. Since we want to achieve a second order de-biasing we have to again recurse on the terms that have a $\mathcal{O}\left(\frac{1}{n}\right)$ dependency. However, we do not have to recurse on the terms that have a $\mathcal{O}\left(\frac{1}{n^2}\right)$ dependency, since any recursion would increase the power of the n-dependency. Therefore a debiased estimator up to order $\mathcal{O}(n^2)$ is:

$$\begin{aligned} r_{corr} := r\Bigg(1 - \frac{1}{n}\left(r_b^* - r_a^*\right) - \frac{1}{n^2}\Bigg(&\frac{(\widehat{\mathrm{Cov}(X^2,Y)} - 2\widehat{\mu_X}\widehat{\mathrm{Cov}(X,Y)})}{\widehat{\mu_X^2}\widehat{\mu_Y}} \\ &- \frac{(\widehat{\mathrm{Cov}(X^2,X)} - 2\widehat{\mu_X}\widehat{\mathrm{Var}(X)})}{\widehat{\mu_X^3}} - \frac{3\widehat{\mathrm{Var}(X)}\widehat{\mathrm{Cov}(X,Y)}}{\widehat{\mu_X^3}\widehat{\mu_Y}} + \frac{3\widehat{\mathrm{Var}(X)}^2}{\widehat{\mu_X^4}}\Bigg)\Bigg) \end{aligned} \tag{73}$$

where

$$\begin{aligned} r_a^* = \underbrace{\frac{\widehat{\mathrm{Cov}(X,Y)}}{\widehat{\mu_X\mu_Y}}}_{=r_a}\Bigg(1 + \frac{1}{(n-1)}\Bigg(&\frac{\widehat{\mu_Y}\widehat{\mathrm{Cov}(X^2,Y)} + \widehat{\mu_X}\widehat{\mathrm{Cov}(Y^2,X)}}{\widehat{\mathrm{Cov}(X,Y)}\widehat{\mu_X}\widehat{\mu_Y}} - 4\Bigg) \\ &- \frac{1}{(n-1)}\left(\frac{\widehat{\mathrm{Var}(X)}}{\widehat{\mu_X^2}} + \frac{\widehat{\mathrm{Var}(Y)}}{\widehat{\mu_Y^2}} + 2\frac{\widehat{\mathrm{Cov}(X,Y)}}{\widehat{\mu_X}\widehat{\mu_Y}}\right)\Bigg) \end{aligned} \tag{74}$$

$$r_b^* = \underbrace{\frac{\widehat{\mathrm{Var}(X)}}{\widehat{\mu_X^2}}}_{=r_b}\left(1 + \frac{4}{(n-1)}\left(\frac{\frac{1}{2}\widehat{\mathrm{Cov}(X^2,X)}}{\widehat{\mu_X}\widehat{\mathrm{Var}(X)}} - 1\right) - \frac{4}{(n-1)}\frac{\widehat{\mathrm{Var}(X)}}{\widehat{\mu_X^2}}\right). \tag{75}$$

D De-biasing of ratios of squared means

Now consider the problem of finding an unbiased estimator for the ratio of the squared means of x and Y:

$$R = \frac{\mu_Y^2}{\mu_X^2}. \tag{76}$$

Both the numerator and denominator of R can separately be estimated by a second order U-statistics, respectively:

$$r = \hat{R} = \frac{\widehat{\mu_Y^2}}{\widehat{\mu_X^2}} = \frac{\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1\wedge j\neq i}^{n} Y_iY_j}{\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1\wedge j\neq i}^{n} X_iX_j} =: \frac{\bar{Y}_2}{\bar{X}_2}. \tag{77}$$

The subscript 2 in $\bar{X}_2$ should emphasize that we are dealing with a second order U-statistics here. Again, the ratio $\frac{\bar{Y}_2}{\bar{X}_2}$,is a biased estimator. Using the approach with the converging geometric series and neglecting the higher order terms, we obtain:

$$\begin{aligned} r \approx R\Bigg(1 + &\frac{(\bar{Y}_2 - \mu_Y^2)}{\mu_Y^2} - \frac{(\bar{X}_2 - \mu_X^2)}{\mu_X^2} - \frac{(\bar{X}_2 - \mu_X^2)(\bar{Y}_2 - \mu_Y^2)}{\mu_Y^2\mu_X^2} + \frac{(\bar{X}_2 - \mu_X^2)^2}{\mu_X^4} \\ &+ \frac{(\bar{X}_2 - \mu_X^2)^2(\bar{Y}_2 - \mu_Y^2)}{\mu_X^4\mu_Y^2} - \frac{(\bar{X}_2 - \mu_X^2)^3}{\mu_X^6} - \frac{(\bar{X}_2 - \mu_X^2)^3(\bar{Y}_2 - \mu_Y^2)}{\mu_X^6\mu_Y^2} + \frac{(\bar{X}_2 - \mu_X^2)^4}{\mu_X^8}\Bigg) \end{aligned} \tag{78}$$

Identities to compute the terms of the series expansion of r

$$\begin{aligned}
\mathbb{E}[\bar{X}_2 - \mu_X^2] &= \mathbb{E}[\bar{Y}_2 - \mu_Y^2] = 0 &\quad (79)\\
\mathbb{E}[(\bar{X}_2 - \mu_X^2)^2] &= \operatorname{Var}(\bar{X}_2) &\quad (80)\\
\mathbb{E}\Big[\Big(\bar{X}_2 - \mu_X^2\Big)\Big(\bar{Y}_2 - \mu_Y^2\Big)\Big] &= \operatorname{Cov}(\bar{X}_2, \bar{Y}_2) &\quad (81)\\
\mathbb{E}\Big[\Big(\bar{X}_2 - \mu_X^2\Big)^2\Big(\bar{Y}_2 - \mu_Y^2\Big)\Big] &= \operatorname{Cov}(\bar{X}_2^2, \bar{Y}_2) - 2\mu_X^2 \operatorname{Cov}(\bar{X}_2, \bar{Y}_2) &\quad (82)\\
\mathbb{E}\Big[\Big(\bar{X}_2 - \mu_X^2\Big)^3\Big] &= \operatorname{Cov}(\bar{X}_2^2, \bar{X}_2) - 2\mu_X^2 \operatorname{Var}(\bar{X}_2) &\quad (83)\\
\mathbb{E}\Big[\Big(\bar{X}_2 - \mu_X^2\Big)^3\Big(\bar{Y}_2 - \mu_Y^2\Big)\Big] &= \operatorname{Cov}(\bar{X}_2^3, \bar{Y}_2) - 3\mu_X^2 \operatorname{Cov}(\bar{X}_2^2, \bar{Y}_2) + 3\mu_X^4 \operatorname{Cov}(\bar{X}_2, \bar{Y}_2) &\quad (84)\\
\mathbb{E}\Big[\Big(\bar{X}_2 - \mu_X^2\Big)^4\Big] &= \operatorname{Cov}(\bar{X}_2^3, \bar{X}_2) - 3\mu_X^2 \operatorname{Cov}(\bar{X}_2^2, \bar{X}_2) + 3\mu_X^4 \operatorname{Var}(\bar{X}_2) &\quad (85)
\end{aligned}$$

Bias Computing $\mathbb{E}[r]$ using the above identities:

$$\begin{aligned}
\mathbb{E}[r] \approx R\Bigg(1 &- \frac{\operatorname{Cov}(\bar{X}_2, \bar{Y}_2)}{\mu_X^2 \mu_Y^2} + \frac{\operatorname{Var}(\bar{X}_2)}{\mu_X^4} + \left(\frac{\operatorname{Cov}(\bar{X}_2^2, \bar{Y}_2) - 2\mu_X^2 \operatorname{Cov}(\bar{X}_2, \bar{Y}_2)}{\mu_X^4 \mu_Y^2}\right)\\
&- \left(\frac{\operatorname{Cov}(\bar{X}_2^2, \bar{X}_2) - 2\mu_X^2 \operatorname{Var}(\bar{X}_2)}{\mu_X^6}\right) - \left(\frac{\operatorname{Cov}(\bar{X}_2^3, \bar{Y}_2) - 3\mu_X^2 \operatorname{Cov}(\bar{X}_2^2, \bar{Y}_2) + 3\mu_X^4 \operatorname{Cov}(\bar{X}_2, \bar{Y}_2)}{\mu_X^6 \mu_Y^2}\right)\\
&+ \left(\frac{\operatorname{Cov}(\bar{X}_2^3, \bar{X}_2) - 3\mu_X^2 \operatorname{Cov}(\bar{X}_2^2, \bar{X}_2) + 3\mu_X^4 \operatorname{Var}(\bar{X}_2)}{\mu_X^8}\right)\Bigg)
\end{aligned} \quad (86)$$

$$\begin{aligned}
= R\Bigg(1 &- \overbrace{\frac{6\operatorname{Cov}(\bar{X}_2, \bar{Y}_2)}{\mu_X^2 \mu_Y^2}}^{\text{Term (a)}} + \overbrace{\frac{6\operatorname{Var}(\bar{X}_2)}{\mu_X^4}}^{\text{Term (b)}} + \overbrace{\frac{4\operatorname{Cov}(\bar{X}_2^2, \bar{Y}_2)}{\mu_X^4 \mu_Y^2}}^{\text{Term (c)}} - \overbrace{\frac{4\operatorname{Cov}(\bar{X}_2^2, \bar{X}_2)}{\mu_X^6}}^{\text{Term (d)}}\\
&- \underbrace{\frac{\operatorname{Cov}(\bar{X}_2^3, \bar{Y}_2)}{\mu_X^6 \mu_Y^2}}_{\text{Term (e)}} + \underbrace{\frac{\operatorname{Cov}(\bar{X}_2^3, \bar{X}_2)}{\mu_X^8}}_{\text{Term (f)}}\Bigg)
\end{aligned} \quad (87)$$

$$
\begin{aligned}
&= R\Bigg\{1 - \left(\frac{12}{n(n-1)}\frac{\operatorname{Cov}(X,Y)^2}{\mu_X^2\mu_Y^2} + \frac{24}{n}R_a\right) \\
&\quad + \left(\frac{12}{n(n-1)}\frac{\operatorname{Var}(X)^2}{\mu_X^4} + \frac{24}{n}R_b\right) \\
&\quad + \left(\frac{32(n-2)}{n(n-1)^2}\left(\frac{\operatorname{Cov}(X^2,Y)}{\mu_X^2\mu_Y} + \frac{2\operatorname{Cov}(X,Y)(\operatorname{Var}(X)+\mu_X^2)}{\mu_X^3\mu_Y}\right)\right. \\
&\qquad \left. + \frac{4(n-2)(n-3)}{n(n-1)}\left(\frac{8}{n}R_a + \frac{12}{n(n-1)}\frac{\operatorname{Cov}(X,Y)^2}{\mu_X^2\mu_Y^2}\right)\right) \\
&\quad - \left(\frac{32(n-2)}{n(n-1)^2}\left(\frac{\operatorname{Cov}(X^2,X)}{\mu_X^3} + \frac{2\operatorname{Var}(X)(\operatorname{Var}(X)+\mu_X^2)}{\mu_X^4}\right)\right. \\
&\qquad \left. + \frac{4(n-2)(n-3)}{n(n-1)}\left(\frac{8}{n}R_b + \frac{12}{n(n-1)}\frac{\operatorname{Var}(X)^2}{\mu_X^4}\right)\right) \\
&\quad - \left(\frac{24(n-2)(n-3)(n-4)}{n^2(n-1)^3}\left(\frac{\operatorname{Cov}(X^2,Y)}{\mu_Y\mu_X^2} + \frac{4\operatorname{Cov}(X,Y)(\operatorname{Var}(X)+\mu_X^2)}{\mu_Y\mu_X^3}\right)\right. \\
&\qquad \left. + \frac{(n-2)(n-3)(n-4)(n-5)}{n^2(n-1)^2}\left(\frac{12}{n}R_a + \frac{30}{n(n-1)}\frac{\operatorname{Cov}(X,Y)^2}{\mu_X^2\mu_Y^2}\right)\right) \\
&\quad + \left(\frac{24(n-2)(n-3)(n-4)}{n^2(n-1)^3}\left(\frac{\operatorname{Cov}(X^2,X)}{\mu_X^3} + \frac{4\operatorname{Var}(X)(\operatorname{Var}(X)+\mu_X^2)}{\mu_X^4}\right)\right. \\
&\qquad \left. + \frac{(n-2)(n-3)(n-4)(n-5)}{n^2(n-1)^2}\left(\frac{12}{n}R_b + \frac{30}{n(n-1)}\frac{\operatorname{Var}(X)^2}{\mu_X^4}\right)\right)
\end{aligned}
\tag{88}
$$

where $R_a = \frac{\mathrm{Cov(X,Y)}}{\mu_X \mu_Y}$ and $R_b = \frac{\operatorname{Var}(X)}{\mu_X^2}$ and where we have used 97, 101 110, 111, 124 and 126 for terms (a)-(f).

Therefore, an estimator unbiased up to order two is given by:

$$
\begin{aligned}
r_{\text{corr}} = \frac{\widehat{\mu_Y^2}}{\widehat{\mu_X^2}}\Bigg\{1 &+ \left(\frac{12}{n(n-1)}\frac{\widehat{\mathrm{Cov}(X,Y)}^2}{\widehat{\mu_X^2}\widehat{\mu_Y^2}} + \frac{24}{n}r_a^*\right) \\
&- \left(\frac{12}{n(n-1)}\frac{\widehat{\mathrm{Var}(X)}^2}{\widehat{\mu_X^4}} + \frac{24}{n}r_b^*\right) \\
&- \Bigg(\frac{32(n-2)}{n(n-1)^2}\left(\frac{\widehat{\mathrm{Cov}(X^2,Y)}}{\widehat{\mu_X^2}\widehat{\mu_Y}} + \frac{2\widehat{\mathrm{Cov}(X,Y)}(\widehat{\mathrm{Var}(X)} + \widehat{\mu_X^2})}{\widehat{\mu_X^3}\widehat{\mu_Y}}\right) \\
&\quad + \frac{4(n-2)(n-3)}{n(n-1)}\left(\frac{8}{n}r_a^* + \frac{12}{n(n-1)}\frac{\widehat{\mathrm{Cov}(X,Y)}^2}{\widehat{\mu_X^2}\widehat{\mu_Y^2}}\right)\Bigg) \\
&+ \Bigg(\frac{32(n-2)}{n(n-1)^2}\left(\frac{\widehat{\mathrm{Cov}(X^2,X)}}{\widehat{\mu_X^3}} + \frac{2\widehat{\mathrm{Var}(X)}(\widehat{\mathrm{Var}(X)} + \widehat{\mu_X^2})}{\widehat{\mu_X^4}}\right) \\
&\quad + \frac{4(n-2)(n-3)}{n(n-1)}\left(\frac{8}{n}r_b^* + \frac{12}{n(n-1)}\frac{\widehat{\mathrm{Var}(X)}^2}{\widehat{\mu_X^4}}\right)\Bigg) \\
&+ \Bigg(\frac{24(n-2)(n-3)(n-4)}{n^2(n-1)^3}\left(\frac{\widehat{\mathrm{Cov}(X^2,Y)}}{\widehat{\mu_Y}\widehat{\mu_X^2}} + \frac{4\widehat{\mathrm{Cov}(X,Y)}(\widehat{\mathrm{Var}(X)} + \widehat{\mu_X^2})}{\widehat{\mu_Y}\widehat{\mu_X^3}}\right) \\
&\quad + \frac{(n-2)(n-3)(n-4)(n-5)}{n^2(n-1)^2}\left(\frac{12}{n}r_a^* + \frac{30}{n(n-1)}\frac{\widehat{\mathrm{Cov}(X,Y)}^2}{\widehat{\mu_X^2}\widehat{\mu_Y^2}}\right)\Bigg) \\
&- \Bigg(\frac{24(n-2)(n-3)(n-4)}{n^2(n-1)^3}\left(\frac{\widehat{\mathrm{Cov}(X^2,X)}}{\widehat{\mu_X^3}} + \frac{4\widehat{\mathrm{Var}(X)}(\widehat{\mathrm{Var}(X)} + \widehat{\mu_X^2})}{\widehat{\mu_X^4}}\right) \\
&\quad + \frac{(n-2)(n-3)(n-4)(n-5)}{n^2(n-1)^2}\left(\frac{12}{n}r_b^* + \frac{30}{n(n-1)}\frac{\widehat{\mathrm{Var}(X)}^2}{\widehat{\mu_X^4}}\right)\Bigg)\Bigg\},
\end{aligned}
\tag{89}
$$

where we used equations (96), (100):

$$
\begin{aligned}
r_a^* = \underbrace{\frac{\widehat{\mathrm{Cov}(X,Y)}}{\widehat{\mu_X\mu_Y}}}_{=r_a}\Bigg(1 &+ \frac{1}{(n-1)}\left(\frac{\widehat{\mu_Y}\widehat{\mathrm{Cov}(X^2,Y)} + \widehat{\mu_X}\widehat{\mathrm{Cov}(Y^2,X)}}{\widehat{\mathrm{Cov}(X,Y)}\widehat{\mu_X}\widehat{\mu_Y}} - 4\right) \\
&- \frac{1}{(n-1)}\left(\frac{\widehat{\mathrm{Var}(X)}}{\widehat{\mu_X^2}} + \frac{\widehat{\mathrm{Var}(Y)}}{\widehat{\mu_Y^2}} + 2\frac{\widehat{\mathrm{Cov}(X,Y)}}{\widehat{\mu_X}\widehat{\mu_Y}}\right)\Bigg)
\end{aligned}
\tag{90}
$$

$$
r_b^* = \underbrace{\frac{\widehat{\mathrm{Var}(X)}}{\widehat{\mu_X^2}}}_{=r_b}\left(1 + \frac{4}{(n-1)}\left(\frac{\frac{1}{2}\widehat{\mathrm{Cov}(X^2,X)}}{\widehat{\mu_X}\widehat{\mathrm{Var}(X)}} - 1\right) - \frac{4}{(n-1)}\frac{\widehat{\mathrm{Var}(X)}}{\widehat{\mu_X^2}}\right). \tag{91}
$$

D.1 Term (a)

Let us first look at the first term: $\frac{6\,\mathrm{Cov}(\bar{X}_2,\bar{Y}_2)}{\mu_X^2\mu_Y^2}$. Using the expression for the covariance between two second order U-statistics we get:

$$
\frac{6\,\mathrm{Cov}(\bar{X}_2,\bar{Y}_2)}{\mu_X^2\mu_Y^2} = \frac{6}{\mu_X^2\mu_Y^2}\left(\frac{4}{n}\mu_X\mu_Y\,\mathrm{Cov(X,Y)} + \frac{2}{n(n-1)}\,\mathrm{Cov}(X,Y)^2\right) \tag{92}
$$

$$
= \underbrace{\frac{24}{n}\frac{\mathrm{Cov(X,Y)}}{\mu_X\mu_Y}}_{\in\mathcal{O}(n^{-1})} + \underbrace{\frac{12}{n(n-1)}\frac{\mathrm{Cov}(X,Y)^2}{\mu_X^2\mu_Y^2}}_{\in\mathcal{O}(n^{-2})} \tag{93}
$$

Since we know that for every recursion (i.e., geometric series expansion) we will get at least another factor of $\frac{1}{n}$, we don't have to further recurse on term that is of order $\mathcal{O}(n^{-2})$. Consequently, we only expand the following term via a geometric series,

$$R_a = \frac{\mathrm{Cov}(X,Y)}{\mu_X \mu_Y}, \tag{94}$$

since the ratio of the respective unbiased estimators,

$$r_a = \frac{\widehat{\mathrm{Cov}(X,Y)}}{\widehat{\mu_X \mu_Y}}, \tag{95}$$

is biased.
Using the same machinery as before, we obtain a corrected version of r_a:

$$\begin{aligned} r_a^* = \underbrace{\frac{\widehat{\mathrm{Cov}(X,Y)}}{\widehat{\mu_X \mu_Y}}}_{=r_a} \Bigg(1 + \frac{1}{(n-1)} \Bigg(\frac{\widehat{\mu_Y}\widehat{\mathrm{Cov}(X^2,Y)} + \widehat{\mu_X}\widehat{\mathrm{Cov}(Y^2,X)}}{\widehat{\mathrm{Cov}(X,Y)}\widehat{\mu_X}\widehat{\mu_Y}} - 4 \Bigg) \\ - \frac{1}{(n-1)} \Bigg(\frac{\widehat{\mathrm{Var}(X)}}{\widehat{\mu_X^2}} + \frac{\widehat{\mathrm{Var}(Y)}}{\widehat{\mu_Y^2}} + 2\frac{\widehat{\mathrm{Cov}(X,Y)}}{\widehat{\mu_X}\widehat{\mu_Y}} \Bigg) \Bigg) \end{aligned} \tag{96}$$

The complete correction of term (a), $\frac{6\,\mathrm{Cov}(\bar{X}_2,\bar{Y}_2)}{\mu_X^2 \mu_Y^2}$, looks therefore as follows:

$$\begin{aligned} \frac{6\widehat{\mathrm{Cov}(\bar{X}_2,\bar{Y}_2)}}{\widehat{\mu_X^2}\widehat{\mu_Y^2}} = {} & \frac{12}{n(n-1)} \frac{\widehat{\mathrm{Cov}(X,Y)}^2}{\widehat{\mu_X^2}\widehat{\mu_Y^2}} \\ & + \frac{24}{n} \frac{\widehat{\mathrm{Cov}(X,Y)}}{\widehat{\mu_X \mu_Y}} \Bigg(1 + \frac{1}{(n-1)} \Bigg(\frac{\widehat{\mu_Y}\widehat{\mathrm{Cov}(X^2,Y)} + \widehat{\mu_X}\widehat{\mathrm{Cov}(Y^2,X)}}{\widehat{\mathrm{Cov}(X,Y)}\widehat{\mu_X}\widehat{\mu_Y}} - 4 \Bigg) \\ & \qquad - \frac{1}{(n-1)} \Bigg(\frac{\widehat{\mathrm{Var}(X)}}{\widehat{\mu_X^2}} + \frac{\widehat{\mathrm{Var}(Y)}}{\widehat{\mu_Y^2}} + 2\frac{\widehat{\mathrm{Cov}(X,Y)}}{\widehat{\mu_X}\widehat{\mu_Y}} \Bigg) \Bigg) \end{aligned} \tag{97}$$

D.2 Term (b)

The correction of term (b), $\frac{6\,\mathrm{Var}(\bar{X}_2)}{\mu_X^4}$ is analogous to that of term (a). Define

$$R_b = \frac{\mathrm{Var(X)}}{\mu_X^2} \tag{98}$$

$$r_b = \frac{\widehat{\mathrm{Var(X)}}}{\widehat{\mu_X^2}}. \tag{99}$$

Then using the geometric series expansion, a corrected version of r_b is given by

$$r_b^* = \underbrace{\frac{\widehat{\mathrm{Var}(X)}}{\widehat{\mu_X^2}}}_{=r_b} \Bigg(1 + \frac{4}{(n-1)} \Bigg(\frac{\frac{1}{2}\widehat{\mathrm{Cov}(X^2,X)}}{\widehat{\mu_X}\widehat{\mathrm{Var}(X)}} - 1 \Bigg) - \frac{4}{(n-1)} \frac{\widehat{\mathrm{Var}(X)}}{\widehat{\mu_X^2}} \Bigg). \tag{100}$$

The full correction of term (b) is

$$\frac{6\widehat{\mathrm{Var}(\bar{X}_2)}}{\widehat{\mu_X^4}} = \frac{12}{n(n-1)} \frac{\widehat{\mathrm{Var}(X)}^2}{\widehat{\mu_X^4}} + \frac{24}{n} \frac{\widehat{\mathrm{Var}(X)}}{\widehat{\mu_X^2}} \Bigg(1 + \frac{4}{(n-1)} \Bigg(\frac{\frac{1}{2}\widehat{\mathrm{Cov}(X^2,X)}}{\widehat{\mu_X}\widehat{\mathrm{Var}(X)}} - 1 - \frac{\widehat{\mathrm{Var}(X)}}{\widehat{\mu_X^2}} \Bigg) \Bigg) \tag{101}$$

D.3 Term (c)

In this section we want to find an expression for term (c):

$$\frac{4\,\mathrm{Cov}(\bar{X}_2^2,\bar{Y}_2)}{\mu_X^4\mu_Y^2}. \tag{102}$$

To this end, we first need a convenient representation of $\bar{X}_2^2$ in terms of other U-statistics:

$$\bar{X}_2^2 = \left(\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} X_iX_j\right)^2 = \frac{2}{n(n-1)}U_\alpha + \frac{4(n-2)}{n(n-1)}U_\beta + \frac{(n-2)(n-3)}{n(n-1)}\bar{X}_4, \tag{103}$$

with the U-statistics:

$$U_\beta = \frac{1}{n(n-1)(n-2)}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\sum_{\substack{k=1\\ k\neq j\\ k\neq i}}^{n} X_i^2X_jX_k \tag{104}$$

$$U_\alpha = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} X_i^2X_j^2 \tag{105}$$

$$\bar{X}_4 = \frac{1}{n(n-1)(n-2)(n-3)}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\sum_{\substack{k=1\\ k\neq j\\ k\neq i}}^{n}\sum_{\substack{l=1\\ l\neq k\\ l\neq j\\ l\neq i}}^{n} X_iX_jX_kX_l. \tag{106}$$

Hence, term (c) becomes:

$$\frac{4\,\mathrm{Cov}(\bar{X}_2^2,\bar{Y}_2)}{\mu_X^4\mu_Y^2} = \underbrace{\frac{8}{n(n-1)}\frac{\mathrm{Cov}(U_\alpha,\bar{Y}_2)}{\mu_X^4\mu_Y^2}}_{\text{First term}} + \underbrace{\frac{16(n-2)}{n(n-1)}\frac{\mathrm{Cov}(U_\beta,\bar{Y}_2)}{\mu_X^4\mu_Y^2}}_{\text{Second term}} + \underbrace{\frac{4(n-2)(n-3)}{n(n-1)}\frac{\mathrm{Cov}(\bar{X}_4,\bar{Y}_2)}{\mu_X^4\mu_Y^2}}_{\text{Third term}} \tag{107}$$

All all the covariances in the above equation are covariances between U-statistics which are $\mathcal{O}\left(\frac{1}{n}\right)$. Therefore, the first term, which already has an explicit $\mathcal{O}\left(\frac{1}{n^2}\right)$ dependence, can be neglected entirely. The second term has an explicit $\mathcal{O}\left(\frac{1}{n}\right)$, combined with the $\mathcal{O}\left(\frac{1}{n}\right)$ from the covariance this is in total a $\mathcal{O}\left(\frac{1}{n^2}\right)$ dependency. Hence, we have to find an estimator for that term but do not have to recurse on it. On the last term, we do have to recurse, however, we have derived the recursion already in equation (96). We can rewrite the above equation using the symmetrized U-statistics

$$U_\beta = \frac{1}{n(n-1)(n-2)}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\sum_{\substack{k=1\\ k\neq j\\ k\neq i}}^{n}\frac{1}{3}\Big(X_i^2X_jX_k + X_iX_j^2X_k + X_iX_jX_k^2\Big). \tag{108}$$

$$\begin{aligned}\frac{4\,\mathrm{Cov}(\bar{X}_2^2,\bar{Y}_2)}{\mu_X^4\mu_Y^2} \approx{} & \underbrace{\frac{32(n-2)}{n(n-1)^2}\left(\frac{\mathrm{Cov}(X^2,Y)}{\mu_X^2\mu_Y} + \frac{2\,\mathrm{Cov}(X,Y)(\mathrm{Var}(X)+\mu_X^2)}{\mu_X^3\mu_Y}\right)}_{\text{Second term}} \\ & + \underbrace{\frac{4(n-2)(n-3)}{n(n-1)}\left(\frac{8}{n}\frac{\mathrm{Cov}(X,Y)}{\mu_X\mu_Y} + \frac{12}{n(n-1)}\frac{\mathrm{Cov}(X,Y)^2}{\mu_X^2\mu_Y^2}\right)}_{\text{Third term}}\end{aligned} \tag{109}$$

Taking the recursion of the third term into account, the total correction of term (c) is:

$$\begin{aligned}\frac{4\widehat{\mathrm{Cov}(\bar{X}_2^2,\bar{Y}_2)}}{\widehat{\mu_X^4}\widehat{\mu_Y^2}} \approx{} & \frac{32(n-2)}{n(n-1)^2}\left(\frac{\widehat{\mathrm{Cov}(X^2,Y)}}{\widehat{\mu_X^2}\widehat{\mu_Y}} + \frac{2\widehat{\mathrm{Cov}(X,Y)}(\widehat{\mathrm{Var}(X)}+\widehat{\mu_X^2})}{\widehat{\mu_X^3}\widehat{\mu_Y}}\right) \\ & + \frac{4(n-2)(n-3)}{n(n-1)}\left(\frac{8}{n}r_a^* + \frac{12}{n(n-1)}\frac{\widehat{\mathrm{Cov}(X,Y)^2}}{\widehat{\mu_X^2}\widehat{\mu_Y^2}}\right)\end{aligned} \tag{110}$$

D.4 Term (d)

The computation of the correction of term (d), $\frac{4\,\mathrm{Cov}(\bar{X}_2^2,\bar{X}_2)}{\mu_X^6}$, is similar to that of term (c). Hence, we only present the resulting correction:

$$\begin{aligned}\frac{4\widehat{\mathrm{Cov}(\bar{X}_2^2,\bar{X}_2)}}{\widehat{\mu_X^4}\widehat{\mu_Y^2}} \approx \frac{32(n-2)}{n(n-1)^2}&\left(\frac{\widehat{\mathrm{Cov}(X^2,X)}}{\widehat{\mu_X^3}} + \frac{2\widehat{\mathrm{Var}(X)}(\widehat{\mathrm{Var}(X)}+\widehat{\mu_X^2})}{\widehat{\mu_X^4}}\right)\\ &+\frac{4(n-2)(n-3)}{n(n-1)}\left(\frac{8}{n}r_b^* + \frac{12}{n(n-1)}\frac{\widehat{\mathrm{Var}(X)^2}}{\widehat{\mu_X^4}}\right)\end{aligned} \tag{111}$$

D.5 Term (e)

Term (e) is:

$$\frac{\mathrm{Cov}(\bar{X}_2^3,\bar{Y}_2)}{\mu_X^6\mu_Y^2} \tag{112}$$

To be able to compute that term, we reexpress the numerator in terms of several U-statistics:

$$\begin{aligned}\bar{X}_2^3 = &\frac{4}{n^2(n-1)^2}U_I + \frac{24(n-2)}{n^2(n-1)^2}U_{II} + \frac{8(n-2)}{n^2(n-1)^2}U_{III} + \frac{8(n-2)(n-3)}{n^2(n-1)^2}U_{IV}\\ &+\frac{30(n-2)(n-3)}{n^2(n-1)^2}U_V + \frac{12(n-2)(n-3)(n-4)}{n^2(n-1)2}U_{VI} + \frac{(n-2)(n-2)(n-4)(n-5)}{n^2(n-1)^2}\bar{X}_6,\end{aligned} \tag{113}$$

where

$$U_I := \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} X_i^3X_j^3, \tag{114}$$

$$U_{II} := \frac{1}{n(n-1)(n-2)}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\sum_{\substack{k=1\\ k\neq j\\ k\neq i}}^{n} X_i^3X_j^2X_k, \tag{115}$$

$$U_{III} = \frac{1}{n(n-1)(n-2)}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\sum_{\substack{k=1\\ k\neq j\\ k\neq i}}^{n} X_i^2,X_j^2X_k^2 \tag{116}$$

$$U_{IV} := \frac{1}{n(n-1)(n-2)(n-3)}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\sum_{\substack{k=1\\ k\neq j\\ k\neq i}}^{n}\sum_{\substack{l=1\\ l\neq k\\ l\neq j\\ l\neq i}}^{n} X_i^3X_jX_kX_l, \tag{117}$$

$$U_V := \frac{1}{n(n-1)(n-2)(n-3)}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\sum_{\substack{k=1\\ k\neq j\\ k\neq i}}^{n}\sum_{\substack{l=1\\ l\neq k\\ l\neq j\\ l\neq i}}^{n} X_i^2X_j^2X_kX_l, \tag{118}$$

$$U_{VI} := \frac{1}{n(n-1)(n-2)(n-3)(n-4)}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\sum_{\substack{k=1\\ k\neq j\\ k\neq i}}^{n}\sum_{\substack{l=1\\ l\neq k\\ l\neq j\\ l\neq i}}^{n}\sum_{\substack{p=1\\ p\neq l\\ p\neq k\\ p\neq j\\ p\neq i}}^{n} X_i^2X_jX_kX_lX_p, \tag{119}$$

$$\bar{X}_6 = \frac{1}{n(n-1)(n-2)(n-3)(n-4)(n-5)} \sum_{i=1}^{n} \sum_{\substack{j=1\\ j\neq i}}^{n} \sum_{\substack{k=1\\ k\neq j\\ k\neq i}}^{n} \sum_{\substack{l=1\\ l\neq k\\ l\neq j\\ l\neq i}}^{n} \sum_{\substack{p=1\\ p\neq l\\ p\neq k\\ p\neq j\\ p\neq i}}^{n} \sum_{\substack{q=1\\ q\neq p\\ q\neq l\\ q\neq k\\ q\neq j\\ q\neq i}}^{n} X_i X_j X_k X_l X_p X_q, \tag{120}$$

Hence term (e) can be written as:

$$\begin{aligned}\frac{\operatorname{Cov}(\bar{X}_2^3, \bar{Y}_2)}{\mu_X^6 \mu_Y^2} &= \underbrace{\frac{4}{n^2(n-1)^2}}_{\in\mathcal{O}\left(\frac{1}{n^4}\right)} \frac{\operatorname{Cov}(U_I, \bar{Y}_2)}{\mu_X^6 \mu_Y^2} + \underbrace{\frac{24(n-2)}{n^2(n-1)^2}}_{\in\mathcal{O}\left(\frac{1}{n^3}\right)} \frac{\operatorname{Cov}(U_{II}, \bar{Y}_2)}{\mu_X^6 \mu_Y^2} + \underbrace{\frac{8(n-2)}{n^2(n-1)^2}}_{\in\mathcal{O}\left(\frac{1}{n^3}\right)} \frac{\operatorname{Cov}(U_{III}, \bar{Y}_2)}{\mu_X^6 \mu_Y^2}\\ &+ \underbrace{\frac{8(n-2)(n-3)}{n^2(n-1)^2}}_{\in\mathcal{O}\left(\frac{1}{n^2}\right)} \frac{\operatorname{Cov}(U_{IV}, \bar{Y}_2)}{\mu_X^6 \mu_Y^2} + \underbrace{\frac{30(n-2)(n-3)}{n^2(n-1)^2}}_{\in\mathcal{O}\left(\frac{1}{n^2}\right)} \frac{\operatorname{Cov}(U_V, \bar{Y}_2)}{\mu_X^6 \mu_Y^2}\\ &+ \underbrace{\frac{12(n-2)(n-3)(n-4)}{n^2(n-1)^2}}_{\in\mathcal{O}\left(\frac{1}{n}\right)} \frac{\operatorname{Cov}(U_{VI}, \bar{Y}_2)}{\mu_X^6 \mu_Y^2} + \underbrace{\frac{(n-2)(n-3)(n-4)(n-5)}{n^2(n-1)^2}}_{\in\mathcal{O}(1)} \frac{\operatorname{Cov}(\bar{X}_6, \bar{Y}_2)}{\mu_X^6 \mu_Y^2}.\end{aligned} \tag{121}$$

At this point, we can immediately discard the first three terms as they are at least $\mathcal{O}\left(\frac{1}{n^3}\right)$ and so can directly be neglected for a second order correction. In addition, as we are dealing with covariances between U-statistics they add another $\mathcal{O}\left(\frac{1}{n}\right)$. Therefore, the fourth and fifth term are actually $\mathcal{O}\left(\frac{1}{n}\right)\mathcal{O}\left(\frac{1}{n^2}\right) = \mathcal{O}\left(\frac{1}{n^3}\right)$, so they can be neglected as well. Only the last and the second to last term remain:

$$\frac{\operatorname{Cov}(\bar{X}_2^3, \bar{Y}_2)}{\mu_X^6 \mu_Y^2} \approx \underbrace{\frac{12(n-2)(n-3)(n-4)}{n^2(n-1)^2} \frac{\operatorname{Cov}(U_{VI}, \bar{Y}_2)}{\mu_X^6 \mu_Y^2}}_{\text{Sixth term}} + \underbrace{\frac{(n-2)(n-3)(n-4)(n-5)}{n^2(n-1)^2} \frac{\operatorname{Cov}(\bar{X}_6, \bar{Y}_2)}{\mu_X^6 \mu_Y^2}}_{\text{Seventh term}}. \tag{122}$$

Re-expressing the covariances between U-statistics as covariances between random variables X and Y (and using the symmetrized version of U_{VI}), we obtain:

$$\begin{aligned}\frac{\operatorname{Cov}(\bar{X}_2^3, \bar{Y}_2)}{\mu_X^6 \mu_Y^2} &\approx \underbrace{\frac{24(n-2)(n-3)(n-4)}{n^2(n-1)^3}\left(\frac{\operatorname{Cov}(X^2, Y)}{\mu_Y \mu_X^2} + \frac{4\operatorname{Cov}(X, Y)(\operatorname{Var}(X) + \mu_X^2)}{\mu_Y \mu_X^3}\right)}_{\text{Sixth term}}\\ &+ \underbrace{\frac{(n-2)(n-3)(n-4)(n-5)}{n^2(n-1)^2}\left(\frac{12}{n}\frac{\operatorname{Cov}(X, Y)}{\mu_X \mu_Y} + \frac{30}{n(n-1)}\frac{\operatorname{Cov}(X, Y)^2}{\mu_X^2 \mu_Y^2}\right)}_{\text{Seventh term}}\end{aligned} \tag{123}$$

Since the term $\frac{12}{n}\frac{\operatorname{Cov}(X,Y)}{\mu_X \mu_Y}$ is in $\mathcal{O}\left(\frac{1}{n}\right)$ we have to recurse on it. However, we already have derived its correction in equation (96). Therefore, the total correction of term (e) comes down to:

$$\begin{aligned}\frac{\widehat{\operatorname{Cov}(\bar{X}_2^3, \bar{Y}_2)}}{\widehat{\mu_X^6}\widehat{\mu_Y^2}} &= \frac{24(n-2)(n-3)(n-4)}{n^2(n-1)^3}\left(\frac{\widehat{\operatorname{Cov}(X^2, Y)}}{\widehat{\mu_Y}\widehat{\mu_X^2}} + \frac{4\widehat{\operatorname{Cov}(X, Y)}(\widehat{\operatorname{Var}(X)} + \widehat{\mu_X^2})}{\widehat{\mu_Y}\widehat{\mu_X^3}}\right)\\ &+ \frac{(n-2)(n-3)(n-4)(n-5)}{n^2(n-1)^2}\left(\frac{12}{n} r_a^* + \frac{30}{n(n-1)}\frac{\widehat{\operatorname{Cov}(X, Y)}^2}{\widehat{\mu_X^2}\widehat{\mu_Y^2}}\right)\end{aligned} \tag{124}$$

D.6 Term (f)

Term (f) is:

$$\frac{\mathrm{Cov}(\bar{X}_2^3, \bar{X}_2)}{\mu_X^8} \tag{125}$$

The procedure to obtain its correction is analogous to that of term (e), hence we only present the result:

$$\begin{aligned}\frac{\widehat{\mathrm{Cov}(\bar{X}_2^3, \bar{X}_2)}}{\widehat{\mu_X^8}} &= \frac{24(n-2)(n-3)(n-4)}{n^2(n-1)^3}\left(\frac{\widehat{\mathrm{Cov}(X^2, X)}}{\widehat{\mu_X^3}} + \frac{4\widehat{\mathrm{Var}(X)}(\widehat{\mathrm{Var}(X)} + \widehat{\mu_X^2})}{\widehat{\mu_X^4}}\right) \\ &+ \frac{(n-2)(n-3)(n-4)(n-5)}{n^2(n-1)^2}\left(\frac{12}{n} r_b^* + \frac{30}{n(n-1)}\frac{\widehat{\mathrm{Var}(X)^2}}{\widehat{\mu_X^4}}\right)\end{aligned} \tag{126}$$

E Top-label calibration

Following standard practice in related work on calibration, we report the L_1 ECE^{bin} for top-label (also called confidence) calibration on CIFAR-10/100. ECE^{bin} was calculated using 15 bins and an adaptive width binning scheme, which determines the bin sizes so that an equal number of samples fall into each bin [Nguyen and O'Connor, 2015, Mukhoti et al., 2020]. The 95% confidence intervals for ECE^{bin} are obtained using 100 bootstrap samples, as in Kumar et al. [2019]. In all experiments with calibration regularized training, the biased version of ECE^{KDE} was used.

Table 5 summarizes our evaluation of the efficacy of KDE-XE in lowering the calibration error over the baseline XE on CIFAR-10 and CIFAR-100. The best performing λ coefficient for KDE-XE is shown in the brackets. The results show that KDE-XE consistently reduces the calibration error, without dropping the accuracy. Figure 5 depicts the L_2 ECE^{bin} for several choices of the λ parameter for KDE-XE, using ResNet-110 (SD) on CIFAR-10/100. Figure 6 shows reliability diagrams with 10 bins for top-label calibration on CIFAR-100 using ResNet and Wide-ResNet. Comapared to XE, we notice that KDE-XE lowers the overconfident predictions, and obtains better calibration than MMCE ($\lambda = 2$) and FL-53 on average, as summarized by the ECE value in the gray box.

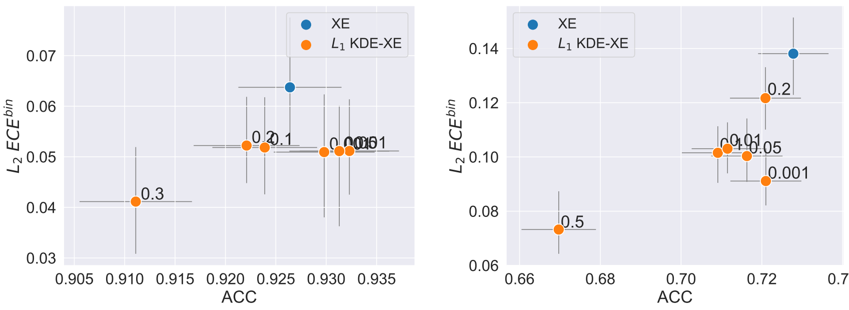
Table 5: Top-label L_1 adaptive-width ECE^{bin} and accuracy for XE and KDE-XE for various architectures on CIFAR-10/100. Best ECE values are marked in bold. The value in the brackets represent the value of the λ parameter.

Dataset	Model	ECE^{bin}		Accuracy	
		XE	KDE-XE	XE	KDE-XE
CIFAR-10	ResNet-110	3.890 ± 0.602	**3.093** ± 0.604 (0.001)	0.925 ± 0.005	0.930 ± 0.005
	ResNet-110 (SD)	3.555 ± 0.623	**2.778** ± 0.468 (0.01)	0.926 ± 0.005	0.932 ± 0.005
CIFAR-100	ResNet-110	12.769 ± 0.784	**8.969** ± 1.047 (0.2)	0.700 ± 0.009	0.696 ± 0.009
	ResNet-110 (SD)	11.175 ± 0.642	**7.828** ± 0.814 (0.001)	0.728 ± 0.009	0.721 ± 0.009
	Wide-ResNet-28-10	7.279 ± 0.876	**3.703** ± 1.086 (0.5)	0.762 ± 0.008	0.770 ± 0.008
	DenseNet-40	9.196 ± 0.881	**8.016** ± 1.079 (0.01)	0.756 ± 0.008	0.756 ± 0.008

F Relationship between ECE^{bin} and ECE^{KDE}

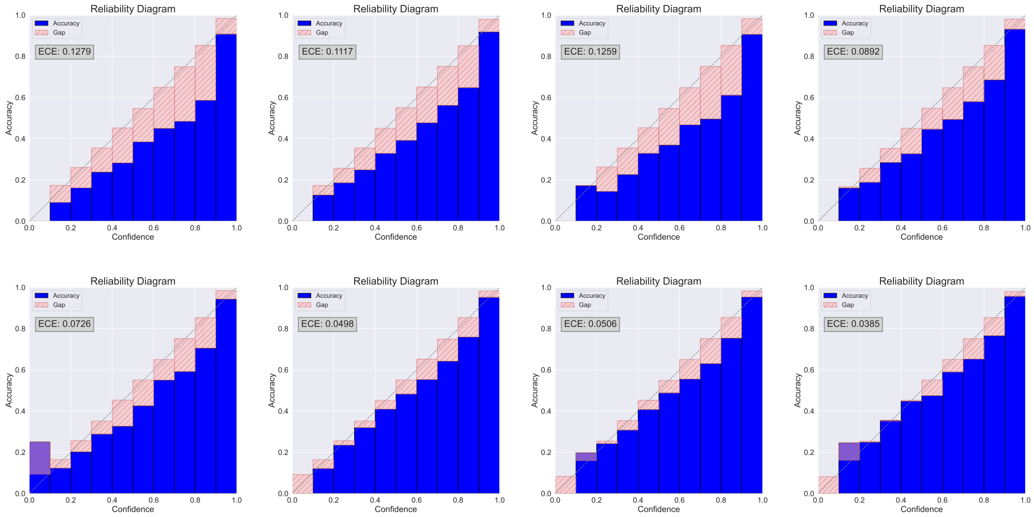
In the following two sections, we investigate further the relationship between ECE^{bin}, as the most widely used metric, and our ECE^{KDE} estimator. For the three types of calibration, ECE^{bin} is calculated with equal-width binning scheme. The values for the bandwidth in ECE^{KDE} and the number of bins per class for ECE^{bin} are chosen with leave-one-out maximum likelihood procedure and Doane's formula [Doane, 1976], respectively.

Figure 7 shows an example of ECE^{bin} in a three-class setting on CIFAR-10. The points are mostly concentrated at the edges of the histogram, as can be seen from Figure 7b. The surface of the corresponding Dirichlet KDE is given in 7c.



(a) CIFAR-10 (b) CIFAR-100

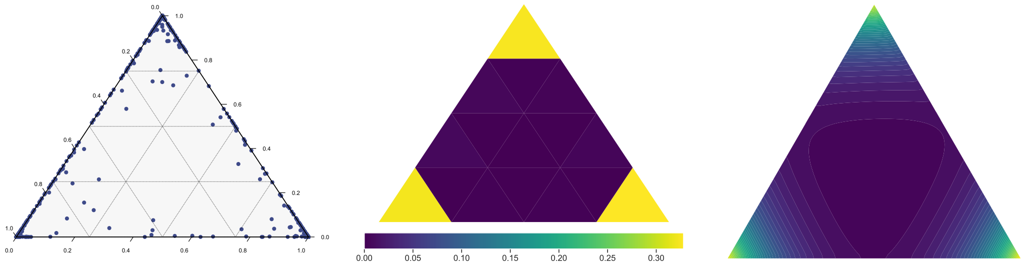
Figure 5: L_2 ECE^{bin} for top-label calibration using ResNet (SD).



XE MMCE FL-53 L_1 KDE-XE

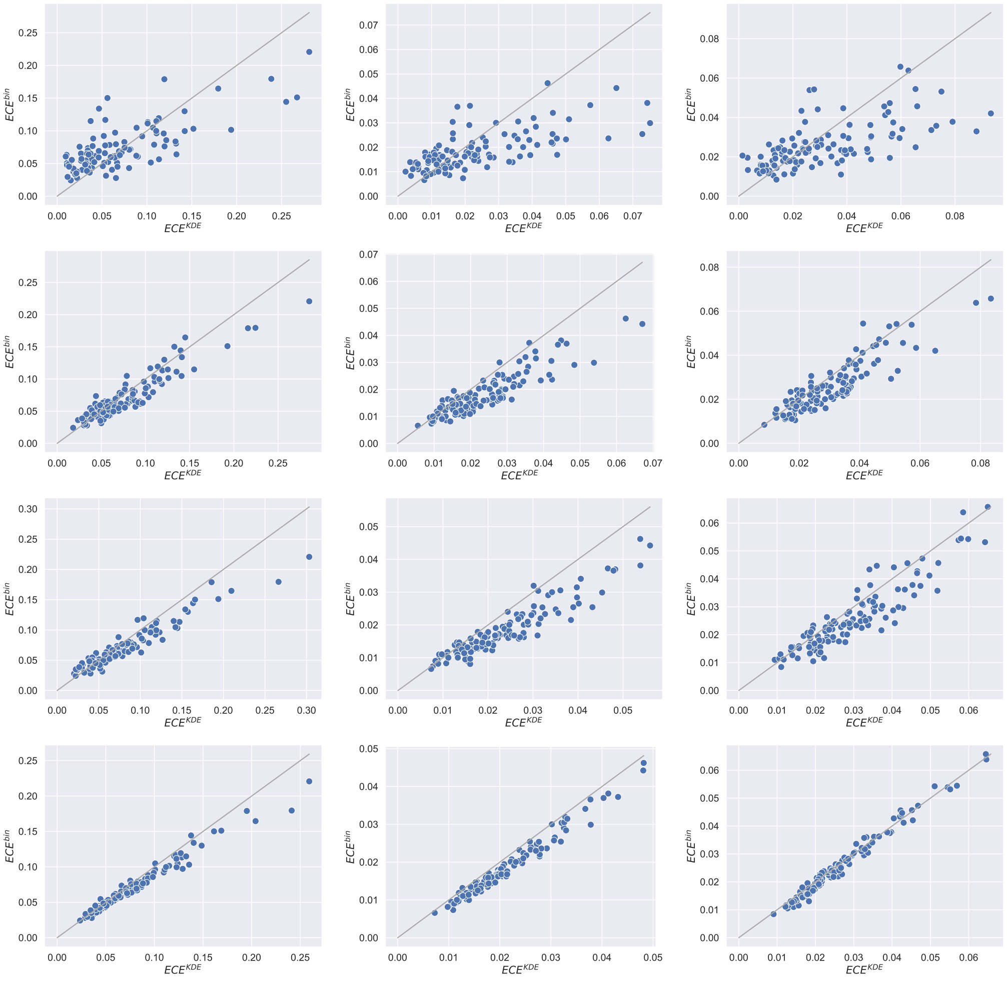
Figure 6: Reliability diagrams for top-label calibration on CIFAR-100 using ResNet (top row) and Wide-ResNet (bottom row) for each of the considered baselines.

Figure 8 shows the relationship between ECE^{bin} and ECE^{KDE}. The points represent a trained Resnet-56 model on a subset of three classes from CIFAR-10. In every row, a differnt number of points was used to estimate the ECE^{KDE}. We notice the ECE^{KDE} estimates of the three types of calibration closely correspond to their histogram-based approximations.



(a) Splitting the simplex in 16 bins (b) Corresponding histogram (c) Corresponding Dirichlet KDE

Figure 7: An example of a simplex binned estimator and kernel-density estimator for CIFAR-10



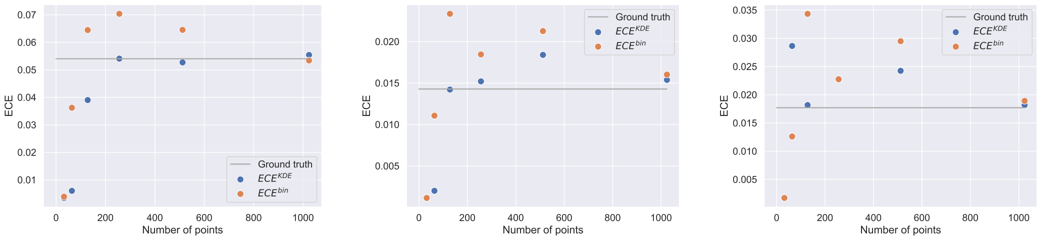
Canonical
Marginal
Top-label

Figure 8: Relationship between ECE^{bin} and ECE^{KDE} for the three types of calibration: canonical (first column), marginal (second column) and top-label (third column). In every row top to bottom, different number of points (100, 500, 1000 and all points, respectively) are used to approximate ECE^{KDE}. Each point represents a ResNet-56 model trained on a subset of three classes from CIFAR-10. The number of bins per class (13) is selected using Doane's formula [Doane, 1976], while the bandwidth is selected using a leave-one-out maximum likelihood procedure (typical chosen values are 0.001 for 100 points and 0.0001 otherwise).

G Bias and convergence rates

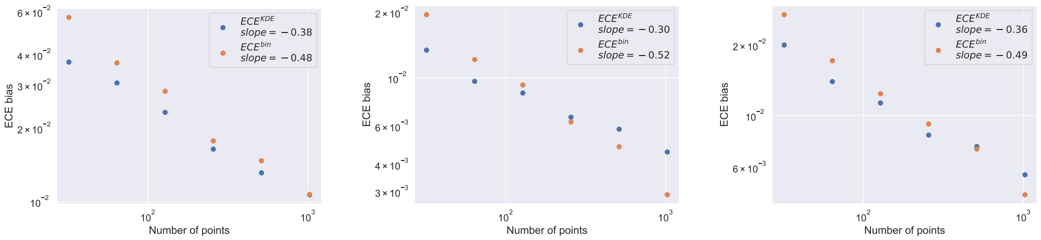
Figure 9 shows a comparison of ECE^{KDE} and ECE^{bin} estimated with a varying number of points. The ground truth is computed from 3000 test points with ECE^{KDE}. The used model is a ResNet-56, trained on a subset of three classes from CIFAR-10. The figure shows that the two estimates are comparable and both are doing a reasonable job in a three-class setting.

Figure 10 shows the absolute difference between the ground truth and estimated ECE using ECE^{KDE} and a ECE^{bin} with varying number of points. The results are averaged over 120 ResNet-56 models trained on a subset of three classes from CIFAR-10. Both estimators are biased and have some variance, and the plot shows that the combination of the two is in the same order of magnitude. The empirical convergence rates (slope of the log-log plot) is given in the legend and is shown to be close to the theoretically expected value of -0.5. We observe that ECE^{KDE} has similar statistical properties in terms of bias and convergence as ECE^{bin}.



(a) Canonical (b) Marginal (c) Top-label

Figure 9: ECE^{KDE} estimates and their corresponding binned approximations on the three types of calibration for varying number of points used for the estimation. The ground truth is calculated using 3000 probability scores of the test set using ECE^{KDE}. Optimal number of bins and bandwidth are chosen with Doane's formula and LOO MLE, respectively. Typical chosen number of bins is 6-11, and common values for the bandwidth are 0.0001 and 0.001.



(a) Canonical (b) Marginal (c) Top-label

Figure 10: Absolute difference between ground truth and estimated ECE for varying number of points used for the estimation. The ground truth is calculated using 3000 probability scores of the test set. Note that the axes are on a log scale.