1. [35pts] Support Vector Machine

(1) Recall that the soft margin support vector machine solves the problem:

$$\min \quad \frac{1}{2}w^\top w + C\sum_i \varepsilon_i$$

$$\text{s.t. } \; y_i(w^\top x_i + b) \geq 1 - \varepsilon_i, \quad \varepsilon_i \geq 0.$$

a) [10pts] Derive its dual problem using the method of Lagrange multipliers.

b) [10pts] Further simplify the dual problem when at its saddle point to prove

$$\max_{\alpha} \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^\top x_j$$

$$\text{s.t. } C \geq \alpha_i \geq 0, \quad \sum_i \alpha_i y_i = 0,$$

is equivalent to the primal problem.

a) We can write down its dual problem by applying the method of Lagrange multipliers easily

$$L(w,b,\epsilon,\alpha,\ \mu) = \frac{1}{2}w^\top w + C\sum_i \epsilon_i - \sum_i \alpha_i\,[y_i(w^\top x_i + b) - 1 + \epsilon_i] - \sum_i \mu_i \epsilon_i \tag{1}$$

Then from Lagrange duality, we know dual problem is as follows

$$\max_{\alpha,\mu} \min_{w,b,\epsilon} L(w,b,\epsilon,\alpha,\mu) \tag{2}$$

$$s.t.\ \alpha_i \geq 0,\ \mu_i \geq 0 \tag{3}$$

b) In order to solve the dual problem, we need to find the minimum of $L(w,b,\epsilon,\alpha,\mu)$ to w, b,ϵ first, and then find the maximum to α.
Find the partial derivatives of L with respect to w, b and ϵ_i respectively, and set them equal to zero

$$\frac{\partial L}{\partial w} = w - \sum_i \alpha_i y_i x_i = 0$$

$$\frac{\partial L}{\partial b} = -\sum_i \alpha_i y_i = 0$$

$$\frac{\partial L}{\partial \epsilon_i} = C - \alpha_i - \mu_i = 0$$

Obtain that

$$w = \sum_i \alpha_i y_i x_i \tag{4}$$

$$\sum_i \alpha_i y_i = 0 \tag{5}$$

$$C - \alpha_i - \mu_i = 0 \tag{6}$$

Substitute (4) (5) (6) them into (2), obtain

$$\min_{w,b,\epsilon} L(w,b,\epsilon,\alpha,\mu) = -\frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i^{\mathsf{T}} x_j + \sum_i \alpha_i \tag{7}$$

The dual problem can be obtained by finding the maximum value of α for $\min_{w,b,\epsilon} L(w,b,\epsilon,\alpha,\mu)$,

$$\max \sum_i \alpha_i - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i^{\mathsf{T}} x_j \tag{8}$$

$$s.t. \sum_i \alpha_i y_i = 0 \tag{9}$$

$$C - \alpha_i - \mu_i = 0 \tag{10}$$

$$\alpha_i \geq 0, \mu_i \geq 0 \tag{11}$$

From (11) (10), we know that $0 \leq \alpha_i = C - \mu_i \leq C$. Finally,

$$\max \sum_i \alpha_i - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i^{\mathsf{T}} x_j$$

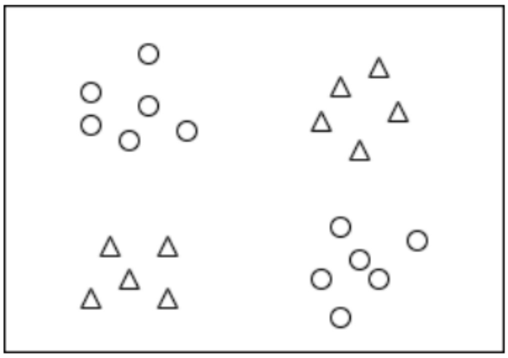
$$s.t. \sum_i \alpha_i y_i = 0$$

$$0 \leq \alpha_i \leq C_i$$

Is proved to be equivalent to the primal problem.

(2) [15pts] Given the XOR sample points as below, we train an SVM with a quadratic kernel, i.e. our kernel function is a polynomial kernel of degree 2: $\kappa(x_i, x_j) = (x_i^T x_j)^d, d = 2$.

(a) [5pts] what is the corresponding mapping function $\phi(x)$?



(b) [5pts] Use the following code to generate XOR data, and according to the answer of (a), map the data with $\phi(x)$ to see if it can be linearly separable.

(c) [5pts] Could we get a reasonable model with hard margin? If yes, draw the decision boundary in the figure (original feature space), otherwise state reasons.

```
import numpy as np
import matplotlib.pyplot as plt
#创建数据
X_xor = np.random.randn(40,2)
y_xor = np.logical_xor(X_xor[:,0]>0, X_xor[:,1]>0)
y_xor = np.where(y_xor, 1, -1)
#绘制散点图
plt.scatter(x=X_xor[y_xor==1,0]),   # 横坐标
            y=X_xor[y_xor==1,1]),   # 纵坐标
            color='g', marker='x', label='1')
plt.scatter(x=X_xor[y_xor==-1,0]),
            y=X_xor[y_xor==-1,1]),
            color='b', marker='o', label='-1')
plt.legend() #显示图例
plt.show()
```

(a) Polynomial kernel is $\kappa(x_i, x_j) = (\gamma \cdot x_i^T x_j + r)^d$, and here, we specify that $d = 2$. Now we need to find the corresponding feature space $\mathcal{H}$ and mapping $\phi(x): R^2 \rightarrow \mathcal{H}$.

Let $x_i = \left(x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(n)}\right)^{\mathrm{T}}$ $x_j = \left(x_j^{(1)}, x_j^{(2)}, \cdots, x_j^{(n)}\right)^{\mathrm{T}}$, then

$$(\gamma \cdot x_i^{\mathsf{T}} x_j + r)^2 = \gamma^2 (x_i^{\mathsf{T}} x_j)^2 + 2r\gamma x_i^{\mathsf{T}} x_j + r^2$$

$$= \gamma^2 \sum_m \sum_n x_i^{(m)} x_j^{(n)} x_i^{(m)} x_j^{(n)} + 2r\gamma \sum_k x_i^{(k)} x_j^{(k)} + r^2$$

$$= \left(\gamma {x_i^{(1)}}^2, \gamma x_i^{(1)} x_i^{(2)}, \gamma x_i^{(1)} x_i^{(3)} \cdots, {x_i^{(n)}}^2, \sqrt{2r\gamma} x_i^{(1)}, \sqrt{2r\gamma} x_i^{(2)}, \cdots, \sqrt{2r\gamma} x_i^{(n)}, r\right)^{\mathsf{T}}$$

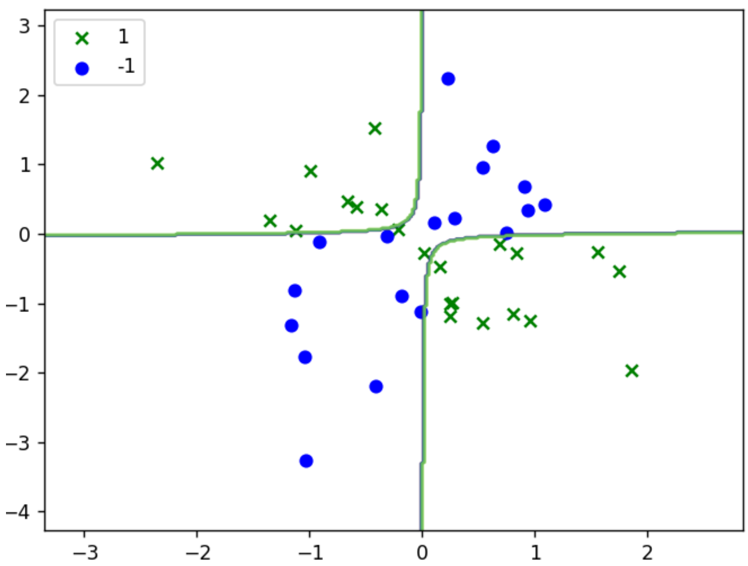
$$\cdot \left(\gamma {x_j^{(1)}}^2, \gamma x_j^{(1)} x_j^{(2)}, \gamma x_j^{(1)} x_j^{(3)} \cdots, {x_j^{(n)}}^2, \sqrt{2r\gamma} x_j^{(1)}, \sqrt{2r\gamma} x_j^{(2)}, \cdots, \sqrt{2r\gamma} x_j^{(n)}, r\right)$$

$$= \phi(x_i) \cdot \phi(x_j)$$

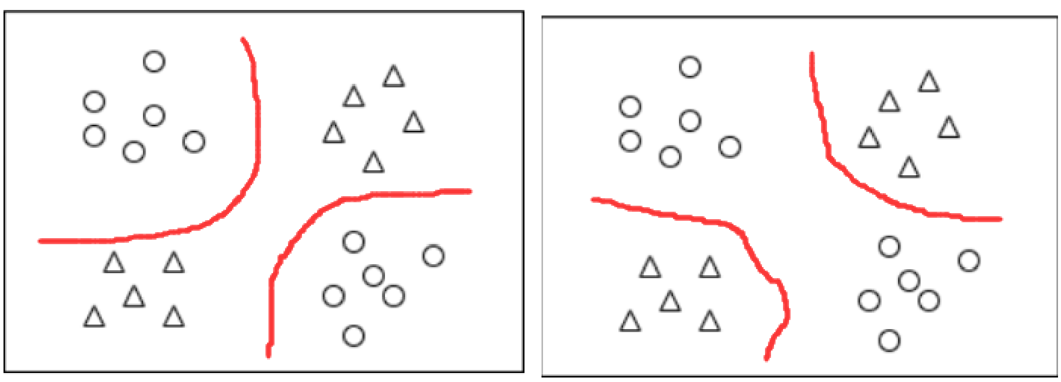
Therefore,

$$\phi(x) = (\gamma {x^{(1)}}^2, \gamma x^{(1)} x^{(2)}, \gamma x^{(1)} x^{(3)} \cdots, {x^{(n)}}^2, \sqrt{2r\gamma} x^{(1)}, \sqrt{2r\gamma} x^{(2)}, \cdots, \sqrt{2r\gamma} x^{(n)}, r)$$

(b) After applying the mapping function $\phi(x)$ above, we can obtain the following figure which shows that $\phi(x)$ is linearly separable.



(c) Yes, we can get a reasonable hard margin. Here are two possible instances.



2. [30pts] Kernel Functions

(1) **[15 pts]** 对于 $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^N$，考虑函数 $\kappa(x,y) = \tanh(a\boldsymbol{x}^\top \boldsymbol{y} + b)$，其中 a, b 是任意实数。试说明 $a \geq 0, b \geq 0$ 是 κ 为核函数的必要条件。

(2) **[15 pts]** 考虑 $\mathbb{R}^N$ 上的函数 $\kappa(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}^\top \boldsymbol{y} + c)^d$，其中 c 是任意实数，d, N 是任意正整数。试分析函数 κ 何时是核函数，何时不是核函数，并说明理由。

说明：该核函数是多项式核的更一般的形式。

（第（3）小题是 extra 部分，可选）

(3) [10 pts] 当上一小问中的函数是核函数时，考虑 $d = 2$ 的情况，此时 κ 将 N 维数据映射到了什么空间中？具体的映射函数是什么？更一般的，对 d 不加限制时，κ 将 N 维数据映射到了什么空间中？(本小问的最后一问可以只写结果)

(1) Recall the definition of a positive definite kernel

Let $X \subset R^n$,$\kappa(x,z)$ be a symmetric function defined on $X \times X$. If the Gram matrix

$$K = \left[\kappa(x_i, x_j)\right]_{m \times m}$$

corresponding to any $x_i \in X$,$i = 1,2,\cdots,m$. $\kappa(x,z)$ is a semi-positive definite matrix, then $\kappa(x,z)$ is a positive definite kernel.

Now, consider $m = 1$ and $\boldsymbol{x} = (x, 0, \cdots, 0)$, then Gram matrix is $K = [\tanh ax^2 + b]$. Since $\tanh y \geq 0$, if and only if $y \geq 0$, we know that the necessary condition for κ be kernel function is $ax^2 + b \geq 0$. When $a = 0$, if and only if $b \geq 0$, the original equality holds. When $a \neq 0$, note that $\Delta = -4ab$, hence if and only if $a > 0$ and $b \geq 0$, the original equality holds. In conclusion, $a \geq 0$ and $b \geq 0$ is necessary condition for κ be kernel function.

(2) $\kappa(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}^\top \boldsymbol{y} + c)^d = (\boldsymbol{x}^\top \boldsymbol{y} + c) \cdots (\boldsymbol{x}^\top \boldsymbol{y} + c)$
When $c \geq 0$, recall some conclusions:

- If κ_1 and κ_2 are kernel functions, then $\kappa_1 \otimes \kappa_2 = \kappa_1(x,z) \cdot \kappa_2(x,z)$ is kernel function.
- If κ_1 and κ_2 are kernel functions, then for any $\gamma_1, \gamma_2 > 0$, their linear combination is kernel function.

$\boldsymbol{x}^\top \boldsymbol{y}$ is linear kernel and c is constant kernel, hence $\boldsymbol{x}^\top \boldsymbol{y} + c$ is kernel function, hence $\kappa(\boldsymbol{x}, \boldsymbol{y})$ is kernel function.

When $c < 0$, consider $m = 2$ and $\boldsymbol{x} = \left(\sqrt{-2c}, 0, \cdots, 0\right), \boldsymbol{y} = \left(-\sqrt{-2c}, 0, \cdots, 0\right)$, then Gram matrix is

$$K = \begin{bmatrix} (-c)^d & (3c)^d \\ (3c)^d & (-c)^d \end{bmatrix}$$

Since $|K| = (1 - 3^{2d})c^{2d} < 0$, we know K is not semi positive definite, thus κ is not a kernel function.

(3) When $d=2$, κ maps N-dimension data into $\binom{N+2}{2}$. Let $\boldsymbol{x}=(x_1,\cdots,x_N)$, then mapping function is

$$\phi(x)=(x_1^2,x_1x_2,x_1x_3\cdots,x_n^2,\sqrt{2c}x_1,\sqrt{2c}x_2,\cdots,\sqrt{2c}x_n,c)$$

More generally, κ maps N-dimension data into $\binom{N+d}{d}$.

3. [35 pts] Kernel Methods

请给出 kernel PCA 的推导过程。

核函数：$\kappa: R^N \times R^N \to R$,输入两个 N 维向量得到一个数值，可以看为两个变换后向量的内积，即$\kappa(x_i,x_j)=\phi(x_i)\cdot\phi(x_j)$。定义去均后的向量为：$\tilde{\phi}(x)=\phi(x)-\frac{1}{N}\sum_i\phi(x_i)$.

则协方差矩阵为:

$$\boldsymbol{C}=\frac{1}{N}\sum_i\tilde{\phi}(x_i)\tilde{\phi}(x_i)^\top=\frac{1}{N}\widetilde{\boldsymbol{\phi}}(\boldsymbol{x})\widetilde{\boldsymbol{\phi}}(\boldsymbol{x})^\top \tag{1}$$

kernel-PCA 指的是在进行映射后的空间内进行主成分分析，即求解

$$\boldsymbol{CW}=\lambda\boldsymbol{W} \tag{2}$$

中的$W\in R^{d\times N'}$作为子空间的N'个正交单位基向量。而由定理：空间中的任一向量都可以由该空间中的所有样本线性表示，所以可以记$\boldsymbol{W}=\sum_i\alpha_i\widetilde{\boldsymbol{\phi}}(x_i)=\widetilde{\boldsymbol{\phi}}(\boldsymbol{x})\boldsymbol{A}$，再代入(2)中，得

$$\boldsymbol{C}\widetilde{\boldsymbol{\phi}}(\boldsymbol{x})\boldsymbol{A}=\lambda\widetilde{\boldsymbol{\phi}}(\boldsymbol{x})\boldsymbol{A} \tag{3}$$

进一步，(3)式两边同时左乘$\boldsymbol{\phi}(\boldsymbol{x})^\top$并将 C 展开

$$\frac{1}{N}\widetilde{\boldsymbol{\phi}}(\boldsymbol{x})^\top\widetilde{\boldsymbol{\phi}}(\boldsymbol{x})\widetilde{\boldsymbol{\phi}}(\boldsymbol{x})^\top\widetilde{\boldsymbol{\phi}}(\boldsymbol{x})\boldsymbol{A}=\lambda\widetilde{\boldsymbol{\phi}}(\boldsymbol{x})^\top\widetilde{\boldsymbol{\phi}}(\boldsymbol{x})\boldsymbol{A} \tag{4}$$

将$\kappa=\widetilde{\boldsymbol{\phi}}(\boldsymbol{x})^\top\widetilde{\boldsymbol{\phi}}(\boldsymbol{x})$代入到（4）式当中，得到

$$\frac{1}{N}\kappa^2\boldsymbol{A}=\lambda\kappa\boldsymbol{A} \tag{5}$$

为了求解（5）式，我们需要求解

$$\frac{1}{N}\kappa\boldsymbol{A}=\lambda\boldsymbol{A} \tag{6}$$

而（6）式就是 PCA 的基本形式。

4. [extra, 30pts] Surrogate Function in SVM & Bayesian Optimal Classifier

（本题可选）

在软间隔支持向量机问题中，我们的优化目标为

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m} \ell_{0/1}\left(y_i\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b\right) - 1\right). \tag{1}$$

然而 $\ell_{0/1}$ 数学性质不太好，它非凸、非连续，使得式 (1) 难以求解。实践中我们通常会将其替换为"替代损失"，替代损失一般是连续的凸函数，且为 $\ell_{0/1}$ 的上界，比如 hinge 损失，指数损失，对率损失。下面我们证明在一定的条件下，这样的替换可以保证最优解不变。

我们考虑实值函数 $h: \mathcal{X} \to \mathbb{R}$ 构成的假设空间，其对应的二分类器 $f_h: \mathcal{X} \to \{+1, -1\}$ 为

$$f_h(x) = \begin{cases} +1 & \text{if } h(x) \geq 0 \\ -1 & \text{if } h(x) < 0 \end{cases}$$

h 的期望损失为 $R(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[I_{f_h(x)\neq y}\right]$，其中 I 为指示函数。设 $\eta(x) = \mathbb{P}(y = +1|x)$，则贝叶斯最优分类器当 $\eta(x) \geq \frac{1}{2}$ 时输出 1，否则输出 -1。因此可以定义贝叶斯得分 $h^*(x) = \eta(x) - \frac{1}{2}$ 和贝叶斯误差 $R^* = R(h^*)$。

设 $\Phi: \mathbb{R} \to \mathbb{R}$ 为非减的凸函数且满足 $\forall u \in \mathbb{R}, 1_{u\leq 0} \leq \Phi(-u)$。对于样本 (x, y)，定义函数 h 在该样本的 Φ-损失为 $\Phi(-yh(x))$，则 h 的期望损失为 $\mathcal{L}_\Phi(h) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\Phi(-yh(x))]$。定义 $L_\Phi(x, u) = \eta(x)\Phi(-u) + (1 - \eta(x))\Phi(u)$，设 $h_\Phi^*(x) = \mathop{\mathrm{argmin}}_{u\in[-\infty,+\infty]} L_\Phi(x, u)$，$\mathcal{L}_\Phi^* = \mathcal{L}_\Phi(h_\Phi^*(x))$。

我们考虑如下定理的证明：

若对于 Φ，存在 $s \geq 1$ 和 $c > 0$ 满足对 $\forall x \in \mathcal{X}$ 有

$$|h^*(x)|^s = \left|\eta(x) - \frac{1}{2}\right|^s \leq c^s\left[L_\Phi(x, 0) - L_\Phi\left(x, h_\Phi^*(x)\right)\right] \tag{2}$$

则对于任何假设 h，有如下不等式成立

$$R(h) - R^* \leq 2c\left[\mathcal{L}_\Phi(h) - \mathcal{L}_\Phi^*\right]^{\frac{1}{s}} \tag{3}$$

(1) **[5 pts]** 请证明

$$\Phi\left(-2h^*(x)h(x)\right) \leq L_\Phi(x, h(x)) \tag{4}$$

(2) **[10 pts]** 请证明

$$R(h) - R^* = 2\mathop{\mathbb{E}}_{x\sim\mathcal{D}_x}\left[|h^*(x)|\, 1_{h(x)h^*(x)\leq 0}\right] \tag{5}$$

提示：先证明

$$R(h) = \mathop{\mathbb{E}}_{x\sim\mathcal{D}_x}\left[2h^*(x)1_{h(x)<0} + (1 - \eta(x))\right]$$

(3) **[10 pts]** 利用式 (4) 和式 (5) 完成定理的证明。

(4) **[5 pts]** 请验证对于 Hinge 损失 $\Phi(u) = \max(0, 1 + u)$，有 $s = 1, c = \frac{1}{2}$。

(1)

$$\Phi\left(-2h^*(x)h(x)\right) = \Phi\left(\left(1 - 2\eta(x)\right)h(x)\right)$$

$$= \Phi\left(\eta(x)\big(-h(x)\big) + \big(1-\eta(x)\big)h(x)\right)$$
$$\le \eta(x)\Phi\big(-h(x)\big) + \big(1-\eta(x)\big)\Phi\big(h(x)\big)$$
$$= L_\Phi(x, h(x))$$

(2) First, we need to prove $R(h) = \mathbb{E}_{x\sim\mathcal{D}_x}\left[2h^*(x)I_{h(x)<0} + \big(1-\eta(x)\big)\right]$

$$
\begin{aligned}
R(h) &= \mathbb{E}_{x\sim\mathcal{D}_x}\left[I_{f_h(x)\neq y}\right] \\
&= \mathbb{E}_{x\sim\mathcal{D}_x}\left[\eta(x)I_{h(x)<0} + \big(1-\eta(x)\big)I_{h(x)\ge 0}\right] \\
&= \mathbb{E}_{x\sim\mathcal{D}_x}\left[\eta(x)I_{h(x)<0} + \big(1-\eta(x)\big)(1-I_{h(x)<0})\right] \\
&= \mathbb{E}_{x\sim\mathcal{D}_x}\left[(2\eta(x)-1)I_{h(x)<0} + \big(1-\eta(x)\big)\right] \\
&= \mathbb{E}_{x\sim\mathcal{D}_x}\left[2h^*(x)I_{h(x)<0} + \big(1-\eta(x)\big)\right]
\end{aligned}
$$

Then

$$
\begin{aligned}
R(h) - R(h^*) &= \mathbb{E}_{x\sim\mathcal{D}_x}\left[2h^*(x)(I_{h(x)<0} - I_{h^*(x)<0}\right] \\
&\le 2\mathbb{E}_{x\sim\mathcal{D}_x}\left[|h^*(x)|I_{h^*(x)\le 0}\right]
\end{aligned}
$$

(3)

$$R(h) - R(h^*) = \mathbb{E}_{x\sim\mathcal{D}_x}\left[|2\eta(x)-1|I_{h(x)h^*(x)\le 0}\right]$$

$$\le \left[\mathbb{E}_{x\sim\mathcal{D}_x}[|2\eta(x)-1|]^s I_{h(x)h^*(x)\le 0}\right]^{\frac{1}{s}} \qquad (Jensen\ unequality)$$

$$\le 2c\left[\mathbb{E}_{x\sim\mathcal{D}_x}\left[\Phi(0) - L_\Phi\big(x, h_\Phi^*(x)\big)\right]I_{h(x)h^*(x)\le 0}\right]^{\frac{1}{s}} \qquad (assumption)$$

$$\le 2c\left[\mathbb{E}_{x\sim\mathcal{D}_x}\left[\Phi\big(-2h^*(x)h(x)\big) - L_\Phi\big(x, h_\Phi^*(x)\big)\right]I_{h(x)h^*(x)\le 0}\right]^{\frac{1}{s}}$$

$$\le 2c\left[\mathbb{E}_{x\sim\mathcal{D}_x}\left[L_\Phi\big(x, h(x)\big) - L_\Phi\big(x, h_\Phi^*(x)\big)\right]I_{h(x)h^*(x)\le 0}\right]^{\frac{1}{s}}$$

$$\le 2c\left[\mathbb{E}_{x\sim\mathcal{D}_x}\left[L_\Phi\big(x, h(x)\big) - L_\Phi\big(x, h_\Phi^*(x)\big)\right]\right]^{\frac{1}{s}}$$

$$= 2c[\mathcal{L}_\Phi(h) - \mathcal{L}_\Phi^*]^{\frac{1}{s}}$$

(4) When $\Phi(u) = \max(0, 1+u)$,

$$
\begin{aligned}
\mathcal{L}_\Phi(x,u) &= \eta(x)\max(0, 1-u) + \big(1-\eta(x)\big)\max(0, 1+u) \\
&= \begin{cases} (1-u)\eta(x) & u < -1 \\ 1 + \big(1+2\eta(x)\big)u & -1 \le u < 1 \\ (1+u)\big(1-\eta(x)\big) & u \ge 1 \end{cases}
\end{aligned}
$$

Since $h_\Phi^*(x) = argmin_{u\in[-\infty,+\infty]}\mathcal{L}_\Phi(x,u)$, when $\eta(x) > \frac{1}{2}$, $h_\Phi^* = 1$, $L_\Phi\big(x, h_\Phi^*(x)\big) = 2\big(1-\eta(x)\big)$; when $\eta(x) \le \frac{1}{2}$, $h_\Phi^* = -1$, $L_\Phi\big(x, h_\Phi^*(x)\big) = 2\eta(x)$. Moreover, $\mathcal{L}(x, 0) = \Phi(0) = 1$, so $s = 1, c = \frac{1}{2}$.