E0 370 Statistical Learning Theory **Lecture 13 (Sep 24, 2013)**

Consistency of Surrogate Risk Minimization Methods for Binary Classification using Strongly Proper Losses

Lecturer: *Shivani Agarwal* Scribe: *Rohit Vaish*

1 Introduction

Recall from last lecture ('Consistency of Nearest Neighbour Methods') that the regret (or excess-error) of a plug-in classifier h_S can be bounded in terms of how good an approximation of η is η_S. Formally,

Theorem 1.1. The excess-error for a plug-in classifier $h_S(x) = \text{sign}\left(\eta_S(x) - \frac{1}{2}\right)$ w.r.t a distribution $\mathcal{D}$ on $\mathcal{X} \times \{\pm 1\}$ and 0-1 loss can be upper-bounded as:

$$\text{er}_{\mathcal{D}}^{0\text{-}1}[h_S] - \text{er}_{\mathcal{D}}^{0\text{-}1,*} \leq 2\mathbf{E}_X\left[\,|\eta_S(X) - \eta(X)|\,\right] \leq 2\sqrt{\mathbf{E}_X\left[\,(\eta_S(X) - \eta(X))^2\,\right]}$$

We learnt that under certain conditions on weights, a weighted-average plug-in classifier (or any learning algorithm that outputs such a classifier for the same training sample) is universally Bayes consistent w.r.t 0-1 loss. One might wonder for what other learning algorithms can similar statements be made. Can some of the other commonly studied/used learning algorithms be shown to be Bayes consistent w.r.t. 0-1 loss? We've already seen results on Bayes consistency of the ERM algorithm w.r.t 0-1 loss at the expense of computational feasibility. At the other end of the spectrum, we have algorithms like SVM, Logistic Regression etc. that are ubiquitous and computationally feasible but do not directly operate on the 0-1 loss. A natural desideratum in such a situation would be 'the best of both worlds' i.e. can we somehow use the minimization of 'surrogate' regret by commonly employed learning algorithms as a proxy for minimizing the 0-1 regret?

In this lecture we'll consider a specific class of surrogate loss functions, namely *proper losses*. Specifically, we'll see a *surrogate regret bound* result - i.e. bounding the regret of a function w.r.t. the target loss of interest (in this case the 0-1 loss) in terms of the regret w.r.t. a surrogate loss on which the (feasible) algorithm of choice operates. Bayes consistency w.r.t surrogate loss would then imply Bayes consistency w.r.t 0-1 loss.

1.1 Preliminaries

For label and prediction spaces $\mathcal{Y}$ and $\mathcal{T}$ resp., let the *target loss* function be denoted by $\boldsymbol{\ell} : \mathcal{Y} \times \mathcal{T} \mapsto \mathbb{R}_+$. In the binary/multiclass classification learning problem, we'll consider $\mathcal{Y}, \mathcal{T}$ are finite spaces and that quite often renders direct learning of prediction models of the form $h : \mathcal{X} \mapsto \mathcal{T}$ computationally difficult. For this reason, prediction is typically performed in a continuous space $\mathcal{C}$ ($f : \mathcal{X} \mapsto \mathcal{C}$) w.r.t some *surrogate loss* $\psi : \mathcal{Y} \times \mathcal{C} \mapsto \mathbb{R}_+$ followed by appropriate mapping via a function $\text{pred} : \mathcal{C} \mapsto \mathcal{T}$. For example, an ERM algorithm operating on the loss ψ and function class $\mathcal{F} \subseteq \mathcal{C}^{\mathcal{X}}$ (called $\psi - \text{ERM}$ algorithm) would return a hypothesis $f_S : \mathcal{X} \mapsto \mathcal{C}$ such that $f_S \in \arg\min_{f \in \mathcal{F}} \hat{\text{er}}_S^{\psi}[f]$ where $\hat{\text{er}}_S^{\psi}[f] = \frac{1}{m}\sum_{i=1}^{m} \psi(y_i, f(x_i))$ and $S = \{(x_i, y_i)\}_{i=1}^m$. Finally, $h_S = \text{pred} \circ f_S$.

For example, in binary classification, $\mathcal{Y} = \{\pm 1\}$ and $\mathcal{T} = \{\pm 1\}$; often $\mathcal{C} = \mathbb{R}$ (therefore $f_S : \mathcal{X} \mapsto \mathbb{R}$) and a commonly used pred function is the sign based thresholding i.e. pred = sign where $\text{pred} : \mathbb{R} \mapsto \{\pm 1\}$.

We define the following quantities of interest (w.r.t loss ψ and distribution $\mathcal{D}$):

- *Generalization error/ψ-error/ψ-risk of f*: $\mathrm{er}_{\mathcal{D}}^{\psi}[f] = \mathbf{E}_{(X,Y)\sim\mathcal{D}}\big[\psi\left(Y, f(X)\right)\big]$
- *Generalization error of function class $\mathcal{F}$*: $\mathrm{er}_{\mathcal{D}}^{\psi}[\mathcal{F}] = \inf_{f\in\mathcal{F}} \mathrm{er}_{\mathcal{D}}^{\psi}[f]$
- *Bayes ψ-error*: $\mathrm{er}_{\mathcal{D}}^{\psi,*} = \inf_{f:\mathcal{X}\mapsto\mathcal{C}} \mathrm{er}_{\mathcal{D}}^{\psi}[f]$

We are now in a position to formalize the ideas we alluded to earlier. Specifically, we put forth the following two questions:

(1) Are there surrogate losses ψ for which ψ-consistency implies $\boldsymbol{\ell}$-consistency i.e. are there loss functions ψ for which $\exists$ pred $: \mathcal{C} \mapsto \mathcal{T}$ such that $\forall\ \mathcal{D}$:

$$\mathrm{er}_{\mathcal{D}}^{\psi}[f_S] \xrightarrow{P} \mathrm{er}_{\mathcal{D}}^{\psi,*} \implies \mathrm{er}_{\mathcal{D}}^{\boldsymbol{\ell}}[\mathrm{pred} \circ f_S] \xrightarrow{P} \mathrm{er}_{\mathcal{D}}^{\boldsymbol{\ell},*} \ ?$$

(2) If so, can we bound $\boldsymbol{\ell}$-regret in terms of ψ-regret i.e. is there an increasing function $g : \mathbb{R}_+ \mapsto \mathbb{R}_+$, continuous at 0, with $g(0) = 0$ such that $\forall\ \mathcal{D}, f$:

$$\mathrm{er}_{\mathcal{D}}^{\boldsymbol{\ell}}[\mathrm{pred} \circ f] - \mathrm{er}_{\mathcal{D}}^{\boldsymbol{\ell},*} \leq g\big(\mathrm{er}_{\mathcal{D}}^{\psi}[f_S] - \mathrm{er}_{\mathcal{D}}^{\psi,*}\big) \ ?$$

Note that answering question (2) above in positive automatically answers (1).

1.2 Conditional Risk, Conditional Bayes Risk and Conditional Regret

Recall that a binary loss function over a prediction space $\hat{\mathcal{Y}}$ and label space $\mathcal{Y} = \{\pm 1\}$ is a function $\boldsymbol{\ell}$ of the form $\boldsymbol{\ell} : \{\pm 1\} \times \hat{\mathcal{Y}} \mapsto \mathbb{R}_+$. For such a loss function, we define:

- *Conditional $\boldsymbol{\ell}$-risk $L_{\boldsymbol{\ell}} : [0,1] \times \hat{\mathcal{Y}} \mapsto \mathbb{R}_+$* as:

  $$L_{\boldsymbol{\ell}}(\eta, \hat{y}) = \mathbf{E}_{Y\sim\eta}[\boldsymbol{\ell}(Y, \hat{y})] = \eta\ \boldsymbol{\ell}(1, \hat{y}) + (1-\eta)\boldsymbol{\ell}(-1, \hat{y})$$

  Here Y is a $\{\pm 1\}$ valued random variable which is $+1$ with probability η.

- *Conditional Bayes $\boldsymbol{\ell}$-risk $H_{\boldsymbol{\ell}} : [0,1] \mapsto \mathbb{R}_+$* as:

  $$H_{\boldsymbol{\ell}}(\eta) = \inf_{\hat{y}\in\hat{\mathcal{Y}}} L_{\boldsymbol{\ell}}(\eta, \hat{y})$$

- *Conditional $\boldsymbol{\ell}$-regret $R_{\boldsymbol{\ell}} : [0,1] \times \hat{\mathcal{Y}} \mapsto \mathbb{R}_+$* as:

  $$R_{\boldsymbol{\ell}}(\eta, \hat{y}) = L_{\boldsymbol{\ell}}(\eta, \hat{y}) - H_{\boldsymbol{\ell}}(\eta)$$

For $f : \mathcal{X} \mapsto \hat{\mathcal{Y}}$, observe that:

$$\mathrm{er}_{\mathcal{D}}^{\boldsymbol{\ell}}[f] = \mathbf{E}_X\big[L_{\boldsymbol{\ell}}\left(\eta(X), f(X)\right)\big]$$

$$\mathrm{er}_{\mathcal{D}}^{\boldsymbol{\ell},*} = \mathbf{E}_X\big[H_{\boldsymbol{\ell}}\left(\eta(X)\right)\big]$$

$$\mathrm{regret}_{\mathcal{D}}^{\boldsymbol{\ell},*} = \mathbf{E}_X\big[R_{\boldsymbol{\ell}}\left(\eta(X), f(X)\right)\big]$$

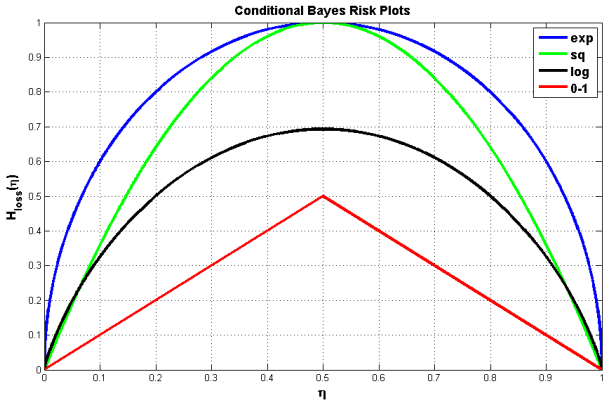

Figure 1: Conditional Bayes Risk $H_{\boldsymbol{\ell}}(\eta)$ plots for $\boldsymbol{\ell}_{0\text{-}1}$, $\boldsymbol{\ell}_{\exp}$, $\boldsymbol{\ell}_{\text{sq}}$ & $\boldsymbol{\ell}_{\log}$ losses

Examples:

1. $\boldsymbol{\ell}_{0\text{-}1} : \{\pm 1\} \times \hat{\mathcal{Y}} \mapsto \{0, 1\}$, $\hat{\mathcal{Y}} = \{\pm 1\}$ where $\boldsymbol{\ell}_{0\text{-}1}(y, \hat{y}) = \mathbf{1}(y \neq \hat{y})$

$$L_{0\text{-}1}(\eta, \hat{y}) = \eta\ \boldsymbol{\ell}_{0\text{-}1}(1, \hat{y}) + (1 - \eta)\boldsymbol{\ell}_{0\text{-}1}(-1, \hat{y}) = \eta\ \mathbf{1}(\hat{y} = -1) + (1 - \eta)\mathbf{1}(\hat{y} = 1)$$

$$H_{0\text{-}1}(\eta) = \inf_{\hat{y} \in \hat{\mathcal{Y}}} L_{0\text{-}1}(\eta, \hat{y}) = \min(\eta, 1 - \eta)$$

2. $\psi_{\exp} : \{\pm 1\} \times \mathbb{R} \mapsto \mathbb{R}_+$ where $\psi_{\exp}(y, \alpha) = e^{-y\alpha}$.

$$L_{\exp}(\eta, \alpha) = \eta\ \psi_{\exp}(1, \alpha) + (1 - \eta)\psi_{\exp}(-1, \alpha) = \eta e^{-\alpha} + (1 - \eta)e^{\alpha}$$

$$H_{\exp}(\eta) = \inf_{\alpha \in \mathbb{R}} L_{\exp}(\eta, \alpha) = 2\sqrt{\eta(1 - \eta)} \qquad \left(\text{at } \alpha^* = \frac{1}{2} \ln \left(\frac{\eta}{1 - \eta}\right)\right)$$

3. $\psi_{\text{sq}} : \{\pm 1\} \times \mathbb{R} \mapsto \mathbb{R}_+$ where $\psi_{\text{sq}}(y, \alpha) = (y - \alpha)^2$.

$$L_{\text{sq}}(\eta, \alpha) = \eta\ \psi_{\text{sq}}(1, \alpha) + (1 - \eta)\psi_{\text{sq}}(-1, \alpha) = \eta(1 - \alpha)^2 + (1 - \eta)(1 + \alpha)^2$$

$$H_{\text{sq}}(\eta) = \inf_{\alpha \in \mathbb{R}} L_{\text{sq}}(\eta, \alpha) = 4\eta(1 - \eta) \qquad (\text{at } \alpha^* = 2\eta - 1)$$

The conditional Bayes risk function is always concave, which we note in the following lemma.

Lemma 1.2. For any prediction space $\hat{\mathcal{Y}} \subseteq \mathbb{R}$ and loss function $\boldsymbol{\ell} : \{\pm 1\} \times \hat{\mathcal{Y}} \mapsto \mathbb{R}_+$, the conditional Bayes-$\boldsymbol{\ell}$ risk $H_{\boldsymbol{\ell}}(\eta)$ is always a concave function of η on $[0, 1]$.

Proof. By def., $H_{\boldsymbol{\ell}}(.)$ is the pointwise infimum of a family of concave functions (conditional $\boldsymbol{\ell}$-risk functions being linear, are concave) and is therefore concave. □

The conditional Bayes risk for certain commonly studied loss functions is shown in Figure 1.

We now define *proper losses* which are useful in class probability estimation and which we also use in developing surrogate regret bounds for 0-1 regret.

2 Proper and Proper Composite Losses

A binary CPE (class probability estimation) loss $\boldsymbol{\ell} : \{\pm 1\} \times [0,1] \mapsto \mathbb{R}_+$ is called a:

- *proper* loss if
$$L_{\boldsymbol{\ell}}(\eta, \hat{\eta}) \geq L_{\boldsymbol{\ell}}(\eta, \eta) \ \forall \ \eta, \hat{\eta} \in [0,1]$$
- *strictly proper* loss if $\boldsymbol{\ell}$ is proper and
$$L_{\boldsymbol{\ell}}(\eta, \hat{\eta}) > L_{\boldsymbol{\ell}}(\eta, \eta) \ \forall \ \hat{\eta} \neq \eta$$
- λ-*strongly proper* loss ($\lambda > 0$) if $\boldsymbol{\ell}$ is proper and
$$L_{\boldsymbol{\ell}}(\eta, \hat{\eta}) - L_{\boldsymbol{\ell}}(\eta, \eta) \geq \frac{\lambda}{2}(\hat{\eta} - \eta)^2 \ \forall \ \eta, \hat{\eta} \in [0,1]$$

A λ-strongly proper loss ($\lambda > 0$) is also strictly proper, but not vice-versa.

Example of proper loss

The logarithmic loss, $\boldsymbol{\ell}_{\log} : \{\pm 1\} \times [0,1] \mapsto \mathbb{R}_+$ defined as:

$$\boldsymbol{\ell}_{\log}(y, \hat{\eta}) = \begin{cases} -\ln \hat{\eta} & y = 1 \\ -\ln(1 - \hat{\eta}) & y = -1 \end{cases}$$

is a proper loss. To see this, note that we can write:

$$L_{\log}(\eta, \hat{\eta}) = \eta \ \boldsymbol{\ell}_{\log}(1, \hat{\eta}) + (1 - \eta)\boldsymbol{\ell}_{\log}(-1, \hat{\eta}) = -\eta \ln(\hat{\eta}) - (1 - \eta) \ln(1 - \hat{\eta})$$

Now, $L'_{\log}(\eta, \hat{\eta}) = -\dfrac{\eta}{\hat{\eta}} + \dfrac{1 - \eta}{1 - \hat{\eta}} = 0 \ \Rightarrow \hat{\eta} = \eta$

Also, $L''_{\log}(\eta, \hat{\eta}) = \dfrac{\eta}{\hat{\eta}^2} + \dfrac{1 - \eta}{(1 - \hat{\eta})^2}$ which gives $L''_{\log}(\eta, \hat{\eta})|_{\hat{\eta} = \eta} = \dfrac{1}{\eta} + \dfrac{1}{1 - \eta} > 0.$

This means $L_{\log}(\eta, \hat{\eta})$ is minimized at $\hat{\eta} = \eta$ and therefore $\boldsymbol{\ell}_{\log}$ is proper. Moreover, since $\hat{\eta} = \eta$ is the unique minimizer for $L_{\log}(\eta, \hat{\eta})$, $\boldsymbol{\ell}_{\log}$ is in fact strictly proper.

Characterizing strictly and strongly proper losses

We now state two important results providing characterization of strictly proper and λ-strongly proper losses.

Theorem 2.1. (Hendrickson and Buehler (1971) [Hendrickson and Buehler, 1971]; Schervish (1989) [Schervish, 1989]) A proper loss $\boldsymbol{\ell} : \{\pm 1\} \times [0,1] \mapsto \mathbb{R}_+$ is strictly proper iff $H_{\boldsymbol{\ell}}$ is strictly concave.

A loss $\boldsymbol{\ell} : \{\pm 1\} \times [0,1] \mapsto \mathbb{R}_+$ is said to be *regular* if $\boldsymbol{\ell}(1, \hat{\eta}) \in \mathbb{R}_+ \ \forall \hat{\eta} \in (0,1]$ and $\boldsymbol{\ell}(-1, \hat{\eta}) \in \mathbb{R}_+ \ \forall \hat{\eta} \in [0,1)$ i.e. if $\boldsymbol{\ell}(y, \hat{\eta})$ is finite for all $y, \hat{\eta}$ except possibly for $\boldsymbol{\ell}(1,0)$ and $\boldsymbol{\ell}(-1,1)$, which could be infinite.

Theorem 2.2. (Agarwal (2013) [Agarwal, 2013]) Let $\lambda > 0$ and let $\boldsymbol{\ell} : \{\pm 1\} \times [0,1] \mapsto \mathbb{R}_+$ be a regular proper loss. Then $\boldsymbol{\ell}$ is λ-strongly proper if and only if $H_{\boldsymbol{\ell}}$ is λ-strongly concave.

Example of strongly proper loss

For the same example,

$$H_{\log}(\eta) = \inf_{\hat{\eta} \in [0,1]} L_{\log}(\eta, \hat{\eta}) = -\eta \ln(\eta) - (1 - \eta) \ln(1 - \eta)$$

Fig.1 shows (and it's also easy to observe) that $H_{\log}(\eta)$ is strictly concave and therefore by Theorem 2.1, $\boldsymbol{\ell}_{\log}$ is strictly proper (as we already checked).

Again, it can be verified that $H_{\log}(\eta)$ is 4-strongly concave and therefore by Theorem 2.2, $\boldsymbol{\ell}_{\log}$ is 4-strongly proper.

Proper Composite Losses

Let $\mathcal{C} \subseteq \mathbb{R}$. A loss $\psi : \{\pm 1\} \times \mathcal{C} \mapsto \mathbb{R}_+$ is *proper composite* if $\exists$ a proper loss $\boldsymbol{\ell} : \{\pm 1\} \times [0,1] \mapsto \mathbb{R}_+$ and a strictly increasing link function $\gamma : [0,1] \mapsto \mathcal{C}$ such that:

$$\psi(y, \alpha) = \boldsymbol{\ell}(y, \gamma^{-1}(\alpha)) \; \forall \; y \in \{\pm 1\}, \; \alpha \in \mathcal{C}$$

Example of proper composite loss

Consider Logistic Loss: $\psi_{\text{logistic}} : \{\pm 1\} \times \mathbb{R} \mapsto \mathbb{R}_+$ where $\psi_{\text{logistic}}(y, \alpha) = \ln\left(1 + e^{-y\alpha}\right)$.

Consider a link function $\gamma(\eta) = \ln\left(\dfrac{\eta}{1-\eta}\right) \; \forall \; \eta \in [0,1]$[1]. Since $\gamma(.)$ is strictly increasing and onto, we can define the inverse function $\gamma^{-1}(\alpha) = \dfrac{1}{1+e^{-\alpha}}$.

We can now write the logistic loss as a composition of the logarithmic loss (a proper loss) and our choice of link function $\gamma(.)$ above in the following manner:

$$\begin{aligned}
\psi_{\text{logistic}}(1, \alpha) &= \ln\left(1 + e^{-\alpha}\right) &&= -\ln\left(\gamma^{-1}(\alpha)\right) &&= \boldsymbol{\ell}_{\log}(1, \gamma^{-1}(\alpha)) \\
\psi_{\text{logistic}}(-1, \alpha) &= \ln\left(1 + e^{\alpha}\right) &&= -\ln\left(1 - \gamma^{-1}(\alpha)\right) &&= \boldsymbol{\ell}_{\log}(-1, \gamma^{-1}(\alpha))
\end{aligned}$$

We also know (from earlier example) that logarithmic loss is 4-strongly proper. Logistic loss, therefore, is a 4-strongly proper composite loss. Similar checks could be performed for other loss functions e.g. exponential, squared etc.

For more on proper (composite) losses, see for example [Buja et al., 2005], [Reid and Williamson, 2009], [Reid and Williamson, 2010] & [Reid and Williamson, 2011].

3 Surrogate Regret Bound in terms of Strongly Proper Composite Losses

We now state the surrogate regret bound result that we set ourselves out for in the beginning. This result implies that if for a strongly composite loss ψ, the ψ-regret of a function f is small, then the 0-1 regret of the plug-in classifier built using f is also small.

Theorem 3.1. Let $\mathcal{C} \subseteq \mathbb{R}$. Let $\psi : \{\pm 1\} \times \mathcal{C} \mapsto \mathbb{R}_+$ be a λ-strongly proper composite loss with underlying proper loss $\boldsymbol{\ell} : \{\pm 1\} \times [0,1] \mapsto \mathbb{R}_+$ and link function $\gamma : [0,1] \mapsto \mathcal{C}$. Then $\forall$ distributions $\mathcal{D}$ on $\mathcal{X} \times \{\pm 1\}$ & $\forall$ functions $f : \mathcal{X} \mapsto \mathcal{C}$:

$$\text{er}_{\mathcal{D}}^{0\text{-}1}\left[\text{sign} \circ \left(\gamma^{-1} \circ f - 1/2\right)\right] - \text{er}_{\mathcal{D}}^{0\text{-}1,*} \leq 2\sqrt{\frac{2}{\lambda}\left(\text{er}_{\mathcal{D}}^{\psi}[f] - \text{er}_{\mathcal{D}}^{\psi,*}\right)}$$

[1] We define $\gamma(.)$ under a codomain of extended reals i.e. $\gamma : [0,1] \mapsto \mathbb{R} \cup \{\pm\infty\}$

Proof.

$$\begin{aligned}
\text{LHS} &\leq 2\sqrt{\mathbf{E}_X\left[\left(\gamma^{-1}(f)-\eta(X)\right)^2\right]} && \text{(Theorem 1.1)}\\
&\leq 2\sqrt{\frac{2}{\lambda}\,\mathbf{E}_X\left[R_{\boldsymbol{\ell}}\left(\eta(X),\gamma^{-1}(f(X))\right)\right]} && \text{(by def. of strongly proper loss)}\\
&= 2\sqrt{\frac{2}{\lambda}\left(\mathrm{er}^{\psi}_{\mathcal{D}}[f]-\mathrm{er}^{\psi,*}_{\mathcal{D}}\right)} = \text{RHS} && \text{(by def. of regret)}
\end{aligned}$$

□

Note that the form of this result is precisely what we demanded in question (2). We also note that a more implicit form of surrogate regret bound using strictly proper losses can be found in [Reid and Williamson, 2009].

4 Extension to Cost-Sensitive Binary Classification

The 0-1 loss penalizes false positives and false negatives are penalized exactly the same way. In some settings (e.g. medical diagnosis), however, the two types of misclassifications might entail different costs and therefore need to be treated differently. A simple extension of the 0-1 loss allows us to capture this idea. In particular, define a *cost-sensitive binary classification loss* $\boldsymbol{\ell}_c : \{\pm 1\} \times \{\pm 1\} \mapsto \mathbb{R}_+$ for a fixed constant $c \in (0,1)$ as:

$$\boldsymbol{\ell}_c(y,\hat{y}) = \begin{cases} c & y=-1,\ \hat{y}=1 \\ 1-c & y=1,\ \hat{y}=-1 \\ 0 & \text{otherwise} \end{cases}$$

Theorem 1.1 stated at the beginning of this lecture can be extended to cost-sensitive classification in the following manner: the cost-sensitive loss based regret of a plug-in classifier (thresholded at c) can be upper bounded, as before, by the extent of closeness of η and η_S.

$$\mathrm{er}^{\boldsymbol{\ell}_c}_{\mathcal{D}}[\mathrm{sign}\circ(\eta_S - c)] - \mathrm{er}^{\boldsymbol{\ell}_c,*}_{\mathcal{D}} \leq \mathbf{E}_X\left[|\eta_S(X)-\eta(X)|\right]$$

The same approach as above then yields that for a λ-strongly proper composite loss $\psi : \{\pm 1\} \times \mathcal{C} \mapsto \mathbb{R}_+$ with link $\gamma : [0,1] \mapsto \mathcal{C}$, we have $\forall\ \mathcal{D}$ and $f : \mathcal{X} \mapsto \mathcal{C}$:

$$\mathrm{er}^{c}_{\mathcal{D}}\left[\mathrm{sign}\circ\left(\gamma^{-1}\circ f - c\right)\right] - \mathrm{er}^{c,*}_{\mathcal{D}} \leq \sqrt{\frac{2}{\lambda}\left(\mathrm{er}^{\psi}_{\mathcal{D}}[f]-\mathrm{er}^{\psi,*}_{\mathcal{D}}\right)}$$

5 Next Lecture

In the next lecture, we'll look at another way to obtain surrogate regret bounds (without having to estimate class conditional probabilities as an intermediate step) for a broader class of loss functions, namely *classification calibrated losses*.

References

[Agarwal, 2013] Agarwal, S. (2013). Surrogate regret bounds for the area under the ROC curve via strongly proper losses. Conference on Learning Theory.

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

[Hendrickson and Buehler, 1971] Hendrickson, A. D. and Buehler, R. J. (1971). Proper scores for probability forecasters. *The Annals of Mathematical Statistics*, pages 1916–1921.

[Reid and Williamson, 2009] Reid, M. D. and Williamson, R. C. (2009). Surrogate regret bounds for proper losses. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 897–904. ACM.

[Reid and Williamson, 2010] Reid, M. D. and Williamson, R. C. (2010). Composite binary losses. *The Journal of Machine Learning Research*, pages 2387–2422.

[Reid and Williamson, 2011] Reid, M. D. and Williamson, R. C. (2011). Information, divergence and risk for binary experiments. *The Journal of Machine Learning Research*, pages 731–817.

[Schervish, 1989] Schervish, M. J. (1989). A general method for comparing probability assessors. *The Annals of Statistics*, 17(4):1856–1879.