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

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

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

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

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)

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

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,

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

[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 $\underline{\mathrm{L}}(\eta) = -(1-\eta)\cdot\left(\frac{\eta}{1-\eta}\cdot\phi'(\frac{\eta}{1-\eta}) - \phi^*\left(\phi'\left(\frac{\eta}{1-\eta}\right)\right)\right)$, which by definition of conjugacy is $\underline{\mathrm{L}}(\eta) = -(1-\eta)\cdot\phi\left(\frac{\eta}{1-\eta}\right)$, 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.