A loss framework for calibrated anomaly detection

Aditya Krishna Menon
Australian National University*
aditya.menon@anu.edu.au

Robert C. Williamson
Australian National University
bob.williamson@anu.edu.au

Abstract

Given samples from a distribution, anomaly detection is the problem of determining if a given point lies in a low-density region. This paper concerns *calibrated* anomaly detection, which is the practically relevant extension where we additionally wish to produce a *confidence* score for a point being anomalous. Building on a classification framework for standard anomaly detection, we show how minimisation of a suitable *proper loss* produces density estimates only for anomalous instances. These are shown to naturally relate to the pinball loss, which provides implicit quantile control. Finally, leveraging a result from point processes, we show how to efficiently optimise a special case of the objective with kernelised scores. Our framework is shown to incorporate a close relative of the one-class SVM as a special case.

1 Calibrated anomaly detection

Given a set of instances with some systematic pattern (e.g., images of household objects), anomaly detection is informally understood as the problem of identifying if a particular instance deviates from this pattern (e.g., an image of a tree). There are several formalisations of this notion, each of which have led to many different algorithms (see, e.g., Chandola et al. [2009], Pimentel et al. [2014] for some recent surveys). For example, in the *density sublevel set* formulation of anomaly detection, given samples from a probability distribution, we wish to determine if an instance belongs to the induced low-density region of the input space [Ripley, 1996, pg. 24–25], [Steinwart et al., 2005], [Chandola et al., 2009, Section 5.2].

A practically relevant extension of this problem, which we term *calibrated* anomaly detection, is where we additionally wish to produce a *confidence* score for an instance being anomalous [Gao and Tan, 2006, Sotiris et al., 2010, Kriegel et al., 2011]. Intuitively, this confidence reflects the instance's level of abnormality, i.e., the *relative* value of its density compared to the other instances. Such information is valuable for subsequent decision making using these predictions [Gao and Tan, 2006].

In this paper, we present a loss function framework for calibrated anomaly detection. This builds upon an extant classification framework for density sublevel set problem [Steinwart et al., 2005], which related the latter to binary classification of samples from the observed distribution against a known "background" distribution. Our contributions are to extend this framework in three ways:

(**C1**) we show that minimising a class of *proper* losses asymptotically estimates the density for anomalous instances, while ignoring the density for non-anomalous instances
(**C2**) we show how to obtain (implicit) *quantile control* by connecting these proper losses to a generalised pinball loss for quantile elicitation [Ehm et al., 2016, Equation 5]
(**C3**) for a specific family of losses, we show how efficient optimisation can be performed with *kernelised* scores using an idea from the field of point processes [Flaxman et al., 2017].

C2 & **C3** are also applicable to standard anomaly detection, and thus may be of independent interest.

*Now at Google Research.

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

Method	Bayes-optimal?	Calibration?	Quantile control?	Tractable optimisation?
Heuristic sampling [Fan et al., 2001]	×	×	×	✓
Cost-sensitive loss [Steinwart et al., 2005]	✓	×	×	∼
Minimum volume set [Scott and Nowak, 2006]	✓	✓	✓	×
One-class SVM [Schölkopf et al., 2001]	∼	∼	✓	✓
Ours	✓	✓	✓*	✓
	§4.2	§4.2	§5	§6

Table 1: Comparison of various loss-based approaches to anomaly detection. For each method, we indicate whether they asymptotically produce a Bayes-optimal solution; can produce calibrated anomaly scores; offer quantile control; and afford tractable optimisation. The approach of Steinwart et al. [2005] is tractable only when sampling to approximate a high-dimensional integral. For the one-class SVM, Bayes-optimality and calibration are only achieved with certain approximations; see §6.2. For our method, we use a qualified ✓* since our quantile control is implicit; see §5.

Broadly, our aim in this paper is not to propose a new "best" method for (calibrated) anomaly detection. Rather, akin to Steinwart et al. [2005], we wish to explicate what precisely the target object for the problem is, and characterise the family of losses which yield this object upon risk minimisation. Optimisation-specific considerations are then built upon this foundation, allowing for a separation of statistical and computational concerns.

As an illustration, a special case of our framework closely resembles the one-class SVM (OC-SVM) [Schölkopf et al., 2001]. This algorithm has previously been noted to produce estimates of the tail density in a certain limiting regime, owing to the regulariser acting as both a model penalty, and a risk approximator [Vert and Vert, 2006]. Our framework allows one to separate these statistical and computational roles, and ascribe their influence to individual terms in the OC-SVM objective. Table 1 compares our framework to this, and various other loss-based approaches to anomaly detection.

2 Background on anomaly detection

Fix a probability distribution P over a measurable space $\mathcal{X}$ (typically $\mathbb{R}^n$ with the Borel sigma-algebra), and a reference measure μ (typically Lebesgue) for which the density $p \doteq \frac{\mathrm{d}P}{\mathrm{d}\mu}$ exists.

Standard anomaly detection. Given samples from P, there are two ways we can pose low-density anomaly detection. The first is the *density sublevel set* formulation [Ripley, 1996, pg. 24–25], [Steinwart et al., 2005]. Here, we are given $\alpha > 0$, and wish to produce $c\colon \mathcal{X} \to \{\pm 1\}$ satisfying

$$c(x) = 2 \cdot [\![p(x) > \alpha]\!] - 1. \tag{1}$$

That is, instances with low density are deemed to be anomalous. For simplicity, we assume $P(\{x \in \mathcal{X}\colon p(x) = \alpha\}) = 0$, and thus disregard ties. The second is the *minimum volume set* or *p-value* formulation [Scott and Nowak, 2006, Zhao and Saligrama, 2009, Chen et al., 2013]. Here, we are given $q \in (0, 1)$, and wish to produce $c\colon \mathcal{X} \to \{\pm 1\}$ satisfying

$$c(x) = 2 \cdot [\![p(x) > \alpha_q]\!] - 1,$$

where α_q is the qth quantile of the random variable of density scores $\mathsf{P} \doteq p(\mathsf{X})$, where $\mathsf{X} \sim P$. That is, instances with low density *relative to other instances* are deemed to be anomalous. The appeal of this is one does not need to *a priori* know the range of density values to pick a suitable threshold α.

Both formulations of anomaly detection have been the subject of considerable study, and algorithmic development [Chandola et al., 2009]. These include the one-class SVM [Schölkopf et al., 2001, Tax and Duin, 2004], neighbourhood-based approaches [Breunig et al., 2000, Zhao and Saligrama, 2009], and classification-based approaches [Fan et al., 2001, Steinwart et al., 2005, Cheema et al., 2016].

Calibrated anomaly detection. In *calibrated anomaly detection*, we wish to output a confidence in an instance being anomalous, based on the density quantiles: we are given $q \in (0, 1)$, and wish to produce $\eta\colon \mathcal{X} \to [0, 1]$ satisfying

$$\eta(x) \in \begin{cases} \{F_{\mathsf{P}}(p(x))\} & \text{if } p(x) < \alpha_q \\ [q, 1] & \text{if } p(x) > \alpha_q, \end{cases} \tag{2}$$

where F_{P} denotes the cumulative distribution function of $\mathsf{P} = p(\mathsf{X})$. Intuitively, $F_{\mathsf{P}}(p(x))$ is the density quantile corresponding to $p(x)$, so that $p(x) > \alpha_q \iff F_{\mathsf{P}}(p(x)) > q$. The goal is to model this quantile only for the anomalous instances, i.e., those with $p(x) < \alpha_q$; for non-anomalous instances, we need only stipulate that the outputs are larger than all anomalous ones.

Calibrated anomaly detection may be contrasted to the problem of density estimation [Wasserman, 2006, Chapter 6], which seeks to accurately model the density p, rather than $F_{\mathsf{P}} \circ p$; since the range of the density will be unknown *a priori*, the raw values will not be obviously interpretable as a measure of confidence. It may also be contrasted to the problem of anomaly ranking [Clémençon and Jakubowicz, 2013, Goix et al., 2015], which seeks to produce *any* strictly increasing transformation of p; the resulting scores thus may not be meaningful probabilities. Nonetheless, given a scorer $f\colon \mathcal{X} \to \mathbb{R}$ which preserves the order of the densities, and instances $\{x_n\}_{n=1}^N \sim P^N$, one may compute the empirical p-value [Zhao and Saligrama, 2009, Chen et al., 2013]

$$\hat{\eta}(x; \{x_n\}) \doteq \frac{1}{N} \sum_{n=1}^{N} [\![f(x_n) < f(x)]\!] \,. \tag{3}$$

3 Losses for density (sublevel-set) estimation

We review a result of Steinwart et al. [2005], which showed how to cast density sublevel set estimation as loss minimisation. We then show how by tweaking the loss, we can cast full density estimation in the same framework. This will motivate tuning the loss to interpolate between these problems.

3.1 A loss framework for density (sublevel set) estimation

Let $\ell\colon \{\pm 1\} \times \mathbb{R} \to \mathbb{R}$ be a binary label loss, such as the logistic loss $\ell(y, v) = \log(1 + e^{-yv})$. Let $f\colon \mathcal{X} \to \mathbb{R}$ be a measurable, real-valued scorer. For P, μ as in §2, suppose $\ell(+1, f(\cdot))$ is P-integrable, and $\ell(-1, f(\cdot))$ is μ-integrable. Define the *risk* of f as

$$R(f; P, \mu, \ell) \doteq \mathop{\mathbb{E}}_{\mathsf{X} \sim P} [\ell(+1, f(\mathsf{X}))] + \int_{\mathcal{X}} \ell(-1, f(x)) \, \mathrm{d}\mu(x). \tag{4}$$

Where clear from context, we drop the dependence of the risk on (P, μ, ℓ). When $\mu(\mathcal{X}) < +\infty$, μ is a scaled probability measure, and $R(f)$ is the risk for a binary classification problem of distinguishing "positive" observations (drawn from P) from a "negative" background (drawn from μ). We now see that minimising $R(f)$ can be used for density (sublevel) set estimation, by appropriately choosing ℓ.

Density sublevel set estimation as classification. Steinwart et al. [2005] observed that for suitable *cost-weighted* losses, one may use (4) to solve density sublevel set estimation. Specifically, let $\ell\colon \{\pm 1\} \times \mathbb{R} \to \mathbb{R}$ be any classification-calibrated loss (e.g., hinge or logistic) in the sense of [Zhang, 2004, Bartlett et al., 2006]. Given a cost parameter $c \in (0, 1)$, define the cost-weighted loss

$$(\forall v \in \mathbb{R})\, \ell^{(c)}(+1, v) \doteq (1 - c) \cdot \ell(+1, v) \qquad \ell^{(c)}(-1, v) \doteq c \cdot \ell(-1, v). \tag{5}$$

As an example, for the 0-1 loss ℓ_{01}, we obtain the *cost-sensitive loss* $\ell_{01}^{(c)}$.

It turns out that minimising (4) with $\ell^{(c)}$ yields density sublevel set estimates. One can see this by studying the *Bayes-optimal scorers* for the risk, i.e., the theoretical minimisers of the risk over the set $\mathcal{M}(\mathcal{X}, \mathbb{R})$ of all measurable scorers. The following is a trivial generalisation of [Steinwart et al., 2005, Proposition 5], which was for ℓ being the 0-1 loss ℓ_{01}.

Proposition 1: *Pick any $\alpha > 0$, and classification-calibrated ℓ. For $c_\alpha \doteq \frac{\alpha}{1+\alpha}$ and $\ell^{(c_\alpha)}$ per (5), let*

$$\begin{aligned} f^* &\in \mathop{\mathrm{Argmin}}_{f \in \mathcal{M}(\mathcal{X}, \mathbb{R})} R(f; P, \mu, \ell^{(c_\alpha)}) \\ &= \mathop{\mathrm{Argmin}}_{f \in \mathcal{M}(\mathcal{X}, \mathbb{R})} \mathop{\mathbb{E}}_{\mathsf{X} \sim P} [\ell(+1, f(\mathsf{X}))] + \alpha \cdot \int_{\mathcal{X}} \ell(-1, f(x)) \, \mathrm{d}\mu(x). \end{aligned}$$

Then, for μ-almost every $x \in \mathcal{X}$, $f^*(x) > 0 \iff p(x) > \alpha$.

Proposition 1 says that minimising (4) asymptotically recovers the target object for density sublevel set estimation, i.e., (1). Following Steinwart et al. [2005], we may also derive excess-risk bounds for $\ell_{01}^{(c)}$ in terms of $\ell^{(c)}$. This justifies reducing sublevel set estimation to classification of P versus μ.

Density estimation as class-probability estimation. We now observe that (4) can be used for estimating full densities as well. To do so, we move from classification-calibrated losses to the subset of *strictly proper composite* losses [Reid and Williamson, 2010]. These are the basic losses of *class-probability estimation* [Buja et al., 2005], and satisfy, for invertible *link* $\Psi\colon (0,1) \to \mathbb{R}$,

$$(\forall \theta \in (0,1)) \operatorname*{argmin}_{v \in \mathbb{R}} \mathop{\mathbb{E}}_{\mathsf{Y} \sim \mathrm{Bern}(\theta)} [\ell(\mathsf{Y}, v)] = \Psi(\theta). \tag{6}$$

In words, (6) stipulates that when using ℓ to distinguish positive and negative samples, it is optimal to predict (an invertible transformation of) the positive class-probability. A canonical example is the logistic loss $\ell(y,v) = \log(1 + e^{-yv})$ with link the logit function $\Psi\colon u \mapsto \log \frac{u}{1-u}$.

Applying strictly proper composite ℓ to distinguish P from μ in (4), the Bayes-optimal solution is a transform of the underlying density. For ℓ being the logistic loss, this observation has been made previously to motivate reducing unsupervised to supervised learning [Hastie et al., 2009, Section 14.2.4]. The generalisation is trivial, but will prove to have useful further implications.

Proposition 2: *Let ℓ be a strictly proper composite loss with link Ψ. If $f^* \in \operatorname{Argmin}_f R(f; P, \mu, \ell)$, then for μ-almost every $x \in \mathcal{X}$, $f^*(x) = \Psi_{\mathrm{rat}}(p(x))$, for the "ratio transformed" link function*

$$\Psi_{\mathrm{rat}}\colon z \mapsto \Psi\left(\frac{z}{1+z}\right). \tag{7}$$

Example 3: Consider the strictly proper composite loss

$$\ell(+1, v) = -v \qquad \ell(-1, v) = \frac{1}{2} \cdot v^2, \tag{8}$$

with link $\Psi(u) = u/(1-u)$, and corresponding $\Psi_{\mathrm{rat}}(z) = z$. Following Kanamori et al. [2009], we term this the "LSIF" loss. Minimising its risk performs least squares minimisation of f versus p:

$$R(f) = \mathop{\mathbb{E}}_{\mathsf{X} \sim P} [-f(\mathsf{X})] + \frac{1}{2} \cdot \int_{\mathcal{X}} f(x)^2 \, \mathrm{d}\mu(x) = \frac{1}{2} \cdot \int_{\mathcal{X}} [f(x) - p(x)]^2 \, \mathrm{d}\mu(x) + \mathrm{const}. \tag{9}$$

It is thus Bayes-optimal to predict $f^* = p$, i.e., recover the underlying density. This fact has been noted in the context of density ratio estimation [Kanamori et al., 2009].

3.2 Towards calibrated anomaly detection

It is not hard to show that if ℓ is strictly proper composite, then so is the loss $\ell^{(c)}$ of (5) [Menon and Ong, 2016, Lemma 5]. Further, every such ℓ is also classification-calibrated [Reid and Williamson, 2010, Theorem 16]. Propositions 1 and 2 thus imply that if f^* is the risk minimiser for $\ell^{(c)}$, computing $\mathrm{sign}(f^*)$ yields the target for density sublevel set estimation, while f^* by itself yields the (transformed) target for density estimation. The hardness of density estimation discourages against the latter.[2] However, the fact that the risk $R(\cdot)$ accommodates two extremes of the problem space offers hope in using it to address the "intermediate" problem of calibrated anomaly detection per (2).

Tackling calibrated anomaly detection poses several challenges, however. We need to

(i) obtain density p-values $F_{\mathsf{P}}(p(x))$, rather than raw density values $p(x)$,
(ii) focus attention on density values lower than a threshold, and
(iii) (ideally) have the threshold correspond to a specified quantile $q \in (0,1)$ of $p(\mathsf{X})$.

For (i), we may start with some base scorer $f\colon \mathcal{X} \to \mathbb{R}$, and compute a non-parametric estimate of the p-value via (3). For (ii), this scorer must preserve the order of the tail values of the density. This suggests we modify (2) and focus on the problem of *partial density estimation*, where given $\alpha > 0$, the goal is to produce a scorer $f\colon \mathcal{X} \to \mathbb{R}$ satisfying

$$f(x) \in \begin{cases} \{p(x)\} & \text{if } p(x) < \alpha \\ [\alpha, 1] & \text{if } p(x) > \alpha. \end{cases} \tag{10}$$

[2] A nonparametric density estimator will, in a minimax sense, require exponentially many samples to yield a good approximation to the underlying density [Stone, 1980]. However, thresholding potentially rough density estimates may still yield a reasonable sublevel set estimator. This is analogous to the plugin approach to binary classification, where one estimates the underlying class-probability and thresholds it to make a classifier. Under some assumptions, plugin approaches can achieve fast classification rates [Audibert and Tsybakov, 2007].

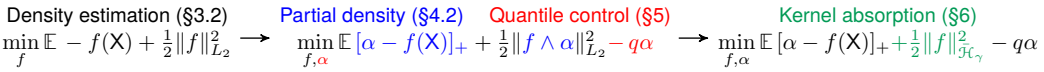


Figure 1: Summary of our approach for the LSIF loss from Example 3. Starting from a loss for full density estimation, we focus on only the tail of the densities by capping the loss (§4.2), add quantile control by adding a linear term related to the pinball loss (§5), and allow for tractable optimisation using a kernel absorption trick of Flaxman et al. [2017] (§6). Here, $\|f\|^2_{L_2} = \int_{\mathcal{X}} f(x)^2 \,\mathrm{d}\mu(x)$.

Finally, for (iii), we need to somehow automatically relate α to the quantile α_q.

We will now see how to suitably modify the loss in (4) to solve both (ii) and, at least implicitly, (iii); Figure 1 summarises. Our basic approach is to interpolate between the losses for density sublevel set and density estimation, noting that the former problem involves extracting a single sublevel set $\{x\colon p(x) < \alpha\}$ for given $\alpha > 0$, and the latter the *entire family* of sublevel sets $\bigcup_{\alpha'}\{x\colon p(x) < \alpha'\}$. We thus seek losses which are suitable for a *partial family* of sublevel sets $\bigcup_{\alpha' \leq \alpha}\{x\colon p(x) < \alpha'\}$. Quantile control is achieved by relating the result to the generalised pinball loss [Ehm et al., 2016].

4 Losses for partial density estimation

In order to obtain the optimal scorer for partial density estimation in (10), one could in theory simply perform full density estimation. However, this would entail solving a harder problem than we require. Can we focus attention on the *tail* of the density, so that no effort is placed on values larger than a threshold $\alpha > 0$? We will show how to do this using the *weight function* view of a proper loss.

4.1 Weight functions and density (sublevel set) estimation

Modelling only the tail of the density via (4) requires moving beyond strictly proper composite losses: from Proposition 2, *all* such losses have the *same* Bayes-optimal solution, up to a transformation. We will construct suitable alternate losses via the *weight function* representation of a proper composite loss: every such loss with invertible link Ψ can be written [Reid and Williamson, 2010, Theorem 6]

$$\ell\colon (y, v) \mapsto \int_0^1 w(c) \cdot \ell^{(c)}_{01}(y, \Psi^{-1}(v)) \,\mathrm{d}c, \tag{11}$$

where the (generalised) function $w\colon [0,1] \to \bar{\mathbb{R}}_{\geq 0}$ is the *weight*, and $\ell^{(c)}_{01}$ is the cost-sensitive loss

$$(\forall u \in [0,1])\; \ell^{(c)}_{01}(+1, u) \doteq (1-c) \cdot [\![u < c]\!] \qquad \ell^{(c)}_{01}(-1, u) \doteq c \cdot [\![u > c]\!]\,. \tag{12}$$

Thus, minimising ℓ equivalently minimises a mixture of cost-sensitive losses for various cost ratios, whose relative importance is determined by w; see Buja et al. [2005] for several examples of such w.[3]

To get some intuition on the role of w, let us re-interpret our previous results through this lens. For $\alpha > 0$, pick $w(c) = \delta(c - c_\alpha)$, a Dirac delta-function centred at $c_\alpha = \alpha/(1+\alpha)$. This weight corresponds (for a suitable link function) to the cost-sensitive loss $\ell^{(c_\alpha)}_{01}$ of (12), which by Proposition 1 estimates the αth density sublevel set. Intuitively, by having w be non-zero only at c_α, we encourage modelling *only* the single sublevel set $\{x \in \mathcal{X} \mid p(x) < \alpha\}$.

If we instead pick $w(c) = (c \cdot (1-c))^{-1}$, then (11) corresponds (for the logistic link function) to the logistic loss [Buja et al., 2005, Equation 19]. Proposition 2 shows this loss estimates the *entire* density. Intuitively, by having w be non-zero everywhere, we encourage modelling of an entire *family* of sublevel sets $\{x \in \mathcal{X} \mid p(x) < \alpha\}$ for every $\alpha > 0$, which is equivalent to modelling the density.

Inspired by this, we now design suitable losses for partial density estimation.

4.2 Partial density estimation via c_α-strictly proper losses

Suppose we only want to model density values below some $\alpha > 0$, per (10). A natural idea is to start with a loss having strictly positive weight function $w\colon [0,1] \to \bar{\mathbb{R}}_{>0}$, which we modify to

$$\bar{w}\colon c \mapsto [\![c \leq c_\alpha]\!] \cdot w(c) \tag{13}$$

[3] One can also interpret w as a prior about cost ratios we will be evaluated on; see Appendix B.

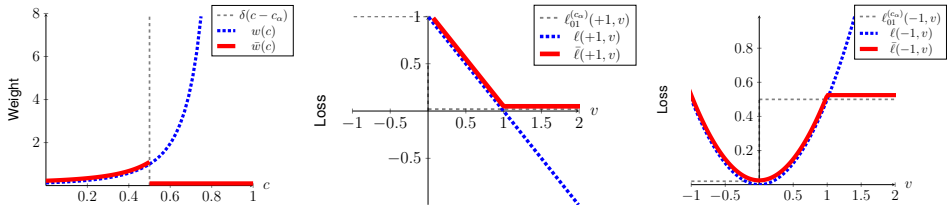


Figure 2: Illustration of weight functions and losses. Given cost threshold $c_\alpha = 0.5$, the weight $\delta(c - c_\alpha)$ gives a cost-sensitive loss suitable for density sublevel set estimation; $w(\cdot)$ gives a strictly proper loss suitable for density estimation; and $\bar{w}$ gives a c_α-strictly proper loss suitable for partial density estimation. The losses corresponding to $\bar{w}$ are saturated versions of those corresponding to w.

for $c_\alpha = \alpha/(1+\alpha)$, so that we place *no* weight on cost ratios above c_α. Intuitively, we will pay no attention to modelling the sublevel sets for thresholds above α, i.e., on modelling densities above α. We will call the loss resulting from this $\bar{w}$ a *c_α-strictly proper loss*.

Given a base loss ℓ, we may explicitly compute the form of $\bar{\ell}$ induced from (13).

Proposition 4: *Pick any $\alpha > 0$ and strictly proper composite ℓ with link $\Psi : (0,1) \to \mathbb{R}$. Suppose $\bar{\ell}$ is the loss with weight function given by (13). Then, for $\rho_\alpha = \Psi_{\mathrm{rat}}(\alpha)$, up to constant translation,*

$$(\forall v \in \mathbb{R})\, \bar{\ell}(+1, v) = \ell(+1, v \wedge \rho_\alpha) \qquad \bar{\ell}(-1, v) = \ell(-1, v \wedge \rho_\alpha). \tag{14}$$

Further, if $f^* \in \operatorname{Argmin}_f R(f; P, \mu, \bar{\ell})$, then for μ-almost every $x \in \mathcal{X}$,

$$p(x) < \alpha \implies f^*(x) = \Psi_{\mathrm{rat}}(p(x)) \quad \textit{and} \quad p(x) \geq \alpha \implies f^*(x) \geq \Psi_{\mathrm{rat}}(\alpha). \tag{15}$$

Figure 2 provides an illustrative example. Observe that if $\ell(+1, \cdot)$ is strictly decreasing, $\bar{\ell}(+1, v) = [\ell(+1, v) - \ell(+1, \rho_\alpha)]_+$ up to constant translation, explicating that the loss *saturates* beyond ρ_α.

From (15), the new Bayes-optimal scorer will only model densities less than α: for larger densities, any prediction larger than α suffices. This additional flexibility may prove useful in the pervasive scenario when one learns with a restricted class of scorers $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$: by ignoring the behaviour of the density above α, we may be able to select a scorer from $\mathcal{F}$ which better models the tail densities.

When $\alpha = \sup_x p(x)$, partial density estimation is identical to full density estimation, which is challenging. As $\alpha \to 0$, the estimation problem becomes less constrained and thus intuitively easier[4], with the limiting case $\alpha = 0$ trivially having every scorer as optimal. For fixed α, the choice of underlying weight w influences the precise range of density values where modelling effort is spent.

Example 5: For the LSIF loss in Example 3, one may show $w(c) = (1-c)^{-3}$ [Menon and Ong, 2016, Table 1]. Unwrapping (14), the c_α-strictly proper version of the loss is

$$\bar{\ell}(+1, v) = -(v \wedge \rho_\alpha) \qquad \bar{\ell}(-1, v) = 1/2 \cdot (v \wedge \rho_\alpha)^2$$

with $\rho_\alpha = \alpha$. Upto translation, $\bar{\ell}(+1, \cdot) = [\rho_\alpha - v]_+$ is the hinge loss. The risk is (c.f. (9))

$$R(f) = \frac{1}{2} \cdot \int_{\mathcal{X}} (p(x) - (f(x) \wedge \rho_\alpha))^2 \, \mathrm{d}\mu(x) + \text{const} \tag{16}$$

so that we perform least-squares minimisation with a *capped* version of our predictor. Evidently, the Bayes-optimal solution f^* is exactly the partial density estimation target (10)

$$p(x) \leq \alpha \implies f^*(x) = p(x) \quad \text{and} \quad p(x) > \alpha \implies f^*(x) \geq \alpha.$$

In practice, one does not expect to be in a Bayes-optimal regime: thus, empirically minimising (16), e.g., will not yield exact tail densities. Nonetheless, the scores should approximately preserve the density ordering, owing to the relation between proper loss minimisation and ranking [Agarwal, 2013]; this suffices for the empirical p-value per (3). Note also that minimising the ranking risk with exponential surrogate equivalently minimises its classification counterpart [Ertekin and Rudin, 2011].

[4]The claim of "easiness" of the problem is informal; we do not claim that changing α affects the minimax rate of convergence. We note however that even for sublevel set estimation, without some smoothness assumptions on p, the minimax rate will have exponential dependence on the dimensionality [Tsybakov, 1997, Theorem 4].

4.3 Partially capped losses

The loss $\bar{\ell}$ of (14) yields Bayes-optimal solutions suitable for density sublevel set estimation, but at price: in general, the loss will be non-convex. This is because a strictly proper composite loss has strictly increasing $\ell(-1,\cdot)$, meaning the capped version $\bar{\ell}(-1,\cdot)$ will saturate, and thus be non-convex. To resolve this, one may naïvely remove the saturation, yielding the *partially capped* loss

$$(\forall v \in \mathbb{R})\ \tilde{\ell}(+1, v) = \ell(+1, v \wedge \rho_\alpha) \qquad \tilde{\ell}(-1, v) = \ell(-1, v). \tag{17}$$

Surprisingly, this seemingly simple-minded correction retains admissible Bayes-optimal solutions: one now obtains a saturated version of the underlying density.

Proposition 6: *Pick any $\alpha > 0$, and let $\tilde{\ell}$ be a partially capped loss as per (17). If $f^* \in \operatorname{Argmin}_f R(f; P, \mu, \tilde{\ell})$, then for μ-almost every $x \in \mathcal{X}$, $f^*(x) = \Psi_{\mathrm{rat}}(p(x) \wedge \alpha)$.*

Observe that when $p(x)$ is above α, the Bayes-optimal scorers no longer have complete flexibility in their predictions: instead, they are clamped at exactly α. Intuitively, the problem of estimating such capped densities is easier than full density estimation, but harder than partial density estimation. Thus, if non-convexity is not an issue, it may be preferable to simply use c_α-strictly proper losses. On the other hand, we shall shortly see that partially proper losses can provide some quantile control.

Example 7: Unwrapping (17), the modified version of the LSIF loss in Example 3 has

$$\tilde{\ell}(+1, v) = [\rho_\alpha - v]_+ \qquad \tilde{\ell}(-1, v) = {}^1\!/_2 \cdot v^2, \tag{18}$$

where again $\rho_\alpha = \alpha$. Observe that $\tilde{\ell}(-1,\cdot)$ is not capped. The risk is (c.f. (16))

$$\begin{aligned} R(f) &= \mathop{\mathbb{E}}_{\mathsf{X}\sim P}[\rho_\alpha - f(\mathsf{X})]_+ + \frac{1}{2}\cdot\int_{\mathcal{X}} f(x)^2 \,\mathrm{d}\mu(x) \\ &= \frac{1}{2}\cdot\int_{\mathcal{X}} \left\{[\![f(x) \le \rho_\alpha]\!]\cdot(p(x) - f(x))^2 + [\![f(x) \ge \rho_\alpha]\!]\cdot f(x)^2\right\} \mathrm{d}\mu(x) + \mathrm{const}, \end{aligned} \tag{19}$$

so that we perform least-squares minimisation with a *capped* version of our predictor, with an additional penalisation term. The Bayes-optimal solution is the capped density $f^*(x) = p(x) \wedge \alpha$.

5 Quantile control via the pinball loss

We consider the last challenge highlighted in §3.2: achieving quantile control. Concretely, given a desired quantile level $q \in (0,1)$, how can we automatically infer (a bound on) the suitable density threshold α_q? A natural idea is to incorporate the *pinball loss* [Steinwart and Christmann, 2008, Equation 2.26] into our objective. Recall that for $q \in (0,1)$, this is the asymmetric linear loss $\phi(z; q) \doteq (1-q)\cdot[z]_+ + q\cdot[-z]_+$. The distributional minimiser of this loss produces the qth quantile of a distribution [Steinwart and Christmann, 2008, Proposition 3.9]: for a real-valued $\mathsf{F} \sim F$, any $\rho^* \in \operatorname{argmin}_{\rho\in\mathbb{R}} \mathbb{E}\left[\phi(\rho - \mathsf{F}; q)\right]$ is the qth quantile of F.

One would like to apply this loss to the distribution of our scorer's predictions, i.e., the distribution of $\mathsf{F} \doteq f(\mathsf{X})$, where $\mathsf{X} \sim P$. Naïvely, one might think to first estimate a scorer f, and to then find the quantile. However, one cannot do this in conjunction with modelling only the density tail: to even construct a c_α-strictly proper (or partially proper) loss, we require prior specification of the density threshold $\alpha > 0$, which by Proposition 4 maps to a threshold $\rho_\alpha = \Psi(\frac{\alpha}{1+\alpha})$ on the scores.

Fortunately, it is simple to jointly learn the scorer f and threshold ρ_{α_q} with partially proper losses. The key is the following elementary observation, relating the pinball and capped linear loss.

Lemma 8: *For any $q \in (0,1)$ and $z \in \mathbb{R}$, $\phi(z; q) = [z]_+ - q\cdot z$.*

To see why this is useful, observe that minimising the pinball loss on scores $\mathsf{F} = f(\mathsf{X})$ is equivalently

$$\min_\rho \mathop{\mathbb{E}}_{\mathsf{X}\sim P}\left[\phi\left(\rho - f(\mathsf{X}); q\right)\right] = \min_\rho \left[\mathop{\mathbb{E}}_{\mathsf{X}\sim P}[\rho - f(\mathsf{X})]_+ - q\cdot\rho\right] + q\cdot\mathop{\mathbb{E}}_{\mathsf{X}\sim P}\left[f(\mathsf{X})\right].$$

The second term on the right hand side is independent of ρ, and may be ignored. The first term has the partially proper loss $\tilde{\ell}(+1,\cdot)$ of (18) with parameter ρ, and an additional term $(-q\cdot\rho)$ independent

of f. Thus, if we add this term to our risk, it does not affect the optimisation over f; however, if we now optimise over ρ as well, we will get a quantile for $f(\mathsf{X})$.

This idea can be generalised to any partially capped loss $\tilde{\ell}$ (17) that builds on some strictly proper composite ℓ.[5] Concretely, we take such a loss with parameter ρ and construct the modified risk

$$\begin{aligned} R_q(f, \rho; \ell) &\doteq R(f; \tilde{\ell}) + q \cdot \ell(+1, \rho) \\ &= \mathop{\mathbb{E}}_{\mathsf{X} \sim P} [\ell(+1, f(\mathsf{X})) - \ell(+1, \rho)]_+ + \int_{\mathcal{X}} \ell(-1, f(\mathsf{X})) \, \mathrm{d}\mu(x) + q \cdot \ell(+1, \rho), \end{aligned} \tag{20}$$

which we now optimise over both the scorer f and threshold ρ. The following guarantee relies on relating the first and third terms to the generalised pinball loss functions for quantile estimation [Ehm et al., 2016, Equation 5]; effectively, one just transforms the inputs to the pinball loss through $\ell(+1, \cdot)$.

Proposition 9: *Pick any $q \in (0, 1)$ and strictly proper composite loss ℓ. If $(f^*, \rho^*) \in \operatorname{Argmin}_{f,\rho} R_q(f, \rho; q)$ then ρ^* is the qth quantile of $f^*(\mathsf{X})$, for $\mathsf{X} \sim P$.*

A subtlety with the above is that one obtains the qth quantile of $f^*(\mathsf{X})$; however, this is *not* the same as obtaining the qth quantile of $p(\mathsf{X})$. The reason is simple: the optimal $f^*(\mathsf{X})$ is itself a capped version of $p(\mathsf{X})$, and so the quantiles of the two quantities will not coincide. Nonetheless, we are guaranteed to get a quantity that is *bounded* by the qth quantile of $p(\mathsf{X})$, and the above parametrisation thus offers an intuitive control knob; see Appendix C for more discussion.

Example 10: For the LSIF loss in Example 3, the modified risk over f *and* ρ is (c.f. (19))

$$R_q(f, \rho) = \mathop{\mathbb{E}}_{\mathsf{X} \sim P} [\rho - f(\mathsf{X})]_+ + \int_{\mathcal{X}} \frac{1}{2} \cdot f(x)^2 \, \mathrm{d}\mu(x) - q \cdot \rho. \tag{21}$$

Remark 11: We can equally use (20) for density sublevel set estimation: we can minimise (20) and construct $\operatorname{sign}(f^*(x))$. This will produce an estimate of the sublevel set where $p(x) < \rho^*$ (by Proposition 4), and ρ^* corresponds to a quantile of f^* (by Proposition 9).

6 Optimisation without integral quadrature

The risk R_q of (20) asymptotically recovers capped density estimates, and provides implicit quantile control. But given samples $\{x_n\}_{n=1}^N \sim P$, how do we practically optimise R_q? A natural strategy is to replace the expectation over P with its empirical counterpart, and minimise

$$\hat{R}_q(f, \rho) = \frac{1}{N} \cdot \sum_{n=1}^{N} [\ell(+1, f(x_n)) - \ell(+1, \rho)]_+ + \int_{\mathcal{X}} \ell(-1, f(x)) \, \mathrm{d}\mu(x) + q \cdot \ell(+1, \rho). \tag{22}$$

While the first term is a standard empirical risk, the second term is more problematic: when $\mathcal{X}$ is high-dimensional, approximating the integral, e.g. via quadrature, will require a large number of samples. Fortunately, when using kernelised scorers, a simple trick lets us deal with this issue.

6.1 A kernel absorption trick

We now show how a specific choice of loss lets us avoid explicit integration. This uses a trick developed in the context of point processes [McCullagh and Møller, 2006, Flaxman et al., 2017, Walder and Bishop, 2017]. Suppose that $\mathcal{X}$ is compact with $\mu(\mathcal{X}) < +\infty$. Suppose we choose f from an RKHS $\mathcal{H}$ with continuous kernel $k \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. For $\gamma > 0$, define the regularised empirical risk

$$\hat{R}_q(f, \rho; \gamma) \doteq \hat{R}_q(f, \rho) + \frac{\gamma}{2} \cdot \|f\|_{\mathcal{H}}^2.$$

Now suppose $\ell(-1, v) = {}^1\!/_2 \cdot v^2$. Then, Flaxman et al. [2017] showed that by Mercer's theorem,

$$\int_{\mathcal{X}} \frac{1}{2} \cdot f(x)^2 \, \mathrm{d}\mu(x) + \frac{\gamma}{2} \cdot \|f\|_{\mathcal{H}}^2 = \frac{1}{2} \cdot \|f\|_{\bar{\mathcal{H}}(\gamma, \mu)}^2, \tag{23}$$

[5] We cannot apply this to a c_α-strictly proper loss, as here $\bar{\ell}(-1, \cdot)$ will also depend on ρ.

where $\bar{\mathcal{H}}(\gamma, \mu)$ is a *modified* RKHS, corresponding to the integral operator $T_{\bar{k}} \doteq T_k(T_k + \gamma I)^{-1}$. While somewhat abstract, one can explicitly compute the corresponding kernel $\bar{k}$ when the Mercer expansion of k is known, and use numerical approximations otherwise [Flaxman et al., 2017, Section 4]. Consequently, for $\ell(-1, v) = 1/2 \cdot v^2$ and $f \in \mathcal{H}$, the regularised empirical risk is *equivalent* to

$$\hat{R}_q(f, \rho; \gamma) = \frac{1}{N} \sum_{n=1}^{N} [\ell(+1, f(x_n)) - \ell(+1, \rho)]_+ + q \cdot \ell(+1, \rho) + \frac{1}{2} \cdot \|f\|^2_{\bar{\mathcal{H}}(\gamma,\mu)}.$$

One may now appeal to the representer theorem, and perform optimisation without any quadrature.

Example 12: For the LSIF loss in Example 3, the regularised risk with a kernelised f is (c.f. (21))

$$R_q(f, \rho; \gamma) = \mathop{\mathbb{E}}_{\mathsf{X} \sim P} [\rho - f(\mathsf{X})]_+ - q \cdot \rho + \frac{1}{2} \cdot \|f\|^2_{\bar{\mathcal{H}}(\gamma,\mu)}. \tag{24}$$

One subtlety with the empirical minimiser $\hat{f}^*$ of (24) is that $\hat{f}^*(x)$ could be negative for some x; this is inadmissible if we wish to interpret $\hat{f}^*(x)$ as an anomaly score. Following Kanamori et al. [2009], one may post-process scores to enforce non-negativity. More generally, one may start with an ℓ whose link Ψ has image the entire $\bar{\mathbb{R}}$ rather than $\bar{\mathbb{R}}_+$, e.g., the logistic loss.

Remark 13: We can equally use the above trick to ease optimisation for density sublevel set estimation, when employing $\ell(-1, v) = 1/2 \cdot v^2$ or its cost-weighted version.

6.2 Relation to one-class SVMs

We may contrast our objective in (24) to the one-class SVM (OC-SVM) [Schölkopf et al., 2001]. Fix an RKHS $\mathcal{H}$, and let $q \in (0, 1)$ be given. The primal OC-SVM objective is

$$R_q(f, \rho) = \mathop{\mathbb{E}}_{\mathsf{X} \sim P} [\rho - f(\mathsf{X})]_+ - q \cdot \rho + \frac{q}{2} \cdot \|f\|^2_{\mathcal{H}}.$$

This objective is almost identical to (24). The differences are that (24) regularises in a modified Hilbert space $\bar{\mathcal{H}}(\gamma, \mu)$, and that the regularisation strength is not tied to q. In (24), the regulariser $\|f\|^2_{\bar{\mathcal{H}}(\gamma,\mu)}$ conceptually plays two roles: it acts as a penalty on model complexity (owing to the $\|f\|^2_{\mathcal{H}}$ ingredient), and as a background loss (owing to the $\int_{\mathcal{X}} f(x)^2 \, \mathrm{d}\mu(x)$ ingredient). Since $\bar{\mathcal{H}}(\gamma, \mu)$ is transparently derived from $\mathcal{H}$ given $\gamma > 0$, one has complete control over the model complexity for a given choice of $q \in (0, 1)$. By contrast, in the OC-SVM, fixing q also fixes the model complexity.

Interestingly, Vert and Vert [2006] also related the Hilbert space norm $||f||^2_{\mathcal{H}}$ to $\int_{\mathcal{X}} f(x)^2 \, \mathrm{d}\mu(x)$, in the limit of zero bandwidth for a Gaussian kernel. This was then used to analyse the Bayes-optimal solutions for a limiting version of the OC-SVM. The trick from Flaxman et al. [2017] in the previous section shows how to more generally relate these two norms through $\bar{\mathcal{H}}(\gamma, \mu)$.

More broadly, our framework offers a different perspective on the individual ingredients in a OC-SVM style objective: the hinge loss arises from capping a linear loss $\ell(+1, \cdot)$, so as to compute tail density estimates; the term linear in ρ arises from re-expressing the hinge loss in terms of the pinball loss; and the regulariser arises from absorbing a model penalty plus a squared loss on the background distribution. Compared to the usual derivation of the OC-SVM, our objective arises from an explicit binary discrimination task, i.e., of P versus μ. We emphasise also that (24) is merely one special case of our framework, where the base ℓ is the LSIF loss (8). One may obtain different objectives by modifying ℓ: e.g., using the logistic loss, we obtain a form of "one-class logistic regression".

7 Concluding remarks

We presented a loss function framework for calibrated anomaly detection, built upon an extant classification framework for density level-set problem [Steinwart et al., 2005]. We extended this framework to a class of proper losses, incorporated quantile control, and discussed means of avoiding explicit quadrature for optimisation. The framework also produced a close relative of the one-class SVM as a special case, giving a different perspective on the individual components of this method.

While our focus has been mostly conceptual, some illustrative experiments are presented in Appendix D. More broadly, viewing calibrated anomaly detection as loss minimisation also lets one interpret the problem as *estimating an entropy*; see Appendix E for further discussion.

Acknowledgements

This work was supported by the Australian Research Council. It was motivated in part from a discussion with Zahra Ghafoori, whom we thank.

References

Shivani Agarwal. Surrogate regret bounds for the area under the ROC Curve via Strongly Proper Losses. In *Conference on Learning Theory (COLT)*, pages 338–353, 2013.

Jean-Yves Audibert and Alexandre B. Tsybakov. Fast learning rates for plug-in classifiers. *Ann. Statist.*, 35(2):608–633, 04 2007.

Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. LOF: Identifying density-based local outliers. In *ACM SIGMOD International Conference on Management of Data*, SIGMOD '00, pages 93–104, 2000.

Andreas Buja, Werner Stuetzle, and Yi Shen. Loss functions for binary class probability estimation and classification: Structure and applications. Technical report, UPenn, 2005.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3):15:1–15:58, July 2009. ISSN 0360-0300.

Prasad Cheema, Nguyen Lu Dang Khoa, Mehrisadat Makki Alamdari, Wei Liu, Yang Wang, Fang Chen, and Peter Runcie. On structural health monitoring using tensor analysis and support vector machine with artificial negative data. In *ACM International on Conference on Information and Knowledge Management*, CIKM '16, pages 1813–1822, 2016.

Yuting Chen, Jing Qian, and Venkatesh Saligrama. A new one-class SVM for anomaly detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*, pages 3567–3571, 2013.

Stéphan Clémençon and Jérémie Jakubowicz. Scoring anomalies: a M-estimation formulation. In *International Conference on Artificial Intelligence and Statistics*, pages 659–667, 2013.

Werner Ehm, Tilmann Gneiting, Alexander Jordan, and Fabian Krüger. Of quantiles and expectiles: consistent scoring functions, Choquet representations and forecast rankings. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(3):505–562, 2016.

Şeyda Ertekin and Cynthia Rudin. On equivalence relationships between classification and ranking algorithms. *Journal of Machine Learning Research*, 12:2905–2929, Oct 2011.

Wei Fan, M. Miller, S. J. Stolfo, Wenke Lee, and P. K. Chan. Using artificial anomalies to detect unknown and known network intrusions. In *Proceedings 2001 IEEE International Conference on Data Mining*, pages 123–130, 2001.

Seth R. Flaxman, Yee Whye Teh, and Dino Sejdinovic. Poisson intensity estimation with reproducing kernels. In *International Conference on Artificial Intelligence and Statistics*, pages 270–279, 2017.

Jing Gao and Pang-Ning Tan. Converting output scores from outlier detection algorithms into probability estimates. In *International Conference on Data Mining*, ICDM '06, pages 212–221, Washington, DC, USA, 2006. IEEE Computer Society.

Nicolas Goix, Anne Sabourin, and Stéphan Clémençon. On anomaly ranking and excess-mass curves. In *International Conference on Artificial Intelligence and Statistics*, 2015.

David J. Hand and Veronica Vinciotti. Local versus global models for classification problems: Fitting models where it matters. *The American Statistician*, 57(2):124–131, 2003. ISSN 00031305.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer, 2nd edition, 2009.

Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10:1391–1445, 2009.

Hans-Peter Kriegel, Peer Kroger, Erich Schubert, and Arthur Zimek. Interpreting and unifying outlier scores. In *SIAM International Conference on Data Mining*, pages 13–24, 2011.

Peter McCullagh and Jesper Møller. The permanental process. *Advances in Applied Probability*, 38 (4):873–888, 2006.

Aditya Krishna Menon and Cheng Soon Ong. Linking losses for density ratio and class-probability estimation. In *International Conference on Machine Learning*, pages 304–313, 2016.

Marco A.F. Pimentel, David A. Clifton, Lei Clifton, and Lionel Tarassenko. A review of novelty detection. *Signal Processing*, 99:215 – 249, 2014. ISSN 0165-1684.

Mark D. Reid and Robert C. Williamson. Composite binary losses. *Journal of Machine Learning Research*, 11:2387–2422, 2010.

Mark D. Reid and Robert C. Williamson. Information, divergence and risk for binary experiments. *Journal of Machine Learning Research*, 12:731–817, Mar 2011.

Brian D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge, 1996.

Bernhard Schölkopf, John C. Platt, John Shawe-Taylor, Alexander J. Smola, and Robert C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13 (7):1443–1471, 2001.

Clayton D. Scott and Robert D. Nowak. Learning minimum volume sets. *Journal of Machine Learning Research*, 7:665–704, December 2006. ISSN 1532-4435.

V. A. Sotiris, P. W. Tse, and M. G. Pecht. Anomaly detection through a Bayesian support vector machine. *IEEE Transactions on Reliability*, 59(2):277–286, June 2010. ISSN 0018-9529.

Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer, 2008. ISBN 9780387772417.

Ingo Steinwart, Don R. Hush, and Clint Scovel. A classification framework for anomaly detection. *Journal of Machine Learning Research*, 6:211–232, 2005.

Ingo Steinwart, Chloé Pasin, Robert C. Williamson, and Siyu Zhang. Elicitation and identification of properties. In *Conference on Learning Theory (COLT)*, pages 482–526, 2014.

Charles J. Stone. Optimal rates of convergence for nonparametric estimators. *Ann. Statist.*, 8(6): 1348–1360, 11 1980.

David M. J. Tax and Robert P. W. Duin. Support vector data description. *Machine Learning*, 54(1): 45–66, Jan 2004. ISSN 1573-0565.

Alexandre B. Tsybakov. On nonparametric estimation of density level sets. *Ann. Statist.*, 25(3): 948–969, 06 1997.

Régis Vert and Jean-Philippe Vert. Consistency and convergence rates of one-class SVMs and related algorithms. *Journal of Machine Learning Research*, 7:817–854, 2006.

Christian J. Walder and Adrian N. Bishop. Fast Bayesian intensity estimation for the permanental process. In *International Conference on Machine Learning*, ICML, pages 3579–3588, 2017.

Larry Wasserman. *All of Nonparametric Statistics*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387251456.

Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, 32(1):56–85, 2004.

Manqi Zhao and Venkatesh Saligrama. Anomaly detection with score functions based on nearest neighbor graphs. In *Advances in Neural Information Processing Systems 22*, pages 2250–2258, 2009.

Supplementary material for "A loss framework for calibrated anomaly detection"

A Proofs

We first present proofs of some helper results, before presenting those of results in the main body.

A.1 Helper results

In the body, we considered losses corresponding to weights restricted to be non-zero on an interval. To generalise this idea, suppose $w\colon [0,1] \to \bar{\mathbb{R}}_+$ is the weight for some proper loss. Now for some $c_\alpha \in (0,1)$, consider a weight function $\bar{w}\colon [0,1] \to \bar{\mathbb{R}}_+$ is of the form

$$\bar{w}\colon c \mapsto \begin{cases} w(c) & \text{if } c \leq c_\alpha \\ a \cdot w(c_\alpha) & \text{else,} \end{cases} \tag{25}$$

for fixed $c \in (0,1)$, $a \in [0,1]$. Intuitively, as $a \to 0$, we place less emphasis on cost ratios larger than c_α. We may explicitly compute the form of the corresponding proper loss $\bar{\ell}$.

Lemma 14: *Let $\lambda\colon \{\pm 1\} \times [0,1] \to \bar{\mathbb{R}}_+$ be a strictly proper loss with weight function $w\colon [0,1] \to \bar{\mathbb{R}}_+$. Pick any $c_\alpha \in (0,1)$ and $a \in [0,1]$. Let $\bar{\lambda}$ be the proper loss with weight function given by Equation 25. Then, for each $u \in (0,1)$,*

$$\begin{aligned} \bar{\lambda}(-1,u) &= a \cdot \lambda(-1,u) + (1-a) \cdot \lambda(-1, u \wedge c_\alpha) \\ \bar{\lambda}(+1,u) &= a \cdot \lambda(+1,u) + (1-a) \cdot (\lambda(+1, u \wedge c_\alpha) - \lambda(+1, c_\alpha)). \end{aligned} \tag{26}$$

One may easily verify that when $a = 1$, we get the original loss ℓ.

Proof of Lemma 14. By Shuford's integral formula [Reid and Williamson, 2010, Theorem 6],

$$\begin{aligned} \bar{\lambda}(-1,u) &= \int_0^1 \bar{\lambda}^{\mathrm{CS}(c)}(-1,u) \cdot \bar{w}(c) \,\mathrm{d}c \\ &= \int_0^u c \cdot \bar{w}(c) \,\mathrm{d}c \\ &= \int_0^{u \wedge c_\alpha} c \cdot \bar{w}(c) \,\mathrm{d}c + \int_{u \wedge c_\alpha}^u c \cdot \bar{w}(c) \,\mathrm{d}c \\ &= \int_0^{u \wedge c_\alpha} c \cdot w(c) \,\mathrm{d}c + \int_{u \wedge c_\alpha}^u a \cdot c \cdot w(c) \,\mathrm{d}c \\ &= \lambda(-1, u \wedge c_\alpha) + a \cdot (\lambda(-1,u) - \lambda(-1, u \wedge c_\alpha)) \\ &= a \cdot \lambda(-1,u) + (1-a) \cdot \lambda(-1, u \wedge c_\alpha) \end{aligned}$$

$$\begin{aligned} \bar{\lambda}(+1,u) &= \int_0^1 \bar{\lambda}^{\mathrm{CS}(c)}(+1,u) \cdot \bar{w}(c) \,\mathrm{d}c \\ &= \int_u^1 (1-c) \cdot \bar{w}(c) \,\mathrm{d}c \\ &= \int_u^{u \vee c_\alpha} (1-c) \cdot \bar{w}(c) \,\mathrm{d}c + \int_{u \vee c_\alpha}^1 (1-c) \cdot \bar{w}(c) \,\mathrm{d}c \\ &= \int_u^{u \vee c_\alpha} (1-c) \cdot w(c) \,\mathrm{d}c + a \cdot \int_{u \vee c_\alpha}^1 (1-c) \cdot w(c) \,\mathrm{d}c \\ &= \lambda(+1,u) - \lambda(+1, u \vee c_\alpha) + a \cdot \lambda(+1, u \vee c_\alpha) \\ &= a \cdot \lambda(+1,u) + (1-a) \cdot (\lambda(+1,u) - \lambda(+1, u \vee c_\alpha)) \\ &= a \cdot \lambda(+1,u) + (1-a) \cdot (\lambda(+1, u \wedge c_\alpha) - \lambda(+1, c_\alpha)), \end{aligned}$$

where the last line is because $f(x \wedge y) + f(x \vee y) = f(x) + f(y)$. □

A.2 Results in the body

Proof of Proposition 1. Since $p = \frac{\mathrm{d}P}{\mathrm{d}\mu}$, we may write

$$\begin{aligned} R(f; P, \mu, \ell) &= \mathop{\mathbb{E}}_{\mathsf{X} \sim P} [\ell(+1, f(\mathsf{X}))] + \int_{\mathcal{X}} \ell(-1, f(x)) \, \mathrm{d}\mu(x) \\ &= \int_{\mathcal{X}} [p(x) \cdot \ell(+1, f(\mathsf{X})) + \ell(-1, f(x))] \, \mathrm{d}\mu(x). \end{aligned}$$

For μ-almost every $x \in \mathcal{X}$, the Bayes-optimal scorer for loss $\ell^{(c_\alpha)}$ is

$$\begin{aligned} f^*(x) &\in \mathop{\mathrm{Argmin}}_v \left[p(x) \cdot \ell^{(c_\alpha)}(+1, v) + \ell^{(c_\alpha)}(-1, v) \right] \\ &= \mathop{\mathrm{Argmin}}_v \left[r(x) \cdot \ell^{(c_\alpha)}(+1, v) + (1 - r(x)) \cdot \ell^{(c_\alpha)}(-1, v) \right] \text{ for } r(x) \doteq \frac{p(x)}{1 + p(x)} > 0 \\ &= \mathop{\mathrm{Argmin}}_v \left[(1 - c_\alpha) \cdot r(x) \cdot \ell(+1, v) + c_\alpha \cdot (1 - r(x)) \cdot \ell(-1, v) \right] \\ &= 2 \cdot [\![(1 - c_\alpha) \cdot r(x) > c_\alpha \cdot (1 - r(x))]\!] - 1 \text{ by classification-calibration of } \ell \\ &= 2 \cdot [\![r(x) > c_\alpha]\!] - 1 \\ &= 2 \cdot \left[\!\!\left[p(x) > \frac{c_\alpha}{1 - c_\alpha} \right]\!\!\right] - 1 \text{ by definition of } r(x) \\ &= 2 \cdot [\![p(x) > \alpha]\!] - 1 \text{ by definition of } c_\alpha. \end{aligned}$$

□

Proof of Proposition 2. Following the proof of Proposition 1, for μ-almost every $x \in \mathcal{X}$, the Bayes-optimal scorer is

$$\begin{aligned} f^*(x) &\in \mathop{\mathrm{Argmin}}_v \left[r(x) \cdot \ell(+1, v) + (1 - r(x)) \cdot \ell(-1, v) \right] \text{ for } r(x) \doteq \frac{p(x)}{1 + p(x)} > 0 \\ &= \Psi(r(x)) \text{ by definition of a strictly proper composite loss (6)} \\ &= \Psi\left(\frac{p(x)}{1 + p(x)} \right) \text{ by definition of } r(x). \end{aligned}$$

□

Proof of Proposition 4. Applying Lemma 14 with $a = 0$, the proper loss corresponding to $\bar{w}$ is

$$\begin{aligned} \bar{\lambda}(-1, u) &= \lambda(-1, u \wedge c_\alpha) \\ \bar{\lambda}(+1, u) &= \lambda(+1, u \wedge c_\alpha) - \lambda(+1, c_\alpha). \end{aligned}$$

Since $\bar{\ell} = \bar{\lambda} \circ \Psi^{-1}$ and $c_\alpha = \Psi(\rho_\alpha)$,

$$\begin{aligned} \bar{\ell}(-1, u) &= \ell(-1, v \wedge \rho_\alpha) \\ \bar{\ell}(+1, u) &= \ell(+1, v \wedge \rho_\alpha) - \ell(+1, \rho_\alpha). \end{aligned}$$

Note that $\ell(+1, \rho_\alpha)$ is a constant, which plays no role in optimisation and thus may be safely ignored.

For any $\eta \in [0, 1]$, define the conditional risk $L_\eta(v) \doteq \eta \cdot \ell(+1, v) + (1 - \eta) \cdot \ell(-1, v)$. Following the proof of Proposition 1, for μ-almost every $x \in \mathcal{X}$, the Bayes-optimal scorer is

$$\begin{aligned} f^*(x) &\in \mathop{\mathrm{Argmin}}_v \left[r(x) \cdot \bar{\ell}(+1, v) + (1 - r(x)) \cdot \bar{\ell}(-1, v) \right] \text{ for } r(x) \doteq \frac{p(x)}{1 + p(x)} > 0 \\ &= \mathop{\mathrm{Argmin}}_v \left[r(x) \cdot (\ell(+1, v \wedge \rho_\alpha) - \ell(+1, \rho_\alpha)) + (1 - r(x)) \cdot \ell(-1, v \wedge \rho_\alpha) \right] \\ &= \mathop{\mathrm{Argmin}}_v \left[r(x) \cdot \ell(+1, v \wedge \rho_\alpha) + (1 - r(x)) \cdot \ell(-1, v \wedge \rho_\alpha) \right] \\ &= \mathop{\mathrm{Argmin}}_v L_{r(x)}(v \wedge \rho_\alpha) \end{aligned}$$

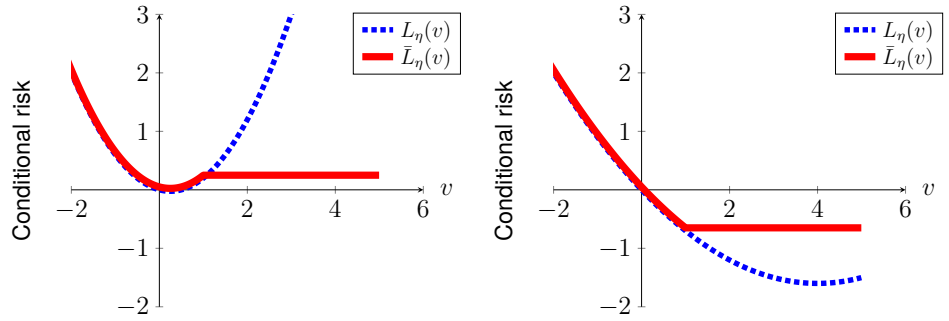


Figure 3: Illustration of conditional risk for original loss $\ell(+1, v) = -v$ and $\ell(-1, v) = \frac{1}{2} \cdot v^2$, and its c_α-strictly proper version $\bar{\ell}(+1, v) = -(v \wedge \rho)$ and $\ell(-1, v) = \frac{1}{2} \cdot (v \wedge \rho)^2$. We choose $\alpha = 1$, which corresponds to $\rho = \Psi(\frac{\alpha}{1+\alpha}) = 1$ as well. In the left plot, we show the conditional risks for $\eta = 0.2$, for which $\Psi(\eta) < \rho$. Confirming the theory, the minimiser for $\bar{L}$ is exactly $\Psi(\eta) = \frac{1}{4}$. In the right plot, we show the conditional risks for $\eta = 0.8$, for which $\Psi(\eta) > \rho$. Confirming the theory, the minimiser for $\bar{L}$ is any value in $[\rho, +\infty)$.

$$= \operatorname*{Argmin}_v \bar{L}_{r(x)}(v),$$

where $\bar{L}_\eta(v) \doteq L_\eta(v \wedge \rho_\alpha)$.

For any strictly proper loss, L_η is strictly quasi-convex for any $\eta \in [0, 1]$ [Reid and Williamson, 2010, Theorem 23]. Further, it has a unique minimum at $v^* = \Psi(\eta)$, by definition of strict properness. Consequently, L_η must be strictly decreasing on $[0, v^*]$ and strictly increasing on $[v^*, 1]$. Observe that $\bar{L}_\eta$ and L_η coincide on $[0, \rho_\alpha]$, while $\bar{L}_\eta$ remains constant on $[\rho_\alpha, +\infty)$. Thus, $\bar{L}_\eta$ must also be strictly decreasing on $[0, \rho_\alpha \wedge v^*]$, and constant on $[\rho_\alpha, +\infty)$.

Consider two cases for the minimiser of $\bar{L}_\eta$:

(a) suppose $v^* < \rho_\alpha$. Then, since $\bar{L}_\eta$ and L_η coincide on $[0, \rho_\alpha]$, we have $\bar{L}_\eta(v^*) < \bar{L}_\eta(v)$ for any $v \in [0, \rho_\alpha] - \{v^*\}$. But we know $\bar{L}_\eta$ is constant on $[\rho_\alpha, +\infty)$, and so $\bar{L}_\eta(v^*) < \bar{L}_\eta(v)$ for any v on this interval as well. Thus, the minimiser must occur at $v = v^*$.

(b) suppose $v^* \geq \rho_\alpha$. Then, we must have that L_η and $\bar{L}_\eta$ are strictly decreasing on $[0, \rho_\alpha]$. This means that ρ_α is the minimiser of $\bar{L}_\eta$ on $[0, \rho_\alpha]$. Now, since $\bar{L}_\eta$ is constant on $[\rho_\alpha, +\infty)$, it follows that any point in this interval is also a minimiser.

Thus, the optimal solution is $\bar{v}^* = v^*$ when $v^* < \rho_\alpha$, and $\bar{v}^* \geq \rho_\alpha$ when $v^* \geq \rho_\alpha$. Since $v^* = \Psi(\eta) = \Psi(r(x)) = \Psi_{\text{rat}}(p(x))$, the result follows.

As an example, we illustrate the behaviour of the conditional risk in Figure 3.

□

Proof of Proposition 6. Following the proof of Proposition 4, for μ-almost every $x \in \mathcal{X}$, the Bayes-optimal scorer is

$$\begin{aligned} f^*(x) &\in \operatorname*{Argmin}_v \left[r(x) \cdot \bar{\ell}(+1, v) + (1 - r(x)) \cdot \bar{\ell}(-1, v) \right] \text{ for } r(x) \doteq \frac{p(x)}{1 + p(x)} > 0 \\ &\in \operatorname*{Argmin}_v \left[r(x) \cdot \ell(+1, v \wedge \rho_\alpha) + (1 - r(x)) \cdot \ell(-1, v) \right] \\ &\in \operatorname*{Argmin}_v \bar{L}_\eta(v), \end{aligned}$$

where $\eta \doteq r(x)$ and $\bar{L}_\eta(v) \doteq \eta \cdot \ell(+1, v \wedge \rho_\alpha) + (1 - \eta) \cdot \ell(-1, v)$.

Recall that the conditional risk for ℓ is the strictly quasi-convex quantity $L_\eta \colon v \mapsto \eta \cdot \ell(+1, v) + (1 - \eta) \cdot \ell(-1, v)$. Further, it has a unique minimum at $v^* = \Psi(\eta)$, by definition of strict properness. We thus have that L_η is strictly decreasing on $[0, v^*]$ and strictly increasing on $[v^*, 1]$.

We have $\bar{L}_\eta(v) = L_\eta(v)$ for $v \leq \rho_\alpha$. For $v > \rho_\alpha$, we saturate the effect of the second term. Consider two cases for the minimiser of $\bar{L}_\eta$:

(a) suppose $v^* \leq \rho_\alpha$. Then, since v^* is the minimiser of L_η, and $\bar{L}_\eta(v) = L_\eta(v)$ on $[0, \rho_\alpha]$, it must also be the minimiser of $\bar{L}_\eta$ on $[0, \rho_\alpha]$. In particular, $\bar{L}_\eta(v^*) \leq \bar{L}_\eta(\rho_\alpha)$. Further, the partial loss $\ell(-1, \cdot)$ is strictly increasing for any strictly proper composite loss; thus, $\bar{L}_\eta$ must be strictly increasing on $[\rho_\alpha, +\infty)$. Consequently, the minimiser must be v^*.

(b) suppose $v^* > \rho_\alpha$. Then, L_η as well as $\bar{L}_\eta$ must be strictly decreasing on $[0, \rho_\alpha]$. Per the previous case, $\bar{L}_\eta$ must be strictly increasing on $[\rho_\alpha, +\infty)$. Consequently, the minimiser must be ρ_α.

The minimiser is thus $v^* \wedge \rho_\alpha$. Since $v^* = \Psi(\eta) = \Psi(r(x)) = \Psi_{\mathrm{rat}}(p(x))$, the result follows.

As an example, we illustrate the behaviour of the conditional risk in Figure 4. Compared to Figure 4, observe that in the right plot, we do *not* have saturation of the modified conditional risk $\bar{L}_\eta$. As a result, the optimal solution is capped at ρ_α, regardless of the value of v^*. □

Proof of Lemma 8. We have

$$\begin{aligned}
[z]_+ &= \begin{cases} z & \text{if } z \geq 0 \\ 0 & \text{else} \end{cases} \\
&= \begin{cases} (1-q) \cdot z + q \cdot z & \text{if } z \geq 0 \\ 0 & \text{else} \end{cases} \\
&= \begin{cases} (1-q) \cdot z & \text{if } z \geq 0 \\ -q \cdot z & \text{else} \end{cases} + q \cdot z \\
&= \phi(z; q) + q \cdot z.
\end{aligned}$$

□

Proof of Proposition 9. For any strictly decreasing function $h \colon \mathbb{R} \to \mathbb{R}$, and any $q \in (0, 1)$,

$$\begin{aligned}
h(z \wedge \rho) - h(\rho) &= [h(z) - h(\rho)]_+ \\
&= \phi(h(z) - h(\rho); q) + q \cdot (h(z) - h(\rho)) \text{ by Lemma 8} \\
&= \phi(h(z) - h(\rho); q) + q \cdot h(z) - q \cdot h(\rho).
\end{aligned}$$

For $\alpha > 0$ and strictly decreasing $\ell(+1, \cdot)$, let $\tilde{\ell}(+1, \cdot)$ be the partially proper loss with $\rho \doteq \rho_\alpha = \Psi(c_\alpha)$. Consider the modified risk

$$\begin{aligned}
R_q(f, \rho) &\doteq R(f) + q \cdot \ell(+1, \rho) \\
&= \mathop{\mathbb{E}}_{\mathsf{X} \sim P} [\ell(+1, f(\mathsf{X}) \wedge \rho) - \ell(+1, f(\mathsf{X}))] + q \cdot \ell(+1, \rho) + \int_{\mathcal{X}} \ell(-1, f(x)) \, \mathrm{d}\mu(x) \\
&= \mathop{\mathbb{E}}_{\mathsf{X} \sim P} [\ell(+1, f(\mathsf{X})) - \ell(+1, \rho)]_+ + q \cdot \ell(+1, \rho) + \int_{\mathcal{X}} \ell(-1, f(x)) \, \mathrm{d}\mu(x) \\
&= \mathop{\mathbb{E}}_{\mathsf{X} \sim P} [\phi(\ell(+1, f(\mathsf{X})) - \ell(+1, \rho); q)] + q \cdot \mathop{\mathbb{E}}_{\mathsf{X} \sim P} [\ell(+1, f(\mathsf{X}))] + \int_{\mathcal{X}} \ell(-1, f(x)) \, \mathrm{d}\mu(x),
\end{aligned}$$

which we now optimise over both f and ρ. For fixed ρ, the minimisation over f is unchanged; for fixed f, the second and third terms are irrelevant, and so the minimisation over ρ is

$$\min_\rho \mathop{\mathbb{E}}_{\mathsf{X} \sim P} [\phi(\ell(+1, f(\mathsf{X})) - \ell(+1, \rho); q)] \,. \tag{27}$$

The resulting minimiser will again be the qth quantile of the distribution of scores. This is owing to the calibration property of a generalisation of the pinball loss [Steinwart et al., 2014, Equation 15], [Ehm et al., 2016, Equation 5]: for strictly increasing $g \colon \mathbb{R} \to \mathbb{R}$, let

$$\begin{aligned}
\bar{\phi}(\hat{y}, y; g, q) &\doteq |\, [\![y \leq \hat{y}]\!] - q| \cdot |g(\hat{y}) - g(y)| \\
&= \begin{cases} (1-q) \cdot (g(\hat{y}) - g(y)) & \text{if } y \leq \hat{y} \\ q \cdot (g(y) - g(\hat{y})) & \text{else} \end{cases} \\
&= \phi(g(\hat{y}) - g(y), q).
\end{aligned} \tag{28}$$

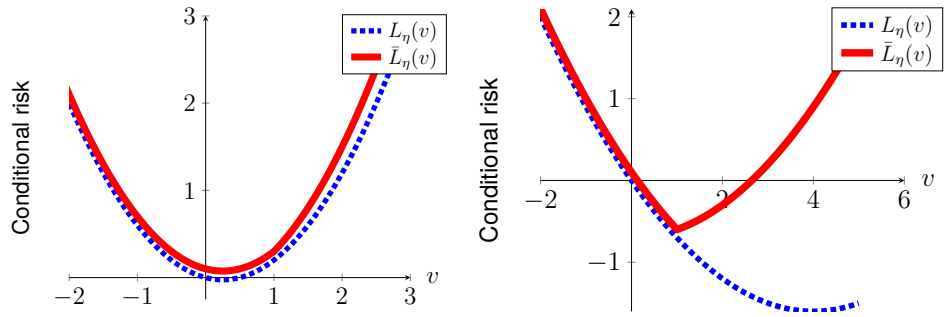


Figure 4: Illustration of conditional risk for original loss $\ell(+1, v) = -v$ and $\ell(-1, v) = \frac{1}{2} \cdot v^2$, and its "partially saturated" version $\bar{\ell}(+1, v) = -(v \wedge \rho)$ and $\ell(-1, v) = \frac{1}{2} \cdot v^2$. We choose $\alpha = 1$, which corresponds to $\rho = \Psi(\frac{\alpha}{1+\alpha}) = 1$ as well. In the left plot, we show the conditional risks for $\eta = 0.2$, for which $\Psi(\eta) < \rho$. Confirming the theory, the minimiser for $\bar{L}$ is exactly $\Psi(\eta) = \frac{1}{4}$. In the right plot, we show the conditional risks for $\eta = 0.5$, for which $\Psi(\eta) = \rho$. Confirming the theory, the minimiser for $\bar{L}$ is exactly ρ.

As before, if for a distribution F over $\mathbb{R}$, we pick

$$\rho^* \in \operatorname*{Argmin}_{\rho \in \mathbb{R}} \mathop{\mathbb{E}}_{\mathsf{F} \sim F} \left[\bar{\phi}(\rho, \mathsf{F}; g, q) \right],$$

then ρ^* is the qth quantile of F. For our problem, we simply need to set $g(v) = -\ell(+1, v)$ (recalling that $\ell(+1, \cdot)$ is strictly decreasing) to conclude that Equation 27 produces the qth quantile of $\mathsf{F} \doteq f^*(\mathsf{X})$.

Let us remark here that the optimisation for ρ is always quasi-convex, but will only be convex for $g(z) = z$, i.e., for $\ell(+1, v) = -v$ [Steinwart et al., 2014, Corollary 11]. However, one could define $\bar{\rho} \doteq g(\rho)$ and optimise over $\bar{\rho}$ instead, taking care to compute $g^{-1}(\bar{\rho})$ for the final prediction. □

B Interpreting the weight function as a prior

Given infinite data, and an arbitrarily powerful function class, there is no reason to favour one strictly proper loss over another: every such loss results in the same Bayes-optimal solution, namely, a transform of the underlying class-probability. Neither of these assumptions holds in practice, of course. In this case, minimising different losses will result in different solutions. How then might one choose amongst different losses?

To obtain one possible answer, consider the following scenario. Suppose there is a distribution D over $\mathcal{X} \times \{\pm 1\}$, representing a binary classification problem. One wishes to design a good classifier for this problem, as evaluated by a cost-sensitive loss $\ell_{01}^{(c)}$ for some $c \in (0,1)$. *However*, there is uncertainty as to what this cost ratio will be: one must first design a model, and only then get informed as to what the cost-ratio is.

It is natural to express this uncertainty over cost-ratios in terms of a (possibly improper) prior distribution π over $[0,1]$. In this case, the natural strategy for the learner is to minimise the *expected cost-sensitive loss* under costs drawn from π. Concretely, we would find a scorer $f \colon \mathcal{X} \to \bar{\mathbb{R}}$ and link $\Psi \colon \bar{\mathbb{R}} \to [0,1]$ to minimise

$$\mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y})\sim D}\left[\mathop{\mathbb{E}}_{\mathsf{C}\sim\pi}\left[\ell_{01}^{(\mathsf{C})}(\mathsf{Y}, \Psi^{-1}(f(\mathsf{X})))\right]\right] = \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y})\sim D}\left[\ell(\mathsf{Y}, f(\mathsf{X}))\right],$$

where

$$\ell(y,v) = \int_0^1 \ell_{01}^{(c)}(y, \Psi^{-1}(v)) \cdot \pi(c)\,\mathrm{d}c.$$

From (11), we identify this as a proper composite loss corresponding to weight function π. Thus, this suggests that we should choose our weight function to reflect our prior belief as to which cost-ratios we expect to be evaluated on. (Observe that, trivially, if one's true belief were π, it would not be rational to use a loss corresponding to $\pi' \neq \pi$; this would necessarily yield a solution that is sub-optimal loss according to cost-ratios drawn from our belief π.)

A related idea is to tune the weight function based on cost-ratios of interest, e.g., [Buja et al., 2005, Section 14], [Hand and Vinciotti, 2003].

In the context of anomaly detection, recall from the proof of Proposition 9 that our choice of loss $\ell(+1,\cdot)$ relates to a particular choice of increasing function $g(\cdot)$ parametrising a generalised quantile elicitation loss (28). Such losses also have an integral representation [Ehm et al., 2016, Theorem 1a],

$$\bar{\ell}_{1-q}^{\mathrm{pin}}(y,\hat{y};g) = \int_{-\infty}^{+\infty} \ell_{1-q}^{\mathrm{prim}}(y,\hat{y};v) \cdot h(v)\,\mathrm{d}v,$$

where $h(v) = g'(v)$, and the "primitive" quantile loss is

$$\ell_{1-q}^{\mathrm{prim}}(y,\hat{y};v) = (\llbracket y < \hat{y} \rrbracket - q) \cdot (\llbracket v < \hat{y} \rrbracket - \llbracket v < y \rrbracket).$$

These primitive losses are analogues of the cost-sensitive loss, and measure whether the prediction and ground truth are on the same side of a fixed threshold of v. For a strictly proper composite ℓ with weight π, we have [Reid and Williamson, 2010, Theorem 10]

$$-\ell'(+1,v) = \frac{\Psi^{-1}(v) - 1}{\Psi'(\Psi^{-1}(v))} \cdot \pi(\Psi^{-1}(v)).$$

Now note that the weighting over thresholds is $h(v) = g'(v) = -\ell'(+1,v)$. Consequently, the weighting π over costs c implicitly translates to a weighting h over thresholds v. We may thus pick π and Ψ to reflect the portion of the distribution F we wish to focus attention on.

C On the quantile-control parameter

In §5, we proposed a pinball-loss based approach to estimate a suitable density threshold α for a given control parameter $q \in (0,1)$. We expand upon the precise nature of the optimal α^*, explicating that while it will *not* be the qth quantile of $p(\mathsf{X})$ in general, we nonetheless get some meaningful control on the fraction of instances classified as anomalous.

Recall that for the underlying loss ℓ of Example 3, and parameter $q \in (0,1)$, our "quantile-controlled" objective is to jointly optimise over a scorer f and threshold α to minimise

$$R(f,\alpha) \doteq \mathop{\mathbb{E}}_{\mathsf{X}\sim P}[\alpha - f(\mathsf{X})]_+ + \frac{1}{2}\cdot \mathop{\mathbb{E}}_{\mathsf{X}\sim Q}\left[\frac{1}{2}\cdot f(\mathsf{X})^2\right] - q\cdot\alpha,$$

where for clarity, we use α instead of ρ (since this loss has $\rho_\alpha = \alpha$), and assume $\mu(\mathcal{X}) = 1$ so that it corresponds to some probability distribution Q.

Explicit quantile control: negative result In Proposition 9, we argued that the optimal α^* will be the qth quantile of $f^*(\mathsf{X})$, for $\mathsf{X}\sim P$, since the objective can be related to the pinball loss. However, this does *not* mean that α^* will be the qth quantile of $p(\mathsf{X})$. To see this, recall from Proposition 6 that the optimal scorer $f^*(x) = p(x)\wedge\alpha^*$. Thus, by definition of a quantile,

$$P(p(\mathsf{X})\wedge\alpha^* < \alpha^*) \le q \le P(p(\mathsf{X})\wedge\alpha^* \le \alpha^*).$$

This simplifies to

$$P(p(\mathsf{X}) < \alpha^*) \le q \le 1.$$

When $p(\mathsf{X})$ has continuous, invertible cumulative distribution function F, this implies $\alpha^* \le F^{-1}(q)$. Consequently, α^* is merely a *lower bound* on the qth quantile of $p(\mathsf{X})$.

Implicit quantile control: positive result. While lacking explicit quantile control, our objective nonetheless provides *implicit* quantile control in the following sense: sweeping over the entire range of q is equivalent to sweeping over the entire range of α^* values. But for fixed q, the corresponding α^* does not correspond to the qth quantile for the density.

To understand better the role of changing q, let us restrict attention to f of the form $f(x) = p(x)\wedge\alpha$, and consider a univariate optimisation over α alone. (Clearly, this choice of f will result in the same optimal solution for α.) Now consider the objective with respect to α alone:

$$\begin{aligned} R(\alpha) &\doteq \mathop{\mathbb{E}}_{\mathsf{X}\sim P}[\alpha - (p(\mathsf{X})\wedge\alpha)]_+ + \frac{1}{2}\cdot\mathop{\mathbb{E}}_{\mathsf{X}\sim Q}\left[\frac{1}{2}\cdot(p(\mathsf{X})\wedge\alpha)^2\right] - q\cdot\alpha \\ &= \mathop{\mathbb{E}}_{\mathsf{X}\sim Q}\left[p(\mathsf{X})\cdot[\alpha - p(\mathsf{X})]_+ + \frac{1}{2}\cdot(p(\mathsf{X})\wedge\alpha)^2\right] - q\cdot\alpha. \end{aligned}$$

Observe now that for any $p \ge 0$,

$$\begin{aligned} g(\alpha;p) &\doteq p\cdot[\alpha - p]_+ + \frac{1}{2}\cdot(p\wedge\alpha)^2 \\ &= \begin{cases} \frac{1}{2}\cdot\alpha^2 & \text{if } \alpha \le p \\ \alpha\cdot p - \frac{1}{2}\cdot p^2 & \text{else} \end{cases} \\ &= \begin{cases} \frac{1}{2}\cdot(\alpha-p)^2 & \text{if } \alpha \le p \\ 0 & \text{else} \end{cases} + \alpha\cdot p - \frac{1}{2}\cdot p^2 \\ &= \frac{1}{2}\cdot([p-\alpha]_+)^2 + \alpha\cdot p - \frac{1}{2}\cdot p^2. \end{aligned}$$

This function is evidently convex and differentiable with respect to α. Since R is constructed by integrating $g(\alpha;p)$ over p, it must also be convex and differentiable. More explicitly,

$$R(\alpha) = \mathop{\mathbb{E}}_{\mathsf{X}\sim Q}\frac{1}{2}\cdot([p(\mathsf{X})-\alpha]_+)^2 + (1-q)\cdot\alpha + \text{constant}.$$

Consequently, the first-order gradient condition on R implies that at optimality,

$$R'(\alpha^*) = 0 = \mathop{\mathbb{E}}_{\mathsf{X}\sim Q} -[p(\mathsf{X})-\alpha^*]_+ + 1 - q.$$

Consequently, for $G(\alpha) \doteq \mathop{\mathbb{E}}_{\mathsf{X}\sim Q}[p(\mathsf{X}) - \alpha]_+$, we have

$$G(\alpha^*) = 1 - q.$$

Now, the function $G\colon \mathbb{R}_+ \to \mathbb{R}_+$ is an integral over x of a family of functions that are continuous and strictly decreasing on $[0, p(x)]$. Consequently, G is continuous and strictly decreasing on $[0, F^{-1}(1))$, where F is the cumulative distribution function of $p(\mathsf{X})$. Further, at the endpoints we have $G(0) = 1$ and $G(F^{-1}(1)) = 0$.

As a result, there exists a continuous inverse G^{-1} such that for any $q \in [0, 1]$, we have $\alpha^* = G^{-1}(1 - q)$. It follows that α^* varies continuously as q is varied, and vice-versa. Further, when $q = 0$, $\alpha^* = 0$, while when $q \to 1$, $\alpha^* \to F^{-1}(1)$. Thus, at the boundaries, α^* coincides with the ordinary quantile of $p(\mathsf{X})$.

Example 15: To make the above concrete, consider the case where P has density $p(x) = 2 \cdot x$ with respect to the uniform measure over $\mathcal{X} = [0, 1]$. The objective of interest is

$$\begin{aligned}
R(\alpha) &= \mathop{\mathbb{E}}_{\mathsf{X}\sim Q}\left[p(\mathsf{X}) \cdot [\alpha - p(\mathsf{X})]_+ + \frac{1}{2} \cdot (p(\mathsf{X}) \wedge \alpha)^2\right] - q \cdot \alpha \\
&= \mathop{\mathbb{E}}_{\mathsf{X}\sim Q}\left[2 \cdot \mathsf{X} \cdot [\alpha - 2 \cdot \mathsf{X}]_+ + \frac{1}{2} \cdot (2 \cdot \mathsf{X} \wedge \alpha)^2\right] - q \cdot \alpha \\
&= \int_0^1 2 \cdot x \cdot [\alpha - 2 \cdot x]_+ + \frac{1}{2} \cdot (2 \cdot x \wedge \alpha)^2 \, \mathrm{d}x - q \cdot \alpha \\
&= \frac{\alpha^3}{12} + \frac{\alpha^2}{2} - \frac{\alpha^3}{6} - q \cdot \alpha \text{ when } \alpha \in [0, 2] \\
&= -\frac{\alpha^3}{12} + \frac{\alpha^2}{2} - q \cdot \alpha \text{ when } \alpha \in [0, 2].
\end{aligned}$$

This function is convex, with

$$R'(\alpha) = -\frac{\alpha^2}{4} + \alpha - q.$$

Thus, the optimal choice of α is

$$\alpha^* = 2 \cdot (1 - \sqrt{1 - q}).$$

Clearly, α^* varies from 0 to 2 as q is varied from 0 to 1. We can further explicitly compute the quantile of $p(\mathsf{X})$ corresponding to this choice of α^* as

$$P(p(\mathsf{X}) \leq \alpha^*) = P(\mathsf{X} \leq \alpha^*/2) = \left(\frac{\alpha^*}{2}\right)^2 = 2 - q - 2 \cdot \sqrt{1 - q}.$$

That is, as q is varied, we capture in a nonlinear manner a suitable quantile of $p(\mathsf{X})$.

D Experimental illustration

We present experiments illustrating the behaviour of methods introduced in the body of the paper.

D.1 Visualisation of optimal scorers

For our first experiment, we visualise the optimal scorers for various methods in a simple one-dimensional example. Specifically, we consider the Bayes-optimal solutions for:

- full density estimation, minimising the LSIF loss ℓ of Example 3
- partial density estimation with the "partially capped" loss $\tilde{\ell}$ of Example 7

D.1.1 Experimental setup

We consider a 1D distribution on $\mathcal{X} = [0, 1]$, with density $p(x) = 2 \cdot x$ with respect to the Lebesgue measure. We draw $n = 10^4$ samples from the distribution. For a given quantile level $q \in (0, 1)$, we minimise the empirical, quantile-corrected risk using Gaussian-kernelised scorers. Since the Gaussian kernel is univeral, the optimal solution is expected to mimic the Bayes-optimal solution. The bandwidth of the kernel was chosen by cross-validation, so as to optimise the log-likelihood of standard kernel density estimation. For ease of optimisation, we use the Nyström approximation to the kernel with $k = 500$ prototypes. For simplicity, we do not employ the kernel absorption trick; rather, we draw n samples from the uniform distribution over $[0, 1]$ to approximate $\int_{\mathcal{X}} \tilde{\ell}(-1, f(x)) \, \mathrm{d}\mu(x)$.

D.1.2 Experimental results

Figure 5 shows the optimal kernelised solution. We observe that, as predicted by the theory:

- full density estimation essentially recovers the underlying p
- partially capped loss minimisation recovers a capped version of the density

D.2 Calibration performance on synthetic data

In our next experiment, we validate that minimising a partially proper loss yields good density tail estimates, and that on finite samples these estimates tend to be superior to those produced by full density estimation.

To control all sources of error, we perform experiments on a number of synthetically generated distributions:

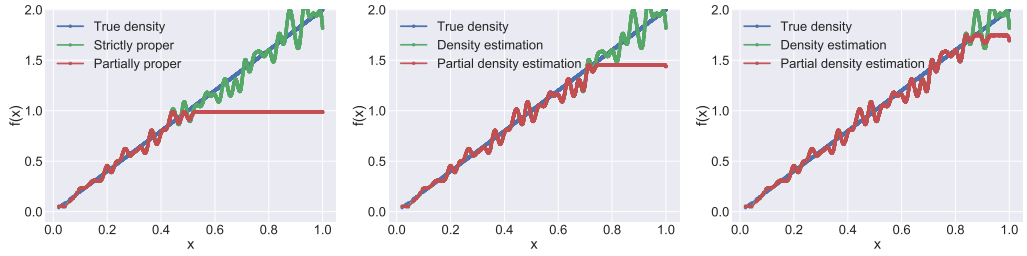
- `linear`, a distribution on $[0, 1]$ with density $p(x) = 2 \cdot x$.
- `arcsin`, an arcsine distribution on $[0, 1]$ with density $p(x) = (\pi \cdot \sqrt{x \cdot (1 - x)})^{-1}$.
- `truncnorm`, a truncated standard Gaussian distribution on $[0, 1]$.

Per the previous section, we draw n samples from each distribution. For a given quantile level $q \in (0, 1)$, we compute the empirical risk minimiser for both the strictly proper composite loss of Example 3, and its "partially capped", quantile-corrected version of Example 7. We use Gaussian-kernelised scorers, combined with the Nyström approximation to the kernel with $k = 500$ prototypes.

One we have an empirical risk minimiser f_n^*, we estimate its regret with respect to the partially capped loss with parameter α_q, where α_q is the (explicitly computable) qth quantile of the density distribution $p(\mathsf{X})$. That is, we estimate $R(f_n^*) - R(f^*)$, where f^* is the Bayes-optimal scorer $f^*(x) = p(x) \wedge \alpha_q$. This estimation is done via a separate empirical sample of instances drawn from p, and a uniform background.

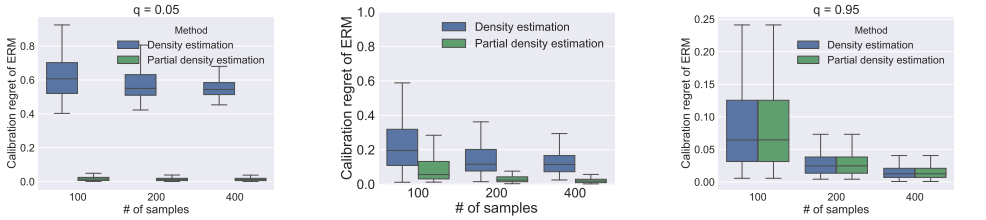
Figures 6 and 7 show the regret of the empirical risk minimiser for various values of $q \in (0, 1)$. We observe that:

- as expected, as the number of samples increases, the regret decreases
- the regrets are generally higher for $q = 0.95$ than all other cases; this is consistent with our intuition that full density estimation is more challenging than partial density estimation

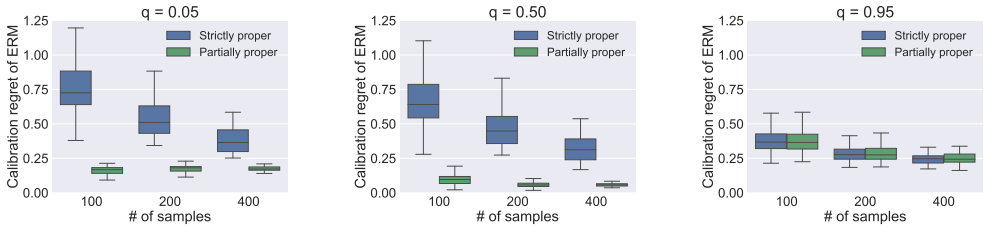


(a) $q = 0.25$. (b) $q = 0.50$. (c) $q = 0.75$.

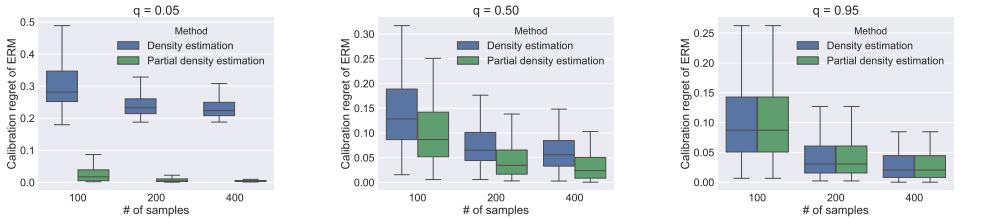
Figure 5: Optimal solution for kernelised scorers, and various quantile levels $q \in (0, 1)$.



(a) `linear`



(b) `arcsin`



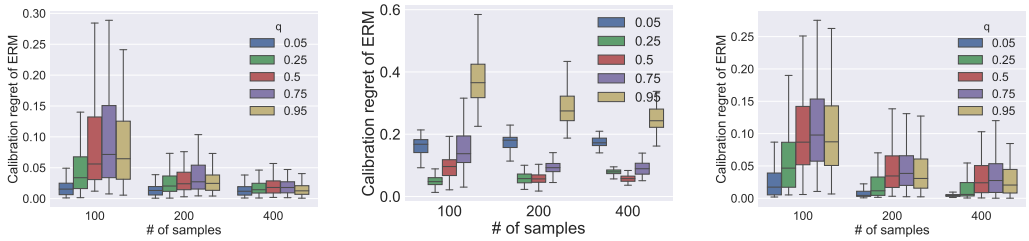
(c) `truncnorm`

Figure 6: Calibration regret of empirical risk minimiser for the strictly proper and partially proper composite loss, as a function of the number of training samples. Note that the vertical scales vary across the rows.

- the regrets are generally higher for $q = 0.05$ than $q = 0.50$; this possibly indicates that while the former involves an "easier" target object, one has to grapple with fewer available samples from the tail of the distribution

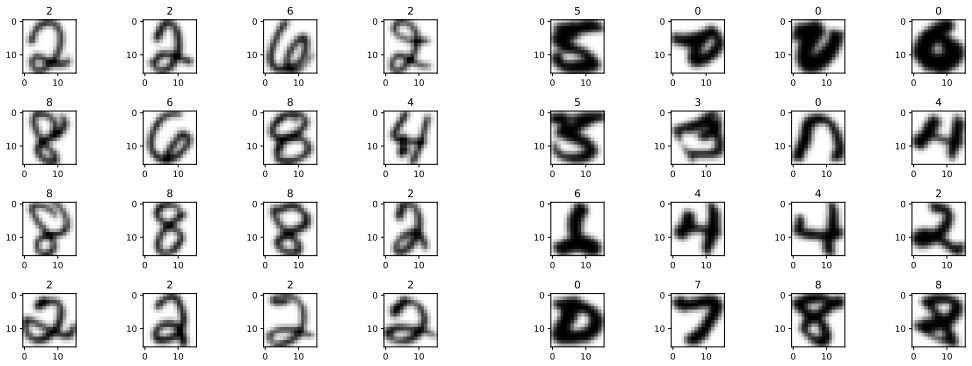
D.3 Anomaly detection performance on real-world datasets

We finally evaluate the anomaly detection performance of various approaches on real-world datasets. We emphasise that standard (as opposed to calibrated) anomaly detection is *not* the primary goal of our partially proper loss approach; while we can use our predictions for this task, we expect there to be a slight dip in performance compared to bespoke approaches. Nonetheless, we illustrate here that our method is consistently competitive.



(a) `linear` (b) `arcsin` (c) `truncnorm`

Figure 7: Calibration regret of empirical risk minimiser for the partially proper composite loss, as a function of the number of training samples. On each plot, we show the regrets for different values of q, controlling the desired quantile. Note that the vertical scales vary across the rows.



(a) Highest score. (b) Lowest score.

Figure 8: Samples of instances with highest and lowest score according to partially proper loss solution, `usps` dataset.

D.3.1 Qualitative analysis: anomalous digits

To begin, we perform a qualitative evaluation of anomaly detection performance on the `usps` dataset. Following Schölkopf et al. [2001], we work with the provided test set, and augment each instance with a binary one-hot encoding of the corresponding digit. We fit our method on the resulting instances, and then examine those which are assigned lowest scores. Intuitively, these instances are expected to be visually anomalous, as they lie in a region of low density.

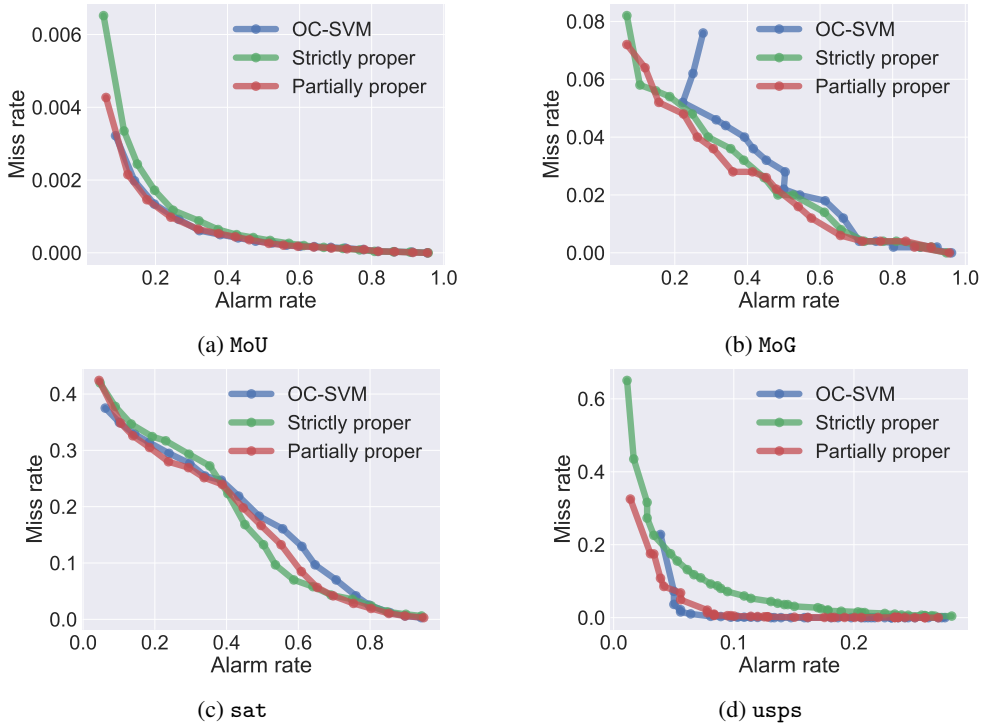
Figure 8 shows a sample of the instances with highest and lowest score. Each instance is annotated with the corresponding class label. The results are largely intuitive: most instances with low score are indeed visually anomalous.

D.3.2 Quantitative analysis

We now provide a quantitative analysis of anomaly detection performance. For several real-world datasets, we treat specific instances as drawn from an anomalous distribution P, and others as drawn from a nominal (non-anomalous) distribution μ. We provide as input to various methods a training sample of anomalous instances, and a desired quantile level $q \in (0, 1)$. The learned scorer $f \colon \mathcal{X} \to \mathbb{R}$ is applied to a testing sample comprising a mixture of anomalous and nominal instances. We evaluate the false negative and false positive rates of f for the resulting binary classification task of distinguishing the nominal (positive) from anomalous (negative) instances, i.e., $P(f(\mathsf{X}) < 0)$ and $\mu(f(\mathsf{X}) > 0)$. Following Steinwart et al. [2005], we term these the *alarm* and *miss rates* respectively. Note that the provided quantile q is meant to control the alarm rate.

We consider the following datasets:

- `MoU`. This synthetically generated data was constructed per Steinwart et al. [2005]. Here, $\mathcal{X} \subset \mathbb{R}^{10}$, with $\mathsf{X} = \mathbf{A}\mathsf{U}$ for $\mathsf{U} \sim \mathrm{Uniform}([0, 1]^{27})$, and $\mathbf{A} \in \mathbb{R}^{10 \times 27}$. For each row of $\mathbf{A}$,



(a) `MoU` (b) `MoG`

(c) `sat` (d) `usps`

Figure 9: Miss-alarm curves for various real-world datasets. Note that the scales vary across plots. Note also that the `usps` dataset has a different range of alarm rates.

we choose a uniformly random number $m \in \{2, 3, 4, 5\}$ of non-zero entries with value $\frac{1}{m}$. We draw nominal samples from $\mathrm{Uniform}([0, 1]^{10})$.

- `MoG`. This synthetically generated data was constructed per Chen et al. [2013]. Here,

$$\mathsf{X} \sim 0.2 \cdot \mathcal{N}\left(\begin{bmatrix}5\\0\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 9\end{bmatrix}\right) + 0.8 \cdot \mathcal{N}\left(\begin{bmatrix}-5\\0\end{bmatrix}, \begin{bmatrix}9 & 0\\0 & 1\end{bmatrix}\right).$$

  The reference measure μ is the uniform distribution over $[-18, +18]^2$, from which we draw nominal samples.
- `satellite`. This dataset is taken from the UCI repository. Following Chen et al. [2013], we treat the three smallest classes as anomalous, and all other classes as nominal.
- `usps`. This dataset is taken from the UCI repository. Following Schölkopf et al. [2001], we treat the digit zero as anomalous, and all other digits as nominal.

Figure 9 summarises how the test set miss rate changes as we request different alarm rates, corresponding to the quantile levels q. We see that in general, our partial density estimation approach can achieve competitive or lower miss rates than full density estimation or the OC-SVM. We emphasise that, per §6.2, our framework when instantiated with the LSIF loss is in fact closely related to the OC-SVM; thus, the similar performance of the two methods is to be expected.

E Calibrated anomaly detection and entropy

The problem of calibrated anomaly detection has a surprising relationship to the estimation of a particular entropy of the distribution P. We make use of the following result [Reid and Williamson, 2011, Theorems 9, 18]: Suppose $f\colon \mathbb{R}_{\geq 0} \to \mathbb{R}$ is a convex function normalised such that $f(1) = 0$, and P and Q be two probability distributions on $\mathcal{X}$. Let $I_f(P,Q) = \int_{\mathcal{X}} f\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}Q$ denote the Csizár f-divergence between P and Q. Let $\Delta\underline{\mathbb{L}}_\ell(\pi, P, Q)$ denote the *statistical information* (with respect to ℓ) of the binary experiment with class conditional distributions P and Q and prior probability π (i.e. the difference between the prior and posterior risk; see Reid and Williamson [2011]). Then for any given ℓ, there exists an f such that for all P, Q

$$I_f(P,Q) = \Delta\underline{\mathbb{L}}_\ell(1/2, P, Q). \tag{29}$$

The particular f, given ℓ can be expressed conveniently in terms of the corresponding weight functions: w for ℓ and γ for f, where

$$\gamma(c) = \frac{1}{c^3} \cdot f''\left(\frac{1-c}{c}\right), \quad c \in (0,1), \tag{30}$$

or equivalently

$$f(s) = \int_0^s \underbrace{\left(\int_0^t \frac{1}{(\tau+1)^3} \gamma\left(\frac{1}{\tau+1}\right) \mathrm{d}\tau\right)}_{\phi(t)} \mathrm{d}t. \tag{31}$$

The correspondence we need is simply

$$\gamma(c) = \frac{1}{16} \cdot w(1-c), \quad c \in [0,1]. \tag{32}$$

When considering the anomaly detection problem, we replace Q by μ and consider it fixed. Thus consider the function

$$H_f^\mu(P) := I_f(P,\mu), \tag{33}$$

which is known as the Csiszár f-entropy of P with respect to reference measure μ.

As explicated in Reid and Williamson [2011], the f-divergence $I_f(P,\mu)$ can be computed in a variational form:

$$I_f(P,\mu) = \sup_{\rho\colon \mathcal{X}\to\mathbb{R}} \mathop{\mathbb{E}}_{P}[\rho] - \mathop{\mathbb{E}}_{\mu}[f^* \circ \rho], \tag{34}$$

where f^* is the Legendre-Fenchel conjugate of f, and the supremum is over all measurable functions from $\mathcal{X}$ to $\mathbb{R}$.

In the main body of the paper we constructed a c_α-strictly proper loss $\bar{\ell}$ from a strictly proper loss ℓ by transforming the weight functions according to

$$\bar{w}(c) = [\![c \leq c_\alpha]\!] \cdot w(c), \quad c \in [0,1]. \tag{35}$$

If ℓ corresponds to f via its weight function γ according to (32), then the weight function for $\bar{f}$ corresponding to $\bar{\ell}$ is given by

$$\bar{\gamma}(c) = [\![c > 1 - c_\alpha]\!] \cdot \gamma(c). \tag{36}$$

Combining (36) with (31) we see that the corresponding $\bar{\phi}$ satisfies

$$\bar{\phi}(s) = \phi(s) \wedge \phi(s^*) \tag{37}$$

and thus $\bar{f}(s) = \bar{f}_1(s) + \bar{f}_2(s)$ for $s \in \mathbb{R}_{\geq 0}$, where

$$\bar{f}_1(s) = [\![s \leq s^*]\!] \cdot f(s) \tag{38}$$

$$\bar{f}_2(s) = [\![s > s^*]\!] \cdot (f'(s^*) \cdot s + f(s^*) - f'(s^*)s^*), \tag{39}$$

where $s^* = \frac{1}{1-c_\alpha} - 1$. That is $\bar{f}$ equals f on the interval $[0, s^*]$ and then affinely continues. Consequently

$$H_{\bar{f}}^\mu(P) = \int_{\mathcal{X}} f\left(\frac{\mathrm{d}P}{\mathrm{d}\mu}\right) \mathrm{d}\mu$$

$$= \int\limits_{\{x\in\mathcal{X}\,|\,\frac{\mathrm{d}P}{\mathrm{d}\mu}(x)<f'(s^*)\}} f\left(\frac{\mathrm{d}P}{\mathrm{d}\mu}\right)\mathrm{d}\mu + \int\limits_{\{x\in\mathcal{X}\,|\,\frac{\mathrm{d}P}{\mathrm{d}\mu}(x)\geq f'(s^*)\}} \left(f'(s^*)\frac{\mathrm{d}P}{\mathrm{d}\mu}(x) + f(s^*) - f'(s^*)s^*\right)\mathrm{d}\mu(x)$$

$$= \int_{\mathcal{X}} f\left(\widetilde{\frac{\mathrm{d}P}{\mathrm{d}\mu}}\right)\mathrm{d}\mu + \underbrace{f'(s^*) + f(s^*) - f'(s^*)s^*}_{\text{constant}},$$

where

$$\widetilde{\frac{\mathrm{d}P}{\mathrm{d}\mu}}(x) = \begin{cases} \frac{\mathrm{d}P}{\mathrm{d}\mu}(x), & \frac{\mathrm{d}P}{\mathrm{d}\mu}(x) < f'(s^*) \\ 0 & \text{otherwise} \end{cases}$$

is the density of the low density region only.

By properties of the Legendre-Fenchel conjugate, we have

$$\bar{f}^*(v) = \begin{cases} f^*(v) & v < f'(s^*) \\ +\infty, & v^* \geq f'(s^*) \end{cases}$$

and thus the variational representation (34) can be rewritten as

$$H^{\mu}_{\bar{f}}(P) = \sup_{\rho\colon \mathcal{X}\to(-\infty,f'(s^*)]} \mathop{\mathbb{E}}_{P}[\rho] - \mathop{\mathbb{E}}_{\mu}[f^*\circ\rho]\,. \tag{40}$$

That is, the reweighting from f to $\bar{f}$ is implemented by merely restricting the range of functions over which one optimises. The argmax in (40) corresponds via simple transformation to the Bayes optimal hypothesis in the risk minimization problem (Proposition 4).

We have thus seen that the problem of calibrated anomaly detection of P relative to μ using $\bar{\ell}$ is equivalent to the determination of the $\bar{f}$-entropy $H^{\mu}_{\bar{f}}(P)$ of P relative to μ.