SUPPLEMENTARY FILE OF "COVARIATE BALANCING PROPENSITY SCORE BY TAILORED LOSS FUNCTIONS"

APPENDIX A: TECHNICAL PROOFS

A.1. Proof of Proposition 2. The same result can be found in Buja et al. (2005, Section 15). For completeness we give a direct proof here. Denote $p = l^{-1}(f) \in (0,1)$. Since $v = v_{\alpha,\beta}$, we have

$$\frac{G''(p)}{l'(p)} = p^{\alpha}(1-p)^{\beta}$$

Therefore, by the chain rule and the inverse function theorem,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}f} S(l^{-1}(f),1) &= (1-p)G''(p)(l^{-1})'(f) = p^{\alpha}(1-p)^{\beta+1}, \\
\frac{\mathrm{d}}{\mathrm{d}f} S(l^{-1}(f),0) &= -pG''(p)(l^{-1})'(f) = -p^{\alpha+1}(1-p)^{\beta}, \text{ and} \\
\frac{\mathrm{d}^2}{\mathrm{d}f^2} S(l^{-1}(f),1) &= \alpha p^{\alpha}(1-p)^{\beta+2} - (\beta+1)p^{\alpha+1}(1-p)^{\beta+1}, \\
\frac{\mathrm{d}^2}{\mathrm{d}f^2} S(l^{-1}(f),0) &= -(\alpha+1)p^{\alpha+1}(1-p)^{\beta+1} + \beta p^{\alpha+2}(1-p)^{\beta}.
\end{aligned}$$

The conclusions immediate follow by letting the second order derivatives be less than or equal to 0.

A.2. Proof of Theorem 2. First we list the technical assumptions in Hirano et al. (2003):

ASSUMPTION 1. *(Distribution of $\boldsymbol{X}$) The support of $\boldsymbol{X}$ is a Cartesian product of compact intervals. The density of $\boldsymbol{X}$ is bounded, and bounded away from 0.*

ASSUMPTION 2. *(Distribution of $Y(0)$, $Y(1)$) The second moments of $Y(0)$ and $Y(1)$ exist and $g(\boldsymbol{X},0) = \mathrm{E}[Y(0)|\boldsymbol{X}]$ and $g(\boldsymbol{X},1) = \mathrm{E}[Y(1)|\boldsymbol{X}]$ are continuously differentiable.*

ASSUMPTION 3. *(Propensity score) The propensity score $p(\boldsymbol{X}) = \mathrm{P}(T = 1|\boldsymbol{X})$ is continuously differentiable of order $s \geq 7d$ where d is the dimension of $\boldsymbol{X}$, and $p(\boldsymbol{x})$ is bounded away from 0 and 1.*

ASSUMPTION 4. *(Sieve estimation) The nonparametric sieve logistic regression uses a power series with* $m = O(n^\nu)$ *for some* $1/(4(s/d-1)) < \nu < 1/9$.

The proof is a simple modification of the proof in Hirano et al. (2003). In fact, Hirano et al. (2003) only proved the convergence of the estimated propensity score up to certain order. This essentially suggests that the semiparametric efficiency of $\hat{\tau}$ does not heavily depend on the accuracy of the sieve logistic regression.

To be more specific, only three properties of the maximum likelihood rule $S = S_{0,0}$ are used in Hirano et al. (2003, Lemmas 1,2):

1. $\tilde{\theta} = \arg\max_\theta S(p_\theta, p_{\tilde{\theta}})$ (line 5, page 19), this is exactly the definition of a strictly proper scoring rule (1);
2. The Fisher information matrix
$$\frac{\partial^2}{\partial\theta\partial\theta^T} S(p_\theta, p_{\tilde{\theta}}) = \mathrm{E}_{\tilde{\theta}} \left\{ \left[\frac{\mathrm{d}^2}{\mathrm{d}f^2} S(l^{-1}(f), T)\Big|_{f=\phi(\boldsymbol{X})^T\boldsymbol{\theta}} \right] \phi(\boldsymbol{X})\phi(\boldsymbol{X})^T \right\}$$
has all eigenvalues uniformly bounded away from 0 for all $\boldsymbol{\theta}$ and $\tilde{\boldsymbol{\theta}}$ in a compact set in $\mathbb{R}^m$, where the expectation on the right hand side is taken over $\boldsymbol{X}$ and $T|\boldsymbol{X} \sim p_{\tilde{\boldsymbol{\theta}}}$.
3. As $m \to \infty$, with probability tending to 1 the observed Fisher information matrix
$$\frac{\partial^2}{\partial\theta\partial\theta^T} \frac{1}{n}\sum_{i=1}^n S(p_\theta(\boldsymbol{X}_i), T_i) = \frac{1}{n}\sum_{i=1}^n \left[\frac{\mathrm{d}^2}{\mathrm{d}f^2} S(l^{-1}(f), T_i)\Big|_{f=\phi(\boldsymbol{X}_i)^T\boldsymbol{\theta}} \right] \phi(\boldsymbol{X}_i)\phi(\boldsymbol{X}_i)^T$$
has all eigenvalues uniformly bounded away from 0 for all $\boldsymbol{\theta}$ in a compact set of $\mathbb{R}^m$ (line 7–9, page 21).

Because the approximating functions ϕ are obtained through orthogonalizing the power series, we have $\mathrm{E}[\phi(\boldsymbol{X})\phi(\boldsymbol{X})^T] = \boldsymbol{I}_m$ and one can show its finite sample version has eigenvalues bounded away from 0 with probability going to 1 as $n \to \infty$. Therefore a sufficient condition for the second and third properties above is that $S(l^{-1}(f), t)$ is strongly concave for $t = 0, 1$. In Proposition 2 we have already proven the strong concavity for all $-1 \le \alpha, \beta \ge 1$ except for $\alpha = -1, \beta = 0$ and $\alpha = 0, \beta = -1$. In these two boundary cases, among $S(l^{-1}(f), 0)$ and $S(l^{-1}(f), 1)$ one score function is strongly concave and the other score function is linear in f. One can still prove the second and third properties by using Assumption 3 that the propensity score is bounded away from 0 and 1.

A.3. Proof of Proposition 3. The conclusion is trivial for $a = 1$. Denote

$$h(f,t) = \frac{\mathrm{d}}{\mathrm{d}f} S(l^{-1}(f),t) \text{ and } h'(f,t) = \frac{\mathrm{d}}{\mathrm{d}f} h(f,t),\ t = 0,1.$$

Because $S(l^{-1}(f),t)$ is concave in f, we have $h'(f,t) < 0$ for all f. The first-order optimality condition of (14) is given by

$$\frac{1}{n}\sum_{i=1}^{n} h(\hat{\boldsymbol{\theta}}_\lambda^T \boldsymbol{\phi}(\boldsymbol{X}_i), T_i)\boldsymbol{\phi}_k(\boldsymbol{X}_i) + \lambda|(\hat{\boldsymbol{\theta}}_\lambda)_k|^{a-1}\mathrm{sign}((\hat{\boldsymbol{\theta}}_\lambda)_k) = 0,\ k = 1,\ldots,m.$$

Let $\nabla\hat{\boldsymbol{\theta}}_\lambda$ be the gradient of $\hat{\boldsymbol{\theta}}_\lambda$ with respect to λ. By taking derivative of the identity above, we get

$$\left[\frac{1}{n}\sum_{i=1}^{n} h'(\hat{\boldsymbol{\theta}}_\lambda^T\boldsymbol{\phi}_i, T_i)\boldsymbol{\phi}_i\boldsymbol{\phi}_i^T + \lambda(a-1)\mathrm{diag}(|\hat{\boldsymbol{\theta}}_\lambda|^{a-2})\right]\nabla\hat{\boldsymbol{\theta}}_\lambda = -|\hat{\boldsymbol{\theta}}_\lambda|^{a-1}\mathrm{sign}(\hat{\boldsymbol{\theta}}_\lambda),$$

where we used the abbreviation $\boldsymbol{\phi}_i = \boldsymbol{\phi}(X_i)$ and $\boldsymbol{\theta}^a = (\theta_1^a,\ldots,\theta_m^a)$. For brevity, let's denote

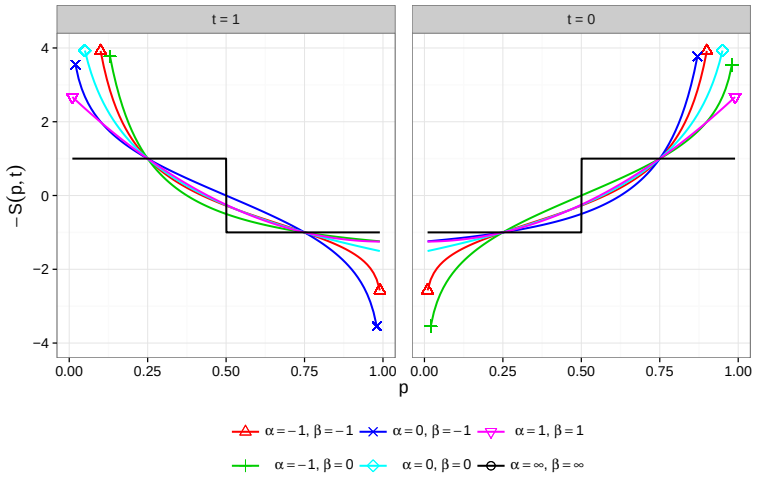
$$\boldsymbol{H} = \frac{1}{n}\sum_{i=1}^{n} h'(\hat{\boldsymbol{\theta}}_\lambda^T\boldsymbol{\phi}_i, T_i)\boldsymbol{\phi}_i\boldsymbol{\phi}_i^T \prec 0 \text{ and } \boldsymbol{G} = \lambda(a-1)\mathrm{diag}(|\hat{\boldsymbol{\theta}}_\lambda|^{a-2}).$$

For $a > 1$, the result is proven by showing the derivative of $\lambda\|\hat{\boldsymbol{\theta}}_\lambda\|_a^{a-1}$ is greater than 0.

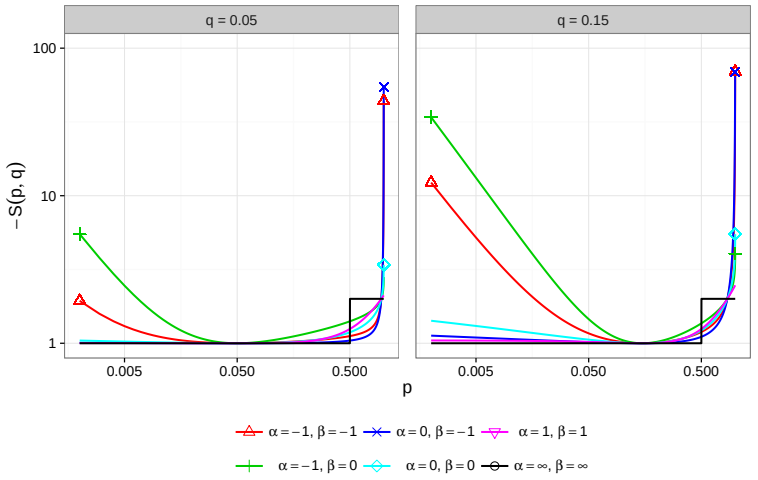
$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\lambda\|\hat{\boldsymbol{\theta}}_\lambda\|_a^{a-1}\right) &= \|\hat{\boldsymbol{\theta}}_\lambda\|_a^{a-1} + \lambda\frac{\mathrm{d}}{\mathrm{d}\lambda}\left[\sum_{j=1}^{m}\left|(\hat{\boldsymbol{\theta}}_\lambda)_k\right|^a\right]^{(a-1)/a} \\
&= \|\hat{\boldsymbol{\theta}}_\lambda\|_a^{a-1} + \lambda(a-1)\|\hat{\boldsymbol{\theta}}_\lambda\|_a^{-1}\sum_{j=1}^{m}\left|(\hat{\boldsymbol{\theta}}_\lambda)_k\right|^{a-1}(\nabla\hat{\boldsymbol{\theta}}_\lambda)_k\,\mathrm{sign}((\hat{\boldsymbol{\theta}}_\lambda)_k) \\
&= \|\hat{\boldsymbol{\theta}}_\lambda\|_a^{a-1} - \lambda(a-1)\|\hat{\boldsymbol{\theta}}_\lambda\|_a^{-1}(|\hat{\boldsymbol{\theta}}_\lambda|^{a-1})^T(\boldsymbol{H}+\boldsymbol{G})^{-1}|\hat{\boldsymbol{\theta}}_\lambda|^{a-1} \\
&> \|\hat{\boldsymbol{\theta}}_\lambda\|_a^{a-1} - \lambda(a-1)\|\hat{\boldsymbol{\theta}}_\lambda\|_a^{-1}(|\hat{\boldsymbol{\theta}}_\lambda|^{a-1})^T\boldsymbol{G}^{-1}|\hat{\boldsymbol{\theta}}_\lambda|^{a-1} \\
&= 0.
\end{aligned}$$

APPENDIX B: A CLOSER LOOK AT THE BETA FAMILY

Figure 1 plots the scoring rules $S_{\alpha,\beta}$ for some combinations of α and β. The top panels show the score function $S(p,0)$ and $S(p,1)$ for $0 < p < 1$,



(a) Loss functions $-S_{\alpha,\beta}(p,t)$ for $t = 0, 1$.



(b) Loss functions $-S_{\alpha,\beta}(p,q)$ for $q = 0.05$ and 0.15.

Fig 1: Graphical illustration of the Beta-family of scoring rules defined in (3).

which are normalized so that $S(1/4,1) = S(3/4,0) = -1$ and $S(1/4,0) = S(3/4,1) = 1$. By a change of variable, one can show $S_{\alpha,\beta}(p,1) = S_{\beta,\alpha}(1-p,0)$. This is the reason that the two subplots in Figure 1a are essentially reflections of each other. The bottom panels show the induced scoring rule $S(p,q)$ defined by section 2.1 or more specifically $S(p,q) = qS(p,1) + (1-q)S(p,0)$ at two different values of $q = 0.05, 0.15$. For aesthetic purposes, the scoring rules in Figure 1b are normalized such that $-S(p,q) = 1$ and $-S(p,1-q) = 2$.

Figure 1 shows that the scoring rules $S_{\alpha,\beta}$, when $-1 \le \alpha, \beta \le 0$, are highly sensitive to small differences of small probabilities. For example, in Figure 1a the loss function $-S_{\alpha,\beta}(p,1)$ is unbounded above when $\alpha, \beta \in \{-1, 0\}$, hence a small change of p near 0 may have a big impact on the score. In Figure 1b, the averaged scoring rules $S_{\alpha,\beta}(p,q)$, when $(\alpha,\beta) = (-1,-1)$ or $(-1,0)$, are also unbounded near $p = 0$. Due to this reason, Selten (1998, Section 2.6) argued that these scoring rules are inappropriate for probability forecast problems.

On the contrary, the unboundedness is actually a desirable feature for propensity score estimation, as the goal is to avoid extreme probabilities. Consider the standard inverse probability weights (IPW)

$$\hat{w}_i = \begin{cases} \hat{p}_i^{-1} & \text{if } T_i = 1, \\ (1-\hat{p}_i)^{-1} & \text{if } T_i = 0, \end{cases} \tag{1}$$

where $\hat{p}_i = p_{\hat{\theta}}(X_i)$ is the estimated propensity score for the i-th data point. This corresponds to $\alpha = \beta = -1$ in the Beta family and estimates ATE. Several previous articles (e.g. Robins and Wang, 2000, Kang and Schafer, 2007, Robins et al., 2007) have pointed out the hazards of using large inverse probability weights. For example, if the true propensity score is $p(\boldsymbol{X}_i) = q = 0.05$ and it happens that $T_i = 1$, we would want $\hat{p}_i$ not too close to 0 so $\hat{w}_i$ is not too large. Conversely, we also want $\hat{p}_i$ not too close to 1, so in the more likely event that $T_i = 0$ the weight $\hat{w}_i$ is not too large either. In an *ad hoc* attempt to mitigate this issue, Lee et al. (2011) studied weight truncation (e.g. truncate the largest 10% weights). They found that the truncation can reduce the standard error of the estimator $\hat{\tau}$ but also increases the bias.

The covariate balancing scoring rules provide a more systematic approach to avoid large weights. For example, the scoring rule $S_{-1,-1}$ precisely penalizes large inverse probability weights as $-S_{-1,-1}(p,q)$ is unbounded above when p is near 0 or 1 (see the left plot in Figure 1b). Similarly, when estimating the ATUT $\tau_{-1,0}$, the weighting scheme would put $\hat{w}_i \propto (1-\hat{p}_i)/\hat{p}_i$ if $T_i = 1$ and $\hat{w}_i \propto 1$ if $T_i = 0$. Therefore we would like $\hat{p}_i$ to be not close

to 0, but it is acceptable if $\hat{p}_i$ is close to 1. As shown in in Figure 1b, the curve $-S_{-1,0}(p,q) = q/p + (1-q)\log(p/(1-p))$ precisely encourages this behavior, as it is unbounded above when p is near 0 and grows slowly when p is near 1.

REFERENCES

Buja, A., W. Stuetzle, and Y. Shen (2005). Loss functions for binary class probability estimation and classification: Structure and applications. *Working draft*.

Hirano, K., G. W. Imbens, and G. Ridder (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica 71*(4), 1161–1189.

Kang, J. D. and J. L. Schafer (2007). Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science 22*(4), 523–539.

Lee, B. K., J. Lessler, and E. A. Stuart (2011). Weight trimming and propensity score weighting. *PloS ONE 6*(3), e18174.

Robins, J., M. Sued, Q. Lei-Gomez, and A. Rotnitzky (2007). Comment: Performance of double-robust estimators when inverse probability weights are highly variable. *Statistical Science 22*(4), 544–559.

Robins, J. M. and N. Wang (2000). Inference for imputation estimators. *Biometrika 87*(1), 113–124.

Selten, R. (1998). Axiomatic characterization of the quadratic scoring rule. *Experimental Economics 1*(1), 43–62.