Supplementary material for ‘Evaluating model calibration in classification’

A Theoretical results on calibration evaluation

A.1 Additional examples

The following example shows that perfect calibration according to Guo et al. (2017) does not imply calibrated marginal predictions.

Example 3 (No calibrated marginal predictions)
Suppose $\mathcal{Y} = \{1, 2, 3\}$. Let g be a probabilistic classifier that predicts only the two probability distributions in the first column of Table 2 with equal probability and assume that the true conditional distribution $\mathbb{P}[Y \in \cdot \,|\, g(X)]$ is given by the second column. The model is perfectly calibrated according to Guo et al. (2017). However, all marginal predictions are uncalibrated and additionally g is not reliable since (1) is not satisfied. Figure 5 provides an illustration of these observations.

$g(X)$	$\mathbb{P}[Y \in \cdot \,\vert\, g(X)]$
(0.6, 0.1, 0.3)	(0.7, 0.2, 0.1)
(0.4, 0.6, 0.0)	(0.3, 0.5, 0.2)

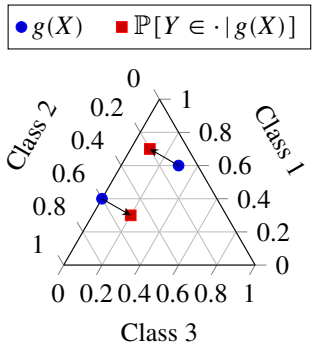


Table 2 & Figure 5: Probabilistic classifier g for $\mathcal{Y} = \{1, 2, 3\}$ with two uniformly distributed predictions.

Using a modification of our framework, one can express the calibration evaluation by Kendall and Gal (2017) by an induced calibration function.

Example 4 (Marginalized calibration)
Kendall and Gal (2017) evaluated model calibration by estimating the calibration function

$$r_\psi(u) = \mathbb{P}[\psi_u(Y, \mu) \in \cdot \,|\, \pi_{\psi_u}(g(X)) = u\delta_1 + (1-u)\delta_0]$$

induced by $\psi_u(y, \mu) = \mathbb{1}_{\{\mu(\{y\})=u\}}$, where $u \in [0, 1]$ and δ_a denotes the Dirac measure at a.

The following example shows that the biases of the estimators of expected miscalibration can differ significantly even for calibrated models.

Example 5 (Different bias)
Consider a binary classification problem with $\mathcal{Y} = \{1, 2\}$ and $p(Y = 1) = p(Y = 2) = 1/2$ that is clearly separable, i.e., with $p(Y = 0 \,|\, X = x) \in \{0, 1\}$ for all $x \in \mathcal{X}$. Both the optimal model g^{opt} and the constant model $g \equiv (1/2, 1/2)$ are calibrated and hence the expected miscalibration η_{TV} is zero for both models.

Let us define an estimator $\hat{\eta}_{\text{TV}}$ of expected miscalibration with only one bin on the whole probability simplex and a single data point (X, Y). Then the bias of the estimator is $\mathbb{E}[\hat{\eta}_{\text{TV}}] - \eta_{\text{TV}} = 0$ for the perfect model since the expected miscalibration estimate is always zero. However, the constant model yields a bias of $\mathbb{E}[\hat{\eta}_{\text{TV}}] - \eta_{\text{TV}} = 1/2$ since in that case the expected miscalibration estimate is always $1/2$.

A.2 Proofs

Proposition 1 (Many calibrated classifiers). *For any measurable function $h\colon \mathcal{X} \to \mathcal{Z}$, where $\mathcal{Z}$ is some measurable space, the map g defined by*

$$g(X) := \mathbb{P}[Y \in \cdot \,|\, h(X)]$$

is a calibrated probabilistic classifier.

Proof. Let $y \in \mathcal{Y}$. We have

$$g(X)(\{y\}) = \mathbb{E}[g(X)(\{y\}) \,|\, g(X)] = \mathbb{E}[\mathbb{E}[\mathbb{1}_{\{y\}}(Y) \,|\, h(X)] \,|\, g(X)].$$

Since by definition $g(X)$ is a function of $h(X)$ it follows from the tower property that

$$g(X)(\{y\}) = \mathbb{E}[\mathbb{1}_{\{y\}}(Y) \,|\, g(X)] = \mathbb{P}[Y = y \,|\, g(X)].$$

Hence $g(X) = \mathbb{P}[Y \in \cdot \,|\, g(X)]$, and therefore the probabilistic classifier g is calibrated. □

Theorem 1. *Suppose that the calibration function r is Lipschitz continuous and $d\colon \Delta_{m-1} \times \Delta_{m-1} \to [0, \infty)$ is continuous and uniformly continuous in the first argument. Let $\{\mathbf{\Phi}_N\}_{N\in\mathbb{N}}$ be a sequence of finite data-independent partitions of $A \subseteq \Delta_{m-1}$ such that $\lim_{N\to\infty} \max_{S\in\mathbf{\Phi}_N} \operatorname{diam} S = 0$, where $\operatorname{diam} S := \sup_{x,y\in S} \|x - y\|_2$. Then*

$$\lim_{N\to\infty} \lim_{n\to\infty} \hat{\eta}_{d,N} = \eta_d,$$

with limits in the almost sure sense, where estimator $\hat{\eta}_{d,N}$ is defined according to (6) for each $N \in \mathbb{N}$.

Proof. To keep the notation simple we provide a proof for $A = \Delta_{m-1}$. The case $A \subsetneq \Delta_{m-1}$ follows in the same way by conditioning on $g(X) \in A$.

For $N \in \mathbb{N}$ let $\mathbf{\Phi}_N = \{\Phi_N^{(i)}\}_{i=1}^{l_N}$ be a finite data-independent partition of Δ_{m-1} such that $\lim_{N\to\infty} \max_i \operatorname{diam} \Phi_N^{(i)} = 0$. We define $\hat{r}_N^{(i)}$, $\hat{g}_N^{(i)}$, $\hat{p}_N^{(i)}$, and $\hat{\eta}_{d,N}$ analogously to the notation in Section 4.1. Similarly we denote the average output distribution, the average predicted distribution, and the proportion of predictions in subset $\Phi_N^{(i)}$ by $\overline{r}_N^{(i)} := \mathbb{E}[r(g(X)) \,|\, g(X) \in \Phi_N^{(i)}]$, $\overline{g}_N^{(i)} := \mathbb{E}[g(X) \,|\, g(X) \in \Phi_N^{(i)}]$, and $\overline{p}_N^{(i)} := \mathbb{P}[g(X) \in \Phi_N^{(i)}]$, respectively.

From the continuous mapping theorem it follows that for all $N \in \mathbb{N}$

$$\lim_{n\to\infty} \sum_{i=1}^{l_N} \left|\hat{p}_N^{(i)} d\left(\hat{r}_N^{(i)}, \hat{g}_N^{(i)}\right) - \overline{p}_N^{(i)} d\left(r(\overline{g}_N^{(i)}), \overline{g}_N^{(i)}\right)\right| = \sum_{i=1}^{l_N} \overline{p}_N^{(i)} \left|d\left(\overline{r}_N^{(i)}, \overline{g}_N^{(i)}\right) - d\left(r(\overline{g}_N^{(i)}), \overline{g}_N^{(i)}\right)\right|, \tag{8}$$

with limit in the almost sure sense.

Let $K \geq 0$ be a Lipschitz constant for calibration function r. Hence for all $N \in \mathbb{N}$ and $i \in \{1, \ldots, l_N\}$ we have

$$\begin{aligned} \left\|\overline{r}_N^{(i)} - r(\overline{g}_N^{(i)})\right\|_2 &= \left\|\mathbb{E}\left[r(g(X)) - r(\overline{g}_N^{(i)}) \,|\, g(X) \in \Phi_N^{(i)}\right]\right\|_2 \leq \mathbb{E}\left[\|r(g(X)) - r(\overline{g}_N^{(i)})\|_2 \,|\, g(X) \in \Phi_N^{(i)}\right] \\ &\leq K\mathbb{E}\left[\|g(X) - \overline{g}_N^{(i)}\|_2 \,|\, g(X) \in \Phi_N^{(i)}\right] \leq K\mathbb{E}\left[\operatorname{diam} \Phi_N^{(i)} \,|\, g(X) \in \Phi_N^{(i)}\right] = K \operatorname{diam} \Phi_N^{(i)} \\ &\leq K \max_{S\in\mathbf{\Phi}_N} \operatorname{diam} S. \end{aligned} \tag{9}$$

Let $\epsilon > 0$. Since by assumption distance function d is uniformly continuous in the first argument, there exists $\delta > 0$ such that for all $x, y, z \in \Delta_{m-1}$ with $\|x - y\|_2 < \delta$ inequality $|d(x, z) - d(y, z)| < \epsilon$ holds. From (9) and the assumption $\lim_{N\to\infty} \max_{S\in\mathbf{\Phi}_N} \operatorname{diam} S = 0$ we know that there exists $N_0 \in \mathbb{N}$ such that for all $N \geq N_0$ and all $i \in \{1, \ldots, l_N\}$ we have $\|\overline{r}_N^{(i)} - r(\overline{g}_N^{(i)})\|_2 < \delta$. Hence together with (8) we obtain for all $N \geq N_0$

$$\lim_{n\to\infty} \sum_{i=1}^{l_N} \left|\hat{p}_N^{(i)} d\left(\hat{r}_N^{(i)}, \hat{g}_N^{(i)}\right) - \overline{p}_N^{(i)} d\left(r(\overline{g}_N^{(i)}), \overline{g}_N^{(i)}\right)\right| < \sum_{i=1}^{l_N} \overline{p}_N^{(i)} \epsilon = \epsilon,$$

with limit in the almost sure sense. Since ϵ was chosen arbitrarily this implies

$$\lim_{N\to\infty} \lim_{n\to\infty} \sum_{i=1}^{l_N} \left|\hat{p}_N^{(i)} d\left(\hat{r}_N^{(i)}, \hat{g}_N^{(i)}\right) - \overline{p}_N^{(i)} d\left(r(\overline{g}_N^{(i)}), \overline{g}_N^{(i)}\right)\right| = 0, \tag{10}$$

with limits in the almost sure sense.

For all $N \in \mathbb{N}$ the triangle inequality yields, with limits taken in the almost sure sense,

$$\begin{aligned} \left|\eta_d - \lim_{n\to\infty} \hat{\eta}_{d,N}\right| &= \left|\mathbb{E}\left[d\left(r(g(X)), g(X)\right)\right] - \lim_{n\to\infty} \sum_{i=1}^{l_N} \hat{p}_N^{(i)} d\left(\hat{r}_N^{(i)}, \hat{g}_N^{(i)}\right)\right| \\ &\leq \left|\mathbb{E}\left[d\left(r(g(X)), g(X)\right)\right] - \sum_{i=1}^{l_N} \overline{p}_N^{(i)} d\left(r(\overline{g}_N^{(i)}), \overline{g}_N^{(i)}\right)\right| \\ &\quad + \left|\lim_{n\to\infty} \sum_{i=1}^{l_N} \hat{p}_N^{(i)} d\left(\hat{r}_N^{(i)}, \hat{g}_N^{(i)}\right) - \sum_{i=1}^{l_N} \overline{p}_N^{(i)} d\left(r(\overline{g}_N^{(i)}), \overline{g}_N^{(i)}\right)\right|. \end{aligned} \tag{11}$$

From the definition of the Riemann-Stieltjes integral it follows that

$$\lim_{N\to\infty}\left|\mathbb{E}\left[d\left(r(g(X)),g(X)\right)\right]-\sum_{i=1}^{l_N}\overline{p}_N^{(i)}d\left(r(\overline{g}_N^{(i)}),\overline{g}_N^{(i)}\right)\right|=0, \tag{12}$$

and (10) implies that

$$\lim_{N\to\infty}\left|\lim_{n\to\infty}\sum_{i=1}^{l_N}\hat{p}_N^{(i)}d\left(\hat{r}_N^{(i)},\hat{g}_N^{(i)}\right)-\sum_{i=1}^{l_N}\overline{p}_N^{(i)}d\left(r(\overline{g}_N^{(i)}),\overline{g}_N^{(i)}\right)\right|$$
$$\leq\lim_{N\to\infty}\lim_{n\to\infty}\sum_{i=1}^{l_N}\left|\hat{p}_N^{(i)}d\left(\hat{r}_N^{(i)},\hat{g}_N^{(i)}\right)-\overline{p}_N^{(i)}d\left(r(\overline{g}_N^{(i)}),\overline{g}_N^{(i)}\right)\right|=0, \tag{13}$$

with limits in the almost sure sense. Thus all in all, from Equations (11) to (13) we obtain

$$\left|\eta_d-\lim_{N\to\infty}\lim_{n\to\infty}\hat{\eta}_{d,N}\right|=\lim_{N\to\infty}\left|\eta_d-\lim_{n\to\infty}\hat{\eta}_{d,N}\right|\leq 0+0=0,$$

with limits in the almost sure sense. □

Theorem 2. *Let $d\colon\Delta_{m-1}\times\Delta_{m-1}\to[0,\infty)$ be a continuous convex function and let Φ be a finite data-independent partition of $A\subseteq\Delta_{m-1}$. Then*

$$\lim_{n\to\infty}\hat{\eta}_d\leq\eta_d, \tag{7}$$

with limit in the almost sure sense. Moreover, if d can be written as $d(p,p')=f(p-p')$ for a convex function f, then equality holds if and only if for every $\Phi^{(i)}$ there exists a set $S\subseteq\mathbb{R}^m$ such that $\mathbb{P}[r(g(X))-g(X)\in S\,|\,g(X)\in\Phi^{(i)}]=1$ and f coincides almost surely with an affine function on the convex hull of S.

Proof. To keep the notation simple we provide a proof for $A=\Delta_{m-1}$. The case $A\subsetneq\Delta_{m-1}$ follows in the same way by conditioning on $g(X)\in A$.

Let $\Phi=\{\Phi^{(i)}\}_{i=1}^{l}$ be a finite data-independent partition of Δ_{m-1}. Then we have

$$\eta_d=\mathbb{E}[d(r(g(X)),g(X))]=\sum_{i=1}^{l}\mathbb{P}[g(X)\in\Phi^{(i)}]\mathbb{E}[d(r(g(X)),g(X))\,|\,g(X)\in\Phi^{(i)}] \tag{14}$$
$$\geq\sum_{i=1}^{l}\mathbb{P}[g(X)\in\Phi^{(i)}]d\left(\mathbb{E}[r(g(X))\,|\,g(X)\in\Phi^{(i)}],\mathbb{E}[g(X)\,|\,g(X)\in\Phi^{(i)}]\right), \tag{15}$$

where (14) follows from the law of total probability and (15) from Jensen's inequality. Hence by the continuous mapping theorem we get

$$\eta_d\geq\lim_{n\to\infty}\sum_{i=1}^{l}\hat{p}^{(i)}d(\hat{r}^{(i)},\hat{g}^{(i)})=\lim_{n\to\infty}\hat{\eta}_d,$$

with limit in the almost sure sense. The exact equality condition is obtained by unwrapping the equality condition in Jensen's inequality. □

B Binning schemes

With enough data available the calibration function can be approximated by partitioning the probability simplex into bins and calculating the observed empirical frequencies of realized outcomes associated with the bins, e.g., by using the histogram regression of Nobel (1996). In the commonly considered binary classification setting the unit interval $[0, 1]$ is typically split into a given number of intervals of equal width (Bröcker 2008) (fixed-width binning). This approach can be extended to multiple dimensions by using symmetric equally-sized higher-dimensional bins but the number of bins grows exponentially with the number of classes. Additionally, predictions of neural networks after training are usually highly non-uniformly distributed, often making accurate estimation of the calibration function in multiclass classification with moderately sized amounts of data practically infeasible in large parts of the probability simplex.

Thus an attractive alternative is to partition the probability simplex into bins with approximately equal number of predictions. Bröcker (2008) suggests this binning procedure even in the case of non-uniformly distributed predictions of binary outcomes. In our study we employed a simple recursive partitioning scheme and split predictions along the mean of the dimension with highest variance as long as the number of predictions per bin was above a given threshold value, which was typically set to 1000 in our experiments. As discussed by Nobel (1996), different data-dependent binning schemes are possible and described in literature.

C Additional visualizations

C.1 Gaussian mixture model

C.1.1 Perfect model

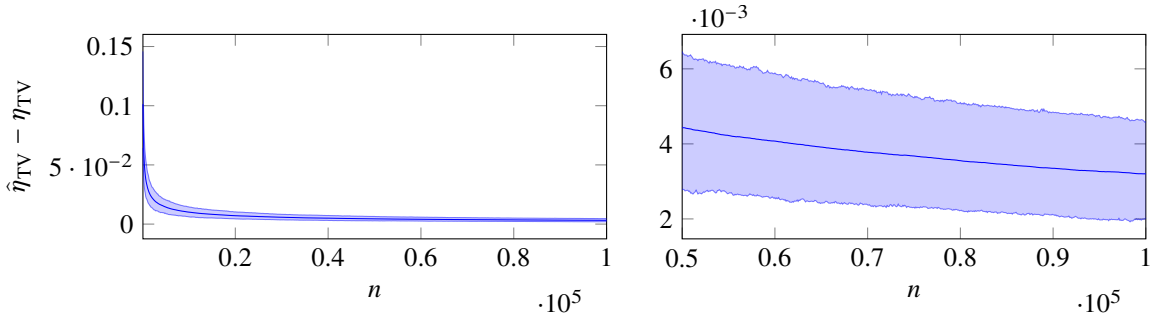


Figure 6: 5th percentile, mean, and 95th percentile of the difference of estimated and expected miscalibration of the perfect model ($\beta_0 = 0$, $\beta_1 = -2$) w.r.t. the total variation distance and 10 equally-sized bins (1000 series of random data).

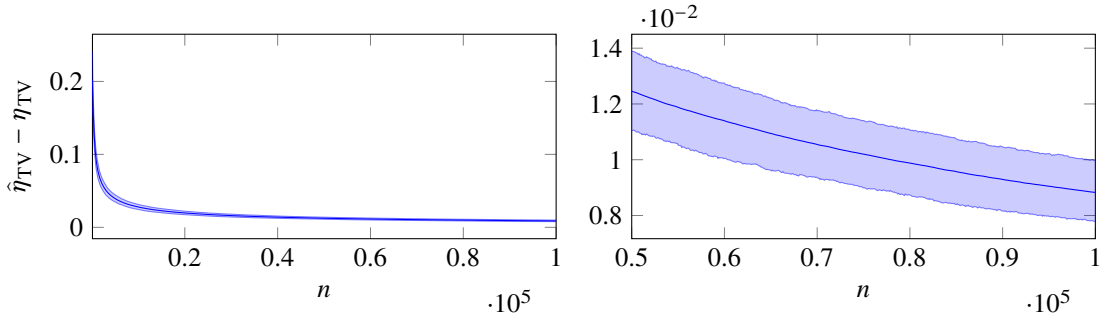
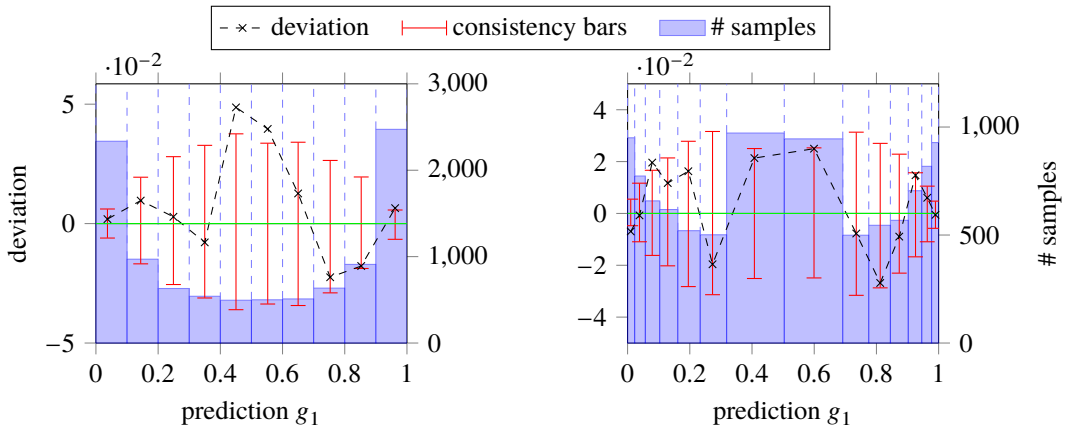


Figure 7: 5th percentile, mean, and 95th percentile of the difference of estimated and expected miscalibration of the perfect model ($\beta_0 = 0$, $\beta_1 = -2$) w.r.t. the total variation distance and 100 equally-sized bins (1000 series of random data).



(a) Equally-sized bins

(b) Data-dependent bins

Figure 8: Reliability diagrams for the perfect model ($\beta_0 = 0$, $\beta_1 = -2$) w.r.t. the total variation distance on a randomly generated test set (10000 inputs). Crosses indicate the deviation of the outcome distribution from the predictions in each bin. Blue bars show the distribution of predictions. Red bars visualize the 5th and 95th percentiles of the deviation in 1000 consistency resamples. The green curve shows the true analytical deviation.

C.1.2 Calibrated constant model

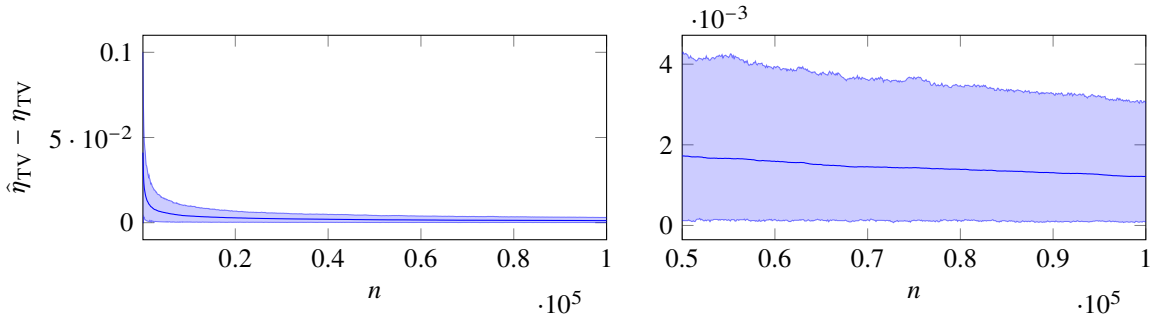


Figure 9: 5th percentile, mean, and 95th percentile of the difference of estimated and expected miscalibration of the calibrated constant model ($\beta_0 = \beta_1 = 0$) w.r.t. the total variation distance and 10 equally-sized bins (1000 series of random data).

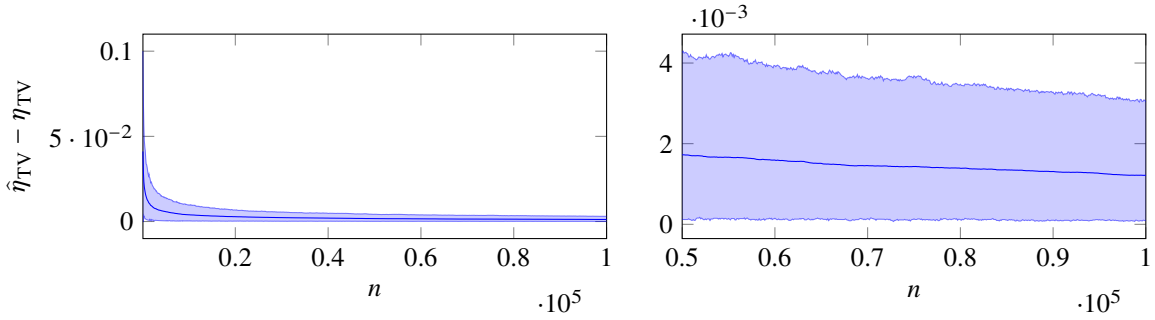
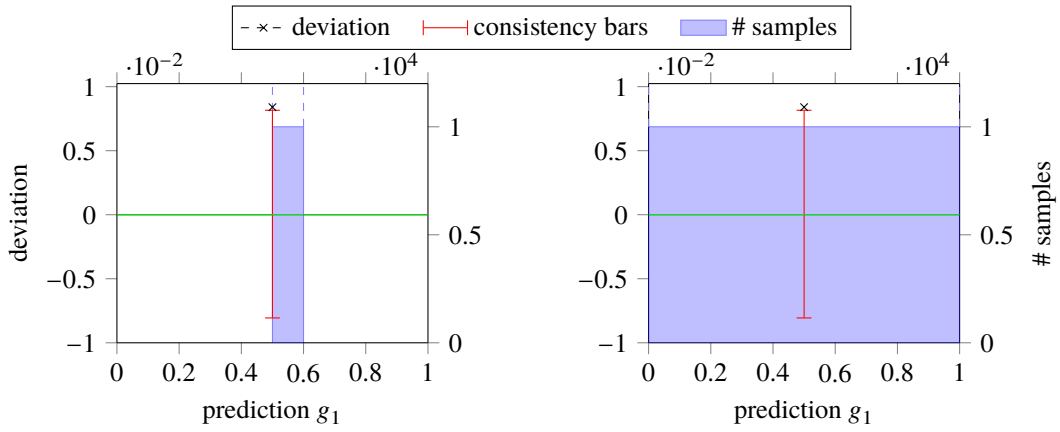


Figure 10: 5th percentile, mean, and 95th percentile of the difference of estimated and expected miscalibration of the calibrated constant model ($\beta_0 = \beta_1 = 0$) w.r.t. the total variation distance and 100 equally-sized bins (1000 series of random data).



(a) Equally-sized bins

(b) Data-dependent bins

Figure 11: Reliability diagrams for the calibrated constant model ($\beta_0 = \beta_1 = 0$) w.r.t. the total variation distance on a randomly generated test set (10000 inputs). Crosses indicate the deviation of the outcome distribution from the predictions in each bin. Blue bars show the distribution of predictions. Red bars visualize the 5th and 95th percentiles of the deviation in 1000 consistency resamples. The green curve shows the true analytical deviation.

C.1.3 Uncalibrated model

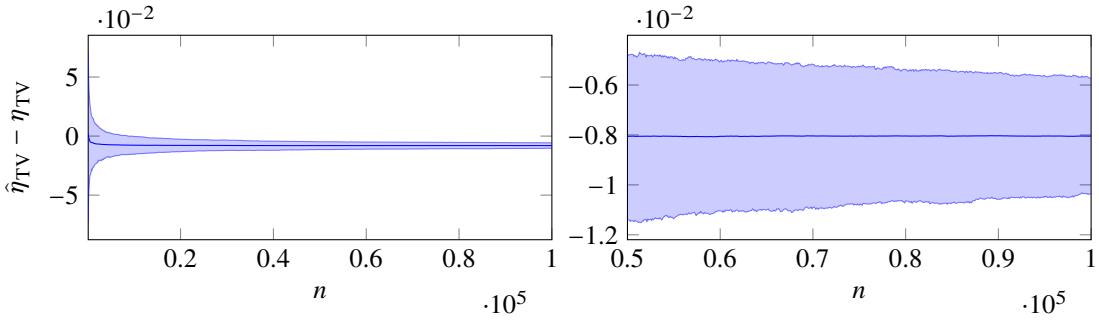


Figure 12: 5th percentile, mean, and 95th percentile of the difference of estimated and expected miscalibration of the uncalibrated model ($\beta_0 = \beta_1 = 1$) w.r.t. the total variation distance and 10 equally-sized bins (1000 series of random data).

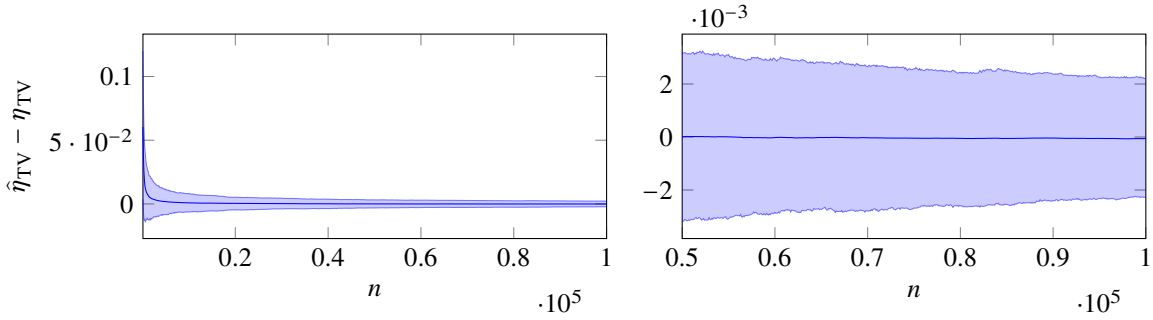
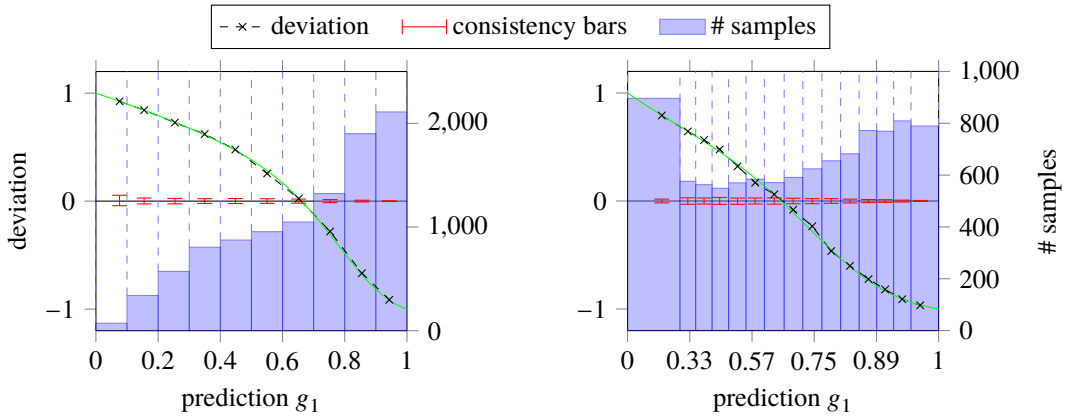


Figure 13: 5th percentile, mean, and 95th percentile of the difference of estimated and expected miscalibration of the uncalibrated model ($\beta_0 = \beta_1 = 1$) w.r.t. the total variation distance and 100 equally-sized bins (1000 series of random data).



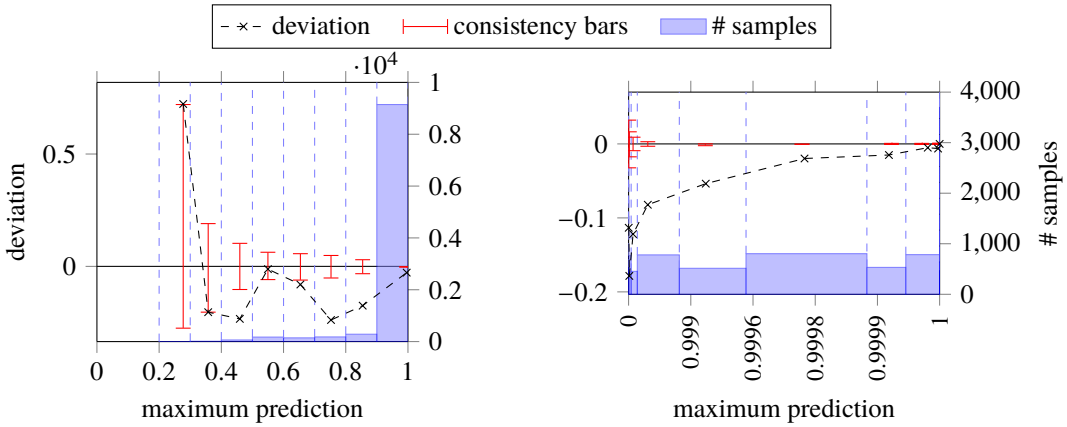
(a) Equally-sized bins

(b) Data-dependent bins

Figure 14: Reliability diagrams for the uncalibrated model ($\beta_0 = \beta_1 = 1$) w.r.t. the total variation distance on a randomly generated test set (10000 inputs). Dots indicate the deviation of the outcome distribution from the predictions in each bin. Blue bars show the distribution of predictions. Red bars visualize the 5th and 95th percentiles of the deviation in 1000 consistency resamples. The green curve shows the true analytical deviation.

C.2 Neural network models

C.2.1 DenseNet on CIFAR-10



(a) Equally-sized bins

(b) Data-dependent bins

Figure 15: Reliability diagrams for the maximum predictions of DenseNet on the CIFAR-10 test set w.r.t. the total variation distance. Crosses indicate the deviation of the outcome distribution from the predictions in each bin. Blue bars show the distribution of predictions. Red bars visualize the 5th and 95th percentiles of the deviation in 1000 consistency resamples.

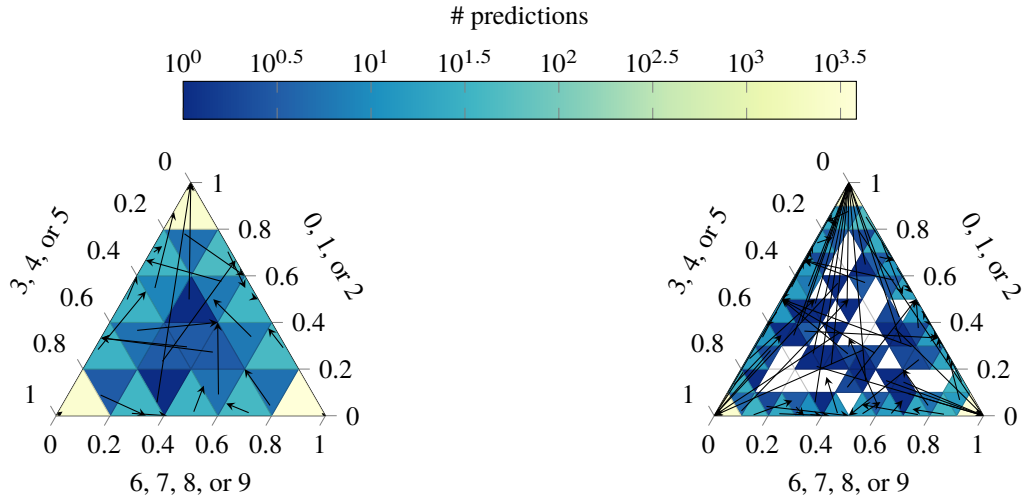
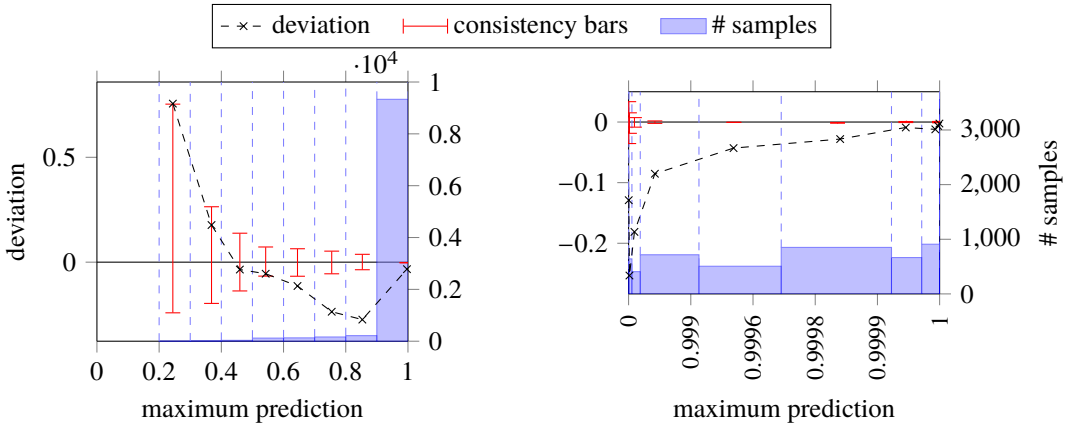


Figure 16: Two-dimensional reliability diagrams for DenseNet on the CIFAR-10 test set with 25 and 100 bins of equals. The predictions are grouped into three groups $\{0, 1, 2\}$, $\{3, 4, 5\}$, and $\{6, 7, 8, 9\}$ of the original classes. Arrows represent the deviation of the estimated calibration function value (arrow head) from the group prediction average (arrow tail) in a bin. The empirical distribution of predictions is visualized by color-coding the bins.

C.2.2 ResNet on CIFAR-10



(a) Equally-sized bins

(b) Data-dependent bins

Figure 17: Reliability diagrams for the maximum predictions of ResNet on the CIFAR-10 test set w.r.t. the total variation distance. Crosses indicate the deviation of the outcome distribution from the predictions in each bin. Blue bars show the distribution of predictions. Red bars visualize the 5th and 95th percentiles of the deviation in 1000 consistency resamples.

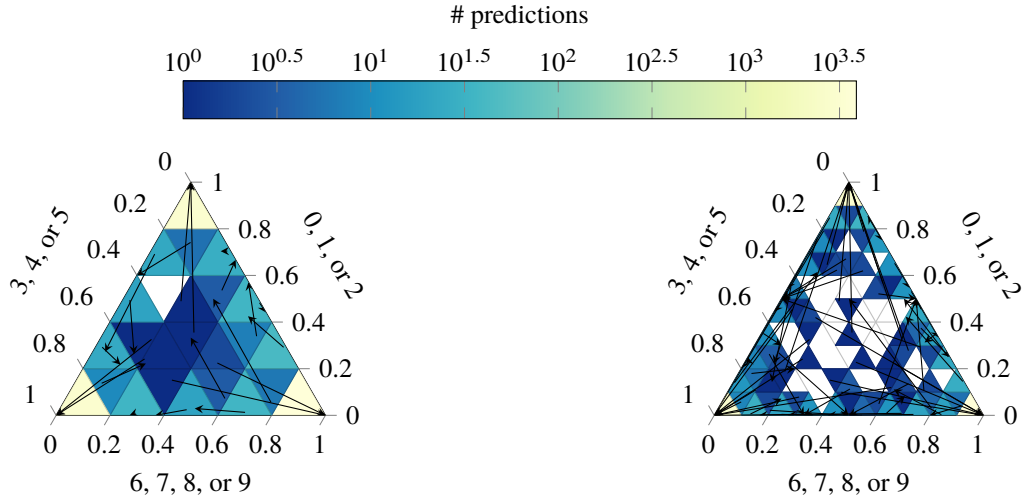


Figure 18: Two-dimensional reliability diagrams for ResNet on the CIFAR-10 test set with 25 and 100 bins of equal size. The predictions are grouped into three groups $\{0, 1, 2\}$, $\{3, 4, 5\}$, and $\{6, 7, 8, 9\}$ of the original classes. Arrows represent the deviation of the estimated calibration function value (arrow head) from the group prediction average (arrow tail) in a bin. The empirical distribution of predictions is visualized by color-coding the bins.

D Expected miscalibration estimates for neural networks

The results presented in Tables 3 to 5 show estimates of the expected miscalibration η for DenseNet, ResNet, and LeNet models trained on CIFAR-10, using different binning schemes (see Appendix B), calibration lenses (see Example 2), and distance functions (see Section 3.2). Estimates with respect to the total variation distance for maximum predictions and equally-sized bins correspond to the expected miscalibration error used by Guo et al. (2017). For comparison, we also approximate estimates of the expected miscalibration under the consistency assumption of a perfectly calibrated model.

Standard deviations are estimated using bootstrapping. The accuracy of the investigated DenseNet, ResNet, and LeNet models is 0.933 ± 0.002, 0.934 ± 0.002, and 0.727 ± 0.004, respectively.

Table 3: Estimates of the expected miscalibration for DenseNet trained on CIFAR-10.

Calibration lens	Distance d	Equally-sized bins		Data-dependent bins	
		$\hat{\eta}_d$	$\hat{\eta}_d^{\mathrm{id}}$	$\hat{\eta}_d$	$\hat{\eta}_d^{\mathrm{id}}$
Canonical	Total variation	0.059 ± 0.002	0.029 ± 0.001	0.041 ± 0.002	0.007 ± 0.001
	Squared Euclidean	0.072 ± 0.003	0.034 ± 0.001	0.046 ± 0.003	0.006 ± 0.001
Maximum	Total variation	0.038 ± 0.002	0.002 ± 0.001	0.038 ± 0.002	0.001 ± 0.001
	Squared Euclidean	0.054 ± 0.003	0.006 ± 0.001	0.053 ± 0.003	0.004 ± 0.001

Table 4: Estimates of the expected miscalibration for ResNet trained on CIFAR-10.

Calibration lens	Distance d	Equally-sized bins		Data-dependent bins	
		$\hat{\eta}_d$	$\hat{\eta}_d^{\mathrm{id}}$	$\hat{\eta}_d$	$\hat{\eta}_d^{\mathrm{id}}$
Canonical	Total variation	0.059 ± 0.002	0.022 ± 0.001	0.042 ± 0.002	0.007 ± 0.001
	Squared Euclidean	0.071 ± 0.003	0.028 ± 0.001	0.047 ± 0.003	0.005 ± 0.001
Maximum	Total variation	0.043 ± 0.002	0.002 ± 0.001	0.043 ± 0.002	0.001 ± 0.001
	Squared Euclidean	0.061 ± 0.003	0.004 ± 0.001	0.061 ± 0.003	0.004 ± 0.001

Table 5: Estimates of the expected miscalibration for LeNet trained on CIFAR-10.

Calibration lens	Distance d	Equally-sized bins		Data-dependent bins	
		$\hat{\eta}_d$	$\hat{\eta}_d^{\mathrm{id}}$	$\hat{\eta}_d$	$\hat{\eta}_d^{\mathrm{id}}$
Canonical	Total variation	0.219 ± 0.003	0.215 ± 0.003	0.027 ± 0.003	0.023 ± 0.002
	Squared Euclidean	0.243 ± 0.004	0.238 ± 0.003	0.024 ± 0.003	0.019 ± 0.002
Maximum	Total variation	0.007 ± 0.003	0.010 ± 0.002	0.009 ± 0.003	0.011 ± 0.002
	Squared Euclidean	0.010 ± 0.004	0.015 ± 0.003	0.013 ± 0.004	0.011 ± 0.003