Two fundamental limits for uncertainty quantification in predictive inference

Felipe Areces FARECES@STANFORD.EDU
Department of Electrical Engineering, Stanford University

Chen Cheng CHENCHENG@STANFORD.EDU
Department of Statistics, Stanford University

John Duchi JDUCHI@STANFORD.EDU
Department of Statistics and Electrical Engineering, Stanford University

Rohith Kuditipudi ROHITHK@STANFORD.EDU
Department of Computer Science, Stanford University

Editors: Shipra Agrawal and Aaron Roth

Abstract

We study the statistical hardness of estimating two basic representations of uncertainty in predictive inference: prediction sets and calibration error. First, we show that conformal prediction sets cannot approach a desired weighted conformal coverage level—with respect to a family of binary witness functions with VC dimension d—at a minimax rate faster than $O(d^{1/2}n^{-1/2})$. We also show that the algorithm in Gibbs et al. (2023) achieves this rate and that extending our class of conformal sets beyond thresholds of non-conformity scores to include arbitrary convex sets of non-conformity scores only improves the minimax rate by a constant factor. Then, under a similar VC dimension constraint on the witness function class, we show it is not possible to estimate the weighted weak calibration error at a minimax rate faster than $O(d^{1/4}n^{-1/2})$. We show that the algorithm in Kumar et al. (2019) achieves this rate in the particular case of estimating the squared weak calibration error of a predictor that outputs d distinct values.

Keywords: Lower Bounds, Conformal Prediction, Calibration, Uncertainty Quantification

1. Introduction

The goal of predictive inference is to produce a model whose output encodes not only a point prediction of a desired target but also an estimate of the prediction's reliability. To achieve this goal, we usually define a measure to quantify how well a model represents its own uncertainty, and develop algorithms guaranteeing that this measure is small. Conformal prediction provides a concrete example: given a training dataset $\{(X_i, Y_i)\}_{i=1}^n$ and a test point (X_{n+1}, Y_{n+1}) drawn i.i.d. from an unknown distribution, we seek to construct prediction sets $C(X_{n+1})$ such that $\mathbb{P}(Y_{n+1} \in C(X_{n+1})) = 1 - \alpha$. In this example, we know from Vovk et al. (2005) and Lei et al. (2018) that if our unknown distribution is continuous, then the intervals obtained via the split-conformal algorithm satisfy

$$|\mathbb{P}(Y_{n+1} \in C(X_{n+1})) - (1 - \alpha)| \leq \frac{1}{n+1}.$$

Gibbs et al. (2023) propose *weighted conformal coverage* as a more general (and more difficult to achieve) notion of coverage: C has valid coverage with respect to a class of binary witness functions $\mathcal{W} \subseteq \mathcal{X} \to \mathbb{R}$ if

$$\mathbb{E}\left[w(X_{n+1})\left(\mathbf{1}\{Y_{n+1} \in C(X_{n+1})\} - (1-\alpha)\right)\right] = 0\,,$$

Calibration error is another example: a binary predictor $f : \mathcal{X} \to [0,1]$ is *calibrated* if $f(X) \approx \mathbb{E}\left[Y \middle| f(X)\right]$ almost surely. There are various ways of quantifying the miscalibration of a predictor f, including the *expected calibration error*

$$\mathrm{ece}(f) := \mathbb{E}\left[\left|\mathbb{E}\left[Y \middle| f(X)\right] - f(X)\right|\right]\,.$$

Lee et al. (2023) and Arrieta-Ibarra et al. (2022) show that estimating the expected calibration error is hard in general, which motivates the relaxed notion of *weak calibration error* with respect to a class of binary witness functions $\mathcal{W} \subseteq \mathcal{X} \to \mathbb{R}$ as

$$\mathrm{CE}(f, \mathcal{W}) := \sup_{w \in \mathcal{W}} \mathbb{E}\left[w(S)(Y - S)\right]\,.$$

Conformal prediction and calibration are the two most widely-adopted frameworks for uncertainty quantification. Our contributions are fundamental lower bounds illustrating the statistical hardness of developing models under each framework. For the former, we provide sample complexity lower bounds on the weighted conformal coverage gap of conformal sets. For the latter, we focus on the complexity of testing calibration through tight lower bounds for the estimation of weak calibration error. Observe testing calibration is in general harder than producing a calibrated predictor, since one can always trivially achieve the latter by designing a predictor that only returns the sample mean of the outputs. In practice, we generally want to find a calibrated model that is somehow close to an existing model, which is fundamentally tied to determining the level of miscalibration of the original model.

1.1. Organization

We split our paper into two main sections corresponding to lower bounds for conformal prediction and (weak) calibration, respectively. For our conformal results, Sections 2.1 and 2.2 provide an overview of the problem and cover related work and definitions. Section 2.3 develops our main lower bound result for quantile sets of non-conformity scores, and Section 2.4 develops a matching upper bound. Finally, Section 2.5 discusses how considering the larger class of convex sets of non-conformity scores impacts the minimax rate.

For weak calibration, Section 3.1 provides an overview of the problem and discusses related work. Section 3.2 contains our main lower bound result for binary function classes with fixed VC dimension. Finally, Sections 3.3 and 3.4 discuss how to obtain a matching upper bound as well as other consequences of our main result.

We defer the technical proofs to the appropriate sections of the Appendix.

2. Conformal prediction

2.1. Overview

Given calibration datapoints $\{(X_i, Y_i)\}_{i=1}^n$ and a confidence level $\alpha \in (0,1)$, conformal prediction methods seek to construct a confidence set mapping $C : \mathcal{X} \rightrightarrows \mathcal{Y}$ so that for a new point

(X_{n+1}, Y_{n+1}) then $Y_{n+1} \in C(X_{n+1})$ with probability $1-\alpha$. The typical approach for constructing this confidence set is to associate a non-conformity score $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ to predictions and return the set of all predictions whose scores fall below some threshold; the idea is to set the threshold so that the coverage of the set always exceeds the conformal guarantee. For example, the original split conformal method Vovk et al. (2005) takes the standard $1-\alpha$ quantile regression estimator

$$\hat{q} = \underset{q}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (1-\alpha)(s(X_i, Y_i) - q)_+ + \alpha(q - s(X_i, Y_i))_+ \,,$$

and makes it slightly more conservative by defining $\hat{q}_c$ as the $\lceil (n+1)(1-\alpha) \rceil / n$ quantile to construct $C(x) = \{y \in \mathcal{Y} \mid s(x,y) \leq \hat{q}_c\}$. This conformal set guarantees $\mathbb{P}\left(Y_{n+1} \in C(X_{n+1})\right) \geq 1-\alpha$ without distributional assumptions and $|\mathbb{P}\left(Y_{n+1} \in C(X_{n+1})\right) - (1-\alpha)| \leq \frac{1}{n}$ when the scores $S_i = s(X_i, Y_i)$ are distinct Lei et al. (2018). The first guarantee is a result of the conformal correction, but the second one is intuitively a consequence of the underlying quantile estimators being – with high probability – very accurate. The split conformal example illustrates how the overall performance of conformal methods is heavily tied to the quality of underlying estimators, so we seek to explore the fundamental limits of these estimation problems beyond the marginal case.

The original split conformal algorithm is very useful in practice but its guarantees turn out to be somewhat unsatisfying as they allow our confidence set to have wildly varying coverage levels for different values of X_{n+1}. As a possible solution we can impose the stronger conditional coverage condition $\mathbb{P}\left(Y_{n+1} \in C(x) \mid X_{n+1} = x\right) \geq 1-\alpha$ for all $x \in \mathcal{X}$, but Barber et al. (2021) show that this goal is impossible to achieve in any meaningful sense when Y has a density as the only sets satisfying this condition have infinite expected Lebesgue measure. A natural middle ground between full conditional coverage a marginal coverage is the idea of group conditional coverage where algorithms seek to guarantee $\mathbb{P}\left(Y_{n+1} \in C(x) \mid X_{n+1} \in G\right) = 1-\alpha$ for G in some set of target groups $\mathcal{G}$. When using quantile sets of non-conformity scores some algorithms also define these groups in terms of the threshold function used to construct the conformal intervals, namely if $C(x) = \{y \mid s(x,y) \leq f(x)\}$ we can define groups $G_\tau = \{x \mid f(x) = \tau\}$ and set $\mathcal{G}_c = \{G_\tau \mid \tau \in \mathbb{R}\}$, we often refer to this type of guarantee as a threshold calibrated guarantee. Using this setup Jung et al. (2023) provides an algorithm to obtain both group conditional coverage guarantees with respect to some finite collection $\mathcal{G}$ and threshold calibrated guarantees simultaneously. Several algorithms such as those in Gupta et al. (2022) and Bastani et al. (2022) also provide similar guarantees in the adversarial setting.

Gibbs et al. (2023) propose a framework encapsulating these approaches via a dual weighted coverage condition where for a class $\mathcal{W} \subseteq \mathcal{X} \to \mathbb{R}$, C has valid coverage if

$$\mathbb{E}\left[w(X_{n+1})\left(\mathbf{1}\{Y_{n+1} \in C(X_{n+1})\} - (1-\alpha)\right)\right] = 0 \,,$$

for all $w \in \mathcal{W}$. Note that using appropriate function classes $\mathcal{W}$ allows us to recover the coverage guarantees discussed earlier, so we can use this general type of guarantee to explore the properties of many techniques used to approach full conditional validity. These refined notions of coverage mitigate some of the issues with the original conformal guarantee but still seem to promote some arguably undesirable sets by allowing averaging over the randomness of C which is usually a function of the random datapoints $\{(X_i, Y_i)\}_{i=1}^n$. As a trivial example, for any calibration split we can simply ignore the data and choose our interval based on a sample of the auxiliary random variable $V \sim \text{Bernoulli}(1-\alpha)$ so that for all $x \in \mathcal{X}$ if $V = 1$ then $C(x) = \mathcal{Y}$ and $C(x) = \emptyset$ otherwise. This satisfies all the notions of coverage discussed so far including conditional coverage and

it's corresponding trivial achievability result (since the proposed interval will have infinite expected Lebesgue measure for $\mathcal{Y} = \mathbb{R}^k$), but is not useful in any practical sense. For this reason we focus on a measure of conformal error that takes this phenomenon into account

$$D(C) = \left|\mathbb{P}\left(Y \in C(X)\right) - (1-\alpha)\right|,$$

and its weighted variant

$$D_{\mathcal{W},p}(C) = \sup_{w \in \mathcal{W}} \mathbb{E}\left[\frac{w(X)}{\|w(X)\|_p}(\mathbf{1}\{Y \in C(X)\} - (1-\alpha))\right],$$

where we take C as fixed, our expectation is only over (X, Y), and $\|\cdot\|_p$ is the L^p norm on the probability space $(\mathcal{X}, \mathcal{B}(\mathcal{X}), P_X)$ for $p \in [1, \infty]$. In fact, it can now be shown that under an i.i.d. assumption (rather than solely exchangeability) and similar regularity conditions, the original split conformal procedure with standard quantile estimators achieves $D(\hat{C}_n) \leq O(n^{-1/2})$ with high probability (an immediate consequence of our main achievability result in Theorem 2). Moreover, this new framework now allows us to derive minimax lower bounds to determine whether these procedures are optimal to approach a desired coverage level.

The general conformal prediction framework does not impose restrictions on the class of allowable sets C, but in practice we usually care about sets that minimize some optimality criterion over all sets with $1-\alpha$ coverage. We will use the notion of *perfectability* to represent this idea, where we say a score is perfectable for a distribution P with respect to some loss φ if the sets defined through quantiles of the score minimize the loss over all sets with $1-\alpha$ coverage for any $\alpha \in (0, 1)$. In fact, if we assume that our optimal set mappings $C_{1-\alpha}$ are nested and inner-semicontinuous for all α we can always construct them using the standard quantile form $C_q(x) = \{y \in \mathcal{Y} \mid s(x, y) \leq q(x)\}$. It is also the case that any score function can be associated with an optimality criterion under which the quantile sets $C_q(x)$ are optimal. In our setup we emulate the second stage of the split conformal procedure so we assume that the score function s is pre-specified, which implies that the corresponding quantile sets are always optimal with respect to some criterion. Therefore, as long as our target set mappings are nested we can limit our analysis to quantile sets without loss of generality as their optimality is solely determined by the choice of score function. For example, if our objective is to produce minimum length intervals as measured by $\text{Leb}(C(x))$ with $X \in \mathbb{R}^k$ and $Y \in \mathbb{R}^m$ then for any distribution on (X, Y) with a density $f(y|x)$ we can use the score $s(x, y) = \exp\left[-f(y|x)\right]$ to produce the optimal intervals through quantiles. For a more precise definition of optimality and perfectability, and further discussion of these results, including proofs, refer to Sections 2.2 and Appendix C.

Restricting our analysis to sets of the form $C_q(x)$ we focus on the case where $\mathcal{W} \subseteq \mathcal{X} \to \{-1, 1\}$ as this is sufficient to approach the conditional coverage guarantee. In this context we provide lower bounds for the weighted conformal error of any estimator q and argue that the underlying estimator in the algorithm proposed by Gibbs et al. (2023) achieves this rate. We also show that the minimax rate remains unchanged even when considering the larger class of convex sets of scores $C_{a,b}(x) = \{y \in \mathcal{Y} \mid s(x, y) \in [a(x), b(x)]\}$, which implies that adding complexity to our intervals in this way does not improve coverage in the minimax sense.

2.2. Formal problem definitions

As described in the previous section we focus on a setting with covariates $X_i \in \mathcal{X}$ and targets $Y_i \in \mathcal{Y}$ where $(X_i, Y_i) \overset{i.i.d.}{\sim} P_{X,Y}$. We also have a pre-specified bounded score function $s : \mathcal{X} \times \mathcal{Y} \to [0, 1]$ that allows us to define $S_i = s(X_i, Y_i)$ so that (X_i, S_i) are still i.i.d. We now define the class of quantile sets of scores

$$\mathcal{C}_1 := \{C_q : \mathcal{X} \rightrightarrows \mathcal{Y} \mid C_q(x) = \{y \in \mathcal{Y} \mid s(x, y) \leq q(x)\}, q : \mathcal{X} \to [0, 1]\}, \tag{1}$$

and the class of convex sets of scores

$$\mathcal{C}_2 := \{C_{a,b} : \mathcal{X} \rightrightarrows \mathcal{Y} \mid C_{a,b}(x) = \{y \in \mathcal{Y} \mid s(x, y) \in [a(x), b(x)]\}, a, b : \mathcal{X} \to [0, 1]\}. \tag{2}$$

In practice, we usually care about problems where the score function is designed to produce reasonably good confidence intervals, so we define the notion of *perfectability* to make this requirement more specific. We say a score s is φ-perfectable with respect to a distribution $P_{X,Y}$ for some loss $\varphi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ if for all $\alpha \in (0, 1)$ there exists $C_{1-\alpha} \in \mathcal{C}_1$ such that

$$\mathbb{P}\left(Y \in C_{1-\alpha}(x) \mid X = x\right) \geq 1 - \alpha,$$

and $C_{1-\alpha}$ minimizes the loss

$$L_\varphi(C) = \int_{\mathcal{X}} \int_{\mathcal{Y}} \varphi(x, y) \mathbb{1}\{y \in C(x)\} dP(x, y)$$

across all set mappings $C : \mathcal{X} \rightrightarrows \mathcal{Y}$ with $\mathbb{P}\left(Y \in C(x) \mid X = x\right) \geq 1 - \alpha$. We note that for any score there exists a φ so that s is perfectable for all distributions $P_{X,Y}$ with no point masses. Conversely, for any nested inner-semicontinuous set mapping minimizing $L_\varphi(C)$ among C with valid conditional coverage, we can find a score function producing those sublevel sets as shown in Appendix C.

We now define the weighted conformal error when sets are constructed through quantiles or intervals as

$$D_{\mathcal{W},p}(q) := \sup_{w \in \mathcal{W}} \mathbb{E}\left[\frac{w(X)}{\|w(X)\|_p}(\mathbf{1}\{S \leq q(X)\} - (1 - \alpha))\right] \tag{3}$$

$$D_{\mathcal{W},p}(a, b) := \sup_{w \in \mathcal{W}} \mathbb{E}\left[\frac{w(X)}{\|w(X)\|_p}(\mathbf{1}\{S \in [a(X), b(X)]\} - (1 - \alpha))\right], \tag{4}$$

respectively. Note that in both cases $\mathcal{W} = \{1, -1\}$ recovers the marginal conformal error, while by choosing $\mathcal{W}$ to be all measurable functions $\mathcal{X} \to \{-1, 1\}$ we obtain the conditional conformal error. The minimax errors with respect to a class of distributions $\mathcal{P}$ now become

$$\mathfrak{M}_n(\mathcal{C}_1) := \inf_{\hat{q}} \sup_{P \in \mathcal{P}} \mathbb{E}\left[D_{\mathcal{W},p}(\hat{q})\right] \tag{5}$$

$$\mathfrak{M}_n(\mathcal{C}_2) := \inf_{\hat{a}, \hat{b}} \sup_{P \in \mathcal{P}} \mathbb{E}\left[D_{\mathcal{W},p}(\hat{a}, \hat{b})\right], \tag{6}$$

where $\hat{q}, \hat{a}, \hat{b}$ are our estimators.

2.3. Minimax lower bound for quantile score sets

Our first result shows how to obtain a lower bound for (5). The first step is to define the class of distributions of interest which must be restricted enough to allow matching achievability results. In this case we focus on the class of distributions $\mathcal{P}$ where $S|X$ has a continuous density with respect to the Lebesgue measure on the unit interval, as well as uniform upper and lower bounds. The key insight in this case is to note that if our function class $\mathcal{W}$ satisfies $\text{VC}(\mathcal{W}) = d$ then we can find $x \in \mathcal{X}^d$ such that for all $v \in \{-1, 1\}^d$ there exists $w \in \mathcal{W}$ with $(w(x_1), \cdots, w(x_d)) = v$. In this case, for any distribution $P \in \mathcal{P}$ such that X is uniformly distributed on x

$$D_{\mathcal{W},p}(q) \geq \frac{1}{d} \sum_{i=1}^{d} \left| \mathbb{P}\left(S \leq q(x_i) \mid X = x_i, q\right) - (1-\alpha) \right|.$$

If we now use p_i^0 to denote the uniform lower bound of the density corresponding to $S|X = x_i$ under P it is clear that

$$D_{\mathcal{W},p}(q) \geq \frac{\min_j p_j^0}{d} \sum_{i=1}^{d} \left| q(x_i) - q_i^\star \right|,$$

where $q_i^\star$ is the true $1-\alpha$ quantile of $S|X = x_i$. We can now simply treat q as a d-dimensional vector $q = [q(x_1), \cdots, q(x_d)]$ – with some notational overloading – to lower bound our weighted conformal error with the estimation error associated with our conditional quantile estimators

$$D_{\mathcal{W},p}(q) \geq \frac{\min_j p_j^0}{d} \left\| q - q^\star \right\|_1,$$

which is significantly easier to handle using classical lower bound techniques.

We would also like to incorporate the notion of optimality we introduced in Section 2.2 and ensure that our lower bounds capture the fundamental complexity of relevant practical conformal problems, rather than relying on a scores and optimality criteria that would never be used in practice. For this reason we construct our distributions so that the quantile sets minimize the score dependent mean loss for $\varphi(x, y) = s(x, y)$ and the distribution dependent mean loss for $\varphi(x, y) = \exp\left[-f(s|x)\right]$ (where $f(s|x)$ is the density of $S|X$), simultaneously. This optimality criterion is reasonable since it is equivalent to minimizing $\text{Leb}(\{s(x, y)|y \in C(x)\})$ and matches the traditional goal of minimizing the size of our sets as measured by $\text{Leb}(C(x))$ in standard setups such as when $Y \in \mathbb{R}$ with score $s(x, y) = |y - f(x)|$ and Y has a unimodal symmetric distribution centered at $f(x)$. However, it is important to note that this is not always equivalent to the goal of minimizing $\text{Leb}(C(x))$, it simply illustrates that our lower bound relies on scores and distributions on those scores that are compatible with standard notions of optimality at least in some instances. We can now present the main theorem of this section.

Theorem 1 *Let $\mathcal{W}$ be a binary function class with $\text{VC}(\mathcal{W}) = d$ and $\mathcal{P}$ be the class of distributions on (X, S) where $S|X$ has a continuous density with uniform upper and lower bounds $f(s|x) \in [\frac{1}{2}, \frac{3}{2}]$, and with respect to which s is $\exp\left[-f(s|x)\right]$-perfectable. In this case*

$$\mathfrak{M}_n(\mathcal{C}_1) \geq c_1 \sqrt{\frac{d\alpha(1-\alpha)}{n}},$$

for $n \geq \frac{cd}{\alpha(1-\alpha)}$ and numerical constants c, c_1.

Proof Refer to Sections B.2 and B.3 in Appendix B. ■

This result implies that even when our score s satisfies a perfectability definition that matches the traditional goal of minimizing $\text{Leb}(C(x))$ in standard setups, our lower bound holds.

2.4. A corresponding upper bound for quantile score sets

It is now important to compare our minimax lower bound with upper bounds for existing algorithms to verify whether it is achievable. In particular, if we assume $\mathcal{W} = \{\Phi(\cdot)^T\beta : \beta \in \mathbb{R}^d\}$ is a class of linear functions over the basis $\Phi : \mathcal{X} \to \mathbb{R}^d$, Gibbs et al. (2023) argue that by appropriately computing

$$\hat{g}_{S_{n+1}} := \underset{g\in\mathcal{W}}{\text{argmin}} \frac{1}{n+1}\sum_{i=1}^{n+1}(1-\alpha)(S_i - g(X_i))_+ + \alpha(g(X_i) - S_i)_+ \,,$$

we can define intervals $C(x) = \{y \in \mathcal{Y} : s(X_{n+1}, y) \leq \hat{g}_{s(X_{n+1},y)}(x)\}$ that satisfy

$$|\mathbb{E}\left[w(X_{n+1})(\mathbb{1}\{Y_{n+1} \in C(X_{n+1})\} - (1-\alpha))\right]| \leq \frac{d}{n+1}\mathbb{E}\left[\max_{1\leq i\leq n+1}|w(X_i)|\right] \,,$$

for all $w \in \mathcal{W}$ when $S|X$ is continuous. This convergence result has the interesting feature of depending on the expectation of a maximum which could potentially be very large, so it is natural to ask if this result is a feature of the underlying estimator or a product of the conformal correction. In fact, we show that the underlying estimator essentially matches the minimax rate for $p \geq 2$ up to higher order terms under mild regularity conditions when $\mathcal{W}$ contains a function class with VC dimension d, so at least in this case it seems unlikely that the convergence result can be improved by using a better underlying estimator.

Theorem 2 *Let $\mathcal{W}$ be a function vector space with the usual inner product*

$$\langle w_1, w_2\rangle = \mathbb{E}\left[w_1(X)w_2(X)\right] = \int_{\mathcal{X}} w_1(x)w_2(x)dP(x) \,,$$

that admits a d-dimensional orthonormal basis $(\tilde{w}_1, \cdots, \tilde{w}_d)$ such that $\forall i : \sup_{x\in\mathcal{X}} \tilde{w}_i(x) \leq M$. Then the estimator

$$\hat{g} = \underset{g\in\mathcal{W}}{\text{argmin}} \frac{1}{n}\sum_{i=1}^{n}(1-\alpha)(S_i - g(X_i))_+ + \alpha(g(X_i) - S_i)_+ \,,$$

satisfies

$$\mathbb{E}\left[D_{\mathcal{W},2}(\hat{g})\right] \leq C_1\sqrt{\frac{d}{n}} + C_2\frac{d^{3/2}M}{n} \,,$$

for numerical constants C_1, C_2 as long as $S|X$ is continuous.

This result always matches the lower bound in Theorem 1 with respect to n – up to higher order terms and a factor of $c\alpha(1-\alpha)$ – and also matches the minimax optimal rate for d when $\mathcal{W}$ contains a function class with VC dimension d.

2.5. Minimax optimality beyond quantile score sets

It is standard for conformal algorithms to focus solely on quantile score sets similar to (1) as these are often easier to compute and analyze than more expressive confidence set classes, while providing robust conformal guarantees. However, the effects of such simplifications on the performance of conformal algorithms is unclear. One natural way to approach this question is using the minimax conformal error as defined in section 2.2 and ask how increasing the complexity of our confidence set class impacts the minimax rate. We believe this is a hard question to answer in general as identifiability issues quickly arise for larger confidence set classes where we can have multiple distinct sets in our class providing $1-\alpha$ coverage for a given distribution, as in such cases the δ-separation condition required by most standard lower bound techniques becomes difficult to enforce. In this section we provide a partial answer to this question by arguing that extending our confidence set classes to include all convex sets of scores – rather than only quantile sets – only improves the minimax rate by a constant factor, even when restricting distributions to those where intervals minimize $\text{Leb}(\{s(x,y)|y \in C(x)\})$ under $1-\alpha$ coverage constraints in analogy to our argument justifying perfectability in the quantile case.

Theorem 3 *Let $\mathcal{W}$ be a binary function class with $VC(\mathcal{W}) = d$ and $\mathcal{P}$ be the class of distributions on (X, S) with respect to which intervals of scores minimize Leb$(\{s(x,y)|y \in C(x)\})$ under $1-\alpha$ coverage constraints, and $S|X$ has a continuous density with uniform upper and lower bounds $f(s|x) \in [\frac{1}{2}, \frac{3}{2}]$. In this case*

$$\mathfrak{M}_n(\mathcal{C}_2) \geq c_2\sqrt{\frac{\alpha(1-\alpha)d}{n}},$$

for $n \geq cd$ and numerical constants c and $c_2 < c_1$.

Proof Refer to Sections B.5 and B.7 in Appendix B. ■

3. Calibration

3.1. Overview

We focus our study of calibration on the setting of binary prediction, where the forecaster's goal is to develop a prediction model $f : \mathcal{X} \to [0,1]$ which for any covariates $x \in \mathcal{X}$ satisfies $f(x) \approx \mathbb{E}[Y|X=x] = \mathbb{P}(Y=1|X=x)$. Achieving this goal is a tall order as such a classifier would of course have perfect predictive accuracy. A weaker but still desireable requirement is that we should at least be able to interpret f as a probability, i.e., f is *calibrated* if $f(X) \approx \mathbb{E}[Y|f(X)]$. Defining the random variable $S = f(X)$, the most natural calibration measure is the expected calibration error defined as

$$\text{ece}(f) := \mathbb{E}\left[|\mathbb{E}[Y|S] - S|\right].$$

This naturally generalizes to arbitrary norms on the $([0,1], \mathcal{B}([0,1]), P_S)$ probability space

$$\text{ece}_{\|\cdot\|}(f) := \|\mathbb{E}[Y|S] - S\|,$$

which more directly point to the dual formulation

$$\text{ece}_{\|\cdot\|}(f) = \sup_{\|w\|_* \leq 1} \mathbb{E}\left[w(S)(Y-S)\right] .$$

This naturally prompts the definition of a – potentially – weaker notion of calibration error, namely the *calibration error relative to a function class* $\mathcal{W} \subseteq \mathcal{X} \to \mathbb{R}$ or weak calibration error,

$$\text{CE}(f, \mathcal{W}) := \sup_{w \in \mathcal{W}} \mathbb{E}\left[w(S)(Y-S)\right] .$$

This quantity will be – for the most part – the focus of our analysis.

At first glance estimating quantities similar to $\text{ece}(f)$ might seem straightforward as we could simply use the naive plug-in estimator

$$\widehat{\text{ece}}_{\text{plug-in}}(f) := \frac{1}{n}\sum_{s \in \mathcal{S}} \left| \sum_{i=1}^{n} (Y_i - s)\mathbf{1}\{S_i = s\} \right| = \sum_{s \in \mathcal{S}} \hat{p}_s |\hat{y}_s - s| ,$$

where $\mathcal{S}$ is the set of observed scores and $\hat{p}_s, \hat{y}_s$ are the standard estimators for $\mathbb{P}(S = s)$ and $\mathbb{E}[Y|S = s]$. The most glaring flaw with this approach is its reliance on the convergence of $\hat{p}_s, \hat{y}_s$ which is evidently impossible if S supported on a set of non-zero Lebesgue measure, as we will rarely observe samples with the same score. This naive estimator arises naturally from the dual witness function definition of $\text{ece}(f)$ by simply replacing the true expectation with an empirical expectation

$$\widehat{\text{ece}}_{\text{plug-in}}(f) = \sup_{\|w\|_\infty \leq 1} \mathbb{E}_{P_n}\left[w(S)(Y-S)\right] ,$$

which is part of the larger family of estimators

$$\widehat{\text{CE}}(f, \widehat{\mathcal{W}}) = \sup_{w \in \widehat{\mathcal{W}}} \mathbb{E}_{P_n}\left[w(S)(Y-S)\right] .$$

In fact, any of these estimators could potentially be used to estimate $\text{ece}(f)$ and the overall error will be bounded by the classic approximation/estimation error decomposition

$$|\widehat{\text{CE}}(f, \widehat{\mathcal{W}}) - \text{ece}(f)| \leq |\widehat{\text{CE}}(f, \widehat{\mathcal{W}}) - \text{CE}(f, \mathcal{W})| + |\text{CE}(f, \mathcal{W}) - \text{ece}(f)| .$$

In most practical applications when estimating $\text{ece}(f)$ the function classes $\widehat{\mathcal{W}}, \mathcal{W}$ are taken to be binning function classes

$$\mathcal{W}_B = \widehat{\mathcal{W}}_B = \left\{ \sum_{i=1}^{d} v_i \mathbb{1}\{s \in B_i\}, v \in \{-1, 1\}^d \right\} ,$$

where $B_1, \cdots, B_d$ are convex sets that partition $[0, 1]$ and more generally

$$\mathcal{W}_{B^p} = \left\{ \sum_{i=1}^{d} v_i \mathbb{1}\{s \in B_i\}, \|v\|_{L_q(P)} \leq 1 \right\}$$
$$\widehat{\mathcal{W}}_{B^p} = \left\{ \sum_{i=1}^{d} v_i \mathbb{1}\{s \in B_i\}, \|v\|_{L_q(P_n)} \leq 1 \right\} ,$$

for q such that $\frac{1}{p} + \frac{1}{q} = 1$ when estimating $\text{ece}_{\|\cdot\|_p}(f)$.

It is usually hard to provide valid upper bounds in this general setup, but when S has a distribution supported on a finite number of scores $\{s_1, \cdots, s_d\}$ and $B = \{\{s_1\}, \cdots, \{s_d\}\}$ standard concentration arguments by Kumar et al. (2019) show that

$$\left|\widehat{\text{CE}}(f, \widehat{\mathcal{W}}_{B^2})^2 - \text{CE}(f, \mathcal{W}_{B^2})^2\right| \leq \tilde{O}\left(\sqrt{\frac{\text{CE}(f, \mathcal{W}_{B^2})^2}{n \min_i \mathbb{P}(S = s_i)}} + \frac{d}{n}\right),$$

with high probability under certain regularity conditions. However, Ferro and Fricker (2012) and Roelofs et al. (2022) point out that $\widehat{\text{CE}}(f, \widehat{\mathcal{W}}_{B^p})$ is a biased estimator and discuss strategies to mitigate this bias, with Kumar et al. (2019) proposing a debiased variant $\hat{\mathcal{E}}^2_{\text{db}}$ that satisfies

$$\left|\hat{\mathcal{E}}^2_{\text{db}} - \text{CE}(f, \mathcal{W}_{B^2})^2\right| \leq \tilde{O}\left(\sqrt{\frac{\text{CE}(f, \mathcal{W}_{B^2})^2}{n \min_i \mathbb{P}(S = s_i)}} + \frac{\sqrt{d}}{n}\right), \tag{7}$$

with high probability under the same conditions. These results suggest that even in this basic setup the estimator has at least two regimes with different rates. If our model is perfectly calibrated the debiased variant achieves a surprisingly fast rate of $\tilde{O}(d^{1/2}n^{-1})$, whereas for any uncalibrated model the rate will drop to $\tilde{O}(d^{1/2}n^{-1/2})$ in the best case. In light of this result, tight lower bounds are of particular interest as they would allow us to determine if these different regimes and rates are a fundamental property of the estimation problem or of this estimator in particular, so we explore this problem in the next section.

3.2. Minimax lower bounds for weak calibration error

Due to the similarities in the formulation of weak calibration problem and weighted conformal objectives, it is not surprising that the first step to develop our lower bounds is to impose a condition that allows us to control the complexity of the function class $\mathcal{W}$, and since $\mathcal{W} \subset \mathcal{X} \to \{-1, 1\}$ is once again sufficient to approximate $\text{ece}(f)$ we restrict our focus to binary function classes. However, in this case we assume that $\mathcal{W}$ can shatter at most d points in $[\epsilon, 1-\epsilon]$ for $\epsilon > 0$, or more explicitly

$$\exists (s_1, \cdots, s_d) \in [\epsilon, 1-\epsilon]^d, \forall v \in \{-1, 1\}^d, \exists w \in \mathcal{W} : (w(s_1), \cdots, w(s_d)) = v\,.$$

This is very similar to the VC dimension requirement we used to control the complexity of the function class in the weighted conformal case, with the crucial difference that the shattered points cannot be too close to the edges of the unit interval. The reason for this correction will become evident in our lower bound construction, but it is intuitively related to the fact that it is easier to test if a model f is perfectly calibrated if it only outputs extreme values. With these constraints in mind we now present the main result of this section.

Theorem 4 *Given a function class $\mathcal{W} \subseteq [0,1] \to \{-1,1\}$ that can shatter at most d points in $[\epsilon, 1-\epsilon]$ for $\epsilon \in (0, \frac{1}{2})$, then*

$$\inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P\left[\left|\text{CE}(f, \mathcal{W}) - \hat{\theta}\right|\right] \geq c_1 \frac{d^{1/4}}{n^{1/2}} \sqrt{\epsilon(1-\epsilon)}\,,$$

for $n \geq c\frac{\sqrt{d}}{\epsilon(1-\epsilon)}$ and numerical constants c, c_1 where $\mathcal{P}$ is the set of distributions on outputs and scores (Y, S).

Proof Refer to Section D.1 in Appendix D. ■

It is important to point out that this theorem not only provides a global lower bound, but also a local lower bound for the perfectly calibrated distribution P_0 as we chose it to be one of the two tilts in our Le Cam two-point construction. Interestingly, we see that this lower bound shares some of the features of the upper bound for the debiased estimator (7) as the optimal rate when our model is perfectly calibrated is $O(d^{1/4}n^{-1/2})$ rather than the standard $O(d^{1/2}n^{-1/2})$. However, our result as presented in Theorem 4 is not compatible with the upper bounds in Kumar et al. (2019) so we provide the following corollary to make the comparison explicit.

Corollary 5 *Let $\mathcal{P}_{\mathcal{S}}$ be the set of distributions on outputs and scores (Y, S) such that S is supported on $\mathcal{S} = \{s_1, \cdots, s_d\}$ with $s_i \in [\epsilon, 1-\epsilon]$ for $\epsilon > 0$, and $B = \{\{s_1\}, \cdots, \{s_d\}\}$ then*

$$\inf_{\hat{\theta}} \sup_{P \in \mathcal{P}_{\mathcal{S}}} \mathbb{E}_P \left[\left| \mathrm{CE}(f, \mathcal{W}_{B^2})^2 - \hat{\theta} \right| \right] \geq c_1 \frac{d^{1/2}}{n} \epsilon(1-\epsilon) \vee c_2 \sqrt{\frac{\epsilon(1-\epsilon)\mathrm{CE}(f, \mathcal{W}_{B^2})^2}{n}},$$

for $n \geq c \max\left\{ \frac{\sqrt{d}}{\epsilon(1-\epsilon)}, \frac{1}{\mathrm{CE}(f, \mathcal{W}_{B^2})^2 \epsilon} \right\}$ and numerical constants c, c_1, c_2, as long as the weak calibration error satisfies $\mathrm{CE}(f, \mathcal{W}_{B^2}) \leq \frac{1}{2} - \epsilon$.

Proof Refer to Section D.2 in Appendix D. ■

It is now clear that our result matches the upper bound for the debiased estimator up to logarithmic factors when the model is perfectly calibrated, but is missing a factor of at least $\sqrt{d}$ when this is not the case. We will see in the next section that this is not in fact looseness of our lower bound, but rather of the existing upper bound. However, before shifting our focus to the achievability result we provide one more consequence of our main theorem.

Note that the witness functions used to obtain the separation condition in our proof of Theorem 4 satisfy $\|w\|_{L_q(P)} = 1$ for $q = \frac{p}{1-p} \in [1, \infty]$ so they also belong to the function class $\mathcal{W}_{B^p}$ and since our lower bound scales with d we can make the task of estimating $\mathrm{ece}_{\|\cdot\|_p}(f)$ arbitrarily hard as shown in the following corollary. The proof for part (i) follows immediately from our construction in Theorem 4 and the remaining two are direct consequences of taking $d \uparrow \infty$. These last two parts provide a new proof for the impossibility of estimating $\mathrm{ece}_{\|\cdot\|_p}(f)$ in general. Lee et al. (2023) use calibration curves to show a related result on the impossibility of testing the null hypothesis of calibration against the alternative hypothesis of ε_0 mis-calibraton, by arguing that any test must have worst case test risk equal to 1.

Corollary 6 *Let $Z_i = (f(X_i), Y_i)$ and define the worst-case test risk for the testing problem between classes $H_0 : P \in \mathcal{P}_0$ and $H_1 : P \in \mathcal{P}_1$ as*

$$R_n(\Psi | \mathcal{P}_0, \mathcal{P}_1) := \sup_{P \in \mathcal{P}_0} P(\Psi(Z_1^n) \neq 0) + \sup_{P \in \mathcal{P}_1} P(\Psi(Z_1^n) \neq 1),$$

and let $P_{p,\delta} = \{P | \mathrm{ece}_{P, \|\cdot\|_p}(f) = \delta\}$.

(i) If there exists $\epsilon \in (0, \frac{1}{2})$ such that $f(\mathcal{X}) \cap [\epsilon, 1-\epsilon]$ has cardinality at least d, then there is a distribution P_0 such that $\mathrm{ece}_{P_0, \|\cdot\|_p}(f) = 0$ and for any $0 \leq \delta \leq \epsilon$

$$\inf_{\Psi} R_n(\Psi | P_{p,0}, P_{p,\delta}) \geq 1 - \frac{n\delta^2}{2\sqrt{d}} \frac{1}{\epsilon(1-\epsilon)}.$$

(ii) If there exists $\epsilon \in (0, \frac{1}{2})$ *such that* $f(\mathcal{X}) \cap [\epsilon, 1-\epsilon]$ *has infinite cardinality then* $\mathcal{P}_0$ *is non-empty and for any* $0 \leq \delta \leq \epsilon$

$$\liminf_n \inf_\Psi R_n(\Psi | P_{p,0}, P_{p,\delta}) = 1\,.$$

(iii) If there exists a neighborhood U *of* $\frac{1}{2}$ *such that* $U \subset f(\mathcal{X})$ *then* $\mathcal{P}_0$ *is non-empty and for any* $\delta < \frac{1}{2}$

$$\liminf_n \inf_\Psi R_n(\Psi | P_{p,0}, P_{p,\delta}) = 1\,.$$

3.3. A corresponding upper bound

As discussed in the previous section our lower bound in corollary 5 does not fully match the upper bounds as presented in Kumar et al. (2019). However, we propose the following refinement of their argument to show that their debiased estimator actually achieves this minimax rate.

Theorem 7 *Let* $p_i = \mathbb{P}(S = s_i) > 0$ *for all* i*. Then the debiased estimator*

$$\hat{\mathcal{E}}^2_{\mathrm{db}} = \sum_{i=1}^{d} \hat{p}_s \left[(s_i - \hat{y}_{s_i})^2 - \frac{\hat{y}_s(1-\hat{y}_s)}{\hat{p}_s n - 1} \right],$$

satisfies

$$|\mathcal{E}^2_{\mathrm{db}} - \mathrm{CE}(f, \mathcal{W}_{B^2})^2| \leq C_1 \sqrt{\frac{\mathrm{CE}(f, \mathcal{W}_{B^2})^2}{n} \log\left(\frac{2}{\delta}\right)} + C_2 \frac{\sqrt{d}}{n} \log\left(\frac{n}{\delta}\right),$$

with probability at least $1 - 4\delta$ *for some numerical constants* C_1, C_2 *and sufficiently large* $n \geq c \log(\frac{d}{\delta})$ *with* $\frac{1}{c} \leq \log\left(\frac{1}{\min_i p_i}\right)$.

Proof Refer to Section D.3 in Appendix D. ■

Combining the results of Theorem 7 and Corollary 5 now shows that even in a basic setting where S has finite support, the problem of estimating $\mathrm{CE}(f, \mathcal{W}_{B^2})^2$ – and thus also $\mathrm{ece}_{\|\cdot\|_2}(f)^2$ – has two different regimes with different rates.

3.4. Some more consequences of our lower bound for weak calibration error

3.4.1. A LESS ADVERSARIAL LOWER BOUND

Theorem 4 lower bounds the calibration error with respect to general binary function classes. In the case of binning function classes $\mathcal{W}_B$, which are a specific instance of a binary function class, the result implies the complexity of the problem in this case scales according to the cardinality of the set B_ϵ of bins contained in $[\epsilon, 1-\epsilon]$. Whereas Theorem 4 gives a worst-case result over the distribution of scores S and outcomes Y, if we restrict our analysis to binning function classes with $|B_\epsilon| = d$ then we can adapt the same ideas to obtain a more fine-grained result which holds instance-wise for any fixed distribution of scores.

Theorem 8 *Given a binning function class $\mathcal{W}_B$ with $|B_\epsilon| = d$ for some $\epsilon > 0$ the for any predefined distribution Q on scores S such that $\min_{B \in B_\epsilon} Q(S \in B) = q_0$ and $q_1 = \max_{B \in B_\epsilon} Q(S \in B)$*

$$\inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P \left[\left| \mathrm{CE}(f, \mathcal{W}_B) - \hat{\theta} \right| \right] \geq c_1 \frac{q_0}{q_1} \cdot \frac{d^{1/4}}{n^{1/2}} \sqrt{\epsilon(1-\epsilon)}\,,$$

for $n \geq \frac{c}{\epsilon(1-\epsilon) q_1 d^{1/2}}$ and numerical constants c, c_1, where $\mathcal{P}$ is the set of distributions on outputs Y.

Proof Refer to Section D.4 in Appendix D. ■

This setup is less general than our original problem (the result applies only to binning function classes) but is still relevant to practical applications and provides a less adversarial perspective that illustrates how the complexity of estimating calibration measures depends on the distribution of the scores S. Perhaps not surprisingly, our result suggests that without further assumptions the problems becomes easier when the probability of each of the bins differs a lot. This is reasonable since even in the case where we have strong regularity conditions that makes estimation of $\mathrm{ece}(f)$ possible, the lower bound is only capturing the difficulty of estimating $\mathrm{CE}(f, \mathcal{W})$ and not how good of a proxy for $\mathrm{ece}(f)$ it is, so we would expect that estimating $\mathrm{CE}(f, \mathcal{W})$ becomes easier as the *effective* number of bins decreases but the result is a progressively worse approximation of $\mathrm{ece}(f)$. In practice, bins are often chosen to have roughly the same probability so our new Theorem 8 is essentially equivalent to Theorem 4.

3.4.2. A LOWER BOUND FOR SMOOTH CALIBRATION

So far we have solely focused on calibration error with respect to binary witness function classes, but many papers such as Blasiok et al. (2023) show interest in the smooth calibration error where the function class $\mathcal{W}_L$ is the set of all $[-1, 1]$ bounded L-Lipschitz functions. It is now natural to wonder if we can obtain sharp lower bounds for estimation in this case, and in fact Theorem 4 already provides a lower bound for this situation by simply noting that $\mathcal{W}_L$ can shatter $\lfloor \frac{L(1-2\epsilon)}{2} \rfloor$ scores in $[\epsilon, 1-\epsilon]$.

Theorem 9 *Let $\mathcal{W}_L$ be the function class containing all $[-1, 1]$ bounded L-Lipschitz functions then*

$$\inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P \left[\left| \mathrm{CE}(f, \mathcal{W}_L) - \hat{\theta} \right| \right] \geq c_1 \frac{L^{1/4}(1-2\epsilon)^{1/4}}{n^{1/2}} \sqrt{\epsilon(1-\epsilon)}\,,$$

for $n \geq c \frac{\sqrt{L(1-2\epsilon)}}{\epsilon(1-\epsilon)}$, numerical constants c, c_1 and $L \geq \frac{4}{1-2\epsilon}$, where $\mathcal{P}$ is the set of distributions on outputs and scores (Y, S).

The smooth calibration algorithm proposed by Blasiok et al. (2023) seems to match the minimax rate with respect to n but not with respect to L, and it is unclear if this mismatch is due to looseness in our lower bound or suboptimality of the algorithm.

4. Discussion

We have shown tight lower bounds illustrating the fundamental hardness of developing uncertainty quantification models using conformal prediction and calibration. Yet, a number of questions remain open.

First, all of our lower bound techniques for both settings rely on bounding the VC dimension of the witness function class; this approach produces tight lower bounds for commonly used witnesses classes such as binning functions, but it may be overly restrictive in general. For example, Theorem 9 provides some insight for the case when our function class satisfies a Lipschitz constraint, but requires Lipschitz constants $L \gg 1$ in general and it is unclear if the minimax rate $O(L^{1/4}n^{-1/2})$ is indeed tight. We believe it might be possible to obtain tighter lower bounds by developing lower bound techniques specialized for Lipschitz function classes.

For the specific case of weighted conformal validity, we discuss in Section 2.5 how allowing convex intervals of non-conformity scores only improves the minimax rate by a constant factor. It is unclear whether this result holds in general; specifically, it would be interesting to further motivate the usual choice of quantile sets of non-conformity scores by showing that they are optimal in some sense with respect to a larger family of confidence set mappings.

Finally, for the case of weak calibration, Blasiok et al. (2023) provide a unifying theory developed to understand calibration via the notion of distance to calibration; it would be interesting to explore what the fundamental limits are for those metrics. For example, one of these metrics is lower-distance to calibration, which Blasiok et al. (2023) show to be equivalent to smooth calibration error with $L = 1$ up to constant factors, and is thus not covered by our lower bounds. Additionally, they propose interval calibration as a binning-based measure of calibration that uses randomized bins, for which our lower bounds are also not applicable. Tight lower bounds for estimation of these quantities could not only improve our understanding of calibration, but also suggest potential improvements to existing estimators to make them more efficient in practice.

Acknowledgments

This work was supported by Office of Naval Research grant N00014-22-12669 and NSF grant IIS-2006777. Felipe Areces acknowledges support from the Mr. and Mrs. Chun Chiu Stanford Graduate Fellowship.

References

Imanol Arrieta-Ibarra, Paman Gujral, Jonathan Tannen, Mark Tygert, and Cherie Xu. Metrics of calibration for probabilistic predictions. *Journal of Machine Learning Research*, 23(351):1–54, 2022.

P. Assouad. Deux remarques sur l'estimation. *Comptes Rendus des Séances de l'Académie des Sciences, Série I*, 296(23):1021–1024, 1983.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference*, 10(2):455–482, 2021.

Osbert Bastani, Varun Gupta, Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Practical adversarial multivalid conformal prediction. *arXiv:2206.01067 [cs.LG]*, 2022.

Jaroslaw Blasiok, Parikshit Gopalan, Lunjia Hu, and Preetum Nakkiran. A unifying theory of distance from calibration. In *Proceedings of the Fifty-Fifth Annual ACM Symposium on the Theory of Computing*, 2023. URL https://arxiv.org/abs/2211.16886.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: a Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Christopher A.T. Ferro and Thomas E. Fricker. A bias-corrected decomposition of the brier score. *Quarterly Journal of the Royal Meteorological Society*, 138(668):1954–1960, 2012.

Isaac Gibbs, John Cherian, and Emmanuel Candès. Conformal prediction with conditional guarantees. *arXiv:2305.12616 [stat.ME]*, 2023.

Varun Gupta, Christopher Jung, Georgy Noarov, Mallesh M. Pai, and Aaron Roth. Online multivalid learning: Means, moments, and prediction intervals. In *Proceedings of the Thirteenth Innovations in Theoretical Computer Science (ITCS)*, 2022.

Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Batch multivalid conformal prediction. *arXiv:2209.15145 [cs.LG]*, 2023.

Ananya Kumar, Percy Liang, and Tengyu Ma. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems 32*, 2019.

Donghwan Lee, Xinmeng Huang, Hamed Hassani, and Edgar Dobriban. T-Cal: An optimal test for the calibration of predictive models. *Journal of Machine Learning Research*, 24(335):1–72, 2023.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113 (523):1094–1111, 2018.

Rebecca Roelofs, Nicholas Cain, Jonathon Shlens, and Michael C. Mozer. Mitigating bias in calibration error estimation. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, pages 4036–4054, 2022.

Vladimir Vovk, Alexander Grammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

Appendix A. Overview of lower bound techniques for weighted conformal prediction

As a starting point to motivate the techniques used to obtain our lower bounds – particularly the one for (6) – we introduce some well known lower bounding methods. For this purpose we define a mapping $\theta : \mathcal{P} \to \Theta$ from distributions to their relevant parameters, a semimetric $\rho : \Theta \times \Theta \to \mathbb{R}_+$ on the space Θ, and a non-decreasing function $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$ with $\Phi(0) = 0$. The most fundamental technique that can be applied to this setup is Le Cam's two point method.

Theorem 10 *[Adapted from Wainwright (2019) (15.4)] Let $P_1, P_2 \in \mathcal{P}$ be two distributions such that*

$$\rho(\theta(P_1), \theta(P_2)) \geq 2\delta\,,$$

for some $\delta > 0$. Then

$$\inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}\left[\Phi\left(\rho(\theta(P), \hat{\theta})\right)\right] \geq \frac{\Phi(\delta)}{2}\left[1 - \|P_1 - P_2\|_{\mathrm{TV}}\right].$$

Interestingly, this simple method is sufficient to obtain a lower bound when $\mathcal{W} = \{-1, 1\}$ (i.e. for the marginal conformal error), but we intuitively know that the problem must become harder as the complexity of $\mathcal{W}$ grows since achieving small conditional conformal error in finite samples is impossible. In fact, Le Cam's two point method's reliance on only two alternatives turns out to be too restrictive to capture the effects of the complexity of $\mathcal{W}$ on the minimax rate, so we turn to Assouad's method for a more powerful approach.

Theorem 11 *[Adapted from Assouad (1983)] Let $\mathcal{V} = \{-1, 1\}^d$ and $\{P_v\}_{v \in \mathcal{V}} \subset \mathcal{P}$ be a family of distributions indexed by the hypercube, then if there exists a function $\hat{v} : \theta(\mathcal{P}) \to \{-1, 1\}^d$ such that*

$$\Phi\left(\rho(\theta, \theta(P_v))\right) \geq 2\delta \sum_{j=1}^{d} \mathbb{1}\{\hat{v}(\theta)_j \neq v_j\},$$

for some $\delta > 0$ then

$$\begin{aligned}\inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}\left[\Phi\left(\rho(\theta(P), \hat{\theta})\right)\right] &\geq \delta \sum_{j=1}^{d} \left[1 - \|P_{+j} - P_{-j}\|_{\mathrm{TV}}\right] \\ &\geq \delta d \left[1 - \max_{d_{\mathrm{ham}}(v, v') \leq 1} \|P_v - P_{v'}\|_{\mathrm{TV}}\right],\end{aligned}$$

where $P_{+j} = 2^{1-d} \sum_{v : v_j = 1} P_v$.

This technique is quite similar to Le Cam's two point method, as it essentially imposes a separation condition that allows us to split the problem into d-dimensions and propose two tilts for each dimension. Intuitively, this is enough to provide a lower bound for (5) as our confidence intervals only have one degree of freedom so we will usually be unable to perfectly cover two distributions simultaneously. However, this is not the case for (6) where we now have 2 degrees of freedom, which will often allow us to cover two distributions simultaneously. As a simple example consider the distributions with densities

$$\begin{aligned}p_1(s) &= \frac{5}{4}\mathbb{1}\{s \in [0, 1/2]\} + \frac{3}{4}\mathbb{1}\{s \in (1/2, 1]\} \\ p_2(s) &= \frac{3}{4}\mathbb{1}\{s \in [0, 1/2]\} + \frac{5}{4}\mathbb{1}\{s \in (1/2, 1]\},\end{aligned}$$

where it is clear that there are unique intervals $[0, q_1], [0, q_2]$ with $q_1 \neq q_2$ that provide $1-\alpha$ coverage for each distribution, but the interval $[\frac{1}{2} - \frac{1-\alpha}{2}, \frac{1}{2} + \frac{1-\alpha}{2}]$ perfectly covers both. This suggests that a valid lower bound for (6) should provide more than one tilt per dimension, so we propose a novel lower bound technique to deal with this situation.

Theorem 12 *Let $\mathcal{V} = \{1, \cdots, K\}^d$ and $\{P_v\}_{v \in \mathcal{V}} \subset \mathcal{P}$. Given loss functions $\mathcal{L}_{P,i} : \theta(\mathcal{P}) \to \mathbb{R}_+$ so that for all $i \in \{1, \cdots, d\}$ and v' in*

$$\mathcal{V}_i' = \{(v_1, \cdots, v_i, \cdots, v_d) | v_i = 0 \text{ and } v_j \in \{1, \cdots, K\} \text{ for } j \neq i\}$$

the following property is satisfied

$$\inf_{\theta \in \theta(\mathcal{P})} \frac{1}{K} \sum_{k=1}^{K} \mathcal{L}_{P_{v'+ke_i},i}(\theta) \geq \delta \,,$$

where e_i is the i-th canonical basis vector, then

$$\inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}\left[\sum_{i=1}^{d} \mathcal{L}_{P,i}(\hat{\theta})\right] \geq \frac{\delta}{K^d} \sum_{i=1}^{d} \sum_{v' \in \mathcal{V}_i'} \left(1 - \sum_{k=2}^{K} \|P_{v'+e_i} - P_{v'+ke_i}\|_{TV}\right) .$$

We defer the proof of this result to a later section in this appendix but note that this technique is very similar to Assouad's method in spirit, with the key difference that our separation condition is much weaker. In essence, our result does not require any two distributions with different tilts in the i-th dimension to be separated, but simply that we cannot simultaneously approximate parameters for all tilts in a given dimension. We will naturally pay a rate penalty because of this weakening – as seen in the following corollary – but this technique will return a non-trivial bound when our estimators can approximate 2 parameters simultaneously, but not arbitrarily many.

Corollary 13 *If the conditions for Theorem 12 are satisfied and for all $i \in \{1, \cdots, d\}, k \in \{2, \cdots, K\}, v' \in \mathcal{V}_i'$*

$$\|P_{v'+e_i} - P_{v'+ke_i}\|_{\mathrm{TV}} \leq \frac{1}{2(K-1)} \,,$$

then

$$\inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}\left[\sum_{i=1}^{d} \mathcal{L}_{P,i}(\hat{\theta})\right] \geq \frac{d\delta}{2K} \,.$$

This corollary nicely illustrates that our weaker requirements cause our lower bound to be strictly looser than Assouad's by a constant factor when $K = 2$, and that we usually have to pay a penalty proportional to $\frac{1}{K^2}$ by allowing K tilts with weak separation.

Appendix B. Deferred proofs for weighted conformal prediction

B.1. Proof of Theorem 12

Start by observing that if we let nature sample V from $\mathcal{V} = \{1, \cdots, K\}^d$ uniformly at random and draw (X, S) according to distribution P_V we get the lower bound

$$\begin{aligned}
\mathfrak{M}_n = \inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}\left[\sum_{i=1}^{d} \mathcal{L}_{P,i}(\hat{\theta})\right] &\geq \inf_{\hat{a},\hat{b}} \frac{1}{K^d} \sum_{v \in \mathcal{V}} \mathbb{E}_{P_v}\left[\sum_{i=1}^{d} \mathcal{L}_{P_v,i}(\hat{a},\hat{b})\right] \\
&\geq \inf_{\hat{a},\hat{b}} \frac{\delta}{K^d} \sum_{v \in \mathcal{V}} \sum_{i=1}^{d} \mathbb{E}_{P_v}\left[\mathbb{1}\left\{\mathcal{L}_{P_v,i}(\hat{a},\hat{b}) \geq \delta\right\}\right]
\end{aligned}$$

We now define $\mathcal{V}_i' = \{(v_1, \cdots, v_i, \cdots, v_d) | v_i = 0 \text{ and } v_j \in \{1, \cdots, K\} \text{ for } j \neq i\}$ so that if e_i is the i-th canonical basis vector

$$\begin{aligned}\mathfrak{M}_n &\geq \inf_{\hat{a},\hat{b}} \frac{\delta}{K^d} \sum_{i=1}^{d} \sum_{v' \in \mathcal{V}_i'} \sum_{k=1}^{K} \mathbb{E}_{P_{v'+ke_i}} \left[\mathbb{1}\left\{ \mathcal{L}_{P_{v'+ke_i},i}(\hat{a},\hat{b}) \geq \delta \right\} \right] \\ &\geq \inf_{\hat{a},\hat{b}} \frac{\delta}{K^d} \sum_{i=1}^{d} \sum_{v' \in \mathcal{V}_i'} \left(\mathbb{E}_{P_{v'+e_i}} \left[\sum_{k=1}^{K} \mathbb{1}\left\{ \mathcal{L}_{P_{v'+ke_i},i}(\hat{a},\hat{b}) \geq \delta \right\} \right] - \sum_{k=2}^{K} \|P_{v'+e_i} - P_{v'+ke_i}\|_{\text{TV}} \right).\end{aligned}$$

It is evident by our separation condition that for any $\hat{a}, \hat{b}$ at least one $k \in \{1, \cdots, K\}$ satisfies $\mathcal{L}_{P_{v'+ke_i},i}(\hat{a},\hat{b}) \geq \delta$ so that our lower bound becomes

$$\inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}\left[\sum_{i=1}^{d} \mathcal{L}_{P,i}(\hat{\theta}) \right] \geq \inf_{\hat{a},\hat{b}} \frac{\delta}{K^d} \sum_{i=1}^{d} \sum_{v' \in \mathcal{V}_i'} \left(1 - \sum_{k=2}^{K} \|P_{v'+e_i} - P_{v'+ke_i}\|_{\text{TV}} \right).$$

B.2. Proof of Theorem 1 without continuous density requirement

Start by defining the sub-class of distributions $\mathcal{P}' \subset \mathcal{P}$ where X is uniformly distributed on the set $\{x_1, \cdots, x_d\}$ shattered by $\mathcal{W}$, and $S|X$ has density uniformly lower bounded by $\frac{1}{2}$. Under this setup it is clear that

$$\mathfrak{M}_n(\mathcal{C}_1) \geq \frac{1}{2d} \inf_{\hat{q}} \sup_{P \in \mathcal{P}'} \mathbb{E}\left[\|\hat{q} - q^*\|_1 \right].$$

We can now define the perturbation function

$$g(s) = \frac{1}{\alpha} \mathbf{1}\{s > (1-\alpha)\} - \frac{1}{1-\alpha} \mathbf{1}\{s \leq (1-\alpha)\},$$

and define a family of distributions in $\mathcal{P}'$ indexed by $v \in \{-1, 1\}^d$ with conditional densities

$$p_v(s|x_i) = 1 + \delta \min\{v_i, 0\} g(s)$$

for $\delta \in (0, \frac{\alpha(1-\alpha)}{2}]$. It is straightforward to verify that these conditional densities are always non-increasing so that our score function s is quantile perfectable with respect to all distributions in P'. Moreover, these have conditional quantiles

$$q^\star_{v_i} = \begin{cases} (1-\alpha) & v_i = 1 \\ (1-\alpha) - \frac{\delta}{1+\delta/(1-\alpha)} & v_i = -1 \end{cases}.$$

The next step is to use the notation in Theorem 11, choosing $\Phi(x) = x, \rho(\theta, \theta') = \|\theta - \theta'\|_1$, and noting that for coordinate-wise testing function

$$\hat{v}_j(q) = \operatorname*{argmin}_{v_j \in \{-1,1\}} |q_j - q^\star_{v_j}|,$$

then

$$\begin{aligned}
2|q_j - q^\star_{v_j}| &\geq |q_j - q^\star_{\hat{v}_j}| + |q_j - q^\star_{v_j}| \\
&\geq |q^\star_{\hat{v}_j} - q^\star_{v_j}| \\
&\geq \mathbf{1}\{\hat{v}_j \neq v_j\} \left(\frac{\delta}{1 + \delta/(1-\alpha)} \right) \\
&\geq \frac{2\delta}{3} \mathbf{1}\{\hat{v}_j \neq v_j\},
\end{aligned}$$

where we have used the fact that $\delta \leq \frac{\alpha(1-\alpha)}{2}$. This result naturally implies that

$$\|q - q^\star\|_1 \geq 2 \cdot \left(\frac{\delta}{6}\right) \sum_{j=1}^{d} \mathbb{1}\{\hat{v}_j(q) \neq v_j\},$$

as required by Assouad's method. The only remaining step is to bound the TV distance, for which we note that if $d_{\mathrm{ham}}(v, v') \leq 1$ then

$$\begin{aligned}
d_{\mathrm{hel}}(P_v, P_{v'})^2 &\leq \frac{1}{d} \left[\left(\sqrt{1 - \frac{\delta}{\alpha}} - 1 \right)^2 \alpha + \left(\sqrt{1 + \frac{\delta}{1-\alpha}} - 1 \right)^2 (1-\alpha) \right] \\
&\leq \frac{1}{d} \left[\frac{\delta^2}{\alpha} + \frac{\delta^2}{1-\alpha} \right] = \frac{\delta^2}{d\alpha(1-\alpha)},
\end{aligned}$$

and since $d_{\mathrm{hel}}(P_v, P'_v) \leq 1/\sqrt{2}$ by our choice of δ we also have that

$$\begin{aligned}
d_{\mathrm{hel}}(P_v^n, P_{v'}^n) &= \sqrt{2 - 2(1 - d_{\mathrm{hel}}(P_v, P_{v'})^2)^n} \\
&\leq \sqrt{2 - 2e^{-2n d_{\mathrm{hel}}(P_v, P_{v'})^2}}.
\end{aligned}$$

We now choose $\delta^2 = \frac{d\alpha(1-\alpha)}{16n}$ to conclude that

$$\|P_v^n - P_{v'}^n\|_{\mathrm{TV}} \leq \frac{1}{2},$$

and by Theorem 11

$$\mathfrak{M}_n(\mathcal{C}_1) \geq \frac{1}{24} \sqrt{\frac{d\alpha(1-\alpha)}{16n}} \geq \frac{1}{96} \sqrt{\frac{d\alpha(1-\alpha)}{n}},$$

as long as $n \geq \frac{d}{4\alpha(1-\alpha)}$ to ensure that $\delta \leq \frac{\alpha(1-\alpha)}{2}$.

B.3. Proof of Theorem 1 with continuous conditional density

In order to enforce a continuous conditional density we will simply take our original perturbation function and modify the sharp transition to occur smoothly, namely

$$g(s) = \begin{cases}
-\frac{\delta}{1-\alpha} & t \leq (1-\alpha) - \frac{(1-\alpha)\delta}{1+\delta/\alpha} \\
\frac{1+\delta/\alpha}{(1-\alpha)^2}(t - (1-\alpha)) & t \in \left((1-\alpha) - \frac{(1-\alpha)\delta}{1+\delta/\alpha}, 1-\alpha \right] \\
\frac{1+\delta/\alpha}{\alpha^2}(t - (1-\alpha)) & t \in \left(1-\alpha, (1-\alpha) + \frac{\alpha\delta}{1+\delta/\alpha} \right] \\
\frac{\delta}{\alpha} & t > (1-\alpha) + \frac{\alpha\delta}{1+\delta/\alpha}
\end{cases}.$$

It is easy to check that this choice of perturbation function still ensures the densities are non-increasing so the quantile perfectability condition still holds. It also has the interesting feature of preserving the quantiles from our original construction, as well as the uniform lower and upper bounds for our conditional density. This implies that our original expression for the separation holds and

$$\|q - q^\star\|_1 \geq 2 \cdot \left(\frac{\delta}{6}\right) \sum_{j=1}^{d} \mathbb{1}\{\hat{v}_j(q) \neq v_j\}\,,$$

for the same function $\hat{v}$. It is also evident that this linear interpolation step can only reduce the pointwise separation between densities when computing the Hellinger distance, so our previous bound must also hold. Finally, combining these two results and applying Assouad's method yields the same bound of

$$\mathfrak{M}_n(\mathcal{C}_1) \geq \frac{1}{96}\sqrt{\frac{d\alpha(1-\alpha)}{n}}\,,$$

for $n \geq \frac{d}{4\alpha(1-\alpha)}$ even when enforcing a continuous conditional density.

B.4. Proof of Theorem 2

The key insight for this proof is to note that using the orthonormal basis our estimator is defined by

$$\begin{aligned}
\hat{\beta} &= \operatorname*{argmin}_{\beta \in \mathbb{R}^d} f_n(\beta) \\
&= \operatorname*{argmin}_{\beta \in \mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} F(\beta, X_i, S_i) \\
&= \operatorname*{argmin}_{\beta \in \mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} (1-\alpha)(S_i - \beta^T\tilde{w}(X_i))_+ + \alpha(\beta^T\tilde{w}(X_i) - S_i)_+\,,
\end{aligned}$$

via $\hat{g} = \hat{\beta}^T\tilde{w}$, which is the stochastic approximation of the true minimizer

$$\beta^\star = \operatorname*{argmin}_{\beta} f(\beta) = \operatorname*{argmin}_{\beta} \mathbb{E}\left[F(\beta, X, S)\right]\,.$$

We can now observe that f_n not always differentiable, but the fact that $S|X$ has a continuous density guarantees that f is differentiable everywhere with

$$\partial_{\beta_i} f(\beta) = \mathbb{E}\left[\tilde{w}_i(X)\left(\mathbb{1}\{S \leq \beta^T\tilde{w}(X)\} - (1-\alpha)\right)\right]\,,$$

so that for any function $w = \gamma^T\tilde{w}$

$$\mathbb{E}\left[\frac{w(X_{n+1})}{\|w(X_{n+1})\|_2}(\mathbf{1}\{S_{n+1} \leq \hat{g}(X_{n+1})\} - (1-\alpha))\middle|\,\hat{g}\right] = \frac{\gamma^T\partial_\beta f(\hat{\beta})}{\|\gamma\|_2}\,.$$

Moreover, since $\hat{\beta}$ is the empirical minimizer we know that for all j

$$\sum_{i:S_i \neq \hat{\beta}^T\tilde{w}(X_i)} \partial_{\beta_j} F(\hat{\beta}, X_i, S_i) + \sum_{i:S_i = \hat{\beta}^T\tilde{w}(X_i)} s_{i,j} = 0\,,$$

for $s_{i,j} \in [M(\alpha-1), M\alpha]$ as F is subdifferentiable but not differentiable when $S_i = \hat{\beta}^T \tilde{w}(X_i)$. Therefore, if we pick a subgradient at random from $[M(\alpha-1), M\alpha]$ when $S_i = \hat{\beta}^T \tilde{w}(X_i)$ we know that

$$\partial_{\beta_j} f_n(\hat{\beta}) \leq \frac{M}{n} \sum_{i=1}^{n} \mathbb{1}\{S_i = \hat{\beta}^T \tilde{w}(X_i)\}\,,$$

and by the proof of Theorem 2 in Gibbs et al. (2023) with probability 1

$$\partial_{\beta_j} f_n(\hat{\beta}) \leq \frac{dM}{n}\,,$$

and

$$\frac{\gamma^T \partial_\beta f(\hat{\beta})}{\|\gamma\|_2} \leq \frac{\gamma^T(\partial_\beta f(\hat{\beta}) - \partial_\beta f_n(\hat{\beta}))}{\|\gamma\|_2} + \frac{d^{3/2}M}{n}\,.$$

We can use this result to bound 3 as

$$D_{\mathcal{W},2}(\hat{g}) \leq \frac{d^{3/2}M}{n} + \max_{v \in \mathbb{B}_2^d} v^T(\partial_\beta f_n(\hat{\beta}) - \partial_\beta f(\hat{\beta})),$$

where $\mathbb{B}_2^d$ is the euclidean ball in d-dimensions. We now take a minimal cover $N_{1/2}$ of this ball in its corresponding norm so that $\forall v \in \mathbb{B}_2^d, \exists u \in N_{1/2} : \|v-u\|_2 \leq 1/2$, so that

$$\max_{v \in \mathbb{B}_2^d} v^T(\partial_\beta f_n(\hat{\beta}) - \partial_\beta f(\hat{\beta})) \leq \max_{u \in N_{1/2}} u^T(\partial_\beta f_n(\hat{\beta}) - \partial_\beta f(\hat{\beta})) + \max_{v \in \frac{1}{2}\mathbb{B}_2^d} v^T(\partial_\beta f_n(\hat{\beta}) - \partial_\beta f(\hat{\beta}))\,,$$

or equivalently

$$\max_{v \in \mathbb{B}_2^d} v^T(\partial_\beta f_n(\hat{\beta}) - \partial_\beta f(\hat{\beta})) \leq 2 \max_{u \in N_{1/2}} u^T(\partial_\beta f_n(\hat{\beta}) - \partial_\beta f(\hat{\beta}))\,,$$

so that our inequality becomes

$$D_{\mathcal{W},2}(\hat{g}) \leq \frac{d^{3/2}M}{n} + 2 \max_{u \in N_{1/2}} u^T(\partial_\beta f_n(\hat{\beta}) - \partial_\beta f(\hat{\beta}))\,.$$

It now only remains to bound this second quantity for which we note that

$$|u^T(\partial_\beta F(\hat{\beta}, x, s) - \partial_\beta f(\hat{\beta}))| \leq 2\sqrt{d}M \|u\|_2 \leq 2\sqrt{d}M\,,$$

and

$$\begin{aligned}
\mathbb{E}\left[|u^T(\partial_\beta F(\hat{\beta}, X, S) - \partial_\beta f(\hat{\beta}))|^2\right] &= \mathbb{E}\left[(u^T \partial_\beta F(\hat{\beta}, X, S))^2\right] - (u^T \partial_\beta f(\hat{\beta}))^2 \\
&\leq \mathbb{E}\left[(u^T \partial_\beta F(\hat{\beta}, X, S))^2\right] \\
&\leq \|u\|_2^2 \leq 1\,.
\end{aligned}$$

This implies that for any $u \in \mathbb{B}_2^d$ the zero mean random variables

$$T_{u,i} = u^T(\partial_\beta F(\hat{\beta}, X_i, S_i) - \partial_\beta f(\hat{\beta}))\,,$$

are $(1, 2\sqrt{d}M)$ sub-exponential and independent, and

$$T_u = \sum_{i=1}^n T_{u,i} = nu^T(\partial_\beta f_n(\hat{\beta}) - f(\hat{\beta}))\,,$$

is $(n, 2\sqrt{d}M)$ sub-exponential. We now have the bound

$$\mathbb{E}\left[\max_{u \in N_{1/2}} T_u\right] \leq \sqrt{2n\log(|N_{1/2}|)} + 2\sqrt{d}M\log(|N_{1/2}|)\,,$$

from Corollary 2.6 in Boucheron et al. (2013). To finish the proof of the upper bound we recall that $|N_{1/2}| \leq 5^d$ and thus

$$\begin{aligned}\mathbb{E}\left[D_{\mathcal{W},2}(\hat{g})\right] &\leq \frac{d^{3/2}M}{n} + \sqrt{\frac{8\log(5)d}{n}} + \frac{4\log(5)d^{3/2}M}{n} \\ &\leq 4\sqrt{\frac{d}{n}} + \frac{8d^{3/2}M}{n}\,.\end{aligned}$$

Observe that this result is always consistent with the lower bound in Theorem 3 since the VC-dimension of the subgraph satisfies $\mathrm{VC}(\mathcal{S}(\mathcal{W})) \leq d$ by Proposition 4.20 in Wainwright (2019) so a vector space of dimension d cannot contain a function class with VC dimension larger than d, and the result is tight in d and n when $\mathcal{W}$ does contain a function class with VC dimension d.

B.5. Proof of Theorem 3 without continuous density requirement

The first steps in this proof are essentially identical to Theorem 1 as we use the VC dimension of $\mathcal{W}$ to argue that we can find points $\{x_1, \cdots, x_d\}$ such that for any distribution $P \in \mathcal{P}' \subset \mathcal{P}$ where X is uniformly distributed on this set

$$D_{\mathcal{W},p}(a,b) \geq \frac{1}{d}\sum_{i=1}^d \left|\mathbb{P}\left(S_{n+1} \in [a(x_i), b(x_i)] | X = x_i, q\right) - (1-\alpha)\right|\,.$$

We now take on the notation of Theorem 12 and define the loss functions

$$\mathcal{L}_{P,i}(a,b) = \left|P(S_{n+1} \in [a(x_i), b(x_i)] | X = x_i, q) - (1-\alpha)\right|\,,$$

so that our lower bound becomes

$$D_{\mathcal{W},p}(a,b) \geq \frac{1}{d}\sum_{i=1}^d \mathcal{L}_{P,i}(a,b)\,,$$

and thus

$$\mathfrak{M}_n(\mathcal{C}_2) \geq \frac{1}{d}\inf_{\hat{a},\hat{b}}\sup_{P\in\mathcal{P}'}\mathbb{E}\left[\sum_{i=1}^d \mathcal{L}_{P,i}(\hat{a},\hat{b})\right]\,.$$

Our construction will now use $K = 5$ and define a family of tilts indexed by $v \in \{1, \cdots, K\}^d$ where X is uniformly distributed on the set $\{x_1, \cdots, x_d\}$ and S has conditional density

$$p_v(s|x_i) = 1 + \delta g_{v_i}(s)\,, \tag{8}$$

for $\delta \leq 1/2$ where the perturbation function $g_k(s)$ is defined as

$$g_k(s) = \begin{cases} D_{k,1} & s \in [0, 1/3) \\ D_{k,2} & s \in [1/3, 2/3) \\ D_{k,3} & s \in [2/3, 1] \end{cases}.$$

using the matrix

$$D = \begin{bmatrix} -1 & 0 & 1 \\ -1 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & -1 \\ 1 & 0 & -1 \end{bmatrix}.$$

It is easy to verify that using these perturbation functions s is interval perfectable with respect to all distributions in our family. We now use the following fact

Fact 14 *Let $D, g_k(s), \{P_v\}$ be as defined in the preceding proof sketch, then for all $i \in \{1, \cdots, d\}$ and v' in $\mathcal{V}_i' = \{(v_1, \cdots, v_i, \cdots, v_d) | v_i = 0$ and $v_j \in \{1, \cdots, K\}$ for $j \neq i\}$*

$$\inf_{\hat{a}, \hat{b}} \sum_{k=1}^{K} \mathcal{L}_{P_{v'+ke_i}, i}(\hat{a}, \hat{b}) \geq \delta \sqrt{\frac{\alpha(1-\alpha)}{7}}\,.$$

In order to apply Theorem 12 it only remains to upper bound the total variation distance for which we note that for $\delta \in [0, 1/2]$

$$\sup_{(b_1, b_2) \in [-1,1]^2} \left(\sqrt{1 + b_1\delta} - \sqrt{1 + b_2\delta}\right)^2 \leq 2\delta^2\,,$$

and thus

$$\begin{aligned} d_{\text{hel}}(P_{v'+e_i} || P_{v'+ke_i})^2 &\leq 2\left[1 - \left(1 - \frac{\delta^2}{d}\right)^n\right] \\ &\leq 2\left[1 - e^{-\frac{2n\delta^2}{d}}\right], \end{aligned}$$

which naturally implies

$$\|P_{v'+e_i} - P_{v'+ke_i}\|_{\text{TV}} \leq \sqrt{2 - 2e^{-\frac{2n\delta^2}{d}}}\,,$$

and if we choose $\delta^2 = \frac{d}{256n}$ then for all $i \in \{1, \cdots, d\}, k \in \{2, \cdots, K\}, v' \in \mathcal{V}_i'$

$$\|P_{v'+e_i} - P_{v'+ke_i}\|_{\text{TV}} \leq \frac{1}{2(K-1)}\,.$$

We now simply apply Corollary 13 to conclude that

$$\mathfrak{M}_n(\mathcal{C}_2) \geq \frac{1}{850}\sqrt{\frac{\alpha(1-\alpha)d}{n}}\,,$$

for $n \geq \frac{d}{64}$.

B.6. Proof of Fact 14

We start by observing that for any x_i the set of allowable estimators $\mathcal{R} = \{(a,b) \in [0,1]^2 | a \le b\}$ can be expressed as

$$\mathcal{R} = \bigcup_{i=1}^{3} \mathcal{R}_i \,,$$

for

$$\begin{aligned}
\mathcal{R}_1 &= \{(a,b) \in [0,2/3]^2 | a \le b\} \\
\mathcal{R}_2 &= \{(a,b) \in [1/3,1]^2 | a \le b\} \\
\mathcal{R}_3 &= \{(a,b) \in [0,1/3] \times [2/3,1]\} \,.
\end{aligned}$$

It is now sufficient to lower bound the optimal function value in each of these 3 regions to obtain a lower bound over $\mathcal{R}$. We now define $D_{\mathcal{I}}$ for any $\mathcal{I} \subseteq \{1,2,3\}$ as the matrix in $\mathbb{R}^{K \times |\mathcal{I}|}$ constructed from the corresponding columns of D and analyze the optimization problem in the 3 distinct regions.
Region 1: If $(\hat{a}(x_i), \hat{b}(x_i)) \in \mathcal{R}_1$ then

$$P_v(S \in [\hat{a}(x_i), \hat{b}(x_i)] | X = x_i) = (1 + \delta D_{v_i,1})\beta_1 + (1 + \delta D_{v_i,2})\beta_2 \,,$$

with

$$\begin{aligned}
\beta_1 &= \left[\left(\frac{1}{3} - \hat{a}(x_i) \right)_+ - \left(\frac{1}{3} - \hat{b}(x_i) \right)_+ \right] \\
\beta_2 &= \left[\left(\hat{b}(x_i) - \frac{1}{3} \right)_+ - \left(\hat{a}(x_i) - \frac{1}{3} \right)_+ \right] ,
\end{aligned}$$

so that for any i

$$\begin{aligned}
\inf_{\hat{a}(x_i), \hat{b}(x_i) \in \mathcal{R}_1} \sum_{k=1}^{K} \mathcal{L}_{P_{v'+ke_i},i}(\hat{a}, \hat{b}) &\ge \inf_{\beta \in \mathbb{R}^2} \left\| (\mathbf{1}_{K \times 2} + \delta D_{\{1,2\}})\beta - (1-\alpha)\mathbf{1}_{K \times 1} \right\|_1 \\
&\ge \inf_{\beta \in \mathbb{R}^2} \left\| (\mathbf{1}_{K \times 2} + \delta D_{\{1,2\}})\beta - (1-\alpha)\mathbf{1}_{K \times 1} \right\|_2 .
\end{aligned}$$

Region 2: If $(\hat{a}(x_i), \hat{b}(x_i)) \in \mathcal{R}_2$ then identical reasoning shows that the optimal function value is also lower bounded by the same convex optimization problem.
Region 3: If $(\hat{a}(x_i), \hat{b}(x_i)) \in \mathcal{R}_3$ then

$$P_v(S \in [\hat{a}(x_i), \hat{b}(x_i)] | X = x_i) = (1 + \delta D_{v_i,1})\beta_1 + \frac{(1 + \delta D_{v_i,2})}{3} + (1 + \delta D_{v_i,3})\beta_2 \,,$$

with

$$\begin{aligned}
\beta_1 &= \frac{1}{3} - \hat{a}(x_i) \\
\beta_2 &= \left(\hat{b}(x_i) - \frac{2}{3} \right) ,
\end{aligned}$$

so that for any i

$$\inf_{\hat{a}(x_i),\hat{b}(x_i)\in\mathcal{R}_3}\sum_{k=1}^{K}\mathcal{L}_{v'+ke_i}(\hat{a},\hat{b}) \geq \inf_{\beta\in\mathbb{R}^2}\left\|(\mathbf{1}_{K\times 2}+\delta D_{\{1,3\}})\beta-\left((1-\alpha)\mathbf{1}_{K\times 1}-\frac{1}{3}D_{\{2\}}\right)\right\|_1$$
$$\geq \inf_{\beta\in\mathbb{R}^2}\left\|(\mathbf{1}_{K\times 2}+\delta D_{\{1,3\}})\beta-\left((1-\alpha)\mathbf{1}_{K\times 1}-\frac{1}{3}D_{\{2\}}\right)\right\|_2 .$$

Combining the 3 results: In this case we can conclude that for all $i \in \{1,\cdots,d\}$ and v' in $\mathcal{V}_i' = \{(v_1,\cdots,v_i,\cdots,v_d)|v_i = 0 \text{ and } v_j \in \{1,\cdots,K\} \text{ for } j \neq i\}$ the optimal objective value is lower bounded by the minimum of two least squares problems

$$\inf_{\hat{a},\hat{b}}\frac{1}{K}\sum_{k=1}^{K}\mathcal{L}_{P_{v'+ke_i},i}(\theta) \geq \frac{1}{K}\inf_{j\in\{1,2\}}\inf_{\beta\in\mathbb{R}^2}\|A_j\beta-y_j\|_2 .$$

Using duality to further lower bound the inner infimum we now obtain

$$\inf_{\hat{a},\hat{b}}\sum_{k=1}^{K}\mathcal{L}_{P_{v'+ke_i},x_i}(\hat{a},\hat{b}) \geq \inf_{j\in\{1,2\}}\sup_{\nu:A_j^T\nu=0}\frac{|\nu^T y_j|}{\|\nu\|_2}$$

and since $\delta \leq 1/3$

$$\nu_1^T = \left[-\tfrac{2}{1-\delta},\tfrac{1}{1-\delta},1,0,0\right] \in \text{Null}(A_1^T)$$
$$\nu_2^T = \left[\tfrac{1}{1-\delta},-\tfrac{2}{1-\delta},1,0,0\right] \in \text{Null}(A_2^T),$$

for all $i \in \{1,\cdots,d\}, v' \in \mathcal{V}_i'$

$$\inf_{\hat{a},\hat{b}}\sum_{k=1}^{K}\mathcal{L}_{P_{v'+ke_i},x_i}(\hat{a},\hat{b}) \geq \delta\sqrt{\frac{\alpha(1-\alpha)}{7}} .$$

B.7. Proof of Theorem 3 with continuous conditional density

In order to extend our result to allow for continuous conditional densities we define the smoothed perturbation function

$$\tilde{h}_k(s) = \begin{cases} D_{k,1} & s\in[0,1/3-\epsilon/2) \\ D_{k,1}\left(1-\frac{s-(1/3-\epsilon/2)}{\epsilon}\right)+D_{k,2}\left(\frac{s-(1/3-\epsilon/2)}{\epsilon}\right) & s\in[1/3-\epsilon/2,1/3+\epsilon/2) \\ D_{k,2} & s\in[1/3+\epsilon/2,2/3-\epsilon/2) \\ D_{k,2}\left(1-\frac{s-(2/3-\epsilon/2)}{\epsilon}\right)+D_{k,3}\left(\frac{s-(2/3-\epsilon/2)}{\epsilon}\right) & s\in[2/3-\epsilon/2,2/3+\epsilon/2) \\ D_{k,3} & s\in[2/3+\epsilon/2,1] \end{cases},$$

where $\epsilon = \frac{32\delta^2\alpha(1-\alpha)}{7K^2}$. It is now important to note that our bound on the total variation distance is still valid, as once again the smoothing can only bring the conditional densities closer together, so

it only remains to analyze the weak separation condition. If we denote $\tilde{P}_v$ as the smoothed version of P_v we can see that

$$\begin{aligned}\sum_{k=1}^{K} \mathcal{L}_{P_{v'+ke_i},x_i}(\hat{a},\hat{b}) &= \sum_{k=1}^{K} \left| P_{v'+ke_i}(S \in [\hat{a}(x_i),\hat{b}(x_i)]|X=x_i) - (1-\alpha) \right| \\ &\leq \sum_{k=1}^{K} \left| \tilde{P}_{v'+ke_i}(S \in [\hat{a}(x_i),\hat{b}(x_i)]|X=x_i) - (1-\alpha) \right| \\ &\quad + \left\| P_{v'+ke_i} - \tilde{P}_{v'+ke_i} \right\|_{\text{TV}} ,\end{aligned}$$

or equivalently

$$\sum_{k=1}^{K} \mathcal{L}_{\tilde{P}_{v'+ke_i},x_i}(\hat{a},\hat{b}) \geq \sum_{k=1}^{K} \mathcal{L}_{P_{v'+ke_i},x_i}(\hat{a},\hat{b}) - \sum_{k=1}^{K} \left\| P_{v'+ke_i} - \tilde{P}_{v'+ke_i} \right\|_{\text{TV}} .$$

We finally bound the total variation distance between the smoothed and non-smoothed variants using the same technique as before

$$\begin{aligned} d_{\text{hel}}(\tilde{P}_{v'+ke_i}||P_{v'+ke_i})^2 &\leq 2\left[1 - \left(1 - \frac{2\delta^2\epsilon}{d}\right)^n\right] \\ &\leq 2\left[1 - e^{-\frac{4n\delta^2\epsilon}{d}}\right] \\ &\leq 2\left[1 - e^{-\frac{\epsilon}{512}}\right] \end{aligned}$$

so that

$$\left\| P_{v'+ke_i} - \tilde{P}_{v'+ke_i} \right\|_{\text{TV}} \leq \frac{1}{16}\sqrt{2\epsilon} = \frac{\delta}{2K}\sqrt{\frac{\alpha(1-\alpha)}{7}} ,$$

and thus

$$\sum_{k=1}^{K} \mathcal{L}_{\tilde{P}_{v'+ke_i},x_i}(\hat{a},\hat{b}) \geq \frac{\delta}{2}\sqrt{\frac{\alpha(1-\alpha)}{7}} ,$$

leading to the minimax lower bound

$$\mathfrak{M}_{C_{a,b}} \geq \frac{1}{1700}\sqrt{\frac{\alpha(1-\alpha)d}{7n}} ,$$

for $n \geq \frac{d}{64}$.

Appendix C. Optimality guarantees in conformal prediction

In Section 2.2 we discussed that choosing the *best* among all sets attaining $1-\alpha$ coverage is a key issue when constructing confidence sets. In fact, there are two somewhat equivalent ways of defining optimal sets: score functions and optimality measures. We prove this equivalence through the following two lemmas.

Lemma 15 *Let $\{C_{1-\alpha}\}$ be φ-perfectable nested sets with conditional coverage $1-\alpha$. If the nested sets $C_{1-\alpha}$ are inner-semicontinuous for every α, namely*

$$\lim_{\alpha' \downarrow \alpha} C_{1-\alpha'} = C_{1-\alpha}\,,$$

then there exists a score function $s : \mathcal{X} \times \mathcal{Y} \to [0,1]$ such that

$$C_{1-\alpha}(x) = \{y | s(x,y) \leq 1-\alpha\}\,.$$

Proof Define

$$s(x,y) = \inf\{\tau | y \in C_\tau(x)\}\,,$$

where it is clear that $s(x,y) \in [0,1]$ since any $y \in \mathcal{Y}$ satisfies $y \in C_1(x)$ and $y \notin C_0(x)$. We now take $y \in \{y \mid s(x,y) \leq 1-\alpha\}$ and note that this implies $y \in C_{1-\alpha'}(x)$ for all $\alpha' > \alpha$ and thus it is clear that $y \in C_{1-\alpha}(x)$ by the nesting and inner-semicontinuity assumption. Conversely, if $y \in C_{1-\alpha}(x)$ then $\inf\{\tau | y \in C_{1-\alpha}(x)\} \leq 1-\alpha$ and thus $y \in \{y \mid s(x,y) \leq 1-\alpha\}$. Therefore, we have found a score so that $C_{1-\alpha}(x) = \{y \mid s(x,y) \leq 1-\alpha\}$. ■

Lemma 16 *For any score function $s : \mathcal{X} \times \mathcal{Y} \to [0,1]$ there exists $\varphi(x,y)$ so that s is φ-perfectable with respect to any distribution $P_{X,Y}$ with no point masses.*

Proof For $\varphi(x,y) = s(x,y)$ we get

$$L_\varphi(C) = \int_{\mathcal{X}} \int_{C(x)} s(x,y) dP(y|x) dP(x)\,.$$

We now define

$$\begin{aligned} q^-_{1-\alpha}(x) &= \inf\{q \in \mathbb{R} \text{ s.t. } \mathbb{P}\left(s(x,Y) \leq q \mid X = x\right) = 1-\alpha\} \\ q^+_{1-\alpha}(x) &= \sup\{q \in \mathbb{R} \text{ s.t. } \mathbb{P}\left(s(x,Y) \leq q \mid X = x\right) = \mathbb{P}\left(s(x,Y) \leq q^-_{1-\alpha}(x) \mid X = x\right)\}\,, \end{aligned}$$

and for any $q(x) \in [q^-_{1-\alpha}(x), q^+_{1-\alpha}(x)]$ construct $C^\star_{1-\alpha}(x) = \{y \mid s(x,y) \leq q(x)\}$. For any other $C(x)$ with $1-\alpha$ conditional coverage define the set E to be all values of x such that $\epsilon(x) := \mathbb{P}\left(Y \in C^\star_{1-\alpha}(x) \setminus C(x) \mid X = x\right) > 0$ so if E has positive measure

$$\int_E \epsilon(x) dP(x) > 0\,.$$

However, as $\mathbb{P}\left(Y \in C^\star_{1-\alpha}(x) \mid X = x\right) = 1-\alpha$ for all $x \in E$ we have

$$\mathbb{P}\left(Y \in C(x) \setminus C^\star_{1-\alpha}(x) \mid X = x\right) \geq \epsilon(x) > 0\,.$$

In this case on $y \in C(x) \setminus C^\star_{1-\alpha}(x)$ we know that $s(x,y) > q(x)$ so

$$\Delta(x) = \int_{C^\star_{1-\alpha}(x)} s(x,y) dP(y|x) - \int_{C(x)} s(x,y) dP(y|x)\,,$$

satisfies

$$\begin{aligned}\Delta(x) &\leq \left[\mathbb{P}\left(C^\star_{1-\alpha}(x) \setminus C(x) \mid X = x\right) - \mathbb{P}\left(C(x) \setminus C^\star_{1-\alpha}(x) \mid X = x\right)\right] q(x) \\ &\leq 0\,.\end{aligned}$$

Thus we have shown that $C^\star_{1-\alpha}(x)$ minimizes the chosen loss for any distribution with no point masses. ■

These lemmas illustrate why we cannot fix both an arbitrary optimality criterion and an arbitrary score when studying conformal prediction without distributional assumptions, as both notions are intimately linked. In practice, practitioners often seek to obtain the shortest possible intervals as measured by $\text{Leb}(C(x))$ so the usual approach is to assume some properties of the observed data distribution and choose scores with quantile sets that provide minimum length intervals under these assumptions. However, it is evident that this approach can only yield minimum length intervals in all cases if the score is chosen in a distribution dependent manner as shown in the following fact.

Fact 17 *For any distribution on $(Y, X) \in (\mathbb{R}^k, \mathbb{R}^m)$ with conditional density $f(y|x)$ the quantile sets $C_q(x) = \{y | s(x, y) \leq q(x)\}$ defined by the distribution dependent score*

$$s(x, y) = \exp\left[-f(y|x)\right]\,,$$

minimize Leb$(C(x))$ over sets with $1 - \alpha$ coverage.

Proof The proof follows immediately from the fact that since a density exists the sets with minimum Lebesgue measure are of the form $C_a(x) = \{y | f(y|x) > a(x)\}$ by definition, and these match the sublevel sets of $s(x, y) = \exp\left[-f(y|x)\right]$. ■

Appendix D. Deferred proofs for weak calibration

D.1. Proof of Theorem 4

The general plan for this proof is to construct a base tilt P_0 and a family of alternate tilts $\{P_{\delta v}\}$ indexed by $v \in \mathcal{V} = \{-1, 1\}^d$ and $\delta \geq 0$, so that these two sets are δ-separated. In this case we can apply the convex hull Le Cam method (Lemma 15.9 in Wainwright (2019)) to conclude that

$$\inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P\left[\left|\text{CE}(f, \mathcal{W}) - \hat{\theta}\right|\right] \geq \frac{\delta}{2}\left[1 - \left\|P_0^n - \bar{P}_\delta^n\right\|_{\text{TV}}\right]\,,$$

where $\bar{P}_\delta^n = \frac{1}{|\mathcal{V}|}\sum_{v \in \mathcal{V}} P_{\delta v}^n$. We start by constructing our perfectly calibrated tilt P_0 where S is uniformly distributed on $\{s_1, \cdots, s_d\}$ and $Y|S = s_i \sim \text{Bernoulli}(s_i)$. To construct the particular members of the alternative family $P_{\delta v}$, for each $j \in \{1, \cdots, d\}$ we define the "tilting" function

$$\phi_j(y, s) := \left(\frac{y}{s_j} - \frac{1 - y}{1 - s_j}\right)\mathbb{1}\{y = 1, s = s_j\}\,.$$

Then $\mathbb{E}_{P_0}\left[\phi_j(Y, S)\right] = 0$ while

$$\text{Var}_{P_0}\left(\phi_j(Y, S)\right) = \frac{1}{s_j d} + \frac{1}{(1 - s_j)d} = \frac{1}{s_j(1 - s_j)d}\,.$$

Note that $|\phi_j(s,y)| \leq \frac{1}{\epsilon}$ as $\epsilon < \frac{1}{2}$ and if we define the vector $\phi(y,s) = (\phi_1(y,s), \cdots, \phi_d(y,s))$, then $\|\phi(y,s)\|_0 \leq 1$, as the number of non-zero entries is at most 1. Now for $\delta \in [0,\epsilon]$ for each $v \in \{-1,1\}^d$ we may define the tilted distribution $P_{\delta v}$ with

$$P_{\delta v}(Y = y, S = s) = (1 + \delta\langle v, \phi(y,s)\rangle)P_0(Y = y, S = s)\,,$$

which is valid whenever $\delta \leq c$ as $|\langle v, \phi(y,s)\rangle| \leq \frac{1}{\epsilon}$. We now compute the calibration error for distributions $P_{\delta v}$ by noting that S is still uniform on $\{s_1, \cdots, s_d\}$ so that

$$\mathbb{E}_{P_{\delta v}}\left[Y \middle| S = s_j\right] = s_j + \delta v_j \mathbb{E}_{P_0}\left[\phi_j(Y, s_j)Y \middle| S = s_j\right] = s_j + \delta v_j\,,$$

and so by our shattering assumption

$$\text{CE}(f, \mathcal{W}) = \sup_{w \in \mathcal{W}} \mathbb{E}_{P_{\delta v}}\left[w(S)(Y - S)\right] \geq \frac{1}{d}\sum_{j=1}^{d} |s_j + v_j\delta - s_j| = \delta\,,$$

so our family $\{P_{\delta v}\}$ is δ-separated from P_0.

Lastly, we compute a bound on testing error. We start by recalling that in the case of our mixture distribution the χ^2-divergence satisfies

$$\begin{aligned} D_{\chi^2}\left(\bar{P}_\delta^n \| P_0^n\right) + 1 &= \frac{1}{2^{2d}}\sum_{v,v'} \mathbb{E}_{P_0}\left[(1 + \delta\langle v, \phi(y,s)\rangle)(1 + \delta\langle v', \phi(y,s)\rangle)\right]^n \\ &= \frac{1}{2^{2d}}\sum_{v,v'} (1 + \delta^2 v^T \text{Cov}_{P_0}(\phi(Y,S))v')^,, \end{aligned}$$

because our sampling is i.i.d. By our variance calculation for ϕ and that each ϕ_j has disjoint support, we have that $\text{Cov}_{P_0}(\phi(Y,S)) = \frac{1}{d}\text{diag}\left(\left[\frac{1}{s_j(1-s_j)}\right]_{j=1}^n\right)$ and so

$$D_{\chi^2}\left(\bar{P}_\delta^n \| P_0^n\right) + 1 = \mathbb{E}\left[\left(1 + \frac{\delta^2}{d}\sum_{j=1}^{d}\frac{V_j V_j'}{s_j(1-s_j)}\right)^n\right] \leq \mathbb{E}\left[\exp\left[\frac{n\delta^2}{d}\sum_{j=1}^{d}\frac{V_j V_j'}{s_j(1-s_j)}\right]\right]\,,$$

where the expectation is over $V, V' \overset{\text{i.i.d.}}{\sim} \text{Uniform}(\{\pm 1\}^d)$. Clearly, since $V_j V_j'$ are i.i.d. random signs they are 1-sub-Gaussian and

$$D_{\chi^2}\left(\bar{P}_\delta^n \| P_0^n\right) + 1 \leq \exp\left[\frac{n^2\delta^4}{2d^2}\sum_{j=1}^{d}\frac{1}{s_j^2(1-s_j)^2}\right] \leq \exp\left[\frac{n^2\delta^4}{2d\epsilon^2(1-\epsilon)^2}\right]\,,$$

because $s_j \in [\epsilon, 1-\epsilon]$. It is now easy to see that for $\delta^4 = \frac{d\epsilon^2(1-\epsilon)^2}{2n^2}$ then $D_{\chi^2}\left(\bar{P}_\delta^n \| P_0^n\right) \leq \frac{1}{2}$ and thus $\left\|P_0^n - \bar{P}_\delta^n\right\|_{\text{TV}} \leq \frac{1}{2}$ so that

$$\inf_{\hat{\theta}} \sup_{P \subseteq \mathcal{P}} \mathbb{E}_P\left[\left|\text{CE}(f, \mathcal{W}) - \hat{\theta}\right|\right] \geq \frac{d^{1/4}}{5n^{1/2}}\sqrt{\epsilon(1-\epsilon)}\,,$$

for $n \geq \frac{1}{\epsilon(1-\epsilon)}\sqrt{\frac{d}{2}}$.

D.2. Proof of Corollary 5

We start by observing that given the conditions in this corollary and under the same construction of our proof for Theorem 4

$$\mathrm{CE}(f, \mathcal{W}_{B^2})^2 = \mathrm{ece}_{\|\cdot\|_2}(f)^2 = \frac{1}{d}\sum_{j=1}^{d} \delta^2 v_j^2 = \delta^2 \,,$$

so by just using the same logic we get a lower bound of

$$\inf_{\hat{\theta}} \sup_{P \subseteq \mathcal{P}} \mathbb{E}_P \left[\left| \mathrm{CE}(f, \mathcal{W})^2 - \hat{\theta} \right| \right] \geq \frac{d^{1/2}}{6n} \epsilon(1-\epsilon) \,,$$

for $n \geq \frac{1}{\epsilon(1-\epsilon)}\sqrt{\frac{d}{2}}$. For the remaining lower bound we use a simple two-point construction with tilts P_0, P_1 such that both are uniformly distributed on $\{s_1, \cdots, s_d\}$ and for $\gamma \in (0, 1/2 - \epsilon), \delta \in (0, \epsilon)$ then

$$P_0(Y = 1 | S = s_i) = \begin{cases} s_i + \gamma & s_i \leq 1/2 \\ s_i - \gamma & s_i > 1/2 \,, \end{cases}$$

and

$$P_1(Y = 1 | S = s_i) = \begin{cases} s_i + \gamma(1+\delta) & s_i \leq 1/2 \\ s_i - \gamma(1+\delta) & s_i > 1/2 \,. \end{cases}$$

In this case it is easy to see that

$$\mathrm{CE}_{P_0}(f, \mathcal{W}_{B^2})^2 = \gamma^2 \qquad\qquad \mathrm{CE}_{P_1}(f, \mathcal{W}_{B^2})^2 = (1+\delta)^2\gamma^2 \,,$$

and thus P_0, P_1 are $2\delta\gamma^2$ separated. It now only remains to bound the probability of testing error, so we note that if $s_i \leq 1/2$

$$\begin{aligned} -\log\left(1 + \frac{\delta\gamma}{s_i + \gamma}\right)(s_i + \gamma) &\leq -\left(\frac{\delta\gamma}{s_i+\gamma} - \frac{\delta^2\gamma^2}{(s_i+\gamma)^2}\right)(s_i + \gamma) \\ &\leq -\delta\gamma + \frac{\delta^2\gamma^2}{s_i + \gamma} \,, \end{aligned}$$

and since $-\frac{\delta\gamma}{1-(s_i+\gamma)} \geq -\frac{1}{2}$ we also know that

$$\begin{aligned} -\log\left(1 - \frac{\delta\gamma}{1-(s_i+\gamma)}\right)(1 - (s_i+\gamma)) &\leq -\left(-\frac{\delta\gamma}{1-(s_i+\gamma)} - \frac{\delta^2\gamma^2}{1-(s_i+\gamma)}\right)(1-(s_i+\gamma)) \\ &\leq \delta\gamma + \frac{\delta^2\gamma^2}{1-(s_i+\gamma)} \,. \end{aligned}$$

An identical argument shows that these bounds also hold when $s_i > 1/2$ so we can conclude that

$$D_{\mathrm{kl}}\left(P_0 \| P_1\right) \leq \frac{1}{d}\sum_{i=1}^{d} \frac{\delta^2\gamma^2}{(1-(s_i+\gamma))(s_i+\gamma)} \leq \frac{\delta^2\gamma^2}{\epsilon(1-\epsilon)}\,, \tag{9}$$

and

$$\|P_0^n - P_1^n\|_{\mathrm{TV}} \leq \sqrt{\frac{n\delta^2\gamma^2}{2\epsilon(1-\epsilon)}}\,.$$

We can now pick $\delta^2 = \frac{\epsilon(1-\epsilon)}{2\gamma^2 n}$ to guarantee that $\|P_0^n - P_1^n\|_{\mathrm{TV}} \leq \frac{1}{2}$ and apply the standard Le Cam two-point method to obtain

$$\inf_{\hat{\theta}} \sup_{P \subseteq \mathcal{P}} \mathbb{E}_P\left[\left|\mathrm{CE}(f,\mathcal{W})^2 - \hat{\theta}\right|\right] \geq \frac{1}{4}\sqrt{\frac{\epsilon(1-\epsilon)\gamma^2}{2n}} \geq \frac{1}{6}\sqrt{\frac{\epsilon(1-\epsilon)\mathrm{CE}(f,\mathcal{W}_{B^2})^2}{n}}\,,$$

for $n \geq \frac{1}{2\mathrm{CE}(f,\mathcal{W}_{B^2})^2\epsilon}$ recalling that $\gamma^2 = \mathrm{CE}(f,\mathcal{W})^2$ for one of our tilts.

D.3. Proof of Theorem 7

Using identical reasoning to the original proof in appendix F of Kumar et al. (2019) we know that the debiased estimator satisfies the decomposition

$$\mathcal{E}_{\mathrm{db}}^2 = \underbrace{\sum_{i=1}^{d} \hat{p}_i e_i^2}_{\text{(C1)}} - \underbrace{2\sum_{i=1}^{d} \hat{p}_i e_i(\hat{y}_i - y_i^*)}_{\text{(C2)}} + \underbrace{\sum_{i=1}^{d} \hat{p}_i\left[(\hat{y}_i - y_i^*)^2 - \frac{\hat{y}_i(1-\hat{y}_i)}{\hat{p}_i n - 1}\right]}_{\text{(C3)}}\,,$$

where $y_i^* = \mathbb{E}\left[Y \middle| S = s_i\right]$, $p_i = \mathbb{P}\left(S_k = s_i\right)$ and $e_i = s_i - y_i^*$. Therefore,

$$|\mathcal{E}_{\mathrm{db}}^2 - \mathrm{CE}(f,\mathcal{W}_{B^2})^2| \leq |(\mathrm{C1}) - \mathrm{CE}(f,\mathcal{W}_{B^2})^2| + |(\mathrm{C2})| + |(\mathrm{C3})|$$

Bounding the first term: Note that

$$\sum_{i=1}^{d} \hat{p}_i e_i^2 = \frac{1}{n}\sum_{i=1}^{d}\sum_{k=1}^{n} \mathbf{1}\{S_k = s_i\} e_i^2 = \frac{1}{n}\sum_{k=1}^{n} E_i^2\,,$$

where $\{E_k^2\}_{k=1}^n$ are the i.i.d. random variables with distribution

$$\mathbb{P}\left(E_k^2 = e_i^2\right) = \mathbb{P}\left(S_k = s_i\right)\,,$$

as they simply return the true conditional calibration error e_i if $S_k = s_i$. It is clear that

$$\mathbb{E}\left[E_k^2\right] = \sum_{i=1}^{d} p_i e_i^2 = \mathrm{CE}(f,\mathcal{W}_{B^2})^2$$

and $|E_k^2| \le 1$ so that $\mathrm{Var}(E_k^2) \le \mathbb{E}\left[E_k^4\right] \le \mathbb{E}\left[E_k^2\right] = \mathrm{CE}(f, \mathcal{W}_{B^2})^2$. We can now use Bernstein's inequality to argue that

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{k=1}^{n} E_i^2 - \mathrm{CE}(f, \mathcal{W}_{B^2})^2\right| \ge t\right) \le 2\exp\left[-n\min\left\{\frac{t^2}{3\,\mathrm{CE}(f, \mathcal{W}_{B^2})^2}, \frac{t}{4}\right\}\right],$$

and thus with probability at least $1-\delta$

$$|(\mathrm{C1}) - \mathrm{CE}(f, \mathcal{W}_{B^2})^2| \le \sqrt{\frac{3\mathrm{CE}(f, \mathcal{W}_{B^2})^2}{n}\log\left(\frac{2}{\delta}\right)} + \frac{4}{n}\log\left(\frac{2}{\delta}\right).$$

Bounding (C2): We start by observing that

$$\mathbb{P}\left(\hat{p}_i = 0\right) = (1-p_i)^n = \exp\left[-n\log\left(\frac{1}{1-p_i}\right)\right] \le \exp\left[-n\log\left(\frac{1}{1-\min_i p_i}\right)\right],$$

so using a union bound since $n \ge c\log(\frac{d}{\delta})$ for $\frac{1}{c} \le \log\left(\frac{1}{\min_i p_i}\right)$

$$\mathbb{P}\left(\exists i : \hat{p}_i = 0\right) \le \delta\,.$$

Therefore, conditioning on the event that we see at least one sample from each score (A_1) and the random variable $Z = (S_1, \cdots, S_n)$ we know that

$$\mathbb{E}\left[\hat{y}_i - y_i^* \,|\, Z, A_1\right] = 0\,.$$

We now observe that given Z the samples $(Y_1, \cdots, Y_n)$ are still independent but not identically distributed as their distribution depends on the corresponding S_k. Moreover, since $\hat{y}_i - y_i^* | Z, A_1$ has absolute value bounded by 1, we know that it is sub-Gaussian with parameter $\frac{1}{4\hat{p}_i n}$. Note that $\hat{p}_i$ is a constant given Z. It is now easy to see that $\hat{p}_i e_i(\hat{y}_i - y_i^*)|Z, A_1$ is also mean zero and sub-Gaussian with parameter $\frac{\hat{p}_i e_i^2}{4n}$, and the full term (C2) conditioned on Z, A_1 also has expected value 0 with sub-Gaussian parameter

$$\sigma^2 = \sum_{i=1}^{d}\frac{\hat{p}_i e_i^2}{4n} = \frac{(\mathrm{C1})}{4n} \le \frac{\mathrm{CE}(f, \mathcal{W}_{B^2})^2 + |(\mathrm{C1}) - \mathrm{CE}(f, \mathcal{W}_{B^2})^2|}{4n}\,.$$

Therefore, if our bound on the previous term holds we can use the sub-Gaussian tail inequality to conclude that with probability at least $1-\delta$

$$|(C2)| \le \sqrt{\frac{2\mathrm{CE}(f, \mathcal{W}_{B^2})^2 + O(n^{-1/2})}{n}\log\left(\frac{2}{\delta}\right)}\,.$$

This holds for any Z conditioned on our bound for the first term and the event A_1 so it must hold after marginalizing over Z.

Bounding (C3): This bound follows from Lemma F.8. in Kumar et al. (2019) with the only modification that the condition $\hat{p}_i > 0$ is achieved via event A_1 in the previous part. Therefore, we can conclude that with probability at least $1-2\delta$

$$|(C3)| \le \frac{3\sqrt{d}}{n}\log\left(\frac{n}{\delta}\right) + \frac{\delta}{n}\,.$$

Combining all our bounds: We now see that for $n \geq c\log(\frac{d}{\delta})$ with $\frac{1}{c} \leq \log\left(\frac{1}{\min_i p_i}\right)$ with probability at least $1-4\delta$

$$|\mathcal{E}^2_{\text{db}} - \text{CE}(f, \mathcal{W}_{B^2})^2| \leq C_1\sqrt{\frac{\text{CE}(f, \mathcal{W}_{B^2})^2}{n}\log\left(\frac{2}{\delta}\right)} + C_2\frac{\sqrt{d}}{n}\log\left(\frac{n}{\delta}\right),$$

for some numerical constants C_1, C_2 and sufficiently large n.

D.4. Proof of Theorem 8

The proof of this theorem follows the logic laid out in our previous proof of Theorem 4. We first construct the perfectly calibrated tilt P_0 where S follows a predefined distribution Q and

$$P_0(Y = 1|S = s) = s\,.$$

For $B_\epsilon = \{B_1, \cdots, B_d\}$ we now define the functions

$$\phi_j(s, y) = \frac{1}{s}\mathbb{1}\{y = 1, s \in B_j\} - \frac{1}{1-s}\mathbb{1}\{y = 0, s \in B_j\}\,.$$

These have mean 0 since $\mathbb{E}_{P_0}\left[\phi_j(s, Y)|\, S = s\right] = 0$ and variance

$$\text{Var}_{P_0}\left(\phi_j(S, Y)\right) = \int_{B_j}\frac{1}{s(1-s)}dQ(s) \leq \frac{q_1}{\epsilon(1-\epsilon)}\,.$$

The supports of our functions ϕ_j are once again disjoint so we can define our family of alternative tilts indexed by $v \in \mathcal{V} = \{-1, 1\}^d$ as

$$dP_{\delta v}(s, y) = (1 + \delta\langle v, \phi(s, y)\rangle)dP_0(s, y)\,,$$

for $0 < \delta \leq \min\{\epsilon, 1-\epsilon\}$. Our assumptions now guarantee that the new construction satisfies

$$\text{CE}(f, \mathcal{W}) \geq dq_0\delta\,,$$

so our family $\{P_{\delta v}\}$ is $\frac{dq_0\delta}{2}$-separated from P_0. It is now easy to see that following the same logic for our previous χ^2-divergence bound we get that for $\delta^4 = \frac{\epsilon^2(1-\epsilon)^2}{2dq_1^2n^2}$

$$\left\|P_0^n - \bar{P}_\delta^n\right\|_{\text{TV}} \leq \frac{1}{2}\,.$$

Le Cam's convex hull method in this case now yields

$$\begin{aligned}\inf_{\hat{\theta}}\sup_{P\subseteq\mathcal{P}}\mathbb{E}_P\left[\left|\text{CE}(f, \mathcal{W}) - \hat{\theta}\right|\right] &\geq \frac{q_0}{\sqrt{q_1}}\cdot\frac{d^{3/4}}{5n^{1/2}}\sqrt{\epsilon(1-\epsilon)} \\ &\geq \frac{q_0 d}{\sqrt{q_1 d}}\cdot\frac{d^{1/4}}{5n^{1/2}}\sqrt{\epsilon(1-\epsilon)} \\ &\geq \frac{q_0}{q_1}\cdot\frac{d^{1/4}}{5n^{1/2}}\sqrt{\epsilon(1-\epsilon)}\,,\end{aligned}$$

for $n \geq \frac{1}{\sqrt{2}\epsilon(1-\epsilon)q_1 d^{1/2}}$.