CS281B/Stat241B (Spring 2006) Statistical Learning Theory **Lecture: 7**

Discrete Loss and Convex Loss

Lecturer: Peter Bartlett *Scribe: James Zhang*

1 Review

Recall from last lecture, the empirical ϕ-risk is given by:

$$\hat{L}_\phi(w) = \frac{1}{n}\sum_{i=1}^{n}\phi(y_i x_i^T w)$$

In the SVM case:

$$\phi(\alpha) = (1-\alpha)_+ = \max(0, 1-\alpha)$$

2 Discrete Loss and Convex Loss

Because we are using a ϕ-loss function to approximate the discrete loss function, we want to show that $L(f)$ decreases when we train a classifier to minimize $L_\phi(f)$.

In other words, we would like to examine the excess risk of $f : \mathcal{X} \to \mathbb{R}$

$$L(f) - L^*, \quad \text{where } L^* = \inf_f L(f)$$

and how it relates to the excess ϕ-risk of $f : \mathcal{X} \to \mathbb{R}$

$$L_\phi(f) - L_\phi^*, \quad \text{where } L_\phi^* = \inf_f L_\phi(f)$$

Before we begin however, we will need several definitions.

Definition.

$$\eta(x) = P(Y = 1 | X = x) \tag{1}$$

Definition.

$$L_\phi(f) = \mathbb{E}[\eta(x)\phi(f(x)) + (1-\eta(x))\phi(-f(x))] \tag{2}$$

(2) is from the following derivation:

$$\begin{aligned} L_\phi &= \mathbb{E}\phi(Yf(X)) \\ &= \mathbb{E}[\,\mathbb{E}[\phi(Yf(X))|X]\,] \\ &= \mathbb{E}\left[\underbrace{\eta(x)\phi(f(x))}_{y=1} + \underbrace{(1-\eta(x))\phi(-f(x))}_{y=-1}\right] \end{aligned}$$

We can minimize L_ϕ by minimizing $\eta(x)\phi(f(x)) + (1-\eta(x))\phi(-f(x))$ at every point x. So when we fix a point x, then $\eta(x) = \eta \in [0,1]$ and $f(x) = \alpha \in \mathbb{R}$.

Definition.

$$H(\eta) = \inf_{\alpha\in\mathbb{R}} \{\eta\phi(\alpha) + (1-\eta)\phi(-\alpha)\} \tag{3}$$

$$\alpha^*(\eta) = \arg \min_{\alpha\in\mathbb{R}\cup\{\pm\infty\}} \{\eta\phi(\alpha) + (1-\eta)\phi(-\alpha)\}$$

Clearly, if ϕ is diffenretiable α^* exists (solve by differentiation).

Example. In the SVM case where we have $\phi(\alpha) = (1-\alpha)_+$, when $\eta \notin \{0, \frac{1}{2}, 1\}$, we can see that $\alpha^*(\eta) = \text{sign}(\eta - \frac{1}{2})$, and correspondingly, $H(\eta) = 2\min(\eta, 1-\eta)$.

Definition.

$$H^-(\eta) = \inf_{\alpha\in\mathbb{R}:\alpha(2\eta-1)\leq 0} \{\eta\phi(\alpha) + (1-\eta)\phi(-\alpha)\} \tag{4}$$

Clearly, $H(\eta) \leq H^-(\eta)$ because $H^-(\eta)$ is an infimum over a proper subset compared to $H(\eta)$. We can view $H^-(\eta)$ as a special case of $H(\eta)$ where α has the "wrong sign". That is, when $\eta > 0$, $\alpha \leq 0$ and when $\eta < 0$, $\alpha \geq 0$. For a convex function ϕ, $H^-(\eta) = \phi(0)$ (for an intuition why this is true, look at Figure 1). $\eta\phi(\alpha)+(1-\eta)\phi(-\alpha)$ is a convex combination of the functions $\phi(\alpha)$ and $\phi(-\alpha)$ and Figure 1 shows the graph when $\eta = 0.7$ and when $\eta = 0.3$ (trivially, when $\eta = \frac{1}{2}$ the minimum occurs at $\alpha^* = 0$). $H^-(\eta)$ is simply the infimum when the function argument is restricted to lie on the wrong side, thus when the function is convex, $H^-(\eta) = \phi(0)$.

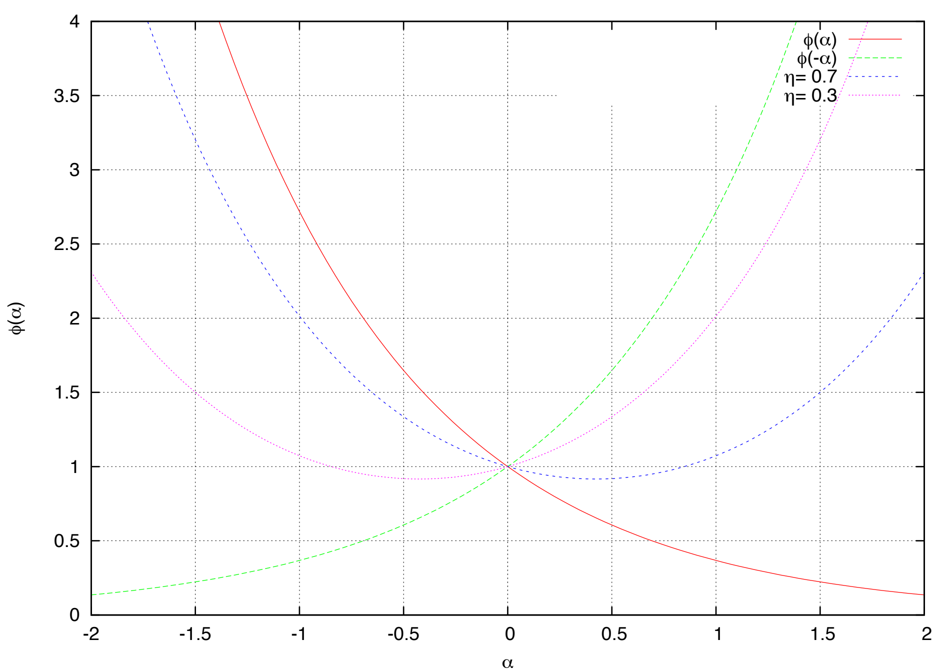

Figure 1: **Convex ϕ where $\phi(\alpha) = e^{-\alpha}$**

Definition. ϕ is *classification-calibrated* if for all $\eta \neq \frac{1}{2}$,

$$H(\eta) < H^-(\eta) \tag{5}$$

Example. Suppose $P_x(x_0) = 1$, we minimize L_ϕ by minimizing $\eta\phi(\alpha) + (1-\eta)\phi(-\alpha)$, where $\eta = \eta(x_0)$ and $\alpha = f(x_0)$. Also $L(f) - L^* = \mathbb{E}[|2\eta(x) - 1|\mathbb{I}[\text{sign}(f(x)) \neq \text{sign}(2\eta - 1)]]$. So for $\eta(x_0) \neq \frac{1}{2}$, we need $\text{sign}(f(x)) = \text{sign}(\eta - \frac{1}{2})$.

Lemma 2.1. A convex cost function ϕ is classification calibrated if and only if:

1. $\phi'(0)$ exists.
2. $\phi'(0) < 1$.

Proof Idea: The *if* direction is easy to show because ϕ is convex. The *only if* direction requires a bit more intuition. If $\phi'(0)$ does not exist then small pertubations of η around $\eta = \frac{1}{2}$ will not move the minimum away from $\phi(0)$. This is illustrated in Figure 2 when $\eta = 0.7$ and when $\eta = 0.3$, $H(\eta) = H^-(\eta) = \phi(0)$, which obviously violates our definition of ϕ being classification calibrated.

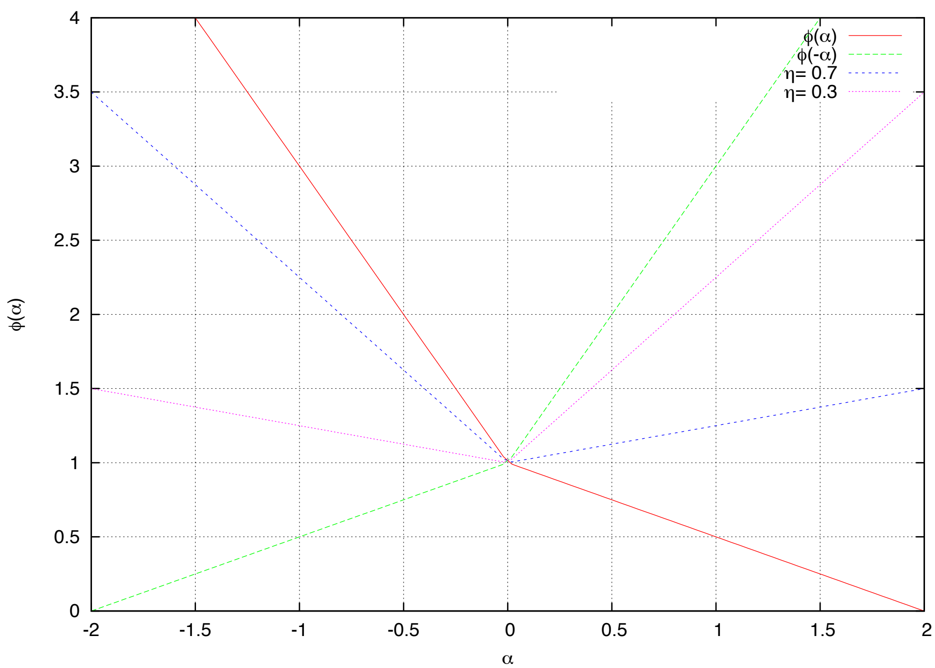

Figure 2: **Convex ϕ where $\phi'(0)$ does not exist.**

Theorem 2.2. For a convex and classification calibrated ϕ-loss function,

$$\psi\left(L(f) - L^*\right) \leq L_\phi(f) - L_\phi^*$$

where

$$\begin{aligned}\psi(\theta) &= H^-\left(\frac{1+\theta}{2}\right) - H\left(\frac{1+\theta}{2}\right)\\ &= \phi(0) - H\left(\frac{1+\theta}{2}\right)\end{aligned}$$

PROOF. H, H^-, and ϕ have the following properties:

1. $H(\eta) = H(1-\eta)$ and $H^-(\eta) = H^-(1-\eta)$
2. H is concave, thus ϕ is convex.
3. $\phi(0) = 0$
4. $H(\eta) < H^-(\eta) = \phi(0)$, when $\eta \neq \frac{1}{2}$ and ϕ convex and classification calibrated. Therefore, for all $\theta > 0$, $\psi(\theta) > 0$.
5. If $\alpha(2\eta - 1) \leq 0$, then $\eta\phi(\alpha) + (1-\eta)\phi(-\alpha) \geq H^-(\eta) = \phi(0)$

Recall Jensen's inequality: for a convex function g, $g(\mathbb{E}[x]) \leq \mathbb{E}[g(x)]$.

Also, recall from Lecture 1:

$$L(f) - L^* = \mathbb{E}\left[\mathbb{I}\left[\text{sign}(f(x)) \neq \text{sign}\left(\eta(x) - \frac{1}{2}\right)\right] |2\eta(x) - 1|\right]$$

Therefore,

$$\begin{aligned}
\psi(L(f) - L)^* &\leq \mathbb{E}\left[\psi\left(\mathbb{I}\left[\text{sign}(f(x)) \neq \text{sign}\left(\eta(x) - \frac{1}{2}\right)\right] |2\eta(x) - 1|\right)\right] \\
&= \mathbb{E}\left[\mathbb{I}\left[\text{sign}(f(x)) \neq \text{sign}\left(\eta(x) - \frac{1}{2}\right)\right] \psi\left(|2\eta(x) - 1|\right)\right] \\
&= \mathbb{E}\left[\mathbb{I}\left[\text{sign}(f(x)) \neq \text{sign}\left(\eta(x) - \frac{1}{2}\right)\right] \left(H^-(\eta(x)) - H(\eta(x))\right)\right]
\end{aligned}$$

For the first equality we used the fact that $\psi(0) = 0$. For the second, we used the definition of ψ and the symmetry of H, H^- about $\frac{1}{2}$.

If $\text{sign}(f(x)) \neq \text{sign}(\eta(x) - \frac{1}{2})$, then $\mathbb{E}[\phi(Yf(x))|X] \geq \inf_{\alpha:\alpha(2\eta(x)-1)\leq 0} \mathbb{E}[\phi(Y\alpha)|X] = H^-(\eta(x))$.

Otherwise, $\mathbb{E}[\phi(Yf(x))|X = x] \geq \inf_\alpha \mathbb{E}[\phi(Y\alpha)|X = x] = H(\eta(x))$

Thus for all x,

$$\begin{aligned}
\mathbb{E}\left[\mathbb{I}\left[\text{sign}(f(x)) \neq \text{sign}\left(\eta(x) - \frac{1}{2}\right)\right] \left(H^-(\eta(x)) - H(\eta(x))\right)\right] &\leq \mathbb{E}\left[\mathbb{E}[\phi(Yf(x))|X] - H(\eta(x))\right] \\
&= L_\phi(f) - L_\phi^*
\end{aligned}$$

Example. In the SVM case, $\phi(\alpha) = (1-\alpha)_+$. We found previously that $H(\eta) = 2\min(\eta, 1-\eta)$, and $H^-(\eta) = \phi(0) = 1$. Therefore,

$$\psi(\theta) = 1 - 2\min\left(\frac{1+\theta}{2}, 1 - \frac{1+\theta}{2}\right) = \theta$$

Thus for SVMs, we have $L(f) - L^* \leq L_\phi(f) - L_\phi^*$.

Example. In AdaBoost, $\phi(\alpha) = e^{-\alpha}$ (see Figure 1). With a bit of calculus we find that $\alpha^*(\eta) = \ln \sqrt{\frac{n}{1-n}}$, $H(\eta) = 2\sqrt{\eta(1-\eta)}$, and as usual, $H^-(\eta) = \phi(0) = 1$. Therefore,

$$\psi(\theta) = 1 - 2\sqrt{\left(\frac{1+\theta}{2}\right)\left(\frac{1-\theta}{2}\right)} = 1 - \sqrt{1-\theta^2}$$