Loss Functions

Robert C. Williamson

Abstract Vapnik described the "three main learning problems" of pattern recognition, regression estimation and density estimation. These are defined in terms of the loss functions used to evaluate performance (0-1 loss, squared loss and log loss respectively). But there are many other loss functions one could use. In this chapter I will summarise some recent work by myself and colleagues studying the theoretical aspects of loss functions. The results elucidate the richness of the set of loss functions and explain some of the implications of their choice.

1 Introduction

If one wishes to give a clear definition of a problem a good starting point is to define what one means by a solution to the problem. Vapnik's "three main learning problems" [25] are so defined via the loss functions used to measure the quality of their solutions. If y is an observed value and $\hat{y}$ is one's estimate, then pattern recognition is defined via the 0-1 loss

$$\ell_{0-1}(y,\hat{y}) = [\![y = \hat{y}]\!] \tag{1}$$

where $[\![p]\!] = 1$ if p is true and equals zero of p is false. Regression is defined in terms of the *squared loss*

$$\ell_{\mathrm{sq}}(y,\hat{y}) = (y-\hat{y})^2 \tag{2}$$

and probability estimation via *log loss*

$$\ell_{\log}(\hat{p}) = -\ln(\hat{p}). \tag{3}$$

Australian National University and NICTA, Canberra, ACT 0200 Australia, e-mail: Bob.Williamson@anu.edu.au

In this chapter I primarily focus on problems where the data (x_i, y_i) is drawn from $\mathscr{X} \times [n]$ where $[n] := \{1, \ldots, n\}$. In this case it is convenient to write the loss as a vector valued map $\ell\colon [n] \to \mathbb{R}^n_+$ (for pattern recognition or "classification") or $\ell\colon \Delta^n \to \mathbb{R}^n_+$ for probability estimation, where Δ^n is the n-dimensional simplex. Thus log-loss is $\ell_{\log}(\hat{p}) = (-\ln(\hat{p}_1), \ldots, -\ln(\hat{p}_n))'$ where the prime denotes transpose and $\hat{p} \in \Delta^n$ is one's estimate of the probability (that is $\hat{p}_i$ is an estimate of the probability that the random variable Y takes on value $i \in [n]$).

These three canonical loss functions that are central to Vapnik's work raise the obvious question of what other loss functions might one choose, and what are the implications for that choice. That is the subject of this chapter which provides an informal overview of some recent results I have (jointly) obtained. While I focus on my own recent work, taking loss functions seriously has been a research topic for some time [4, 11, 12, 3]. Even within the Bayesian framework where it is often claimed one is only interested in "gaining information" from an experiment, ultimately loss functions arise because in the end one will *do something* with the results of the "information" so obtained [18, 6]. Furthermore, when one acknowledges the impracticality of exact computation of Bayesian inference, loss functions do matter even for Bayesians [15].

Formal and precise statements of the various results can be found in the papers to which I refer along with detailed historical references. The emphasis of this chapter is the broader picture and intution and some technical conditions are omitted in the statement of some results.

The rest of this chapter is organised as follows. Section 2 introduces the notion of a proper loss for probability estimation and shows two key representations in terms of Bregman divergences induced by the Bayes risk associated with the loss and a Choquet representation of all possible proper losses in terms weighted combinations of elementary proper losses (0-1 loss is one of these elementary losses). Section 3 introduces the notion of an f-divergence between two distributions and shows the 1:1 correspondence between f-divergences and binary proper losses and explains some of the implications. Section 4 shows how the representations of proper losses make it much simpler to understand surrogate regret bounds (which is a measure of how much one loses from not following Vapnik's advice of solving the problem directly). Section 5 extends the results of section 2 and 3 (which are for $n = 2$) to general n and explains how one can thus define a natural f-divergences between several distributions jointly. Section 6 studies the parametrisation of losses, and in particular looks at the role of "link functions" and explains when a loss can be written as the composition of a proper loss and a link function, and explains the convexity of proper losses in terms of the corresponding Bayes risk. Section 7 summarises the implications of the choice of loss on the rate of convergence obtainable in two distinct learning settings: worst case online sequence prediction and the traditional statistical batch setting as studied by Vapnik. The chapter concludes (Section 8) with some remarks on Vladimir Vapnik's impact on the machine learning community in general, and on my scientific work over 20 years in particular and offers a suggestion of good way forward for the community to effectively build upon his legacy.

2 Proper losses and their representations

Consider the problem of class probability estimation where one receives an iid sample $\wr(x_i,y_i)\int_{i=1}^m$ of points from $\mathscr{X}\times[n]$. The goal is to estimate, for a given x, the probability $p_i=\Pr(Y=1|X=x)$, where (X,Y) are random variables drawn from the same distribution as the sample. Given a loss function $\ell\colon \Delta^n\to\mathbb{R}_+^n$ the *conditional risk* is defined via

$$L\colon \Delta^n\times\Delta^n\ni(p,q)\mapsto L(p,q)=\mathbb{E}_{Y\sim p}\ell_Y(p)=p'\cdot\ell(q)=\sum_{i=1}^n p_i\ell_i(q)\in\mathbb{R}_+.$$

A natural requirement to impose upon ℓ for this problem is that it is *proper*, which means that $L(p,p)\le L(p,q)$ for all $p,q\in\Delta^n$. (It is *strictly proper* if the inequality is strict when $p\neq q$.) The *conditional Bayes risk* $\underline{L}\colon \Delta^n\ni p\mapsto\inf_{q\in\Delta^n}L(p,q)$ and is always concave. If ℓ is proper, $\underline{L}(p)=L(p,p)=p'\cdot\ell(p)$. The full risk $\mathbb{L}(q)=\mathbb{E}_{\mathscr{X}}\mathbb{E}_{\mathscr{Y}|\mathscr{X}}\ell_Y(q(X))$.Examples of proper losses include 0-1 loss, squared loss and log loss.

There are many proper losses. A very convenient way to parametrise them arises from an integral representation. The *cost-sensitive* misclassification loss ℓ_c is a generalisation of 0-1 loss and is defined for $c\in(0,1)$ via

$$\ell_c(q)=(c[\![q>c]\!],(1-c)[\![q\le c]\!])'. \tag{4}$$

A loss $\ell\colon \Delta^2\to\mathbb{R}^+$ is proper if and only if for all $q\in[0,1]$ and $y\in\{1,2\}$

$$\ell_y(q)=\int_0^1\ell_{c,y}(q)w(c)dc, \tag{5}$$

where the *weight function* $w\colon [0,1]\to\mathbb{R}_+$ is given by $w(c)=-\underline{L}''(c)$. The weight function allows a much easier interpretation of the effect of the choice of different proper losses [4, 21]. Examples of weight functions are $w_{0/1}(c)=2\delta(c-1/2)$, $w_{\text{square}}(c)=1$, and $w_{\log}(c)=\frac{1}{c(1-c)}$.

3 Divergences and the bridge to proper losses

An *f-divergence* [5] is a measure of the closeness of two distributions on some space $\mathscr{X}$ and is defined for a convex function $f\colon \mathbb{R}_+\to\mathbb{R}$ (with $f(1)=0$) via

$$\mathbb{I}_f(P,Q)=\int_{\mathscr{X}}f\left(\frac{dP}{dQ}\right)dQ. \tag{6}$$

Examples of f-divergences include Variational, Kullback-Liebler and Hellinger.

The *statistical information* [7] $\Delta\mathbb{L}(\pi,P,Q)$ for a binary decision problem ($\mathscr{Y}=\{1,2\}$) with class conditional probability distributions $P(x)=\Pr(X=x|Y=2)$ and

$Q(x) = \Pr(X = x|Y = 1)$ and prior probability $\pi = \Pr(Y = 2)$ is given by the difference between the prior and posterior Bayes risk (the difference between the best attainable risk when only using π and the distribution on $\mathscr{X}$ and the risk obtainable when using the conditional distributions P and Q).

A key result of [21] is that for any $\pi \in (0,1)$ and any convex f (satisfying $f(1) = 0$), there exists a proper loss ℓ with associated statistical information $\Delta\underline{\mathbb{L}}$ such that for all distributions P, Q, $\mathbb{I}_f(P,Q) = \Delta\underline{\mathbb{L}}(\pi, P, Q)$ and conversely given a proper loss ℓ there exists an f such that the same equivalence holds. Thus in a precise sense, the problem of measuring the divergence between two distributions P and Q is the same as solving the prediction problem relative to some proper loss when P and Q are interpreted as the respective class-conditional distributions. There is also an integral representation for f-divergences [17], and the "weight function" there can be directly related to the weight function for the corresponding proper losses [21].

A *Bregman divergence* is defined in terms of a convex function ϕ as

$$B_\phi(p,q) = \phi(p) - \phi(q) - (p-q)' \cdot D\phi(q) \tag{7}$$

where $D\phi(q)$ is the derivative of ϕ at q. It is also well known that the conditional Bayes risk $\underline{L}$ for a proper loss satisfies $L(p,q) = \underline{L}(q) - (q-p)\underline{L}'(q)$. Consequently the *regret* $L(p,q) - \underline{L}(p) = B_{-\underline{L}}(p,q)$, the Bregman divergence between p and q induced by the convex function $-\underline{L}$. Thus there is an intimate relationship between Bayes risks, f-divergences and Bregman divergences.

4 Surrogate losses

Working with ℓ_c is computationally difficult, so one often uses a convex surrogate. The question then naturally arises: what additional loss is incurred in doing so? One way of answering this question theoretically is via a "surrogate regret bound" [1]. It turns out that by starting with proper losses, and exploiting the representations presented above, one can derive surrogate regret bounds very simply.

Suppose that for some fixed $c \in (0,1)$ we know the regret B_c of predicting with q when the true distribution is p is $B_c(p,q) = \alpha$. Then the regret $B(p,q)$ for any proper loss ℓ satisfies

$$B(p,q) \geq \max(\psi(c,\alpha), \psi(c,-\alpha)) \tag{8}$$

where $\psi(c,\alpha) = \underline{L}(c) - \underline{L}(c-\alpha) + \alpha\underline{L}'(c)$. Furthermore (8) is tight. The above bound can be inverted to obtain the usual results in the literature; see [21, 19] for details.

5 Extension to multiclass losses and divergences

Most of the above results extend to the multiclass setting ($n > 2$), although there are some differences [26, 27]. The advantage of the representation of ℓ in terms of $\underline{L}$ becomes greater when $n > 2$ as can be seen by observing in this case $\ell\colon \Delta^n \to \mathbb{R}^n_+$ whereas $\underline{L}\colon \Delta^n \to \mathbb{R}^+$, a considerable simplification. Technically in the binary case we worked with a projection of Δ^2 onto $[0,1]$. When $n > 2$ we similarly project $p \in \Delta^n$ into $\tilde{p} \in \tilde{\Delta}^n$ via $\tilde{p}_i = p_i$, $i = 1,\ldots,n-1$. The induced set $\tilde{\Delta}^n$ then has an open interior and one can thus differentiate.

Multiclass proper losses still have a Choquet integral representation analogous to (5) since the set of proper losses is convex. However in stark contrast to the case where $n = 2$, the set of primitive losses becomes much larger when $n > 2$ (they are dense in the set of all proper losses!); see [27] for details.

Strictly proper losses are quasi-convex, in the sense that if ℓ is strictly proper, then for all $p \in \Delta^n$, $q \mapsto L(p,q) = p' \cdot \ell(q)$ is quasi-convex. Two proper losses $\ell^1, \ell^2\colon \Delta^n \to \mathbb{R}^n_+$ have the same conditional Bayes risk $\underline{L}$ if and only if $\ell^1 = \ell^2$ almost everywhere. If $\underline{L}$ is differentiable then $\ell^1 = \ell^2$. The proper loss ℓ is continuous at p in the interior of Δ^n if and only if $\underline{L}$ is differentiable at p.

The bridge between proper losses and f-divergences also holds when $n > 2$, but in this case the appropriate defintion of $\mathbb{I}_f(P_{[n]}) = \mathbb{I}_f(P_1,\ldots,P_n)$ was not clear in the literature. It turns out [10] that an exact analog of the binary result holds; for every proper loss $\ell\colon \Delta^n \to \mathbb{R}^n_+$ there exists a convex f such that the induced statistical information for a multiclass experiment with class conditional distributions $P_1,\ldots,P_n$ equals the f divergence $\mathbb{I}_f(P_{[n]})$, and conversely. This multidistribution divergence (which is better understood intuitively as a joint similarity, rather than a joint "distance") satisfies the same properties as traditional (binary) f-divergences. Furthermore, all of these properties are easy to prove by utilising the bridge to loss functions and Bayes risks, and then appealing to the Blackwell-Sherman-Stein theorem.

6 Parametrisations, links and convexity

There are two key factors one needs to take into account in choosing a loss function: the statistical and computational effects. The statistical effect is controlled by the Bayes risk $\underline{L}$ (this assertion is justified more fully in the next section). The computational effect is (to a first approximation) controlled by the convexity of the loss. In this section I will outline how these two factors can be controlled quite independently using the proper composite representation. Details are in [20] (binary case) and [27] (multiclass case).

We have already seen that proper losses $\ell\colon \Delta^n \to \mathbb{R}^n_+$ are characterised by their Bayes risk. Proper losses are a suitable choice when the predictors are probabilities. Oftentimes (e.g. use of linear predictors) the predictor v may live in some other set $\mathscr{V}$, such as $\mathbb{R}^n$. In this case it can make sense to use a *proper composite loss*

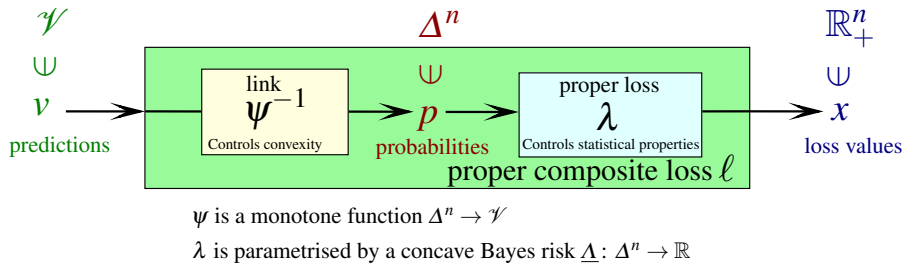

Fig. 1 The idea of a proper composite loss.

$\ell: \mathscr{V} \to \mathbb{R}^n_+$ defined via $\ell(v) = \lambda(\psi^{-1}(v))$ where $\psi: \Delta^n \to \mathscr{V}$ is an invertible *link function* and $\lambda: \Delta^n \to \mathbb{R}^n_+$ is a proper loss. Since the link preserves quasi-convexity, if ℓ has a strictly proper composite representation then ℓ is quasi-convex.

Suppose ℓ is continuous and has a proper composite representations $\ell = \lambda \circ \psi^{-1} = \mu \circ \phi^{-1}$, Then the proper loss is unique ($\lambda = \mu$ almost everywhere). If ℓ is additionally invertible then the link functions are also unique.

Given a loss ℓ, the existence of a proper composite representation for ℓ is governed by the geometry of the image $\ell(\mathscr{V})$. The precise statement is a little subtle, but roughly speaking the existence is controlled by the "Δ^n-convexity" of $\ell(\mathscr{V})$ which means that $\ell(\mathscr{V})$ should "look convex" from the perspective of supporting hyperplans with normal vector in Δ^n; see [27, Section 5.2] for details.

A prediction $v \in \mathscr{V}$ is *admissible* if there is no prediction v_1 better than v in the sense that $\ell(v_1) \leq \ell(v)$ and for some $i \in [n]$, $\ell_i(v_1) < \ell_i(v)$. If $\ell: \mathscr{V} \to \mathbb{R}^n_+$ is continuous, invertible and has a strictly proper composite representation, then for all $v \in \mathscr{V}$, v is admissible.

All continuous strictly proper losses (and strictly proper composite losses) are quasi-convex, but they are not necessarily convex. The quasi convexity means that if ℓ is continuous and has a proper composite representation, then it is *minimax*, meaning that

$$\max_{p \in \Delta^n} \min_{v \in \mathscr{V}} L(p,v) = \min_{v \in \mathscr{V}} \max_{p \in \Delta^n} L(p,v).$$

Note that ℓ need not be convex for the above to hold. The convexity of a proper (or proper composite loss) is readily checked however in terms of the Hessian of $\underline{L}$ and the gradient of ψ (see [27, Section 6.4]). Furthermore, if $\lambda: \Delta^n \to \mathbb{R}^n_+$ is a proper loss which is *not convex*, then one can canonically construct a convex composite proper loss with the same Bayes risk $\underline{\Lambda}$ by composing λ with its *canonical link* $\tilde{\psi}_\lambda(\tilde{p}) := -D\underline{\tilde{L}}(\tilde{p})'$, where $\tilde{p} = (p_1, \ldots, p_{n-1})'$ (it is necessary to use this reparametrisation of the simplex to allow the derivative D to be well defined).

In aggregate, these results on proper composite losses justify the seperation of concerns mentioned at the beginning of this section: the statistical properties are controlled by the proper loss λ (since it controls the Bayes risk); the geometric properites (convexity) of the composite loss are controlled by the link ψ.

7 Effect of losses on statistical convergence

Much of the theoretical literature which analyses the convergence of learning algorithms only looks at loss functions in a crude way. For example, a common trick is to bound the complexity of the class of functions induced by a loss function in terms of the complexity of the hypothesis class by assuming a Lipschitz property of the loss function. Unsurprisingly, such results offer little insight into the consequences of choosing particular loss functions. However in one setting (the online worst case mixture of experts setting [28]) there is a result that shows the effect on the choice of loss (see [9] and references therein for the detailed background). Roughly speaking, in this setting the learner's task is to aggregate expert predictions such that the aggregated predictor has a cumulative loss not much worse than the cumulative loss of the best expert (which is not known to the learner). There is a very precise result which bounds the additional loss that the learner can make in terms of the number of experts and a parameter β_ℓ called the mixability constant of the loss. This parameter depends only on the loss, but until recently its value was only known in certain special cases.

By exploiting the structure of proper losses (and in particular their characterisation in terms of their corresponding Bayes risk), it is possible to determine an exact general formula for β_ℓ when ℓ is continuous, smooth and strictly proper:

$$\beta_\ell = \min_{\tilde{p} \in \tilde{A}^n} \lambda_{\max} \left((H \underline{\tilde{L}}(\tilde{p}))^{-1} \cdot H \underline{\tilde{L}}_{\log}(\tilde{p}) \right),$$

where H denotes the Hessian, $\lambda_{\max}$ the maximum eigenvalue and $\underline{\tilde{L}}_{\log}$ is the Bayes risk for log loss [9]. If ℓ is suitably smooth ([27, Section 6,1]) then mixability of ℓ implies ℓ has a (strictly) proper composite representation.

Furthermore, it turns out that there is a generalisation of the notion of mixability (called stochastic mixability) [8] that in a special case reduces to mixability of ℓ, but applies in general to the standard statistical learning theory setting; stochastic mixability depends upon ℓ, $\mathscr{F}$ and P^* where $\mathscr{F}$ is the hypothesis space and P^* is the underlying distribution of the data. Analogous to the ordinary mixabulity result which characterises when fast learning can occur in the online mixture of experts setting, under some conditions, stochastic mixability of $(\ell, \mathscr{F}, P^*)$ also seems to control when learning with fast rates is possible. (I say "seems to" because the theory of stochastic mixability is still incomplete, although there are certainly cases where it can be applied and does indeed guarantee fast learning, including the special case where $\mathscr{F}$ is convex and ℓ is squared loss [16] .)

Thus we see that by taking a detour to simply understand loss functions better, one can obtain new understanding about one of the key problems in learning which Vapnik made a major contribution to — the bounding of the generalisation performance of a learning algorithm when presented with a finite amount of data.

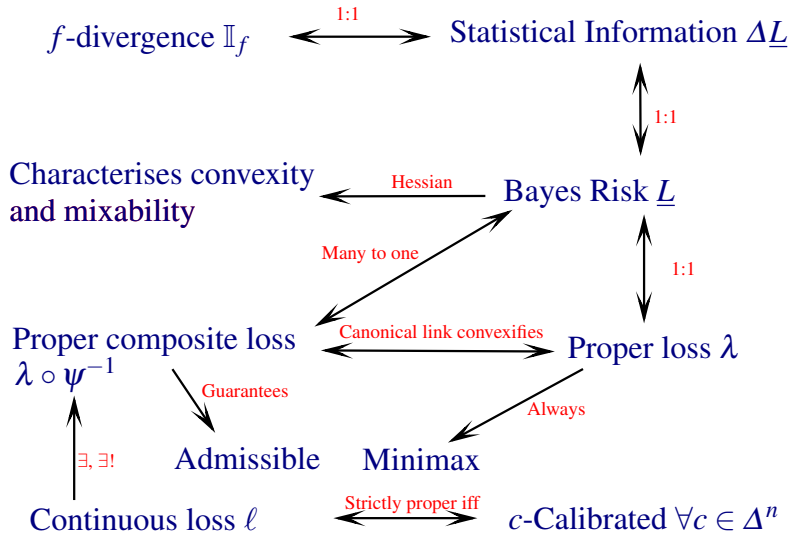

Fig. 2 Connections between key objects discussed in this chapter.

8 Conclusion

Loss functions are central to the definition of machine learning problems. As I hope the above summary shows there is a lot more to loss functions than one normally sees in the theoretical analysis of of learning problems. The results are schematised in figure 2.

Vladimir Vapnik is renowned for his theoretical work in particular the "fundamental theorem of pattern recognition" [25] (which *characterises* the difficulty of pattern recognition problems in terms of a central complexity parameter of the class of hypotheses used) and the support vector machine which vastly expanded the scope of practical machine learning by combining a well posed and mathematically sound optimisation problem as the basis of algorithms, as well as through the device of a kernel which allowed the application of these methods to very diverse data types. More generally he has influenced an entire generation of machine learning researchers in terms of how they go about doing their own research. In my case I can see the clear influence through

Characterisations The characterisation of the learnability of real valued functions in noise in terms of the finiteness of the fat-shattering dimension [2] is analogous to the above mentioned fundamental theorem.

Inductive principles Vapnik formulated the notion of an inductive principle (how to translate the end goal of learning, such as minimizing expected risk) into an empirically implementable scheme. Thinking of such principles more abstractly was one of the main motivations for the notion of "luckiness" [24] [13].

Practical algorithms and their analysis The SVM has had a widespread impact. Making it easier to tune its regularisation parameter [23] generalising the core idea to novelty detection [22], and the online setting [14], and analysing the influence of the choice of kernel on the generalisation performance [30] were all clearly motivated by Vapnik's original contributions.

What of the future? I believe the problem-centric approach that one sees in Vapnik's work will prevail in the long term. Many techniques come and go. But the core problems will remain. Thus I am convinced that a fruitful way foward is to develop *relations* between different problems [29], akin to some of those I sketched in this chapter. This is hardly a new idea in mathematics (confer the viewpoint of functional analysis). But there is a long way yet to go in doing this for machine learning. Success in doing so will help turn machine learning into a mature engineering discipline.

Acknowledgements

This work was supported by the Australian Government through the Australian Research Council and through NICTA, which is co-funded by the Department of Broadband, Communications and the Digital Economy.

References

1. Bartlett, P., Jordan, M., McAuliffe, J.: Convexity, classification, and risk bounds. Journal of the American Statistical Association **101**(473), 138–156 (2006)
2. Bartlett, P.L., Long, P.M., Williamson, R.C.: Fat-shattering and the learnability of real-valued functions. Journal of Computer and System Sciences **52**(3), 434–452 (1996)
3. Berger, J.O.: Statistical Decision Theory and Bayesian Analysis. Springer, New York (1985)
4. Buja, A., Stuetzle, W., Shen, Y.: Loss functions for binary class probability estimation and classification: Structure and applications. Tech. rep., University of Pennsylvania (2005)
5. Csiszár, I.: Information-type measures of difference of probability distributions and indirect observations. Studia Scientiarum Mathematicarum Hungarica **2**, 299–318 (1967)
6. DeGroot, M.: Uncertainty, Information, and Sequential Experiments. The Annals of Mathematical Statistics **33**(2), 404–419 (1962)
7. DeGroot, M.H.: Uncertainty, Information, and Sequential Experiments. The Annals of Mathematical Statistics **33**(2), 404–419 (1962)
8. van Erven, T., Grünwald, P., Reid, M.D., Williamson, R.C.: Mixability in statistical learning. In: Neural Information Processing Systems (2012)
9. van Erven, T., Reid, M.D., Williamson, R.C.: Mixability is Bayes risk curvature relative to log loss. Journal of Machine Learning Research **13**, 1639–1663 (2012)
10. García-García, D., Williamson, R.C.: Divergences and risks for multiclass experiments. In: Conference on Learning Theory (JMLR: W&CP), vol. 23, pp. 28.1–28.20 (2012)
11. Hand, D.: Deconstructing Statistical Questions. Journal of the Royal Statistical Society. Series A (Statistics in Society) **157**(3), 317–356 (1994)
12. Hand, D., Vinciotti, V.: Local Versus Global Models for Classification Problems: Fitting Models Where it Matters. The American Statistician **57**(2), 124–131 (2003)

13. Herbrich, R., Williamson, R.: Algorithmic Luckiness. Journal of Machine Learning Research **3**(2), 175–212 (2002)
14. Kivinen, J., Smola, A., Williamson, R.: Online learning with kernels. IEEE Transactions on Signal Processing **52**(8), 2165–2176 (2004)
15. Lacoste-Julien, S., Huszár, F., Gharamani, Z.: Approximate inference for the loss-calibrated Bayesian. In: Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (2011)
16. Lee, W., Bartlett, P., Williamson, R.: The importance of convexity in learning with squared loss. IEEE Transactions on Information Theory **44**(5), 1974–1980 (1998)
17. Liese, F., Vajda, I.: On divergences and informations in statistics and information theory. IEEE Transactions on Information Theory **52**(10), 4394–4412 (2006)
18. Lindley, D.: On a Measure of the Information Provided by an Experiment. The Annals of Mathematical Statistics **27**(4), 986–1005 (1956)
19. Reid, M.D., Williamson, R.C.: Surrogate regret bounds for proper losses. In: Proceedings of the International Conference on Machine Learning, pp. 897–904 (2009)
20. Reid, M.D., Williamson, R.C.: Composite binary losses. Journal of Machine Learning Research **11**, 2387–2422 (2010)
21. Reid, M.D., Williamson, R.C.: Information, divergence and risk for binary experiments. Journal of Machine Learning Research **12**, 731–817 (2011)
22. Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., Williamson, R.C.: Estimating the Support of a High-Dimensional Distribution. Neural Computation **13**(7), 1443–1471 (2001)
23. Schölkopf, B., Smola, A., Williamson, R.C., Bartlett, P.L.: New support vector algorithms. Neural Computation **12**, 1207–1245 (2000)
24. Shawe-Taylor, J., Bartlett, P., Williamson, R., Anthony, M.: Structural risk minimization over data-dependent hierarchies. IEEE Transactions on Information Theory **44**(5), 1926–1940 (1998)
25. Vapnik, V.N.: Statistical Learning Theory. John Wiley and Sons, New York (1998)
26. Vernet, E., Williamson, R.C., Reid, M.D.: Composite multiclass losses. In: Neural Information Processing Systems (2011)
27. Vernet, E., Williamson, R.C., Reid, M.D.: Composite multiclass losses (2012). URL http://users.cecs.anu.edu.au/∼williams/papers/P189.pdf. Submitted to *Journal of Machine Learning Research*, 42 pages.
28. Vovk, V.: A game of prediction with expert advice. In: Proceedings of the Eighth Annual Conference on Computational Learning Theory, pp. 51–60. ACM (1995)
29. Williamson, R.C.: Introduction. Introductory talk at workshop on Relations Between Machine Learning Probems (2011). URL http://videolectures.net/nipsworkshops2011_williamson_machine/. NIPS2011 workshops, Sierra Nevada
30. Williamson, R.C., Smola, A., Schölkopf, B.: Generalization performance of regularization networks and support-vector machines via entropy numbers of compact operators. IEEE Transactions on Information Theory **47**(6), 2516–2532 (2001)