Linking losses for density ratio and class-probability estimation

Aditya Krishna Menon ADITYA.MENON@DATA61.CSIRO.AU
Cheng Soon Ong CHENGSOON.ONG@ANU.EDU.AU
Data61 and the Australian National University, Canberra, ACT, Australia

Abstract

Given samples from two densities p and q, density ratio estimation (DRE) is the problem of estimating the ratio p/q. In this paper, we formally relate DRE and class-probability estimation (CPE), and theoretically justify the use of existing losses from one problem for the other. In the CPE to DRE direction, we show that essentially *any* CPE loss (e.g. logistic, exponential) minimises a Bregman divergence to the true density ratio, and thus can be used for DRE. We also show how different losses focus on accurately modelling different ranges of the density ratio, and use this to design new CPE losses for DRE. In the DRE to CPE direction, we argue that the least squares importance fitting method has potential use for bipartite ranking of instances with maximal accuracy at the head of the ranking. Our analysis relies on a novel Bregman divergence identity that may be of independent interest.

1. Density ratio estimation

Suppose we have samples from two densities p and q over an instance space $\mathcal{X}$. Density ratio estimation (DRE) is concerned with estimating from these samples the density ratio $r \colon x \mapsto p(x)/q(x)$. A canonical application of these estimates is in the covariate shift problem (Shimodaira, 2000; Sugiyama & Kawanabe, 2012), where one's training and test samples have different marginal distributions over instances: the density ratio between these distributions can be used to re-weight training instances so as to correctly adapt to the test distribution. Other applications of these estimates include outlier detection, independent component analysis, and hypothesis testing (Sugiyama et al., 2012b).

A conceptually simple solution to DRE is to compute kernel density estimates $\hat{p}$ and $\hat{q}$, and then compute the ratio $\hat{p}/\hat{q}$ (Shimodaira, 2000; Sugiyama & Müller, 2005). This suffers from the curse of dimensionality, which has motivated discriminative approaches to the problem, starting with the seminal mean matching work of Huang et al. (2007). Mean matching is however limited by the lack of a principled model selection scheme; two popular subsequent approaches affording model selection are KL importance estimation (KLIEP) (Sugiyama et al., 2008), and least squares importance fitting (LSIF) (Kanamori et al., 2009).

In this paper, we formally relate the DRE and class-probability estimation (CPE) problems, and theoretically justify the use of existing losses from one problem for the other. In the CPE to DRE direction, we first observe (Lemma 1) that KLIEP and LSIF both employ losses belonging to the *proper composite* family, which are the fundamental losses of CPE (Buja et al., 2005; Reid & Williamson, 2010). Motivated by this, we show that essentially *any* CPE loss (e.g. logistic, exponential) minimises a Bregman divergence to the true density ratio (Proposition 3), and thus is suitable for DRE; analyse how different CPE losses focus on accurately modelling different ranges of the density ratio (Lemma 4); and use this to design new CPE losses for DRE (§6.2). In the DRE to CPE direction, we argue that LSIF has potential use in bipartite ranking problems where one desires maximal accuracy at the head of the ranking (§7). Our analysis relies on a novel Bregman identity (Lemma 2) that may be of independent interest.

The basic link between DRE and CPE is not new. Specific CPE methods such as logistic regression have been previously employed for covariate shift adaptation (Bickel et al., 2009, Section 7); the link between the problems has been exploited for semiparametric density estimation, where the form of the density ratio is known (Qin, 1998; Cheng & Chu, 2004); and a variant of LSIF has been applied to the task of class-probability estimation (Sugiyama et al., 2010; Sugiyama, 2010). However, on the DRE side, we are not aware of a formal analysis of the quality of the density ratio estimates produced by a general CPE method in the sense of Proposition 3. On the CPE side, we are unaware of any analysis of the tradeoffs implicit in the LSIF loss, nor any discussion of its potential value for ranking problems.

Proceedings of the 33^{rd} *International Conference on Machine Learning*, New York, NY, USA, 2016. JMLR: W&CP volume 48. Copyright 2016 by the author(s).

2. Background and notation

We review some background material and fix notation.

2.1. Learning from binary labels

Denote by $\mathcal{D}$ a distribution over $\mathcal{X} \times \{\pm 1\}$, with random variables $(\mathsf{X}, \mathsf{Y}) \sim \mathcal{D}$. Any $\mathcal{D}$ may be decomposed into class-conditionals $(P, Q) = (\mathbb{P}(\mathsf{X} \mid \mathsf{Y} = 1), \mathbb{P}(\mathsf{X} \mid \mathsf{Y} = -1))$ and base rate $\pi = \mathbb{P}(\mathsf{Y} = 1)$, or into marginal $M = \mathbb{P}(\mathsf{X})$ and class-probability function $\eta \colon x \mapsto \mathbb{P}(\mathsf{Y} = 1 \mid \mathsf{X} = x)$. We will write $\mathcal{D} = (P, Q, \pi)$ or $\mathcal{D} = (M, \eta)$. The densities of P, Q are assumed to exist and denoted by p, q.

A *loss* is any $\ell \colon \{\pm 1\} \times \mathbb{R} \to \mathbb{R}$; we interchangeably write $\ell(y, \cdot)$ as $\ell_y(\cdot)$. A *scorer* is any $s \colon \mathcal{X} \to \mathbb{R}$. The ℓ-risk for a scorer s wrt $\mathcal{D}$ is $\mathbb{L}(s; \mathcal{D}, \ell) \doteq \mathbb{E}_{(\mathsf{X},\mathsf{Y}) \sim \mathcal{D}} \left[\ell(\mathsf{Y}, s(\mathsf{X})) \right]$; the Bayes-optimal scorer is $s^* \doteq \operatorname{argmin}_s \mathbb{L}(s; \mathcal{D}, \ell)$.

2.2. Class-probability estimation

Class-probability estimation (CPE) is concerned with inferring η. This may be achieved via a suitable loss function. A loss ℓ is *strictly proper composite* with (invertible) *link function* $\Psi \colon [0, 1] \to \mathbb{R}$ if the Bayes-optimal scorer for the ℓ-risk is $s^* = \Psi \circ \eta$ (Buja et al., 2005; Reid & Williamson, 2010). Examples include the logistic loss $\ell(y, v) = \log(1 + e^{-yv})$ with $\Psi(p) = \log p/(1 - p)$, exponential loss $\ell(y, v) = e^{-yv}$ (as in AdaBoost) and square hinge loss $\ell(y, v) = \max(0, 1 - yv)^2$ (as in L2-SVMs). If ℓ is differentiable, it is strictly proper composite with invertible link Ψ iff (Reid & Williamson, 2010, Corollary 12)

$$\Psi^{-1}(v) = \left(1 - \ell_1'(v)/\ell_{-1}'(v)\right)^{-1}. \tag{1}$$

Given a strictly proper composite ℓ, we call $\lambda \doteq \ell \circ \Psi \colon [0, 1] \to \mathbb{R}$ the underlying *proper loss* for ℓ. The negative conditional Bayes risk $f \colon [0, 1] \to \mathbb{R}$ of ℓ is then

$$f(u) \doteq -u \cdot \lambda_1(u) - (1 - u) \cdot \lambda_{-1}(u). \tag{2}$$

Given a scorer s with low ℓ-risk, we can regard $\hat{\eta} \doteq \Psi^{-1} \circ s$ to be an estimate of η. The quality of this estimate can be quantified: the *regret* or excess risk of a scorer s over the Bayes-optimal is (Reid & Williamson, 2010, Corollary 9, Corollary 13)

$$\begin{aligned} \operatorname{reg}(s; \mathcal{D}, \ell) &\doteq \mathbb{L}(s; \mathcal{D}, \ell) - \mathbb{L}(\Psi \circ \eta; \mathcal{D}, \ell) \\ &= \mathbb{E}_{\mathsf{X} \sim M} \left[B_f(\eta(\mathsf{X}), \hat{\eta}(\mathsf{X})) \right] \qquad (3) \\ &= \mathbb{E}_{\mathsf{X} \sim M} \left[\int_0^1 w(c) \cdot \operatorname{reg}_c(\eta(\mathsf{X}), \hat{\eta}(\mathsf{X}))\, dc \right], \qquad (4) \end{aligned}$$

where B_f is a Bregman divergence with generator f, the *weight function* $w = f''$, and $\operatorname{reg}_c(\eta, \hat{\eta}) \doteq |\eta - c| \cdot [\![(\eta - c) \cdot (\hat{\eta} - c) < 0]\!]$ the *cost-sensitive* pointwise regret. Intuitively, Equation 4 says that a loss focusses on accurately modelling the range of η values for which $w(\cdot)$ is large. For example, logistic loss has as f the negative Shannon entropy, and so seeks an $\hat{\eta}$ with minimal KL-divergence to η; further, it has weight function $w(c) = (c \cdot (1 - c))^{-1}$.

2.3. Covariate shift adaptation

In covariate shift problems (Sugiyama & Kawanabe, 2012), we have train and test distributions $\mathcal{D}_{\mathrm{Tr}}$ and $\mathcal{D}_{\mathrm{Te}}$ with $\eta_{\mathrm{Tr}} = \eta_{\mathrm{Te}}$, but $M_{\mathrm{Tr}} \neq M_{\mathrm{Te}}$. Our goal remains to minimise $\mathbb{L}(s; \mathcal{D}_{\mathrm{Te}}, \ell)$. This is a canonical application for DRE because the importance weighting identity

$$\mathbb{L}(s; \mathcal{D}_{\mathrm{Te}}, \ell) = \mathbb{E}_{\mathsf{X} \sim M} \left[r(\mathsf{X}) \cdot \mathbb{E}_{\mathsf{Y} \sim \eta(\mathsf{X})} \left[\ell(\mathsf{Y}, s(\mathsf{X})) \right] \right] \tag{5}$$

for density ratio $r = m_{\mathrm{Te}}/m_{\mathrm{Tr}}$ of the corresponding marginal densities implies that if we estimate r, we can simply re-weight training instances accordingly so as to adapt to the test distribution (Shimodaira, 2000). (On finite samples, importance weighting may actually bring little or *negative* improvement, and thus covariate shift adaptation may require more than simply estimating r (Cortes et al., 2010; Reddi et al., 2015; Swaminathan & Joachims, 2015); this is however beyond the scope of the present paper.)

3. Linking density ratios to class-probabilities

Suppose we wish to estimate a density ratio $r = p/q$, given samples from the respective densities. One can view these samples as arising from a distribution $\mathcal{D}$ over binary labels, where p and q are the densities of the class-conditional distributions, and π is the relative frequency of samples from the densities. (For the covariate shift problem of §2.3, such a $\mathcal{D}$ would encode the discrimination between training and testing instances with $p = m_{\mathrm{Te}}$ and $q = m_{\mathrm{Tr}}$.)

The above suggests that methods for learning from binary labels could be used to estimate r. Indeed, as in Bickel et al. (2009); Smola (2010), Bayes' rule implies:

$$(\forall x \in \mathcal{X})\; \frac{\eta(x)}{1 - \eta(x)} = \frac{p(x)}{q(x)} \cdot \frac{\pi}{1 - \pi}, \tag{6}$$

where $\mathbb{P}(\mathsf{X} = x)$ cancels for both terms in the LHS; thus,

$$(\forall x \in \mathcal{X})\; r(x) \doteq \frac{p(x)}{q(x)} = \Psi_{\mathrm{dr}}(\eta(x)), \tag{7}$$

for the link function

$$\Psi_{\mathrm{dr}}(u) \doteq \frac{1 - \pi}{\pi} \cdot \frac{u}{1 - u}. \tag{8}$$

Intuitively, this elementary fact suggests one should be able to uses CPE losses to perform DRE, and vice-versa. The rest of the paper makes this intuition precise. We devote the next few sections to the usage of CPE losses for DRE, and then return to the usage of DRE losses for CPE in §7.

4. DRE via CPE loss minimisation

We now show how Equation 7 may be used to re-interpret existing DRE approaches as implicitly performing CPE, as well as motivate performing DRE via a general CPE loss. In what follows, we assume that $\pi = 1/2$ [1], so that

$$\Psi_{\mathrm{dr}}(u) = \frac{u}{1-u}, \tag{9}$$

with $\Psi_{\mathrm{dr}}^{-1}(v) = v/(1+v)$ for $v \geq 0$.

4.1. Estimating density ratios with the Ψ_{dr} link

To estimate r, Equation 7 suggests that we minimise the ℓ-risk for proper composite ℓ with link Ψ_{dr}, as the corresponding Bayes-optimal scorer is exactly $s^* = \Psi_{\mathrm{dr}} \circ \eta = r$. Given a scorer s with low ℓ-risk, we can then treat $\hat{r} = s$ as an estimate of the density ratio. We now see how two existing approaches to DRE do precisely this.

The KLIEP loss. Consider the loss

$$\ell_{-1}(v) = a \cdot v \text{ and } \ell_1(v) = -\log v \tag{10}$$

for $a > 0$, with corresponding risk

$$\mathbb{L}(s; \mathcal{D}, \ell) = \mathbb{E}_{\mathsf{X} \sim P}\left[-\log s(\mathsf{X})\right] + a \cdot \mathbb{E}_{\mathsf{X} \sim Q}\left[s(\mathsf{X})\right] \tag{11}$$

defined for any $s \in \mathcal{S} \subseteq \mathbb{R}_+^{\mathcal{X}}$. For suitable a, finding $\min_{s \in \mathcal{S}} \mathbb{L}(s; \mathcal{D}, \ell)$ is equivalent to the constrained problem

$$\min_{s \in \mathcal{S}} \mathbb{E}_{\mathsf{X} \sim P}\left[-\log s(\mathsf{X})\right] : \mathbb{E}_{\mathsf{X} \sim Q}\left[s(\mathsf{X})\right] = 1,$$

which is exactly the objective of the KLIEP method of Sugiyama et al. (2008); we thus call the loss of Equation 10 the *KLIEP loss*. The unconstrained objective was also considered in du Plessis & Sugiyama (2012).

The LSIF loss. Consider the loss

$$\ell_{-1}(v) = {}^{1}\!/\!{}_{2} \cdot v^2 \text{ and } \ell_1(v) = -v, \tag{12}$$

with corresponding risk

$$\mathbb{L}(s; \mathcal{D}, \ell) = \mathbb{E}_{\mathsf{X} \sim P}\left[-s(\mathsf{X})\right] + \mathbb{E}_{\mathsf{X} \sim Q}\left[{}^{1}\!/\!{}_{2} \cdot s(\mathsf{X})^2\right]. \tag{13}$$

The problem $\min_s \mathbb{L}(s; \mathcal{D}, \ell)$ is exactly that considered by the LSIF method of Kanamori et al. (2009); we thus call the loss of Equation 12 the *LSIF loss*. One appealing property of the LSIF loss is that when working with linear scorers $s \colon x \mapsto \langle w, x \rangle$, the risk has a closed form minimiser

$$w^* = \left(\mathbb{E}_{\mathsf{X} \sim Q}\left[\mathsf{X}\mathsf{X}^T\right]\right)^{-1} \cdot \mathbb{E}_{\mathsf{X} \sim P}\left[\mathsf{X}\right]. \tag{14}$$

To analyse the above, consider the family of power losses suggested in Sugiyama et al. (2012a, Section 3.4),

$$\ell_{-1}(v) = \frac{v^{1+\alpha} - 1}{1+\alpha} \text{ and } \ell_1(v) = \frac{1 - v^\alpha}{\alpha} \tag{15}$$

for $\alpha \in (0, 1]$. As $\alpha \to 0^+$, we get the (translated) KLIEP loss for $a = 1$; for $\alpha = 1$, we get the (translated) LSIF loss. It is easy to check that all these losses are strictly proper composite, with link Ψ_{dr}.

Lemma 1. *For any $\alpha \in (0, 1]$, the power loss of Equation 15 is strictly proper composite with link Ψ_{dr}. At the limit $\alpha \to 0^+$, the KLIEP loss with parameter $a > 0$ of Equation 10 is also strictly proper composite with link $a^{-1} \cdot \Psi_{dr}$.*

Proof. Since ℓ is differentiable, with

$$\ell'_{-1}(v) = v^\alpha \text{ and } \ell'_1(v) = -v^{\alpha - 1},$$

by Equation 1 it is strictly proper composite with link

$$\Psi^{-1}(v) = \left(1 + \frac{1}{v}\right)^{-1} = \Psi_{\mathrm{dr}}^{-1}(v).$$

A similar argument applies to the general KLIEP loss. □

The above interpretation of the KLIEP and LSIF losses in terms of CPE evinces why they are suitable for DRE: the optimal scorer for each is exactly the true density ratio.

4.2. Estimating density ratios with *any* link

While the previous section focussed on DRE using the link Ψ_{dr}, nothing prohibits the use of an arbitrary link Ψ. Suppose we have a proper composite ℓ with link Ψ, and Bayes-optimal scorer $s^* = \Psi \circ \eta$. Then, by Equation 7,

$$(\forall x \in \mathcal{X})\, r(x) = \frac{1-\pi}{\pi} \cdot \frac{\Psi^{-1}(s^*(x))}{1 - \Psi^{-1}(s^*(x))}. \tag{16}$$

Given an arbitrary scorer s with low ℓ-risk, it is natural then to use the density ratio estimator

$$\hat{r}(x) \doteq \frac{1-\pi}{\pi} \cdot \frac{\hat{\eta}(x)}{1 - \hat{\eta}(x)}, \tag{17}$$

recalling that $\hat{\eta} = \Psi^{-1} \circ s$. For example, with the logistic loss, for which $\Psi^{-1}(v) = (1 + e^{-v})^{-1}$, we have $\hat{r}(x) = e^{s(x)}$ when $\pi = 1/2$. Precisely such an estimate was previously considered in Bickel et al. (2009, Section 7) in the context of covariate shift adaptation.

The above appears to definitively establish the suitability of CPE losses for DRE. However, while Equation 16 justifies the CPE loss minimisation approach asymptotically, what can be guaranteed about the quality of an *imperfect* estimate $\hat{r}$ as in Equation 17? This issue is more subtle, and is the subject of our next section.

5. A Bregman minimisation view of DRE

Recall from Equation 3 that proper composite loss minimisation is equivalent to minimising a specific Bregman divergence between $\hat{\eta}$ and η. Since r and η are related by

[1] Appendix A covers the more cumbersome case $\pi \neq 1/2$.

a monotone transform (Equation 7), in applications where only the ordering of the density ratio is important, this fact would suffice as guarantee of the quality of $\hat{r}$. In covariate shift modelling, however, one requires a good estimate of r, rather than some monotone transformation of r (other than multipication by a positive scalar). For this application, Equation 3 by itself does not suffice; further, *prima facie*, one might be concerned that errors in the estimate $\hat{\eta}$ are magnified uncontrollably when passed through the transform $u/(1-u)$ to construct $\hat{r}$.

Fortunately, we can show that minimisation of (essentially) *any* strictly proper composite ℓ results in a good $\hat{r}$ in a precise sense: the procedure is equivalent to minimising a specific Bregman divergence of $\hat{r}$ to the true r.

5.1. A novel Bregman identity

To quantify the quality of the density ratio estimates $\hat{r}$, the most natural way to proceed is to re-express Equation 3 in terms of r rather than η. But how do we do this without appealing to specific properties of f or Ψ? The answer is provided by the following Bregman identity, which to our knowledge is novel, and may be of independent interest.

Lemma 2. *For any twice differentiable convex $f\colon [0,1] \to \mathbb{R}$ with Bregman divergence $B_f(\cdot,\cdot)$,*

$$(\forall x, y \in [0,\infty))\, B_f\left(\frac{x}{1+x}, \frac{y}{1+y}\right) = \frac{1}{1+x} \cdot B_{f^{\circledast}}(x,y),$$

where $f^{\circledast}\colon [0,\infty) \to \mathbb{R}$ is given by

$$f^{\circledast}\colon z \mapsto (1+z)\cdot f\left(\frac{z}{1+z}\right) \tag{18}$$

Proof. By (Reid & Williamson, 2009, Equation 12),

$$B_f(x,y) = \int_y^x (x-z)\cdot f''(z)\,dz. \tag{19}$$

Applying this to the LHS,

$$B_f\left(\frac{x}{1+x}, \frac{y}{1+y}\right) = \int_{\frac{y}{1+y}}^{\frac{x}{1+x}} \left(\frac{x}{1+x} - z\right)\cdot f''(z)\,dz.$$

Employing the substitution $z = \frac{u}{1+u}$, with $dz = \frac{du}{(1+u)^2}$,

$$\begin{aligned}
\text{LHS} &= \int_y^x \left(\frac{x}{1+x} - \frac{u}{1+u}\right)\cdot f''\left(\frac{u}{1+u}\right)\cdot \frac{1}{(1+u)^2}\,du \\
&= \int_y^x \frac{x-u}{(1+x)\cdot(1+u)}\cdot f''\left(\frac{u}{1+u}\right)\cdot\frac{1}{(1+u)^2}\,du \\
&= \frac{1}{1+x}\cdot\int_y^x (x-u)\cdot f''\left(\frac{u}{1+u}\right)\cdot\frac{1}{(1+u)^3}\,du \\
&= \frac{1}{1+x}\cdot B_{f^{\circledast}}(x,y),
\end{aligned}$$

where the last line is since by definition of $f^{\circledast}$,

$$(f^{\circledast})'(z) = \frac{1}{1+z}\cdot f'\left(\frac{z}{1+z}\right) + f\left(\frac{z}{1+z}\right)$$

and

$$(f^{\circledast})''(z) = f''\left(\frac{z}{1+z}\right)\cdot\frac{1}{(1+z)^3}. \tag{20}$$

□

Remark 1. One might think to generalise Lemma 2 using a different change of variable in the integral above; however, this in general will not yield another Bregman divergence.

Remark 2. $f^{\circledast}$ is closely related to the perspective transform $f^{\diamond}\colon x \mapsto x\cdot f(1/x)$ of a convex function.

Remark 3. The somewhat awkward form of the arguments in the LHS is to simplify the results in the next section.

5.2. Proper losses minimise a Bregman divergence to r

Using Lemma 2, we can establish that proper composite loss minimisation is equivalent to minimising a Bregman divergence to the true density ratio.

Proposition 3. *Pick any strictly proper composite ℓ with twice differentiable negative Bayes risk f. Then, for any distribution $\mathcal{D} = (P, Q, 1/2)$ and scorer $s\colon \mathcal{X} \to \mathbb{R}$,*

$$\text{reg}(s;\mathcal{D},\ell) = 1/2\cdot\mathbb{E}_{\mathsf{X}\sim Q}\left[B_{f^{\circledast}}\left(r(\mathsf{X}), \hat{r}(\mathsf{X})\right)\right],$$

for $r = \Psi_{dr}\circ\eta$, $\hat{r} = \Psi_{dr}\circ\hat{\eta}$, and $f^{\circledast}$ per Equation 18.

Proof. Letting $R = 2\cdot\text{reg}(s;\mathcal{D},\ell)$, by Equation 3,

$$\begin{aligned}
R &= 2\cdot\mathbb{E}_{\mathsf{X}\sim M}\left[B_f(\eta(\mathsf{X}),\hat{\eta}(\mathsf{X}))\right] \\
&= \mathbb{E}_{\mathsf{X}\sim P}\left[B_f(\eta(\mathsf{X}),\hat{\eta}(\mathsf{X}))\right] + \mathbb{E}_{\mathsf{X}\sim Q}\left[B_f(\eta(\mathsf{X}),\hat{\eta}(\mathsf{X}))\right] \\
&= \mathbb{E}_{\mathsf{X}\sim Q}\left[(1+r(\mathsf{X}))\cdot B_f\left(\eta(\mathsf{X}),\hat{\eta}(\mathsf{X})\right)\right],
\end{aligned}$$

where the last line is because $\mathbb{E}_{\mathsf{X}\sim P}\left[g(\mathsf{X})\right] = \mathbb{E}_{\mathsf{X}\sim Q}\left[r(\mathsf{X})\cdot g(\mathsf{X})\right]$. Now, expressing Lemma 2 as

$$(1+x)\cdot B_f\left(\Psi_{\text{dr}}^{-1}(a), \Psi_{\text{dr}}^{-1}(b)\right) = B_{f^{\circledast}}(a,b),$$

and noting that by Equation 7, $\eta = \Psi_{\text{dr}}^{-1}\circ r$, the result follows by picking $a = r(\mathsf{X})$ and $b = \hat{r}(\mathsf{X})$. □

Remark 4. See Appendix B for an alternate proof using the connection of proper losses to f-divergences (Reid & Williamson, 2011), and Appendix D for a discussion of how the proper loss view sheds some light on existing interpretations of KLIEP and LSIF in terms of f-divergences.

Remark 5. The expectation in Equation 3 is over the marginal M, but above, it is over the class-conditional Q. This is intuitive for covariate shift adaptation (§2.3): here, we estimate the class-conditional density ratio for $\mathcal{D} = (M_{\text{Te}}, M_{\text{Tr}}, 1/2)$, and wish this to be accurate on average for the reweighed training instances i.e. on average under the "negative" class-conditional M_{Tr}.

Remark 6. The need for f'' to exist is why we used the qualifier "essentially" when describing the result in §1.

Remark 7. See Appendix C for the simple, but slightly more cumbersome generalisation to $\pi \neq 1/2$.

Proposition 3 has implicit precedent in three special cases: Sugiyama et al. (2012a, Section 3) showed that logistic regression, KLIEP and LSIF perform Bregman minimisation. Proposition 3 shows that this is simply a manifestation of the fact that they all use proper composite losses, and broadly generalises the result to other ℓ.

Proposition 3 has at least three useful implications. First, it theoretically justifies the reduction of DRE to CPE, as in e.g. Bickel et al. (2009). Second, it opens the door to performing DRE using any other of the standard arsenal of CPE losses (e.g. exponential, square hinge). Third, we can leverage existing analyses for CPE to help us *design* suitable losses for a DRE task. This last point is important: while Proposition 3 implies all proper composite losses are "equally good" for DRE if we have sufficiently many samples and a rich function class, in practice neither of these conditions is expected to hold. It is thus of interest to determine what tradeoffs are imposed by different losses. This is studied in the next section.

6. Designing CPE losses for DRE

We now show that different CPE losses focus on modelling different regions of the density ratio, as specified by an implicit *weight function* whose form we provide. We then discuss how to design CPE losses so as to employ a fixed such weighting, and provide some new losses for DRE.

6.1. A weight function view of losses for DRE

Recall from Equation 4 that every proper composite ℓ has an associated weight function w over cost ratios, so that w specifies the tradeoffs in modelling η implied by a given loss. We can similarly interpret each ℓ as focussing on different ranges of the density ratio r, as specified by a *density ratio weight function* $w_{\mathrm{DR}}(\rho)$ defined below.

Lemma 4. *Pick any strictly proper composite ℓ with twice differentiable f and weight function w. For any scorer s and distribution $\mathcal{D} = (P, Q, 1/2)$,*

$$\mathrm{reg}(s; \mathcal{D}, \ell) = \frac{1}{2} \cdot \mathbb{E}_{\mathsf{X} \sim Q} \left[\int_0^\infty w_{\mathrm{DR}}(\rho) \cdot \mathrm{reg}_\rho(r(\mathsf{X}), \hat{r}(\mathsf{X}))\, d\rho \right],$$

where $\hat{r}$ is as per Equation 17, the density ratio weight function $w_{\mathrm{DR}} \colon [0, \infty) \to \mathbb{R}_+$ is

$$w_{\mathrm{DR}}(\rho) \doteq \frac{1}{(1+\rho)^3} \cdot w\left(\frac{\rho}{1+\rho}\right), \tag{21}$$

and the pointwise regret around threshold ρ is

$$\mathrm{reg}_\rho(r, \hat{r}) \doteq |r - \rho| \cdot [\![(r - \rho) \cdot (\hat{r} - \rho) < 0]\!]. \tag{22}$$

Proof of Lemma 4. By Proposition 3 and Equation 19,

$$\mathrm{reg}(s; \mathcal{D}, \ell) = \frac{1}{2} \cdot \mathbb{E}_{\mathsf{X} \sim Q} \left[\int_0^\infty (f^\circledast)''(\rho) \cdot \mathrm{reg}_\rho(r(\mathsf{X}), \hat{r}(\mathsf{X}))\, d\rho \right].$$

Defining $w_{\mathrm{DR}} = (f^\circledast)''$ and applying Equation 20 (recalling that $w = f''$), the result follows. □

Thus, analogous to our intuition for estimating η, minimising a proper composite loss intuitively focusses on the range of r values for which the corresponding $w_{\mathrm{DR}}(\cdot)$ is large. The relationship between w_{DR} and w (Equation 21) features a non-obvious dependence on $(1+\rho)^{-3}$.

Table 1 summarises the weight functions over cost and density ratios for the DRE losses of §4.1, and for some standard proper composite losses. The latter are seen to place more importance on accurate modelling of smaller density ratios. The power family has a similar trend for $\alpha < 1$, but at $\alpha = 1$ the LSIF loss has uniform weighting over all possible values of the density ratio. Also of interest is that only the LSIF and square losses have $w_{\mathrm{DR}}(0) < +\infty$.

Which of these tradeoffs is the most suitable for DRE? A uniform weighting can be motivated by appeal to generalisation bounds: Cortes et al. (2008) showed that ensuring small expected ℓ_2 distance between the true density ratio r and one's estimate $\hat{r}$ yields guarantees on the excess error from using $\hat{r}$ instead of r as weights to a kernel-based learner. This speaks in favour of the LSIF loss, which from Table 1 employs a uniform weighting. However, it is not the *only* such loss with this property; we now see how one can pair a range of link functions Ψ with a given weight w_{DR}, to generate new proper composite losses for DRE.

6.2. New proper composite losses for DRE

Suppose we fix a weight function w_{DR} we believe suitable for DRE, and desire a proper composite loss that employs this weight. By definition, specifying w_{DR} equally specifies the weight w over cost ratios. This, in turn, specifies the negative Bayes-risk f for any loss employing this weight (up to a linear term). Recall however that the negative Bayes-risk f only specifies the underlying proper loss λ: we still have complete flexibility in choosing a link function Ψ to generate a proper composite loss.

To pick a suitable Ψ, one desiderata is that the resulting composite loss should be convex for ease of optimisation. For a fixed weight w, there is a simple *canonical link function* Ψ_{can} that ensures this, being any link that satisfies

$$\Psi'_{\mathrm{can}}(u) \doteq w(u). \tag{23}$$

The corresponding proper composite loss is (Buja et al., 2005, Section 16), (Reid & Williamson, 2010, Section 6.1)

$$\ell'_{-1}(v) = \Psi^{-1}_{\mathrm{can}}(v) \text{ and } \ell'_1(v) = \Psi^{-1}_{\mathrm{can}}(v) - 1, \tag{24}$$

Loss	$\ell_{-1}(v)$	$\ell_1(v)$	$\Psi^{-1}(v)$	$w(c)$	$w_{\mathrm{DR}}(\rho)$
KLIEP	v	$-\log v$	$\frac{v}{1+v}$	$\frac{1}{c\cdot(1-c)^2}$	$\frac{1}{\rho}$
LSIF	$\frac{1}{2}v^2$	$-v$	$\frac{v}{1+v}$	$\frac{1}{(1-c)^3}$	1
Power$_\alpha$	$\frac{v^{1+\alpha}-1}{1+\alpha}$	$\frac{1-v^\alpha}{\alpha}$	$\frac{v}{1+v}$	$\frac{1}{c^{1-\alpha}\cdot(1-c)^{2+\alpha}}$	$\rho^{\alpha-1}$
Square	$(1+v)^2$	$(1-v)^2$	$2v-1$	8	$\frac{1}{(1+\rho)^3}$
Logistic	$\log(1+e^v)$	$\log(1+e^{-v})$	$\frac{1}{1+e^{-v}}$	$\frac{1}{c\cdot(1-c)}$	$\frac{1}{\rho\cdot(1+\rho)}$
Exponential	e^v	e^{-v}	$\frac{1}{1+e^{-2v}}$	$\frac{1}{2\cdot c^{3/2}\cdot(1-c)^{3/2}}$	$\frac{1}{\rho^{3/2}}$

Table 1: Weights over cost and density ratios for existing DRE (top panel) and CPE (bottom panel) losses; derivations in Appendix E.

which is guaranteed to be convex, with derivative bounded in $[-1, 1]$ i.e. the loss is asymptotically (sub-)linear.

Using the above, we can generate new losses with weights matching those of existing ones. Let us return to the weight function $w(c) = (1-c)^{-3}$ corresponding to $w_{\mathrm{DR}}(\rho) = 1$. The canonical link found by solving Equation 23 is

$$\Psi_{\text{can}}(u) = (1/2)\cdot(1/(1-u)^2 - 1), \tag{25}$$

where we chose the constant of integration such that $\Psi_{\text{can}}(u) \in [0, \infty)$. Applying Equation 24 yields the new "canonical LSIF" loss

$$\ell_{-1}(v) = v - \sqrt{2v+1} \text{ and } \ell_1(v) = -\sqrt{2v+1} \tag{26}$$

with the constraint $v \geq 0$. Compared to the LSIF loss, this loss also has a uniform weighting over density ratios, while having derivatives bounded in $[-1, 1]$.

More generally, recall that LSIF is a special case of the power family (Equation 15) with $\alpha = 0$. Sugiyama et al. (2012a) observed that choosing $\alpha > 1$ may yield a loss that is more robust to outliers (at the possible expense of statistical efficiency), as per the origin of these losses in a different context (Basu et al., 1998). However, the presented losses are only convex for $\alpha \in (0, 1]$. Similar to the above, one can seek to retain the weight w of the power losses, but vary the link to design a convex loss. The canonical link in this case does not possess an analytic inverse, but we can still design convex losses using other links. Appendix F shows that for $\alpha = 2$, $\Psi^{-1}(v) = \sqrt{2v}/(1+\sqrt{2v})$ yields

$$\ell_{-1}(v) = (2\sqrt{2}/3)\cdot v^{3/2} \text{ and } \ell_1(v) = -v \text{ for } v \geq 0. \tag{27}$$

7. A new application of existing DRE losses

The preceding sections have established the virtue of using CPE losses to perform DRE. One might equally wonder whether existing DRE losses are useful in tasks where conventional CPE losses (e.g. logistic regression) are employed. We now explore one such potentially fruitful application: the problem of *bipartite ranking* (Agarwal & Niyogi, 2005). Here, we assume there is some $\mathcal{D}$ over instances and binary labels, and our goal is to find a scorer $s\colon \mathcal{X} \to \mathbb{R}$ that ranks instances well, in the sense of possessing a high area under the ROC curve (AUC). As the Bayes-optimal scorer for the AUC is any monotone transform of η (Clmenon et al., 2008), class-probability estimators such as logistic regression may be successfully employed for this task (Kotlowski et al., 2011; Agarwal, 2014).

In practice, one typically wishes to maximise accuracy at the *head* of the ranked list. This is known as the "ranking the best" (RTB) regime (Clémençon & Vayatis, 2007). While standard bipartite ranking methods are of course viable for RTB, one might hope to do better by explicitly targeting instances x with $\eta(x) \sim 1$. For example, Rudin (2009) proposed a family of pairwise ranking risks that explicitly penalise false-positive errors at the head of the ranked list. This was shown in Ertekin & Rudin (2011) to have the same minimiser as the *p-classification loss*,

$$\ell_{-1}(v) = (1/p)\cdot e^{vp} \text{ and } \ell_1(v) = e^{-v}, \tag{28}$$

where $p \gg 0$ emphasises accurate modelling of the head of the list. In fact, from Equation 1, this loss is easily checked to be strictly proper composite with weight function $w(c) = (c^{1+\alpha}\cdot(1-c)^{2-\alpha})^{-1}$, where $\alpha = 1/(p+1)$ (Menon & Williamson, 2014). Thus, setting $p \gg 0$ yields a CPE loss whose underlying weight emphasises $c \sim 1$, and hence seeks to accurately model η values that are ~ 1.

Interestingly, the weight functions of the KLIEP and LSIF losses (Table 1) also place emphasis on $c \sim 1$; indeed, the KLIEP weight is the limit of the p-classification weight as $p \to \infty$. (The two losses employ rather different link functions, which explains their different forms.) This suggests that these DRE losses may be useful for RTB problems; to our knowledge, this application of these losses has not been previously explored. One strong appeal of the LSIF loss in particular is its closed form solution (Equation 14), which affords highly efficient tuning and training; we believe this motivates the use of this loss as a baseline for RTB tasks.

8. Experimental results

We present experiments[2] evincing three aspects of our analysis: first, that a loss' weight function $w_{\rm DR}(\rho)$ dictates the range of density ratio values it focusses on; second, that existing proper losses are viable for DRE in the context of covariate shift adaptation; third; that the new application of the LSIF loss to "ranking the best" problems holds promise.

8.1. Weight functions and resulting tradeoffs

We study the impact of weight functions using an example of Vinciotti & Hand (2003) that was used in Buja et al. (2005) to illustrate the role of weight functions in CPE. Here, distribution $\mathcal{D}$ has marginal uniform over $[0,1]^2$, and

$$\eta\colon (x_1, x_2) \mapsto (2/\pi)\cdot\cos^{-1}(x_1/\sqrt{x_1^2+x_2^2}).$$

To estimate of the resulting density ratio of the class-conditionals, we consider losses with "step" weights

$$w(c) = [\![c \in [a,b]]\!] + [\![c \notin [a,b]]\!] \cdot h^{-1}, \qquad (29)$$

where a, b, h are tuning parameters. Evidently, the role of a and b is to specify a region of cost ratio which is to be accurately modelled. The role of the parameter h is to determine how much emphasis is placed on this range of cost ratios versus others. The corresponding weight over density ratios similarly emphasises density ratio values in $\left[\frac{a}{1-a}, \frac{b}{1-b}\right]$. We pick three losses from this family: we fix $h = 100$, and choose a, b such that

$$\frac{a}{1-a} = r^* - 0.2 \text{ and } \frac{b}{1-b} = r^* + 0.2 \qquad (30)$$

for $r^* = \{0.5, 1, 2\}$. The weights thus focus on ratios around the corresponding r^* value. The canonical proper composite losses for these weights are easy but tedious to derive; see Appendix G for details.

For each of the three losses above, we learn a linear model to discriminate between $n_p = 50,000$ samples S' from p and $n_q = 50,000$ samples S from q. For the resulting density ratio estimate $\hat{r}$ and fixed $\rho \in [0,\infty)$, we compute the expected pointwise regret around $r(x) = \rho$, $\mathbb{E}_{\mathsf{X}\sim\bar{\mathsf{S}}}\left[\mathrm{reg}_\rho(r(\mathsf{X}), \hat{r}(\mathsf{X}))\right]$, where reg_ρ is as per Equation 22 and $\bar{\mathsf{S}}$ is a fresh sample of $10,000$ points drawn from q. We plot these expected regrets for $\rho \in \{0.01, 0.02, \ldots, 4.0\}$.

Figure 1 shows the regret curves for the three losses, labelled "Step$_{r^*}$" for the corresponding r^* values. In each case, there is a pronounced dip in the pointwise regret curve around the corresponding r^* value, indicating accurate modelling of the density ratios in the surrounding region. This is in keeping with Lemma 4, and illustrates that $w_{\rm DR}(\rho)$ provides some intuition as to the range of the density ratio focussed on by a loss.

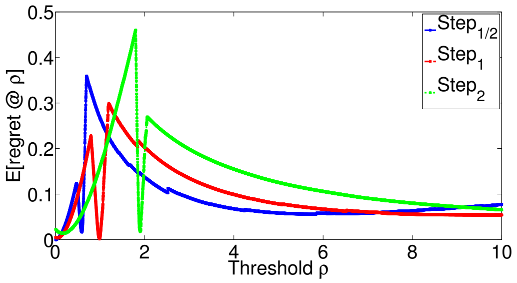

Figure 1: Expected pointwise regret for losses with "step" weights (Equation 29), for three choices of parameters as per Equation 30.

8.2. Proper losses for covariate shift adaptation

We next study the viability of DRE using proper composite losses for covariate shift adaptation. Note that logistic loss has been convincingly demonstrated to be viable in Bickel et al. (2009), and we have seen that the KLIEP and LSIF losses are in fact proper composite; our aim is simply to confirm that *other* standard CPE losses are also viable.

We work with two datasets, both comprising $n_{\rm Src}$ labelled samples $\mathsf{S} = \{(x_i, y_i)\}$ from a source domain, $n_{\rm Tar}$ unlabelled samples $\mathsf{S}' = \{x'_j\}$, $n_{\rm Eval}$ labelled samples $\mathsf{S}'' = \{(x''_k, y''_k)\}$ from a target domain. The marginal distributions for the two domains are assumed different. To correct for the covariate shift, following Equation 5, we estimate the density ratio $\hat{r}$ from $(\mathsf{S}, \mathsf{S}')$ using different approaches, and then solve a weighted least squares problem on S:

$$\min_w \sum_{i=1}^{n_{\rm Src}} \hat{r}_i \cdot (y_i - \langle w, \Phi(x_i)\rangle)^2 + \frac{\lambda_{\rm WLS}}{2}||w||^2,$$

where $\hat{r}_i$ denotes the density estimate for the ith training example, and Φ denotes the feature mapping. Finally, we evaluate prediction performance of this model on S''.

The first dataset (`poly`) follows the example from Shimodaira (2000); Huang et al. (2007). Here, we have instance space $\mathcal{X} = \mathbb{R}$, and conditional label distribution $\mathsf{Y} \mid \mathsf{X} = x \sim \mathcal{N}(-x - x^3, 0.3^2)$. The marginals for the source and target distributions are $M_{\rm Src} = \mathcal{N}(0.5, 0.5^2)$ and $M_{\rm Tar} = \mathcal{N}(0, 0.3^2)$. We set $n_{\rm Src} = 200$, $n_{\rm Tar} = 200$, and $n_{\rm Eval} = 2000$. We use the empirical kernel mapping $\Phi(x) = (\exp(-||x - z||^2))_{z\in\mathsf{S}'}$.

The second dataset (`amazon`) is the real-world Amazon review data from Blitzer et al. (2007); we used the processed data as provided by Chen et al. (2012). Here, we have instance space $\mathcal{X} = \mathbb{R}^{30000}$, being the bag of words representation of Amazon reviews from four different product domains. Each review is endowed with a binary label denoting the sentiment of the review. We train on $n_{\rm Src} = 3000$ samples from the book domain, and test on $n_{\rm Tar} = 3000$, $n_{\rm Eval} = 2000$ samples from the electronics domain. Our feature mapping is the SVD projection to 100 dimensions

[2] Scripts available at first author's webpage.

(a) poly

Loss	NMSE
Uniform	1.2723 ± 0.0302
KLIEP	0.6916 ± 0.0136
LSIF	0.7742 ± 0.0217
uLSIF	0.7038 ± 0.0102
Canonical LSIF	0.7969 ± 0.0288

Loss	NMSE
Power_2	1.1098 ± 0.0408
Logistic	0.6888 ± 0.0108
Square Hinge	0.6995 ± 0.0116
Square	0.6755 ± 0.0060
Exponential	0.6740 ± 0.0064

(b) amazon

Loss	PD
Uniform	0.1582 ± 0.0018
KLIEP	0.1500 ± 0.0018
LSIF	0.1500 ± 0.0019
uLSIF	0.1370 ± 0.0024
Canonical LSIF	0.1517 ± 0.0018

Loss	PD
Power_2	0.1538 ± 0.0018
Logistic	0.1321 ± 0.0021
Square Hinge	0.1567 ± 0.0099
Square	0.1310 ± 0.0023
Exponential	0.2021 ± 0.0048

Table 2: Covariate shift results. For poly, we report normalised MSE (NMSE), *viz.* MSE normalised by variance of the targets. For amazon, we report pairwise disagreement (PD), *viz.* $1-$ AUC. Reported are mean and standard error over 25 trials; lower is better.

(a) german

Loss	AP	PTop
Logistic	0.6087 ± 0.0192	0.0224 ± 0.0083
p-class	0.6121 ± 0.0185	0.0316 ± 0.0084
LSIF	0.6101 ± 0.0196	0.0364 ± 0.0114

(b) magic

Loss	AP	PTop
Logistic	0.8867 ± 0.0018	0.0018 ± 0.0005
p-class	0.8962 ± 0.0017	0.0031 ± 0.0017
LSIF	0.8996 ± 0.0014	0.0095 ± 0.0038

(c) news20-forsale

Loss	AP	PTop
Logistic	0.1487 ± 0.0041	0.0003 ± 0.0003
p-class	0.2817 ± 0.0113	0.0054 ± 0.0018
LSIF	0.2351 ± 0.0106	0.0101 ± 0.0014

Table 3: "Ranking the best" results. We report the average precision (AP), and fraction of positives ranked higher than the first negative (PTop) (Agarwal, 2011). Reported are mean and standard error across 10 random splits; higher is better.

of a TFIDF mapping of the features.

To estimate the density ratio, we use the existing KLIEP and LSIF losses; the "canonical LSIF" loss of Equation 26; the Power_2 loss of Equation 27; and the logistic, exponential, square, and square hinge losses. For each loss, we find

$$\min_{\theta\in\Theta} \frac{1}{n_{\mathrm{Src}}}\sum_{x\in\mathsf{S}} \ell_1(\langle\theta,\Phi(x)\rangle) + \frac{1}{n_{\mathrm{Tar}}}\sum_{x'\in\mathsf{S}'} \ell_1(\langle\theta,\Phi(x')\rangle) + \frac{\lambda_{\mathrm{DR}}}{2}||\theta||_2^2,$$

where Θ is unconstrained for the "standard" CPE losses, and $\Theta = \{\theta \mid (\forall z \in \mathsf{S}\cup\mathsf{S}')\, \langle\theta,\Phi(z)\rangle \geq 0\}$ otherwise. For LSIF, we additionally used a heuristic suggested by Kanamori et al. (2009): we use an unconstrained Θ but *post-hoc* truncate scores at 0. We call this method "uLSIF".

Table 2 summarises results for all losses over 25 random draws of S and S' on both datasets, for $\lambda_{\mathrm{WLS}} = 10^{-6}$ and $\lambda_{\mathrm{DR}} = 10^{-4}$. Nearly all losses are seen to offer significant improvement over assuming a uniform density ratio. Square loss performs well in both tasks; as it has a closed-form solution (like LSIF), it seems worth considering as a simple baseline for covariate shift problems.

8.3. LSIF loss for "ranking the best"

Finally, we consider the viability of the computationally efficient LSIF loss for the "ranking the best" problem, comparing its performance to logistic regression and the p-classification loss (Equation 28) for $p = 4$. We compare these losses on several standard benchmark datasets with binary labels. Each dataset was split in the ratio 2:1, with all instances normalised to lie in the ℓ_2 ball. A regularised linear model trained to score instances, where for each split, we performed 5-fold cross-validation to tune the strength of regularisation from $\lambda \in \{2^{-20}, 2^{-19}, \ldots, 2^{15}\}$.

Table 3 summarises the results over 10 random train–test splits. (See Appendix H for results with more datasets and performance measures.) We find that LSIF consistently performs better than the logistic loss, in keeping with our analysis of the weight function for this loss in §7. LSIF is also competitive with the p-classification loss, sometimes outperforming it on the challenging PTop measure. These results, coupled with the loss' closed form solution, suggest it is worth considering as a baseline for RTB problems.

9. Conclusion

We have shown how existing approaches to discriminative DRE implictly employ CPE losses; how general CPE losses may be employed for DRE, since minimising *any* CPE loss equivalently minimises a Bregman divergence to the true density ratio; and how a specific DRE loss, LSIF (Kanamori et al., 2009), is worth considering for ranking tasks where interest is in accuracy at the head of the list.

Possible directions for future study include more carefully studying the application of LSIF to RTB problems, comparing it to e.g. the recent approach of Li et al. (2014); studying the effects of finite samples on CPE estimates for DRE; and generalising Lemma 2 to show CPE can provably estimate other useful quantities.

Acknowledgements

This work was undertaken at NICTA. NICTA was funded by the Australian Government through the Department of Communications and the Australian Research Council through the ICT Centre of Excellence Program.

References

Agarwal, Shivani. The infinite push: A new support vector ranking algorithm that directly optimizes accuracy at the absolute top of the list. In *SIAM International Conference on Data Mining (SDM)*, pp. 839–850, 2011.

Agarwal, Shivani. Surrogate regret bounds for bipartite ranking via strongly proper losses. *Journal of Machine Learning Research*, 15:1653–1674, 2014.

Agarwal, Shivani and Niyogi, Partha. Stability and generalization of bipartite ranking algorithms. In *Conference on Learning Theory (COLT)*, pp. 32–47. Springer-Verlag, 2005.

Basu, Ayanendranath, Harris, Ian. R, Hjort, Nils L., and Jones, M. C. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 3(85): 549559, 1998.

Bickel, Steffen, Brückner, Michael, and Scheffer, Tobias. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10:2137–2155, December 2009. ISSN 1532-4435.

Blitzer, John, Dredze, Mark, and Pereira, Fernando. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Annual Meeting-Association for Computational Linguistics*, 2007.

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

Chen, Minmin, Weinberger, Kilian Q, and Sha, Fei. Marginalized denoising autoencoders for domain adaptation. In *International Conference on Machine Learning (ICML)*, pp. 767774, 2012.

Cheng, K.F. and Chu, C.K. Semiparametric density estimation under a two-sample density ratio model. *Bernoulli*, 10(4):583–604, 08 2004.

Clémençon, Stéphan and Vayatis, Nicolas. Ranking the best instances. *Journal of Machine Learning Research*, 8:2671–2699, Dec 2007.

Clmenon, Stphan, Lugosi, Gbor, and Vayatis, Nicolas. Ranking and empirical minimization of u-statistics. *The Annals of Statistics*, 36(2):844874, Apr 2008.

Cortes, Corinna, Mohri, Mehryar, Riley, Michael, and Rostamizadeh, Afshin. Sample selection bias correction theory. In *Algorithmic Learning Theory*, volume 5254 of *Lecture Notes in Computer Science*, pp. 38–53. Springer Berlin Heidelberg, 2008. ISBN 978-3-540-87986-2.

Cortes, Corinna, Mansour, Y, and Mohri, M. Learning bounds for importance weighting. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 1–9, 2010.

du Plessis, Marthinus Christoffel and Sugiyama, Masashi. Semi-supervised learning of class balance under class-prior change by distribution matching. In *International Conference on Machine Learning (ICML)*, 2012.

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

Huang, Jiayuan, Smola, Alexander J., Gretton, Arthur, Borgwardt, Karsten M., and Schölkopf, Bernhard. Correcting sample selection bias by unlabeled data. *Advances in Neural Information Processing Systems (NIPS)*, 19:601, 2007.

Kanamori, Takafumi, Hido, Shohei, and Sugiyama, Masashi. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10: 1391–1445, December 2009. ISSN 1532-4435.

Kotlowski, Wojciech, Dembczynski, Krzysztof, and Hüllermeier, Eyke. Bipartite ranking through minimization of univariate loss. In *International Conference on Machine Learning (ICML)*, 2011.

Li, Nan, Jin, R, and Zhou, ZH. Top rank optimization in linear time. *Advances in Neural Information Processing Systems*, pp. 1–9, 2014.

Menon, Aditya Krishna and Williamson, Robert C. Bayes-optimal scorers for bipartite ranking. In *Conference on Learning Theory (COLT)*, pp. 68–106, 2014.

Nguyen, Xuanlong, Wainwright, Martin J., and Jordan, Michael I. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):58475861, 2010. ISSN 00189448.

Qin, Jing. Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, 85(3): 619–630, 1998.

Reddi, Sashank Jakkam, Póczos, Barnabás, and Smola, Alexander J. Doubly robust covariate shift correction. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA.*, pp. 2949–2955, 2015.

Reid, Mark D and Williamson, Robert C. Surrogate regret bounds for proper losses. In *International Conference on Machine Learning (ICML)*, pp. 897–904. ACM, 2009.

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

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

Rockafellar, R T. *Convex Analysis*. Princeton University Press, 1972.

Rudin, Cynthia. The p-norm push: A simple convex ranking algorithm that concentrates at the top of the list. *Journal of Machine Learning Research*, 10:2233–2271, Dec 2009.

Shimodaira, Hidetoshi. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90 (2):227 – 244, 2000. ISSN 0378-3758.

Smola, Alex. Real simple covariate shift correction. http://blog.smola.org/post/4110255196/real-simple-covariate-shift-correction , 2010.

Sugiyama, Masashi. Superfast-trainable multi-class probabilistic classifier by least-squares posterior fitting. *IEICE Transactions*, 93-D(10):2690–2701, 2010.

Sugiyama, Masashi and Kawanabe, Motoaki. *Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation*. The MIT Press, 2012.

Sugiyama, Masashi and Müller, Klaus-Robert. Input-dependent estimation of generalization error under covariate shift. *Statistics & Risk Modeling*, 23(4/2005): 249–279, April 2005.

Sugiyama, Masashi, Suzuki, Taiji, Nakajima, Shinichi, Kashima, Hisashi, von Bünau, Paul, and Kawanabe, Motoaki. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60(4):699–746, 2008. ISSN 0020-3157.

Sugiyama, Masashi, Takeuchi, Ichiro, Suzuki, Taiji, Kanamori, Takafumi, Hachiya, Hirotaka, and Okanohara, Daisuke. Conditional density estimation via least-squares density ratio estimation. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010.

Sugiyama, Masashi, Suzuki, Taiji, and Kanamori, Takafumi. Density-ratio matching under the Bregman divergence: a unified framework of density-ratio estimation. *Annals of the Institute of Statistical Mathematics*, 64(5): 1009–1044, 2012a. ISSN 0020-3157.

Sugiyama, Masashi, Suzuki, Taiji, and Kanamori, Takafumi. *Density Ratio Estimation in Machine Learning*. Cambridge University Press, New York, NY, USA, 1st edition, 2012b.

Swaminathan, Adith and Joachims, Thorsten. The self-normalized estimator for counterfactual learning. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 28*, pp. 3213–3221. Curran Associates, Inc., 2015.

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

Supplementary material for “Linking losses for density ratio and class-probability estimation”

A. The link $\Psi_{\mathbf{dr}}$ when $\pi \neq \frac{1}{2}$

Our discussion in §4.1 assumed that $\pi = \frac{1}{2}$, as the KLIEP and LSIF risks (Equations 11, 13) would otherwise require scaling each of the expectations. But in general, we expect $\pi \neq \frac{1}{2}$. Of course, since the role of π is just to scale the link Ψ_{dr} by a constant, one may legitimately ignore its impact on the Bayes-optimal scorer. But for completeness, the general case may be analysed as follows. A common way of simulating balanced classes is by weighting the loss by the inverse of the class proportions, i.e. constructing

$$\ell_{\text{bal},-1}(v) = (1-\pi)^{-1} \cdot \ell_{-1}(v) \text{ and } \ell_{\text{bal},1}(v) = \pi^{-1} \cdot \ell_1(v).$$

Note that this has risk

$$\mathbb{L}(s; \mathcal{D}, \ell_{\text{bal}}) = \mathbb{E}_{\mathsf{X} \sim P}\left[\ell_1(s(\mathsf{X}))\right] + \mathbb{E}_{\mathsf{X} \sim Q}\left[\ell_{-1}(s(\mathsf{X}))\right].$$

It is easy to check that ℓ_{bal} is also proper composite. This is a consequence of the following elementary fact.

Lemma 5. *Suppose ℓ is differentiable and strictly proper composite with link Ψ. Then, for any $a, b \in \mathbb{R} - \{0\}$, the loss*

$$\tilde{\ell}_{-1}(v) = a \cdot \ell_{-1}(v) \text{ and } \tilde{\ell}_1(v) = b \cdot \ell_1(v)$$

is also strictly proper composite with inverse link

$$\tilde{\Psi}^{-1}(v) = (f_{a,b} \circ \Psi^{-1})(v)$$

for

$$f_{a,b}(z) = \frac{z}{\left(1 - \frac{b}{a}\right) \cdot z + \frac{b}{a}}.$$

Proof. By Equation 1,

$$\frac{\ell_1'(v)}{\ell_{-1}'(v)} = \frac{\Psi^{-1}(v) - 1}{\Psi^{-1}(v)}.$$

We have

$$\tilde{\ell}_{-1}'(v) = a \cdot \ell_{-1}'(v) \text{ and } \tilde{\ell}_1'(v) = b \cdot \ell_1'(v).$$

We can form

$$\begin{aligned}
\tilde{\Psi}^{-1}(v) &= \frac{1}{1 - \frac{b}{a} \cdot \frac{\ell_1'(v)}{\ell_{-1}'(v)}} \\
&= \frac{1}{1 + \frac{b}{a} \cdot \frac{1 - \Psi^{-1}(v)}{\Psi^{-1}(v)}} \\
&= \frac{\Psi^{-1}(v)}{\Psi^{-1}(v) + \frac{b}{a} \cdot (1 - \Psi^{-1}(v))} \\
&= f_{ab}(\Psi^{-1}(v)),
\end{aligned}$$

which is invertible, thus implying that $\tilde{\ell}$ is strictly proper composite with link $\tilde{\Psi}$. □

It is easy to check that

$$f_{a,b}^{-1}(u) = \frac{\frac{b}{a} \cdot u}{1 + \left(\frac{b}{a} - 1\right) \cdot u},$$

so that the link for $\tilde{\ell}$ is

$$\tilde{\Psi}(u) = \Psi(f_{a,b}^{-1}(u)).$$

Now suppose that a loss employs link Ψ_{dr}, as per Equation 9. Then, its corresponding $\tilde{\ell}$ employs the link

$$\begin{aligned}\tilde{\Psi}(u) &= \Psi_{\mathrm{dr}}(f_{a,b}^{-1}(u)) \\ &= \frac{f_{a,b}^{-1}(u)}{1 - f_{a,b}^{-1}(u)} \\ &= \frac{b}{a} \cdot \frac{u}{1-u}.\end{aligned}$$

For the balanced loss with $a = (1-\pi)^{-1}$ and $b = \pi^{-1}$,

$$\Psi_{\mathrm{bal}}(u) = \frac{1-\pi}{\pi} \cdot \frac{u}{1-u},$$

which is exactly the general Ψ_{dr} of Equation 8. Thus, if we minimise ℓ_{bal}, we will have Bayes-optimal scorer exactly the density ratio r. We can view the objectives of both KLIEP and LSIF as doing precisely this, even for general $\pi \neq \frac{1}{2}$.

B. An alternate proof of Proposition 3

Fix some concave differentiable $\underline{\mathrm{L}} \colon [0,1] \to \mathbb{R}$; this will serve as a conditional Bayes risk for the CPE loss

$$\lambda_{-1}(p) = \underline{\mathrm{L}}(p) - p \cdot \underline{\mathrm{L}}'(p) \text{ and } \lambda_1(p) = \underline{\mathrm{L}}(p) + (1-p) \cdot \underline{\mathrm{L}}'(p),$$

which is guaranteed to be proper by Reid & Williamson (2010, Theorem 7). (Unlike the losses in the body, we have $\lambda \colon \{\pm 1\} \times [0,1] \to \mathbb{R}_+$.) Note that any other differentiable proper loss $\tilde{\lambda}$ with same conditional Bayes risk must differ from λ by a linear term; this is because the two losses will have identical weight functions, and so must have identical derivatives by Reid & Williamson (2010, Theorem 1).

The risk for an estimator $\hat{\eta} \colon \mathcal{X} \to [0,1]$ under λ is

$$\begin{aligned} 2 \cdot \mathbb{L}(\hat{\eta}; \mathcal{D}, \lambda) &= \mathbb{E}_{\mathsf{X} \sim P}\left[\lambda_1(\hat{\eta}(\mathsf{X}))\right] + \mathbb{E}_{\mathsf{X} \sim Q}\left[\lambda_{-1}(\hat{\eta}(\mathsf{X}))\right] \\ &= \mathbb{E}_{\mathsf{X} \sim Q}\left[r(\mathsf{X}) \cdot \lambda_1(\hat{\eta}(\mathsf{X})) + \lambda_{-1}(\hat{\eta}(\mathsf{X}))\right] \end{aligned}$$

where the second line uses an importance reweighting for the expectation with respect to Q.

Let $g(z) = -(1+z) \cdot \underline{\mathrm{L}}(\frac{z}{1+z})$. (Evidently, $g = f^{\circledast}$, where $f = -\underline{\mathrm{L}}$.) Now, $g'(z) = \frac{1}{1+z} \cdot -\underline{\mathrm{L}}'(\frac{z}{1+z}) - \underline{\mathrm{L}}(\frac{z}{1+z})$. So,

$$\begin{aligned} g'\left(\frac{\hat{\eta}}{1-\hat{\eta}}\right) &= -(1-\hat{\eta}) \cdot \underline{\mathrm{L}}'(\hat{\eta}) - \underline{\mathrm{L}}(\hat{\eta}) \\ &= -\lambda_1(\hat{\eta}), \end{aligned}$$

and similarly

$$\frac{\hat{\eta}}{1-\hat{\eta}} \cdot g'\left(\frac{\hat{\eta}}{1-\hat{\eta}}\right) - g\left(\frac{\hat{\eta}}{1-\hat{\eta}}\right) = \lambda_{-1}(\hat{\eta}).$$

Further, the Bayes risk for the loss is (Reid & Williamson, 2011, Theorem 9)

$$\begin{aligned} 2 \cdot \mathbb{L}^*(\mathcal{D}, \lambda) &\doteq 2 \cdot \min_{\hat{\eta} \colon \mathcal{X} \to [0,1]} \mathbb{L}^*(\hat{\eta}; \mathcal{D}, \lambda) \\ &= -I_g(P, Q) \\ &= \mathbb{E}_{\mathsf{X} \sim Q}\left[-g(r(\mathsf{X}))\right], \end{aligned}$$

where $I_g(\cdot, \cdot)$ denotes the f-divergence with generator g. Thus,

$$2 \cdot \mathbb{L}(\hat{\eta}; \mathcal{D}, \lambda) = \mathbb{E}_{\mathsf{X} \sim Q}\left[-r(\mathsf{X}) \cdot g'(r(\mathsf{X})) + \hat{r}(\mathsf{X}) \cdot g'\left(\hat{r}(\mathsf{X})\right) - g(\hat{r}(\mathsf{X}))\right],$$

where $\hat{r}(x) = \frac{\hat{\eta}(x)}{1-\hat{\eta}(x)}$. This implies the regret is

$$\begin{aligned} 2 \cdot \mathrm{reg}(\hat{\eta}; \mathcal{D}, \lambda) &= 2 \cdot \mathbb{L}(\hat{\eta}; \mathcal{D}, \lambda) - 2 \cdot \mathbb{L}^*(\mathcal{D}, \lambda) \\ &= \mathbb{E}_{\mathsf{X} \sim Q}\left[-r(\mathsf{X}) \cdot g'(r(\mathsf{X})) + \hat{r}(\mathsf{X}) \cdot g'\left(\hat{r}(\mathsf{X})\right) - g(\hat{r}(\mathsf{X})) + g(r(\mathsf{X}))\right] \\ &= \mathbb{E}_{\mathsf{X} \sim Q}\left[B_g(r(\mathsf{X}), \hat{r}(\mathsf{X}))\right]. \end{aligned}$$

Note now that for any link Ψ and resulting proper composite ℓ, we have $\mathbb{L}(\hat{\eta}; D, \lambda) = \mathbb{L}(s; \mathcal{D}, \ell)$ where $\hat{\eta} = \Psi^{-1} \circ s$, and in particular $\mathrm{reg}(\hat{\eta}; \mathcal{D}, \lambda) = \mathrm{reg}(s; \mathcal{D}, \ell)$. Thus, the statement of the Proposition holds.

C. Proposition 3 when $\pi \neq \frac{1}{2}$

We now show how to generalise the analysis of Proposition 3 when $\pi \neq \frac{1}{2}$. Let

$$\tilde{\ell}_{-1}(v) = 2 \cdot (1 - \pi) \cdot \ell_{-1}(v) \text{ and } \tilde{\ell}_1(v) = 2 \cdot \pi \cdot \ell_1(v).$$

Then, we have the trivial risk equivalence

$$\mathbb{L}(s; (P, Q, \pi), \ell) = \mathbb{L}(s; (P, Q, {}^1\!/\!{}_2), \tilde{\ell}).$$

and so

$$\mathrm{reg}(s; (P, Q, \pi), \ell) = \mathrm{reg}(s; (P, Q, {}^1\!/\!{}_2), \tilde{\ell}).$$

By Lemma 5, the loss $\tilde{\ell}$ is strictly proper composite. So, we can just apply the original statement of Proposition 3 to the right hand side: we get

$$\mathrm{reg}(s; (P, Q, \pi), \ell) = \frac{1}{2} \cdot \mathbb{E}_{\mathsf{X} \sim Q} \left[B_{f_\pi^{\circledast}}(r(\mathsf{X}), \hat{r}(\mathsf{X})) \right],$$

where f_π is the negative Bayes risk of $\tilde{\ell}$. Note that, unlike in the original Proposition 3, the precise divergence being used varies with π. This is somewhat awkward, and hence we favour the presentation of $\pi = \frac{1}{2}$ in the body. Note also that the above may be used to generalise Lemma 4, where again the weight will vary with π.

D. On the f-divergence estimation view of density ratio estimation

Previous work (e.g. (Sugiyama et al., 2012a)) has explicated the relationship between density ratio estimation and the estimation of a suitable f-divergence between the underlying distributions. Recall that for convex ϕ, $\phi(v) = \sup_y yv - \phi^*(y)$, where ϕ^* denotes the convex conjugate of ϕ. Thus, an f-divergence with convex generator ϕ is expressible as (Nguyen et al., 2010)

$$\begin{aligned} I_\phi(P, Q) &= \mathbb{E}_{\mathsf{X} \sim Q}\left[\phi\left(\frac{p(\mathsf{X})}{q(\mathsf{X})}\right)\right] \\ &= \mathbb{E}_{\mathsf{X} \sim Q}\left[\sup_{s \in \mathbb{R}}\left(\frac{p(\mathsf{X})}{q(\mathsf{X})} \cdot s - \phi^*(s)\right)\right] \\ &= \sup_{s \colon \mathcal{X} \to \mathbb{R}} \mathbb{E}_{\mathsf{X} \sim P}[s(\mathsf{X})] - \mathbb{E}_{\mathsf{X} \sim Q}[\phi^*(s(\mathsf{X}))] \\ &= -\inf_{s \colon \mathcal{X} \to \mathbb{R}} \mathbb{E}_{\mathsf{X} \sim P}[-s(\mathsf{X})] + \mathbb{E}_{\mathsf{X} \sim Q}[\phi^*(s(\mathsf{X}))] \\ &= -2 \cdot \inf_{s \colon \mathcal{X} \to \mathbb{R}} \mathbb{E}_{(\mathsf{X},\mathsf{Y}) \sim \mathcal{D}}[\ell(\mathsf{Y}, s(\mathsf{X}))], \end{aligned} \tag{31}$$

for a distribution $\mathcal{D} = (P, Q, \frac{1}{2})$, and a loss ℓ defined by

$$\ell_{-1}(v) = \phi^*(v) \text{ and } \ell_1(v) = -v. \tag{32}$$

For ϕ strictly convex, ϕ^* is differentiable (Rockafellar, 1972, Theorem 26.3), and ϕ' invertible. Thus, the loss of Equation 32 is proper composite with inverse link

$$\begin{aligned} \Psi^{-1}(v) &= \frac{1}{1 - \frac{\ell_1'(v)}{\ell_{-1}'(v)}} \\ &= \frac{(\phi^*)'(v)}{(\phi^*)'(v) + 1} \\ &= \frac{(\phi')^{-1}(v)}{(\phi')^{-1}(v) + 1}. \end{aligned}$$

We thus have

$$\Psi(p) = \phi'\left(\frac{p}{1-p}\right).$$

Consider the Pearson divergence, with $\phi(v) = \frac{1}{2}v^2$. Then, $\Psi = \Psi_{\text{dr}}$ as per Equation 9. Thus, the problem of estimating the Pearson divergence in this manner implicitly involves computing the density ratio p/q for the class-conditionals.

While the above established that the loss $(\phi^*(v), -v)$ is one way to estimate an f-divergence, there is in fact an infinite family of losses that will achieve the same task. All such losses simply modify the underlying link function that is employed. Formally, we can understand Equation 32 in terms of a proper loss λ, defined by

$$\lambda_y(p) = \ell_y(\Psi(p)) = \ell_y\left(\phi'\left(\frac{p}{1-p}\right)\right).$$

In particular,

$$\lambda_{-1}(p) = \phi^*\left(\phi'\left(\frac{p}{1-p}\right)\right) \text{ and } \lambda_1(p) = -\phi'\left(\frac{p}{1-p}\right).$$

Thus, Equation 32 is simply a manifestation of the fact that[3] for such a λ,

$$I_\phi(P, Q) = -2 \cdot \inf_{\hat{\eta} \colon \mathcal{X} \to [0,1]} \mathbb{E}_{(\mathsf{X},\mathsf{Y}) \sim \mathcal{D}}[\lambda(\mathsf{Y}, \hat{\eta}(\mathsf{X}))],$$

[3] This is just a restatement of Reid & Williamson (2011, Theorem 9), which is presented in terms of the underlying Bayes risk for the proper loss. Note that for the given λ, we have conditional Bayes risk $\underline{L}(\eta) = -(1-\eta) \cdot \left(\frac{\eta}{1-\eta} \cdot \phi'(\frac{\eta}{1-\eta}) - \phi^*\left(\phi'\left(\frac{\eta}{1-\eta}\right)\right)\right)$, which by definition of conjugacy is $\underline{L}(\eta) = -(1-\eta) \cdot \phi\left(\frac{\eta}{1-\eta}\right)$, as per Reid & Williamson (2011).

where we now see that we have arbitrary flexibility in terms of the link function Ψ we employ to construct a proper composite loss. In particular, given any ϕ, we can compute the proper loss λ, and then compose it with Ψ_{dr} to get a proper composite loss whose minimisation gives us a density ratio. But we can equally well use some other link function, in which case we will estimate the divergence, but will not directly estimate the density ratio.

E. Weight functions for the family of power losses

See e.g. Reid & Williamson (2010) for the weights and link functions for the standard proper composite losses. For the power loss with parameter $\alpha \in \mathbb{R}_+$,

$$\ell_{-1}(v) = \frac{v^{1+\alpha} - 1}{1+\alpha} \text{ and } \ell_1(v) = \frac{1 - v^\alpha}{\alpha},$$

Lemma 1 established that the loss is proper composite with link Ψ_{dr}. The underlying proper loss $\lambda = \ell \circ \Psi_{\text{dr}}$ is

$$\lambda_{-1}(u) = \frac{1}{1+\alpha} \cdot \left(\left(\frac{u}{1-u} \right)^{1+\alpha} - 1 \right) \text{ and } \lambda_1(u) = \frac{1}{\alpha} \cdot \left(1 - \left(\frac{u}{1-u} \right)^{\alpha} \right). \tag{33}$$

Observe that the partial losses are negatively unbounded; similarly, the negative Bayes risk is unbounded at the endpoints 0 and 1. Thus, this is not a *definite* loss in the sense of Reid & Williamson (2010).

Now, a proper loss satisfies $\lambda'_{-1}(u) = u \cdot w(u)$ and $\lambda'_1(u) = -(1-u) \cdot w(u)$ (Reid & Williamson, 2010, Theorem 1). It is easy to check that

$$\lambda'_{-1}(u) = \left(\frac{u}{1-u} \right)^{\alpha} \cdot \frac{1}{(1-u)^2} \text{ and } \lambda'_1(u) = -\left(\frac{u}{1-u} \right)^{\alpha-1} \cdot \frac{1}{(1-u)^2}.$$

Thus, from either equation, the weight function for the loss is

$$w(c) = \frac{1}{c^{1-\alpha} \cdot (1-c)^{2+\alpha}}.$$

which is an instance of the (α, β) Beta family of weight functions from Buja et al. (2005, Section 11), where $\beta = 1 - \alpha$. By Equation 21, the weight over density ratios is checked to be

$$w_{\text{DR}}(\rho) = \rho^{\alpha-1}.$$

The latter weight relates to a family of power divergences proposed in Basu et al. (1998), as already noted by Sugiyama et al. (2012a). We can explicate this connection in our jargon as follows. Considering the weight $w(c) = c^{\alpha-1}$ over cost ratios, we have corresponding negative Bayes risk

$$f(c) = \int \int w(c) \, dc \, dc = \frac{1}{\alpha \cdot (\alpha+1)} \cdot c^{\alpha+1},$$

assuming for simplicity that $\alpha \notin \{-1, 0\}$. Since $f'(c) = \frac{c^\alpha}{\alpha}$, we have Bregman divergence

$$\begin{aligned} B_f(x,y) &= \frac{1}{\alpha \cdot (\alpha+1)} \cdot (x^{\alpha+1} - y^{\alpha+1} - (\alpha+1) \cdot y^\alpha \cdot (x-y)) \\ &= \frac{1}{\alpha \cdot (\alpha+1)} \cdot (x^{\alpha+1} - y^{\alpha+1} - (\alpha+1) \cdot y^\alpha \cdot x + (\alpha+1) \cdot y^{\alpha+1}) \\ &= \frac{1}{\alpha \cdot (\alpha+1)} \cdot (x^{\alpha+1} - (\alpha+1) \cdot y^\alpha \cdot x + \alpha \cdot y^{\alpha+1}) \\ &= \frac{1}{\alpha+1} \cdot \left(\frac{1}{\alpha} \cdot x^{\alpha+1} - \left(1 + \frac{1}{\alpha} \right) \cdot y^\alpha \cdot x + y^{\alpha+1} \right). \end{aligned}$$

Now consider probability densities p, q over some instance space $\mathcal{X}$. Then,

$$\int_{\mathcal{X}} B_f(p(x), q(x)) \, dx = \frac{1}{\alpha+1} \cdot \int_{\mathcal{X}} \left(\frac{1}{\alpha} \cdot p(x)^{\alpha+1} - \left(1 + \frac{1}{\alpha} \right) \cdot q(x)^\alpha \cdot p(x) + q(x)^{\alpha+1} \right) dx,$$

which is a scaled version of the divergence between p, q proposed in Basu et al. (1998, Equation 2.1).

F. Convex versions of the family of power losses for $\alpha > 1$

Recall from Appendix E that the power family of losses has weight over cost ratios given by

$$w(c) = \frac{1}{c^{1-\alpha} \cdot (1-c)^{2+\alpha}}$$

with underlying proper loss

$$\lambda_{-1}(u) = \frac{1}{1+\alpha} \cdot \left(\left(\frac{u}{1-u} \right)^{1+\alpha} - 1 \right) \text{ and } \lambda_1(u) = \frac{1}{\alpha} \cdot \left(1 - \left(\frac{u}{1-u} \right)^{\alpha} \right).$$

For the proper composite loss $\ell = \lambda \circ \Psi$ to be convex, the weight w and link Ψ must satisfy (Reid & Williamson, 2010, Theorem 29)

$$-\frac{1}{c} \leq \frac{w'(c)}{w(c)} - \frac{\Psi''(c)}{\Psi'(c)} \leq \frac{1}{1-c}.$$

For the power weight, we have

$$w'(c) = \frac{3c + \alpha - 1}{c^{2-\alpha} \cdot (1-c)^{3+\alpha}},$$

and so

$$\frac{w'(c)}{w(c)} = \frac{3c + \alpha - 1}{c \cdot (1-c)}.$$

Thus, any candidate link must satisfy

$$\frac{\Psi''(c)}{\Psi'(c)} \in \left[\frac{2}{1-c} + \frac{\alpha - 1}{c \cdot (1-c)}, \frac{2}{1-c} + \frac{\alpha}{c \cdot (1-c)} \right]. \tag{34}$$

Consider the link function $\Psi_{\text{dr}}(c)$, which we showed was employed by the power losses in Lemma 1. This link has $\Psi'_{\text{dr}}(c) = \frac{1}{(1-c)^2}$, and $\Psi''_{\text{dr}}(c) = \frac{2}{(1-c)^3}$, so that

$$\frac{\Psi''(c)}{\Psi'(c)} = \frac{2}{1-c}.$$

Evidently, this satisfies Equation 34 only if $\alpha < 1$; for $\alpha > 1$, the left hand side of the bound is greater than $\frac{2}{1-c}$. This explains why the power family of losses as presented in §4.1 is only convex for $\alpha \in (0, 1]$.

It is natural to seek the canonical link for the above weight function, so as to guarantee convexity. This is easily checked to be

$$\Psi_{\text{can}}(u) = \frac{1}{\alpha \cdot (\alpha + 1)} \cdot \frac{u^\alpha}{(1-u)^{1+\alpha}} \cdot (\alpha + 1 - u) - \frac{1}{\alpha},$$

with $\Psi_{\text{can}}(u) \to \log \frac{u}{1-u} + \frac{1}{1-u}$ as $\alpha \to 0$. Unfortunately, the link does not possess an analytic inverse for general α. Thus, generating the corresponding proper composite loss is not simple. (For $\alpha = 1$, we presented the canonical link in Equation 25.) We can fortunately generate convex losses for general α by employing a simple non-canonical link. It is easily checked that a Ψ' satisfying the left hand side of the bound in Equation 34 is

$$\Psi'(c) = \frac{1}{c^{1-\alpha} \cdot (1-c)^{1+\alpha}}$$

with corresponding Ψ for $\alpha \neq 0$

$$\Psi(c) = \frac{1}{\alpha} \cdot \left(\frac{c}{1-c} \right)^{\alpha},$$

with the constant of integration chosen such that $\Psi(0) = 0$ and $\Psi(1) = +\infty$. (Choosing the constant so that $\Psi(0) = -1/\alpha$ allows for $\alpha = 0$ to be handled as well, but as this lower bound is somewhat awkward, we will handle that case separately.) This can further be checked to have inverse

$$\Psi^{-1}(v) = \frac{(\alpha \cdot v)^{1/\alpha}}{(\alpha \cdot v)^{1/\alpha} + 1} = \Psi_{\text{dr}}^{-1}((\alpha \cdot v)^{1/\alpha}),$$

defined on $v \in [0, \infty)$. Defining $\ell = \lambda \circ \Psi^{-1}$ for λ as per Equation 33, the corresponding proper composite loss employing this link is then

$$\ell_{-1}(v) = \frac{(\alpha \cdot v)^{\frac{1+\alpha}{\alpha}} - 1}{1 + \alpha} \text{ and } \ell_1(v) = \frac{1 - \alpha \cdot v}{\alpha}$$

for $v \geq 0$.

For example, when $\alpha = 2$, we have the proper composite loss

$$\ell_{-1}(v) = \frac{(2 \cdot v)^{\frac{3}{2}} - 1}{3} \text{ and } \ell_1(v) = \frac{1}{2} - v$$

with link $\Psi^{-1}(v) = \sqrt{2v}/(1 + \sqrt{2v})$ for $v \geq 0$.

For $\alpha = 0$, the admissible link is

$$\Psi(c) = \log \frac{c}{1 - c},$$

being the standard logit function. Note that $\Psi(0) = -\infty$ and $\Psi(1) = +\infty$, so that there is no constraint on the range of scores. The corresponding proper composite loss is

$$\ell_{-1}(v) = e^v \text{ and } \ell_1(v) = -v.$$

G. The canonical loss for the step weight function

Consider the weight given by

$$w(c) = \begin{cases} 1 & \text{if } c \in [a, b] \\ \frac{1}{h} & \text{else.} \end{cases}$$

By definition, the canonical link is

$$\Psi(c) = \begin{cases} \frac{c}{h} & \text{if } c \in [0, a] \\ c + \frac{(1-h)}{h} \cdot a & \text{if } c \in [a, b] \\ \frac{c}{h} + \frac{h-1}{h} \cdot (b - a) & \text{if } c \in [b, 1], \end{cases}$$

with inverse

$$\Psi^{-1}(v) = \begin{cases} h \cdot v & \text{if } v \in [0, \frac{a}{h}] \\ v + \frac{(h-1)}{h} \cdot a & \text{if } v \in [\frac{a}{h}, b + \frac{1-h}{h} \cdot a] \\ h \cdot v + (h-1) \cdot (a - b) & \text{if } c \in [b + \frac{1-h}{h} \cdot a, \frac{1}{h} + \frac{h-1}{h} \cdot (b - a)]. \end{cases}$$

The corresponding canonical proper composite loss is thus the following amalgam of three square losses:

$$\ell_{-1}(v) = \begin{cases} \frac{h}{2} \cdot v^2 & \text{if } v \in [0, \frac{a}{h}] \\ \frac{1}{2} v^2 + \frac{(h-1)}{h} \cdot a \cdot v + \frac{1-h}{2h^2} \cdot a^2 & \text{if } v \in [\frac{a}{h}, b + \frac{1-h}{h} \cdot a] \\ \frac{h}{2} \cdot v^2 + (h-1) \cdot (a-b) \cdot v + (h-1) \cdot \frac{(a-b)^2 \cdot h - 2a^2 + 2ab}{2h} & \text{if } c \in [b + \frac{1-h}{h} \cdot a, \frac{1}{h} + \frac{h-1}{h} \cdot (b - a)], \end{cases}$$

and $\ell_1(v) = \ell_{-1}(v) - v$. In practice, as with square loss, one can allow v to be arbitrary, and then post-hoc truncate scores to lie in $\mathrm{Im}(\Psi^{-1})$; alternately, with a linear model and bounded feature mapping, one can strongly regularise the weight vector to ensure that scores are in the desired range.

H. Additional experiments for “ranking the best” problems

Table 4 summarises the number of samples (n) and dimensions (d) for several benchmark datasets. For the high dimensional `real-sim` and `news20-forsale` datasets, we performed an SVD projection to 100 dimensions.

Dataset	n	d	**Dataset**	n	d
`german`	1000	24	`skin`	245057	3
`spambase`	4601	57	`w8a`	64700	300
`magic`	19020	10	`real-sim`	72309	20958
`news20-forsale`	19928	62061	`nsl-kdd`	148517	119

Table 4: Statistics for datasets used in “ranking the best” experiments.

We use as performance measures the area under the ROC curve (AUC), mean reciprocal rank (MRR), average precision (AP), fraction of positives ranked higher than the first negative (PTop), and Precision@10 (Prec@10) (Agarwal, 2011). Table 5 summarises the results on these datasets, using the methods described in §8.3. We find that the LSIF loss is superior to the logistic loss on all measures but AUC (which is agnostic to the position of a ranking mistake). It is also strongly competitive with the p-classification loss on most measures, although the latter is superior on the DCG and AP measures. We expect that the gap between the two to shrink with a better feature representation or choice of kernel. We again emphasise that we believe the closed form solution of the LSIF loss makes it an appealing choice of baseline.

	Loss	**AUC**	**MRR**	**DCG**	**AP**	**PTop**	**Prec@10**
`german`	Logistic	0.8051 ± 0.0080 (1)	0.0369 ± 0.0016 (2)	0.1846 ± 0.0015 (3)	0.6087 ± 0.0192 (3)	0.0224 ± 0.0083 (3)	0.7000 ± 0.0333 (2)
	p-class	0.8026 ± 0.0084 (2)	0.0392 ± 0.0014 (1)	0.1863 ± 0.0012 (1)	0.6121 ± 0.0185 (1)	0.0316 ± 0.0084 (2)	0.7000 ± 0.0394 (2)
	LSIF	0.8009 ± 0.0098 (3)	0.0392 ± 0.0015 (1)	0.1862 ± 0.0014 (2)	0.6101 ± 0.0196 (2)	0.0364 ± 0.0114 (1)	0.7500 ± 0.0373 (1)
`spambase`	Logistic	0.9658 ± 0.0011 (1)	0.0104 ± 0.0004 (3)	0.1336 ± 0.0004 (2)	0.9337 ± 0.0035 (2)	0.0202 ± 0.0084 (3)	0.9200 ± 0.0249 (3)
	p-class	0.9631 ± 0.0010 (2)	0.0113 ± 0.0001 (1)	0.1344 ± 0.0002 (1)	0.9408 ± 0.0030 (1)	0.0888 ± 0.0345 (1)	0.9800 ± 0.0133 (1)
	LSIF	0.9423 ± 0.0020 (3)	0.0108 ± 0.0003 (2)	0.1335 ± 0.0003 (3)	0.9149 ± 0.0033 (3)	0.0479 ± 0.0162 (2)	0.9500 ± 0.0269 (2)
`magic`	Logistic	0.8418 ± 0.0011 (2)	0.0020 ± 0.0000 (2)	0.0961 ± 0.0000 (3)	0.8867 ± 0.0018 (3)	0.0018 ± 0.0005 (3)	0.9200 ± 0.0133 (2)
	p-class	0.8434 ± 0.0012 (1)	0.0020 ± 0.0000 (2)	0.0962 ± 0.0000 (2)	0.8962 ± 0.0017 (2)	0.0031 ± 0.0017 (2)	0.9100 ± 0.0233 (3)
	LSIF	0.8329 ± 0.0011 (3)	0.0021 ± 0.0000 (1)	0.0963 ± 0.0000 (1)	0.8996 ± 0.0014 (1)	0.0095 ± 0.0038 (1)	0.9500 ± 0.0224 (1)
`news20-forsale`	Logistic	0.8016 ± 0.0033 (3)	0.0035 ± 0.0003 (3)	0.1068 ± 0.0004 (3)	0.1487 ± 0.0041 (3)	0.0003 ± 0.0003 (3)	0.1200 ± 0.0249 (3)
	p-class	0.8456 ± 0.0048 (1)	0.0105 ± 0.0007 (2)	0.1218 ± 0.0012 (1)	0.2817 ± 0.0113 (1)	0.0054 ± 0.0018 (2)	0.5500 ± 0.0637 (1)
	LSIF	0.8178 ± 0.0060 (2)	0.0107 ± 0.0006 (1)	0.1189 ± 0.0013 (2)	0.2351 ± 0.0106 (2)	0.0101 ± 0.0014 (1)	0.5300 ± 0.0616 (2)
`skin`	Logistic	0.9475 ± 0.0003 (1)	0.0002 ± 0.0000 (1)	0.0696 ± 0.0000 (1)	0.9886 ± 0.0001 (1)	0.9146 ± 0.0003 (2)	1.0000 ± 0.0000 (1)
	p-class	0.9466 ± 0.0003 (2)	0.0002 ± 0.0000 (1)	0.0696 ± 0.0000 (1)	0.9884 ± 0.0001 (2)	0.9101 ± 0.0008 (3)	1.0000 ± 0.0000 (1)
	LSIF	0.9461 ± 0.0003 (3)	0.0002 ± 0.0000 (1)	0.0696 ± 0.0000 (1)	0.9884 ± 0.0001 (2)	0.9149 ± 0.0022 (1)	1.0000 ± 0.0000 (1)
`w8a`	Logistic	0.9676 ± 0.0009 (1)	0.0074 ± 0.0003 (3)	0.1232 ± 0.0003 (3)	0.6631 ± 0.0034 (3)	0.0002 ± 0.0002 (3)	0.6500 ± 0.0619 (3)
	p-class	0.9673 ± 0.0010 (2)	0.0106 ± 0.0001 (1)	0.1285 ± 0.0003 (1)	0.7741 ± 0.0029 (1)	0.2206 ± 0.0111 (1)	1.0000 ± 0.0000 (1)
	LSIF	0.9482 ± 0.0014 (3)	0.0102 ± 0.0001 (2)	0.1249 ± 0.0002 (2)	0.6671 ± 0.0049 (2)	0.0702 ± 0.0260 (2)	0.9600 ± 0.0267 (2)
`real-sim`	Logistic	0.9852 ± 0.0001 (1)	0.0013 ± 0.0000 (1)	0.0896 ± 0.0000 (1)	0.9674 ± 0.0003 (2)	0.0927 ± 0.0064 (2)	1.0000 ± 0.0000 (1)
	p-class	0.9842 ± 0.0001 (2)	0.0013 ± 0.0000 (1)	0.0896 ± 0.0000 (1)	0.9675 ± 0.0003 (1)	0.1288 ± 0.0220 (1)	1.0000 ± 0.0000 (1)
	LSIF	0.9805 ± 0.0002 (3)	0.0013 ± 0.0000 (1)	0.0894 ± 0.0000 (2)	0.9568 ± 0.0006 (3)	0.0328 ± 0.0052 (3)	1.0000 ± 0.0000 (1)
`nsl-kdd`	Logistic	0.9810 ± 0.0002 (2)	0.0004 ± 0.0000 (1)	0.0769 ± 0.0000 (2)	0.9803 ± 0.0003 (3)	0.3711 ± 0.0229 (1)	1.0000 ± 0.0000 (1)
	p-class	0.9867 ± 0.0001 (1)	0.0004 ± 0.0000 (1)	0.0770 ± 0.0000 (1)	0.9886 ± 0.0001 (1)	0.2478 ± 0.0654 (3)	1.0000 ± 0.0000 (1)
	LSIF	0.9756 ± 0.0002 (3)	0.0004 ± 0.0000 (1)	0.0769 ± 0.0000 (2)	0.9811 ± 0.0002 (2)	0.2706 ± 0.0563 (2)	1.0000 ± 0.0000 (1)
Average rank	Logistic	1.5000	2.0000	2.2500	2.5000	2.5000	2.0000
	p-class	1.6250	1.2500	1.1250	1.2500	1.8750	1.3750
	LSIF	2.8750	1.2500	1.8750	2.1250	1.6250	1.3750

Table 5: “Ranking the best” results. Reported are mean and standard error across 10 random splits. Higher scores are better.