**EECS 598: Statistical Learning Theory, Winter 2014** **Topic 14**

# Calibrated Surrogate Losses

*Lecturer: Clayton Scott* *Scribe: Efrén Cruz Cortés*

**Disclaimer**: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

## 1 Surrogate Losses for Classification

Recall a loss function is of the form $L : \mathcal{Y} \times \mathbb{R} \to [0, \infty)$. In binary classification ($\mathcal{Y} = \{-1, +1\}$), we usually measure performace with respect to the 0/1 loss $L(y, t) = \mathbf{1}_{\{\text{sign}(t) \neq y\}}$, with the associated risk $R(f) = \mathbb{E}_{XY}\left[\mathbf{1}_{\{\text{sign}(f(X)) \neq Y\}}\right]$. However, the 0/1 loss is neither convex nor differentiable with respect to $t$, which posses computational challenges for (penalized) empirical risk minimization. A **surrogate loss** $L$ is a loss that is used as a proxy for the 0/1 loss, and usually has better computational properties. The associated risk is $R_L(f) = \mathbb{E}_{XY}[L(Y, f(X))]$. Using a surrogate loss raises the question of whether minimizing $R_L(f)$ is still meaningful.

Denote

$$R^* = \inf_f R(f)$$

and

$$R_L^* = \inf_f R_L(f).$$

where in both cases the inf is over all measurable functions $f : \mathcal{X} \to \mathbb{R}$.

In these notes we will give sufficient conditions on $L$ such that

$$R_L(f_n) \to R_L^* \implies R(f_n) \to R^*.$$

To each $L$ we will associate a nondecreasing function $\psi_L : [0, \infty) \to [0, \infty)$ such that for all joint distributions $P_{XY}$ and all $f : \mathcal{X} \to \mathbb{R}$,

$$\psi_L\left(R_L(f) - R_L^*\right) \leq R(f) - R^*.$$

The sufficient conditions on $L$ will imply $\psi_L$ is strictly increasing and therefore invertible.

## 2 Excess Risk Bound

Recall $\eta(x) = \Pr(Y = 1|X = x)$. Then

$$\begin{aligned}
R_L(f) &= \mathbb{E}_{XY}[L(Y, f(X))] \\
&= \mathbb{E}_X \mathbb{E}_{Y|X}[L(Y, f(X))] \\
&= \mathbb{E}_X[\eta(X) L(1, f(X)) + (1 - \eta(X)) L(-1, f(X))] \\
&= \mathbb{E}_X[C_L(\eta(X), f(X))]
\end{aligned}$$

where $C_L(\eta, t) = \eta L(1, t) + (1 - \eta) L(-1, t)$ and $\eta \in [0, 1]$.

Define

$$C_L^*(\eta) := \inf_{t\in\mathbb{R}} C_L(\eta, t)\,,$$
$$C_L^-(\eta) := \inf_{t:t\left(\eta-\frac{1}{2}\right)\leq 0} C_L(\eta, t)$$

and note that $R_L^* = \mathbb{E}_X\left[C_L^*(\eta(X))\right]$. Also define

$$H_L(\eta) = C_L^-(\eta) - C_L^*(\eta)$$

and note that $H_L \geq 0$. For $\epsilon \in [0, 1]$, define

$$\begin{aligned}\nu_L(\epsilon) &:= \min\left\{H_L\left(\frac{1+\epsilon}{2}\right), H_L\left(\frac{1-\epsilon}{2}\right)\right\} \\ &= \min_{\substack{\eta\in[0,1] \\ |2\eta-1|=\epsilon}} H_L(\eta)\,.\end{aligned}$$

Finally, define

$$\begin{aligned}\psi_L &= \nu_L^{**} \\ &= \text{the Fenchel-Legendre biconjugate of } \nu_L \\ &= \text{largest lower semi-continuous function bounded above by } \nu_L\,.\end{aligned}$$

An equivalent definition of the Fenchel-Legendre biconjugate is: given a function $g$, $g^{**}$ is the unique function such that

$$\mathrm{Epi}\, g^{**} = \overline{\mathrm{co}\,\mathrm{Epi}\, g}$$

where $\mathrm{Epi}\, g = \{(r, s)\,|\, g(r) \leq s\}$, co is the convex hull, and the overline indicates set closure. Note that $\nu_L(0) = 0$ (since $\eta = 1/2$ makes $C_L^-$ unconstrained), so $\psi_L(0) = 0$, and that $\psi_L$ is not decreasing because $\nu_L \geq 0$ and $\psi_L$ is convex.

**Example:** Recall the *hinge loss* $L(y, t) = (1 - yt)_+$, where $(a)_+ = \max(0, a)$. Then

$$C_L(\eta, t) = \eta(1-t)_+ + (1-\eta)(1+y)_+\,.$$

Noting that for each $\eta$, $C_L(\eta, t)$ is convex and piecewise linear with breakpoints at $-1, +1$, it is not hard to see that

$$\begin{aligned}C_L^* &= \min_{t\in\mathbb{R}} C_L(\eta, t) \\ &= 2\min\{\eta, 1-\eta\}\,,\end{aligned}$$

and

$$\begin{aligned}C_L^-(\eta) &= \inf_{t:t\left(\eta-\frac{1}{2}\right)\leq 0} C_L(\eta, t) \\ &= C_L(\eta, 0) \\ &= 1\,.\end{aligned}$$

Therefore

$$H_L(\eta) = 1 - 2\min\{\eta, 1-\eta\}\,,$$

which implies

$$\nu_L(\epsilon) = \epsilon = \psi_L(\epsilon)\,.$$

See Figs. 1 and 2.

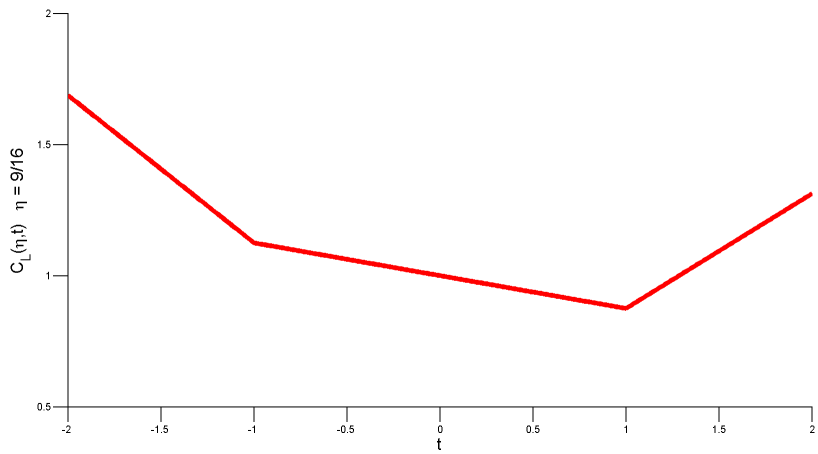


Figure 1: $C_L(\eta, t)$ for Example 1, where $L$ is the hinge loss.

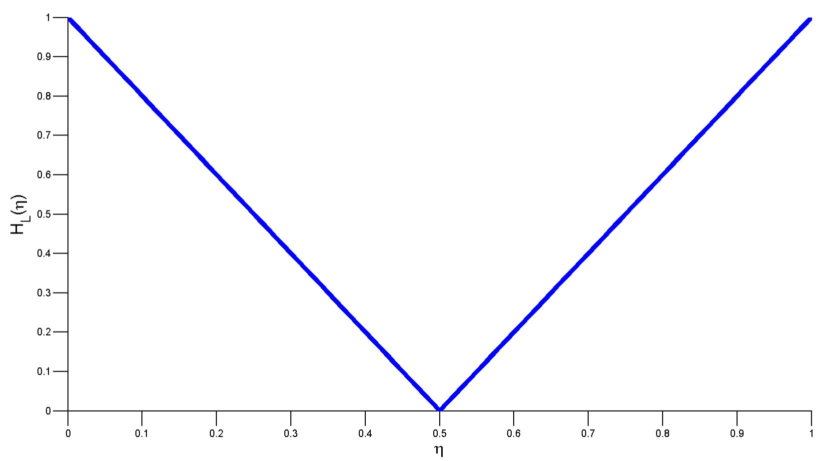


Figure 2: $H_L(\eta)$ for Example 1, where $L$ is the hinge loss.

**Theorem 1.** *For all $L, P_{XY}$, and $f$,*

$$\psi_L\left(R(f) - R^*\right) \leq R_L(f) - R_L^* .$$

*Proof.*

$$\begin{aligned}
\psi_L\left(R(f) - R^*\right) &= \psi_L\left(\mathbb{E}_X\left[|2\eta(X)-1|\,\mathbf{1}_{\{\mathrm{sign}\,(f(X))\neq\mathrm{sign}\,(\eta(X)-\frac{1}{2})\}}\right]\right)\\
&\leq \mathbb{E}_X\left[\psi_L\left(|2\eta(X)-1|\,\mathbf{1}_{\{\mathrm{sign}(f(X))\neq\mathrm{sign}(\eta(X)-\frac{1}{2})\}}\right)\right] \quad \text{(Jensen's)}\\
&\leq \mathbb{E}_X\left[\psi_L\left(|2\eta(X)-1|\,\mathbf{1}_{\{f(X)(\eta(X)-\frac{1}{2})\leq 0\}}\right)\right]\\
&\leq \mathbb{E}_X\left[\nu_L\left(|2\eta(X)-1|\,\mathbf{1}_{\{f(X)(\eta(X)-\frac{1}{2})\leq 0\}}\right)\right]\\
&= \mathbb{E}_X\left[\mathbf{1}_{\{f(X)(\eta(X)-\frac{1}{2})\leq 0\}}\nu_L\left(|2\eta(X)-1|\right)\right] \quad \text{(since } \nu_L(0)=0\text{)}\\
&= \mathbb{E}_X\left[\mathbf{1}_{\{f(X)(\eta(X)-\frac{1}{2})\leq 0\}} \inf_{\substack{\eta\in[0,1]\\ |2\eta-1|=|2\eta(X)-1|}} H_L(\eta)\right]\\
&\leq \mathbb{E}_X\left[\mathbf{1}_{\{f(X)(\eta(X)-\frac{1}{2})\leq 0\}} H_L(\eta(X))\right]\\
&= \mathbb{E}_X\left[\mathbf{1}_{\{f(X)(\eta(X)-\frac{1}{2})\leq 0\}} \inf_{t:t\left(\eta(X)-\frac{1}{2}\right)\leq 0} \left(C_L(\eta(X),t) - C_L^*(\eta(X))\right)\right]\\
&\leq \mathbb{E}_X\left[C_L(\eta(X), f(X)) - C_L^*(\eta(X))\right]\\
&= R_L(f) - R_L^* .
\end{aligned}$$

□

**Corollary 1.** *For the hinge loss,*

$$R(f) - R^* \leq R_L(f) - R_L^*.$$

*Proof.* As we saw above, in this case $\psi_L$ is the identity. □

# 3 Classification Calibrated Losses

**Definition 1.** *We say $L$ is* classification calibrated*, and write $L$ is $CC$, if and only if $H_L(\eta) > 0\ \forall\ \eta \neq 1/2$.*

**Theorem 2.** *$L$ is $CC$ if and only if $\psi_L$ is invertible.*

*Proof sketch.* ( $\Longrightarrow$ ) Suppose $L$ is $CC$. We know $\psi_L(0) = 0$, and $\psi_L$ is convex and nondecreasing, so it suffices to show $\psi_L(\epsilon) > 0\ \forall\ \epsilon \in (0,1]$. So let $\epsilon \in (0,1]$. Then $\nu_L(\epsilon) = \min\left\{H_L\left(\frac{1+\epsilon}{2}\right), H_L\left(\frac{1-\epsilon}{2}\right)\right\} > 0$. Now, $\mathrm{Epi}\,\nu_L$ is closed (since $\nu_L$ is lower semi-continuous, a lemma we will not prove), and so $\mathrm{co}\,\mathrm{Epi}\,\nu_L$ is closed. So if $\psi_L(\epsilon) = 0$, then $(\epsilon, 0)$ is a convex combination of points in $\mathrm{Epi}\,\nu_L$, which is impossible.
( $\Longleftarrow$ ) Similar. Left as an exercise. □

# 4 Convex Margin Losses

Assume $L(y,t) = \phi(yt)$, where $\phi : \mathbb{R} \to [0,\infty)$.

**Theorem 3.** *Suppose $\phi$ is convex and differentiable at $0$. Then $L$ is $CC \iff \phi'(0) < 0$.*

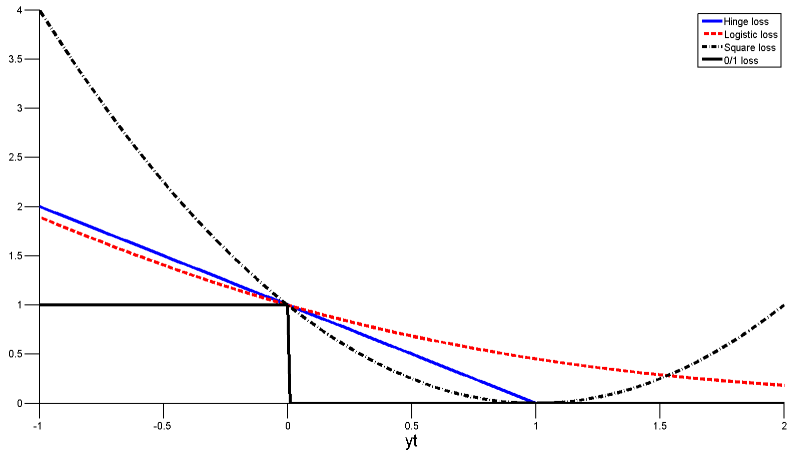


Figure 3: Convex margin losses against the 0/1 loss.

*Proof.* $\phi$ is convex, therefore $C_L(\eta, t) = \eta\phi(t) + (1-\eta)\phi(-t)$ is convex in $t$ for fixed $\eta$. Also note that $\frac{\partial}{\partial t} C_L(\eta, t)\big|_{t=0} = (2\eta - 1)\phi'(0)$. So

$$\begin{aligned}
L \text{ is } CC &\iff \forall\, \eta \neq \frac{1}{2},\ \inf_{t:t(\eta - \frac{1}{2}) \le 0} > C_L^*(\eta) \\
&\iff \forall\, \eta \neq \frac{1}{2},\ \frac{\partial}{\partial t} C_L(\eta, t)\Big|_{t=0} \begin{cases} > 0 & \text{if } \eta < \frac{1}{2} \\ < 0 & \text{if } \eta > \frac{1}{2} \end{cases} \\
&\iff \phi'(0) < 0.
\end{aligned}$$

□

# 5 Further Reading

The material in these notes is based largely on [2]. Other key references include [1, 3]. The idea of calibrated surrogate losses has been extended to other supervised learning problems, including multiclass classification [4], cost-sensitive binary classification [5, 6], and ranking.

# Exercises

1. Consider the exponential loss $L(y, t) = e^{-yt}$. Determine $\psi_L$.
2. Consider the logistic loss $L(y, t) = \log(1 + e^{-yt})$. Determine a closed form expression for $H_L$. Then express $\psi_L(\epsilon)$ as a power series in $\epsilon$, and use this to argue that $\psi_L(\epsilon) \ge \epsilon^2/2$. *Hint:* Use an appropriate Taylor series.

# References

[1] T. Zhang, "Statistical behavior and consistency of classification methods based on convex risk minimization," *The Annals of Statistics*, vol. 32, no. 1, pp. 56-85, 2004.

[2] P. Bartlett, M. Jordan, and J. McAuliffe, "Convexity, classification, and risk bounds," *J. Amer. Statist. Assoc.*, vol. 101, pp. 138-156, 2006.

[3] I. Steinwart "How to compare different loss functions and their risks," *Constructive Approximation*, vol. 26, no. 2, pp. 225-287, 2007.

[4] A. Tewari and P. Bartlett, "On the Consistency of Multiclass Classification Methods," *J. Machine Learning Research*, vol. 8, pp. 1007–1025, 2007.

[5] C. Scott, "Surrogate Losses and Regret Bounds for Cost-Sensitive Classification with Example-Dependent Costs," in *Proceedings of the 28th International Conference on Machine Learning*, L. Getoor and T. Scheffer, Eds., Ominipress, pp. 697-704, 2011.

[6] C. Scott, "Calibrated Asymmetric Surrogate Losses," *Electronic Journal of Statistics*, vol. 6, pp. 958-992, 2012.