

**Solutions 3**  
 Spring 2009

**Solution 3.1**

For part(i), we will use the fact that a convex loss function  $\phi$  is classification-calibrated if and only if it is differentiable at 0 and  $\phi'(0) < 0$ ; for part (ii) we will use the fact that if  $\phi$  is convex and classification-calibrated, then  $\Psi(\theta) = \phi(0) - H\left(\frac{1+\theta}{2}\right)$ ; for part (iii), as shown in the class, the dual program has the form

$$\max_{\alpha \in \mathbb{R}^n} \left\{ -\frac{1}{\lambda} \sum_{i=1}^n \phi^*(-\lambda\alpha_i) - \frac{1}{2} \alpha^T [(yy^T) \odot K] \alpha \right\},$$

where  $\phi^*(s) = \sup_{t \in \mathbb{R}} (st - \phi(t))$  is the conjugate dual function of  $\phi$ .

(a)  $\phi(t) = \log[1 + \exp(-t)]$ .

(i)  $\phi'(t) = -\frac{1}{e^t+1}$  and  $\phi''(t) = \frac{e^t}{(e^t+1)^2} > 0$ , hence  $\phi$  is convex.  $\phi'(0) = \frac{1}{2} < 0$  implies  $\phi$  is classification-calibrated.

(ii) By definition,  $H_\phi(\eta) = \inf_{\alpha \in \mathbb{R}} \{\eta \log(1 + e^{-\alpha}) + (1 - \eta) \log(1 + e^\alpha)\}$ . The infimum is achieved at  $\alpha^*(\eta) = \log\left(\frac{\eta}{1-\eta}\right)$  for  $\eta \in (0, 1)$ . Substituting it back gives us  $H_\phi(\eta) = -\eta \log \eta - (1 - \eta) \log(1 - \eta)$ .

$$\Psi(\theta) = \phi(0) - H\left(\frac{1+\theta}{2}\right) = \frac{1}{2} ((1 - \theta) \log(1 - \theta) + (1 + \theta) \log(1 + \theta)).$$

(iii)  $st - \log(1 + \exp(-t))$  is unbounded if  $s > 0$  or  $s < -1$ . For  $s \in (-1, 0)$ ,  $st - \log(1 + \exp(-t))$  achieves the maximum at  $t = \log\left(-\frac{1+s}{s}\right)$ , so we have  $\phi^*(s) = (1+s) \log(1+s) - s \log(-s)$ . For  $s = 0$  or  $s = -1$ ,  $\phi^*(s) = 0$ . With the interpretation  $0 \log 0 = 0$ , we have

$$\phi^*(s) = \begin{cases} (1+s) \log(1+s) - s \log(-s) & \text{if } s \in [-1, 0] \\ +\infty & \text{otherwise} \end{cases}$$

Therefore, the dual program is

$$\max_{\alpha \in \mathbb{R}^n} -\frac{1}{\lambda} \sum_i^n (1 - \lambda\alpha_i) \log(1 - \lambda\alpha_i) + \lambda\alpha_i \log(\lambda\alpha_i) - \frac{1}{2} \alpha^T [(yy^T) \odot K] \alpha$$

such that  $0 \leq \alpha_i \leq \frac{1}{\lambda} \quad \forall i = 1, \dots, n$

(b)  $\phi(t) = \exp(-t)$ .

- (i)  $\phi'(t) = -e^{-t}$  and  $\phi''(t) = e^{-t} > 0$ , hence  $\phi$  is convex.  $\phi'(0) = -1 < 0$  implies  $\phi$  is classification-calibrated.
- (ii) By definition,  $H_\phi(\eta) = \inf_{\alpha \in \mathbb{R}} \{\eta e^{-\alpha} + (1-\eta)e^\alpha\}$ . The infimum is achieved at  $\alpha^*(\eta) = \frac{1}{2} \log\left(\frac{\eta}{1-\eta}\right)$  for  $\eta \in (0, 1)$ . Substituting it back gives us  $H_\phi(\eta) = 2\sqrt{\eta(1-\eta)}$ .

$$\Psi(\theta) = \phi(0) - H\left(\frac{1+\theta}{2}\right) = 1 - \sqrt{1-\theta^2}.$$

- (iii)  $st - e^{-t}$  is unbounded if  $s > 0$ . For  $s < 0$ ,  $st - e^{-t}$  achieves the maximum at  $t = -\log(-s)$ , so we have  $\phi^*(s) = s - s \log(-s)$ . For  $s = 0$ ,  $\phi^*(s) = 0$ . By convention  $0 \log 0 = 0$ , we have

$$\phi^*(s) = \begin{cases} s - s \log(-s) & \text{if } s \leq 0 \\ +\infty & \text{otherwise} \end{cases}$$

Therefore, the dual program is

$$\begin{aligned} & \max_{\alpha \in \mathbb{R}^n} \sum_i^n (\alpha_i - \alpha_i \log(\lambda \alpha_i)) - \frac{1}{2} \alpha^T [(yy^T) \odot K] \alpha \\ & \text{such that } \alpha_i \geq 0 \quad \forall i = 1, \dots, n \end{aligned}$$

### Solution 3.2

(a)

$$\begin{aligned} \mathbb{P}\left[\min_{i=1, \dots, n} \|X^{(i)} - X\|_\infty > t\right] &= 1 - \mathbb{P}\left[\min_{i=1, \dots, n} \|X^{(i)} - X\|_\infty \leq t\right] \\ &= 1 - \mathbb{P}\left[\bigcup_{i=1, \dots, n} \{\|X^{(i)} - X\|_\infty \leq t\}\right] \\ &\geq 1 - n\mathbb{P}\left[\|X^{(1)} - X\|_\infty \leq t\right] \\ &= 1 - n\mathbb{P}\left[\bigcap_{j=1, \dots, d} \{|X_j^{(1)} - X_j| \leq t\}\right] \\ &= 1 - n\left(\mathbb{P}\left[|X_1^{(1)} - X_1| \leq t\right]\right)^d \\ &= 1 - n\left(1 - (1-t)^2\right)^d \\ &= 1 - n(2t - t^2)^d \\ &\geq 1 - n(2t)^d, \end{aligned}$$

where the union bound is used in the first inequality.

(b) The quantity of interest is  $\mathbb{P} [\min_{i=1,\dots,n} \|X^{(i)} - X\|_\infty \leq 1/4]$ . From part (a), we know that

$$\begin{aligned} \mathbb{P} \left[ \min_{i=1,\dots,n} \|X^{(i)} - X\|_\infty \leq 1/4 \right] &= 1 - \mathbb{P} \left[ \min_{i=1,\dots,n} \|X^{(i)} - X\|_\infty > 1/4 \right] \\ &\leq 1 - \left( 1 - n \left( \frac{1}{2} \right)^d \right) \\ &= n \left( \frac{1}{2} \right)^d. \end{aligned}$$

To guarantee  $\mathbb{P} [\min_{i=1,\dots,n} \|X^{(i)} - X\|_\infty \leq 1/4] \geq 1/2$ , we must make sure  $1/2 \leq n/2^d$ , which is equivalent to  $n \geq 2^{d-1}$ . Therefore, the number of samples must grow exponentially fast as the dimension increases.

(c) By Fubini's theorem,

$$\begin{aligned} \rho_\infty(d, n) &= \int_0^\infty \mathbb{P} \left[ \min_{i=1,\dots,n} \|X^{(i)} - X\|_\infty > t \right] dt \\ &\geq \int_0^\infty \left( 1 - n (2t)^d \right)_+ dt \\ &= \int_0^{\frac{1}{2}n^{-1/d}} \left( 1 - n (2t)^d \right) dt \\ &= \frac{d}{2(d+1)} n^{-1/d}. \end{aligned}$$

(d) The lower bounds for different values of  $d$  and  $n$  is as follows:

$n \backslash d$	1	10	20
100	$2.5 \times 10^{-3}$	0.2868	0.3783
1000	$2.5 \times 10^{-4}$	0.2278	0.3371
10000	$2.5 \times 10^{-5}$	0.1810	0.3005
100000	$2.5 \times 10^{-6}$	0.1437	0.2678

### Solution 3.3

In the follows, we assume  $P_\epsilon = \{\mathbb{B}_\epsilon(x_i), i = 1, \dots, M(\epsilon; S, \rho)\}$  is one of  $\epsilon$ -packings with the largest cardinality.

(a) Let  $C_\epsilon = \{\mathbb{B}_\epsilon(y_j), j = 1, \dots, N(\epsilon; S, \rho)\}$  be an  $\epsilon$ -covering with the smallest cardinality. By definition of  $\epsilon$ -covering, we know that for each  $x_i, i \in \{1, \dots, M(\epsilon; S, \rho)\}$ , there exists some  $y_j, j \in \{1, \dots, N(\epsilon; S, \rho)\}$  such that  $x_i \in \mathbb{B}_\epsilon(y_j)$ . If  $M(\epsilon; S, \rho) > N(\epsilon; S, \rho)$ , then by Pigeonhole principle, there must exist two distinct centers  $x_i, x_{i'}$  for some  $i, i' \in \{1, \dots, M(\epsilon; S, \rho)\}$  and  $y_j$  for some  $j \in \{1, \dots, N(\epsilon; S, \rho)\}$  such that  $x_i \in \mathbb{B}_\epsilon(y_j)$  and  $x_{i'} \in \mathbb{B}_\epsilon(y_j)$ . This is equivalent to  $y_j \in \mathbb{B}_\epsilon(x_i)$  and  $y_j \in \mathbb{B}_\epsilon(x_{i'})$ , which contradicts the fact that  $\mathbb{B}_\epsilon(x_i) \cap \mathbb{B}_\epsilon(x_{i'}) = \emptyset$ . Therefore,  $M(\epsilon; S, \rho) \leq N(\epsilon; S, \rho)$ .

- (b) We claim  $C_{2\epsilon} = \{\mathbb{B}_{2\epsilon}(x_i), i = 1, \dots, M(\epsilon; S, \rho)\}$  is a  $2\epsilon$ -covering of set  $S$ , which implies  $N(2\epsilon; S, \rho) \leq M(\epsilon; S, \rho)$ . If there exists an  $s \in S$  that cannot be covered by  $C_{2\epsilon}$ , then  $\rho(s, x_i) > 2\epsilon$  for all  $i = 1, \dots, M(\epsilon; S, \rho)$ . We can show that  $\mathbb{B}_\epsilon(s)$  is disjoint from all  $\epsilon$ -balls in  $P_\epsilon$ : suppose there exists  $i' \in \{1, \dots, M(\epsilon; S, \rho)\}$  such that  $y \in \mathbb{B}_\epsilon(s) \cap \mathbb{B}_\epsilon(x_{i'})$ , then  $\rho(s, x_{i'}) \leq \rho(s, y) + \rho(y, x_{i'}) \leq 2\epsilon$ . Therefore  $P_\epsilon \cup \mathbb{B}_\epsilon(s)$  is also an  $\epsilon$ -packing, which contradicts that  $P_\epsilon$  is the maximal  $\epsilon$ -packing.

### Solution 3.4

In this problem, the definition of packing number  $M(\epsilon; S, \|\cdot\|_2)$  is the maximum number of points in  $S$  such that the  $\ell_2$  metric between each pair is at least  $\epsilon$  (it differs from problem 3.3 by a factor of 2).

- (a) For each point  $x$  included in the packing, consider an  $\ell_2$  ball centered at  $x$  with radius  $\epsilon/2$ . We will call these  $\ell_2$  balls as packing balls. Because the distance between each pair in the packing is at least  $\epsilon$ , the packing balls are mutually disjoint. Therefore the maximum volumes covered by the packing balls are  $M(\epsilon; S, \|\cdot\|_2)c(\epsilon/2)^d$ , where  $c$  is a scaling constant. Because each point in the packing must belong to  $S$ , all the packing balls are indeed contained in an  $\ell_2$  ball with radius  $1 + \epsilon/2$ . Hence,

$$M(\epsilon; S, \|\cdot\|_2)c(\epsilon/2)^d \leq c(1 + \epsilon/2)^d \implies M(\epsilon; S, \|\cdot\|_2) \leq \left(\frac{2}{\epsilon} + 1\right)^d \leq \left(\frac{4}{\epsilon}\right)^d.$$

- (b) First we prove the concentration bound for chi-squared distributed random variable. Assume  $Z \sim \chi_d^2$ , we notice that  $\mathbb{E}[e^{sZ}] = (1 - 2s)^{-d/2}$  and apply the Chernoff bound,

$$\begin{aligned} \mathbb{P}[Z \geq (1 + \delta)d] &\leq e^{-s(1+\delta)d} (1 - 2s)^{-d/2} \\ &= \exp\{-s(1 + \delta)d - d/2 \log(1 - 2s)\} \\ &\leq \exp\{-d(\delta - \log(1 + \delta))/2\} \\ &\leq \exp(-d\delta^2/16), \end{aligned}$$

where in the second inequality we set  $s = \frac{\delta}{2(1+\delta)}$ . Similarly, we have  $\mathbb{P}[Z \leq (1 - \delta)d] \leq \exp(-d\delta^2/16)$ . Next we write

$$\begin{aligned} \|x_i - x_j\|_2 &= \left\| \frac{z_i}{\|z_i\|_2} - \frac{z_j}{\|z_j\|_2} \right\|_2 \\ &= \left\| \frac{z_i - z_j}{\sqrt{d}} - \frac{z_i}{\sqrt{d}} \left(1 - \frac{\sqrt{d}}{\|z_i\|_2}\right) + \frac{z_j}{\sqrt{d}} \left(1 - \frac{\sqrt{d}}{\|z_j\|_2}\right) \right\|_2 \\ &\geq \frac{\|z_i - z_j\|_2}{\sqrt{d}} - \frac{\|z_i\|_2}{\sqrt{d}} \left|1 - \frac{\sqrt{d}}{\|z_i\|_2}\right| - \frac{\|z_j\|_2}{\sqrt{d}} \left|1 - \frac{\sqrt{d}}{\|z_j\|_2}\right|. \end{aligned}$$

Notice that  $\frac{\|z_i - z_j\|_2^2}{2} \sim \chi_d^2$ ,  $\|z_i\|_2^2 \sim \chi_d^2$ ,  $\|z_j\|_2^2 \sim \chi_d^2$ . Applying the concentration bound derived above, we get

$$\begin{aligned} \mathbb{P} \left[ \frac{\|z_i - z_j\|_2}{\sqrt{d}} \geq \sqrt{2(1-\delta)} \right] &= \mathbb{P} \left[ \frac{\|z_i - z_j\|_2^2}{2} \geq d(1-\delta) \right] \\ &\geq 1 - \exp(-d\delta^2/16) \end{aligned}$$

$$\begin{aligned} \mathbb{P} \left[ \frac{\|z_i\|_2}{\sqrt{d}} \leq \sqrt{1+\delta} \right] &= \mathbb{P} \left[ \|z_i\|_2^2 \leq d(1+\delta) \right] \\ &\geq \mathbb{P} \left[ d(1-\delta) \leq \|z_i\|_2^2 \leq d(1+\delta) \right] \\ &\geq 1 - 2 \exp(-d\delta^2/16) \end{aligned}$$

$$\begin{aligned} \mathbb{P} \left[ \left| 1 - \frac{\sqrt{d}}{\|z_i\|_2} \right| \leq \frac{1}{\sqrt{1-\delta}} - 1 \right] &= \mathbb{P} \left[ 1 - \frac{1}{\sqrt{1-\delta}} \leq \frac{\sqrt{d}}{\|z_i\|_2} - 1 \leq \frac{1}{\sqrt{1-\delta}} - 1 \right] \\ &\geq \mathbb{P} \left[ \frac{1}{\sqrt{1+\delta}} - 1 \leq \frac{\sqrt{d}}{\|z_i\|_2} - 1 \leq \frac{1}{\sqrt{1-\delta}} - 1 \right] \\ &= \mathbb{P} \left[ d(1-\delta) \leq \|z_i\|_2^2 \leq d(1+\delta) \right] \\ &\geq 1 - 2 \exp(-d\delta^2/16) \end{aligned}$$

Therefore, with probability at least  $1 - \exp(-d\delta^2/16) + 2(1 - 2 \exp(-d\delta^2/16)) - 2 = 1 - 5 \exp(-d\delta^2/16)$ ,

$$\|x_i - x_j\|_2 \geq f(\delta) = \sqrt{2(1-\delta)} - 2\sqrt{1+\delta} \left( \frac{1}{\sqrt{1-\delta}} - 1 \right).$$

Now we can set appropriate  $\delta_0$  to let  $f(\delta_0) = 1/4$  to guarantee

$$\mathbb{P} [\|x_i - x_j\|_2 \leq 1/4] \leq 5 \exp(-d\delta_0^2/16).$$

This bound indicates that the probability that the distance between two uniformly random points on the surface of unit  $\ell_2$  ball is less than  $1/4$  is exponentially small.

For part (ii), we use the following probabilistic argument: randomly choose  $M$  points  $x_1, \dots, x_M$  on the surface of unit  $\ell_2$  ball, the probability that they form a  $1/4$ -packing is

$$\begin{aligned} \mathbb{P} \left[ \bigcap_{i \neq j} \left\{ \|x_i - x_j\|_2 > \frac{1}{4} \right\} \right] &= 1 - \mathbb{P} \left[ \bigcup_{i \neq j} \left\{ \|x_i - x_j\|_2 \leq \frac{1}{4} \right\} \right] \\ &\geq 1 - \frac{M(M-1)}{2} \mathbb{P} [\|x_1 - x_2\|_2 \leq 1/4] \\ &> 1 - \frac{M^2}{2} 5 \exp(-d\delta_0^2/16). \end{aligned}$$

So for  $M = \sqrt{2/5} \exp(d\delta_0^2/32)$ , we can guarantee the probability above is strictly greater than 0, which implies there must exist a  $1/4$ -packing of size  $M$ . Hence  $M(1/4; S, \|\cdot\|_2) \geq \sqrt{2/5} \exp(d\delta_0^2/32)$ .