1. (a)

I did this homework with Luning Zhao.
We work on our homework seperately, but we discuss when meeting problems.
Homework is fine.

(b) I certify that all solutions are entirely in my words. and that I have not looked at another student's solutions. I have credited all external sources in this write up.

Siyao Jia

2. (a) The corresponding unconstrained optimization problem is:

$$\min_{\vec{w}, b} \frac{1}{2}\|\vec{w}\|^2 + C\sum_{i=1}^{n} \phi_i \max(1 - y_i(\vec{w}^T x_i - b), 0)$$

(b)

$$\min_{\vec{w}, b, \xi_i} \frac{1}{2}\|\vec{w}\|^2 + C\sum_{i=1}^{n} \xi_i \phi_i$$

s.t. $\xi_i \geq 0 \quad \rightarrow \beta_i$

$\xi_i \geq 1 - y_i(x_i^T w - b) \iff y_i(x_i^T w - b) - (1-\xi_i) \geq 0 \quad \rightarrow \alpha_i$

$$L(\vec{w}, b, \vec{\xi}, \vec{\alpha}, \vec{\beta}) = \frac{1}{2}\|\vec{w}\|^2 + C\sum_{i=1}^{n}\phi_i \xi_i - \sum_{i=1}^{n}\alpha_i [y_i(x_i^T w - b) - (1-\xi_i)] - \sum_{i=1}^{n}\beta_i \xi_i$$

$$= \frac{1}{2}\|\vec{w}\|^2 + \sum_{i=1}^{n}\alpha_i y_i (x_i^T w - b) + \sum_{i=1}^{n}\alpha_i + \sum_{i=1}^{n}(C\phi_i - \alpha_i - \beta_i)\xi_i$$

$\frac{\partial L}{\partial \vec{w}} = 0 \Rightarrow w^T = \sum_{i=1}^{n}\alpha_i y_i x_i^T$ ①

$\frac{\partial L}{\partial b} = 0 \Rightarrow \Sigma \alpha_i y_i = 0$ ②

$\frac{\partial L}{\partial \xi_i} = 0 \Rightarrow C\phi_i - \alpha_i - \beta_i = 0$ ③ where $\alpha_i \geq 0$, $\beta_i \geq 0$, $\xi_i \geq 0$.

Simplify L using ① ② ③.

$$L(\vec{w}, b, \vec{\xi}, \vec{\alpha}, \vec{\beta}) = \frac{1}{2}\|w^*\|^2 - \overset{w^{*T}\ ①}{\cancel{\sum_{i=1}^{n}(\alpha_i^* y_i x_i^T)}} w^* + \overset{0\ ②}{\cancel{\sum_{i=1}^{n}\alpha_i^* y_i}}\, b^*$$

$$+ \Sigma\alpha_i^* + \Sigma\, \overset{0\ ③}{\cancel{(C\phi_i - \alpha_i^* - \beta_i^*)}}\,\xi_i^*$$

$$= \frac{1}{2}\|w^*\|^2 - \|w^*\|^2 + \Sigma \alpha_i^*$$

$$= -\frac{1}{2}\|w^*\|^2 + \Sigma\alpha_i^*$$

$$= -\frac{1}{2}\alpha^T Q \alpha + \alpha^T 1$$

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

$\beta_i = C\phi_i - \alpha_i \geq 0 \Rightarrow 0 \leq \alpha_i \leq C\phi_i$

(c) Q can be kernalized as $Q_{ij} = y_i K(x_i, x_j) y_j$. ϕ_i sets the contrant for α_i.

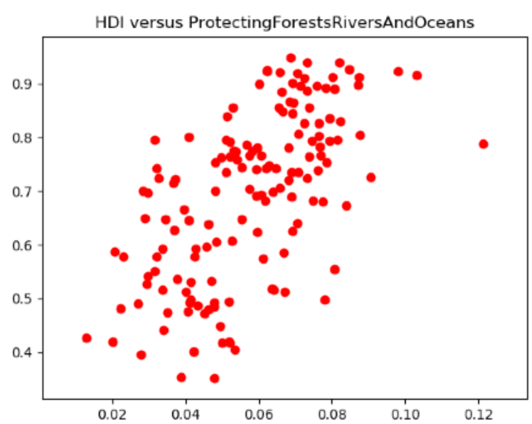
3 (a) take the survey online

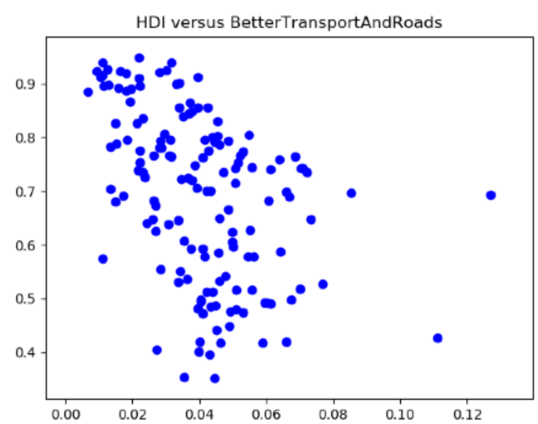
(b)

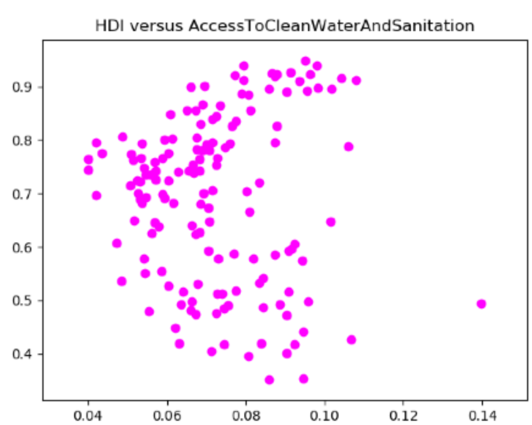
Action taken on climate change 0.473312891543
Better transport and roads -0.439633638622
Support for people who can't work -0.336213236721
Access to clean water and sanitation -0.018169084456
Better healthcare -0.422012359959
A good education -0.303978889772
A responsive government we can trust 0.329445314984
Phone and internet access -0.351604712158
Reliable energy at home -0.285423563836
Affordable and nutritious food 0.195193300786
Protecting forests rivers and oceans 0.613458756271
Protection against crime and violence 0.14331869918
Political freedoms 0.238099006821
Freedom from discrimination and persecution 0.432932375445
Equality between men and women 0.276496043498
Better job opportunities -0.39734452674

Protecting forests rivers and oceans is the most positively correlated with HDI
Better transport and roads is the most negatively correlated with HDI
Access to clean water and sanitation is the least correlated with HDI

(c)





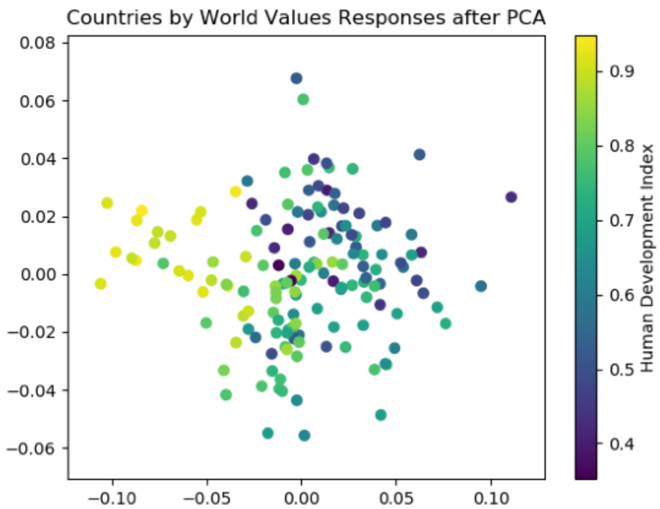


When it’s positively correlated, scatter plot is a relatively tight line with positive slope.

When it’s negatively correlated, scatter plot is a relatively tight line with negative slope.

When it’s least correlated, scatter plot is random.

(d)



(e) Ridge Regression: RMSE: 0.12303337350607803

(f) Lasso Regression: RMSE: 0.12602242808947528

(g) Lasso Coefficients

```
[ 0.1590192  -0.72844929 -0.         -0.85945074 -0.66274144 -0.02556703
  0.33904781 -0.29897158 -0.          0.          3.48536375  0.          0.
  0.87057995  0.32897045 -0.        ]
```

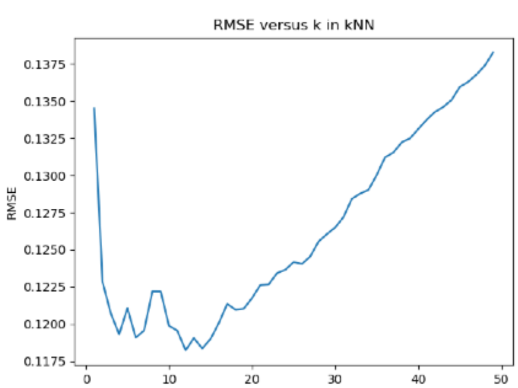
LASSO regression indeed give more 0 weights

(h) Using the average of k nearest neighbors for prediction.

(i)

90	Ireland
61	United Kingdom
37	Belgium
108	Finland
69	Malta
132	Austria
110	France

(j)

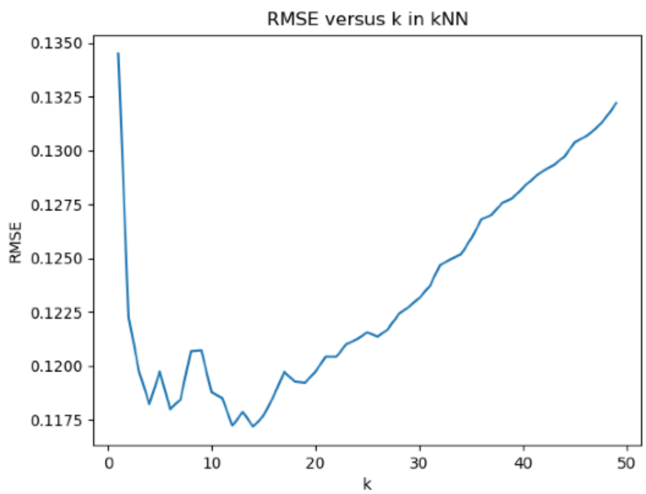


The best value of k is 12

and the RMSE is 0.11824589460776892

(k) When k increases, variance decreases because it's more stable, bias increases because you averaging over a larger area. So we see RMSE first decreases and then increases and there is an optimal k.

(l)



The best value of k is 14

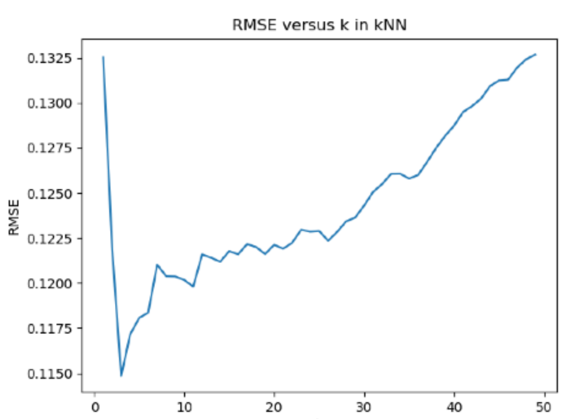
and the RMSE is 0.1171925270311745

(m)

90	Ireland
61	United Kingdom
108	Finland
37	Belgium
69	Malta
110	France
132	Austria

Compared with (i), Belgium and Finland switches; Austria and France switches.

(n)



best value for k is 3

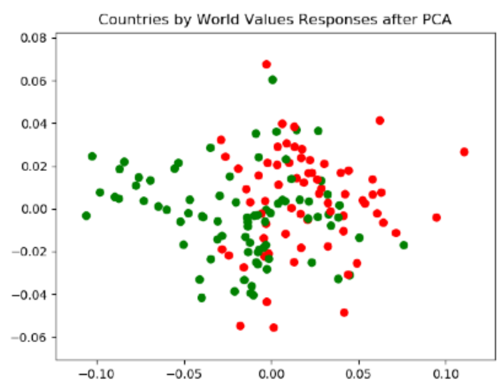
RMSE = 0.11484992771468817

(p)

```
[[ 0.51203818]
 [ 0.62507627]
 [ 0.68910033]
 [ 0.68107546]
 [ 0.7741555 ]
 [ 0.59925476]
 [ 0.68270339]
 [ 0.67685683]
 [ 0.77844755]
 [ 0.80635139]
 [ 0.67370619]
 [ 0.68573173]
 [ 0.75694345]
 [ 0.70599784]
 [ 0.67868664]
 [ 0.53981466]
 [ 0.69859174]
 [ 0.82971783]
 [ 0.70548055]
 [ 0.7756972 ]
 [ 0.63425221]
 [ 0.65054303]
 [ 0.39624097]
 [ 0.73577829]
 [ 0.61700422]
 [ 0.77204031]
 [ 0.43863423]
 [ 0.90431152]
 [ 0.60812481]
 [ 0.63991517]
 [ 0.77361977]
 [ 0.8648769 ]
 [ 0.72916633]
 [ 0.90996579]
 [ 0.91454931]
 [ 0.84350805]
 [ 0.51791246]
 [ 0.78966814]]
```

(q) A classification with k classes, at least we are guaranteed to get 1/k accuracy.

(r)



(s) A linear SVM won’t do a good job of classification because two classes are overlapping with each other.

(t) SVM Classification - Accuracy: 0.75

(u) SVM Classification with PCA and scaling - Accuracy: 0.77027027027

(v) SVM Classification with rbf - Accuracy: 0.689189189189

(w) k Nearest Neighbors Classification - Accuracy: 0.763513513514

k Nearest Neighbors Classification with scaling - Accuracy: 0.77027027027

Scaling helps a little bit.

(x) not too high based on my observation.

(y) We can use the distance to sensor as the feature parameters and average the nearest k objects, use their averaged position as predicted location. I think k nearest neighbors will work pretty well.

(z) I learned that depending on the problem, PCA, Knn, SVM may not work well at some cases, especially when there is overlapping between groups.

Feedback: This problem is very helpful to learn and compare different methods.

4. (a) $E_{S\sim D^n, i\sim U(n), z'\in D}[\ell(\vec{w}(S^{(i)}(z'), z_i) - \ell(\vec{w}(S), z_i)]$

$= E_{S\sim D^n, i\sim U(n), z'\in D}[\ell(\vec{w}(S), z') - \ell(\vec{w}(S), z_i)]$

$= E_{S\sim D^n, z'\in D}[\ell(\vec{w}(S), z')] - E_{S\sim D^n, i\sim U(n)}[\ell(\vec{w}(S), z_i]$

$= E_{S\sim D^n}[E_{z'\in D}[\ell(\vec{w}(S), z')] - E_{S\sim D^n}[\frac{1}{n}\sum_{i=1}^{n}\ell(\vec{w}(S), z_i)]$

$= E_{S\sim D^n} L_D(\vec{w}) - E_{S\sim D^n} L_S(\vec{w})$

$= E_{S\sim D^n}[L_D(\vec{w}(S)) - L_S(\vec{w}(S))]$ $\Rightarrow$ proved Eqn (4)

So if A is ε-stable, $E_{S\sim D^n, z'\in D, i\sim U(n)}[\ell(\vec{w}(S^{(i)}(z'), z_i) - \ell(\vec{w}(S), z_i)] \le \varepsilon_n$

then, $E_{S\sim D^n}[L_D(\vec{w}(S)) - L_S(\vec{w}(S))] \le \varepsilon_n$

$\Rightarrow$ $E_{S\sim D^n}[L_D(\vec{w}(S))] \le E_{S\sim D^n}[L_S(\vec{w}(S))] + \varepsilon_n$

(b) $\mu(S) = \frac{1}{n}\sum_{j=1}^{n} x_j$

$\mu(S^{(i)}(x')) = \frac{1}{n}(\sum_{j\ne i} x_j + x') = \frac{1}{n}\sum_{j=1}^{n}(x_j - x_i + x') = \mu(S) + \frac{1}{n}(x' - x_i)$

$\ell(\mu(S), x_i) = \frac{1}{2}(\mu - x_i)^2$

$\ell(\mu(S^{(i)}(x_i)), x_i) = \frac{1}{2}(\mu + \frac{1}{n}(x' - x_i) - x_i)^2$

So $\ell(\mu(S^{(i)}(x_i)), x_i) - \ell(\mu(S), x_i)$

$= \frac{1}{2}[(\mu - x_i + \frac{1}{n}(x' - x_i))^2 - (\mu - x_i)^2]$

$= \frac{1}{2}[\cancel{(\mu - x_i)^2} + 2(\mu - x_i)\frac{1}{n}(x' - x_i) + (\frac{1}{n}(x' - x_i))^2 - \cancel{(\mu - x_i)^2}]$

$\le \frac{1}{n}(\mu - x_i)(x' - x_i)$

$\le \frac{1}{n}|\mu - x_i|\,|x' - x_i|$

$|\mu| = \frac{1}{n}\sum_j |x_j| \le R$, so $|\mu| \le R$, $|x_i| \le R$, $|x'| \le R$

$\Rightarrow$ $|\mu - x_i| \le 2R$, $|x' - x_i| \le 2R$

So $\ell(\mu(S^{(i)}(x_i), x_i) - \ell(\mu(S), x_i)) \le \frac{CR^2}{n}$

Using Eqa (4), we know

$E_{S\sim D^n}[L_D(\mathcal{A}(S)) - L_S(\mathcal{A}(S))]$

$= E_{S\sim D^n,\, i\sim U(n),\, z'\in D}[\ell(\vec{w}(S^{(i)}(z'), z_i) - \ell(\vec{w}(S), z_i)]$

$\leq \frac{CR^2}{n}$

(c) $\ell(\vec{w}, z_i) = \frac{1}{2}(\vec{w}^T\vec{x}_i - y_i)^2$

$\nabla \ell(\vec{w}, z_i) = \left(\frac{\partial \ell(\vec{w}, z_i)}{\partial \vec{w}}\right)^T$

$= \left((\vec{w}^T\vec{x}_i - y_i)\, \vec{x}_i^{\,T}\right)^T$

$= (\vec{w}^T\vec{x}_i - y_i)\, \vec{x}_i$

$\|\nabla \ell(\vec{v}, z_i) - \nabla \ell(\vec{w}, z_i)\| = \|(\vec{v}^T\vec{x}_i - y_i)\vec{x}_i - (\vec{w}^T\vec{x}_i - y_i)\vec{x}_i\|$

$= \|(\vec{v} - \vec{w})^T \vec{x}_i\, \vec{x}_i\|$

$\leq \|\vec{x}_i\|^2 \quad \|\vec{v} - \vec{w}\|$

$\leq \beta \|\vec{v} - \vec{w}\| \quad$ where $\beta = \max_{x_i} \|x_i\|^2$

(d) $E_{S\sim D^n}[L_D(\vec{w}(S)) - L_S(\vec{w}(S))]$

$= E_{S\sim D^n,\, i\sim U(n),\, z'\in D}[\ell(\vec{w}(S^{(i)}(z'), z_i) - \ell(\vec{w}(S), z_i)]$

$\leq \frac{C\beta}{\lambda n}$

So $E_{S\sim D^n}[L_D(\vec{w}(S))] \leq E_{S\sim D^n}[L_S(\vec{w}(S))] + \frac{C\beta}{\lambda n}$

(e) $f_S(\vec{w}) = L_S(\vec{w}) + \lambda \|\vec{w}\|^2$

$$= \frac{1}{n}\sum_{i=1}^{n} l(\vec{w}, z_i) + \lambda \|\vec{w}\|^2$$

$$f_{S^{(i)}(z')}(\vec{w}) = \frac{1}{n}\Big(\sum_{j\neq i} l(\vec{w}, z_j) + l(\vec{w}, z')\Big) + \lambda\|\vec{w}\|^2$$

$$= \frac{1}{n}\big(\sum_j l(\vec{w}, z_j) - l(\vec{w}, z_i) + l(\vec{w}, z')\big) + \lambda\|\vec{w}\|^2$$

$$= f_S(\vec{w}) + \frac{1}{n} l(\vec{w}, z') - \frac{1}{n} l(\vec{w}, z_i)$$

$\vec{w}(S)$ is the minimizer of $f_S(\vec{w})$

$$f_S(\vec{w}(S^{(i)}(z'))) - f_S(\vec{w}) \geq \lambda \|\vec{w}(S^{(i)}(z')) - \vec{w}\|^2 \quad ①$$

Similarly, $f_{S^{(i)}(z')}(\vec{w}) - f_{S^{(i)}(z')}(\vec{w}(S^{(i)}(z'))) \geq \lambda\|\vec{w}(S^{(i)}(z')) - \vec{w}\|^2$ ②

① + ② $\Rightarrow$ $f_{S^{(i)}(z')}(\vec{w}) - f_S(\vec{w}) - (f_{S^{(i)}(z')}(\vec{w}(S^{(i)}(z'))) - f_S(\vec{w}(S^{(i)}(z')))$

$$= \frac{1}{n}\big(l(\vec{w}, z') - l(\vec{w}, z_i)\big) - \frac{1}{n}\big(l(\vec{w}(S^{(i)}(z'), z') - l(\vec{w}(S^{(i)}(z'), z_i)\big)$$

$$= \frac{1}{n}\big(l(\vec{w}(S^{(i)}(z'), z_i) - l(\vec{w}(S), z_i)\big) + \frac{1}{n}\big(l(\vec{w}(S), z') - l(\vec{w}(S^{(i)}(z')), z')\big)$$

$$\geq \lambda \|\vec{w}(S^{(i)}(z') - \vec{w}(S)\|^2$$

So $\lambda\|\vec{w}(S^{(i)}(z')) - \vec{w}(S)\|^2 \leq \frac{2}{n}\rho\|\vec{w}(S^{(i)}(z')) - \vec{w}(S)\|$

$$\|\vec{w}(S^{(i)}(z')) - \vec{w}(S)\| \leq \frac{2\rho}{n\lambda}$$

$$\Rightarrow l(\vec{w}(S^{(i)}(z')), z_i) - l(\vec{w}(S), z_i) \leq \rho \cdot \frac{2\rho}{n\lambda} = \frac{2\rho^2}{n\lambda}$$

(f) $l(\vec{w}, z_i) = \max(0, 1 - y_i \vec{w}^T \vec{x}_i)$

$l(\vec{w}(S^{(i)}(z')), z_i) = \max(0, 1 - y_i \vec{w}(S^{(i)}(z'))^T \vec{x}_i)$

If $y_i \vec{w}^T \vec{x}_i > 1$, $\quad y_i \vec{w}(S^{(i)}(z'))^T \vec{x}_i > 1$

$$l(\vec{w}(S^{(i)}(z')), z_i) - l(\vec{w}, z_i) = 0 \le \rho \, \| \vec{w}(S^{(i)}(z')), z_i \|$$

If $y_i \vec{w}^T \vec{x}_i < 1$, $\quad y_i \vec{w}(S^{(i)}(z'))^T \vec{x}_i > 1$,

$$l(\vec{w}(S^{(i)}(z')), z_i) - l(\vec{w}, z_i) = 0 - (1 - y_i \vec{w}^T \vec{x}_i) < 0 \quad \le \rho \, \| \vec{w}(S^{(i)}(z'), z_i \|$$

If $y_i \vec{w}^T \vec{x}_i > 1$, $\quad y_i \vec{w}(S^{(i)}(z'))^T \vec{x}_i < 1$,

$$l(\vec{w}(S^{(i)}(z')), z_i) - l(\vec{w}, z_i) = (1 - y_i \vec{w}(S^{(i)}(z'))\vec{x}_i - 0$$

$$\le (1 - y_i \vec{w}(S^{(i)}(z'))\vec{x}_i - (1 - y_i \vec{w}(S)\vec{x}_i)$$

$$\le \| y_i \| \, \| \vec{w}(S^{(i)}(z')) - \vec{w}(S) \| \, \| \vec{x}_i \|$$

$$\le \rho \, \| \vec{w}(S^{(i)}(z'), z_i \|$$

If $y_i \vec{w}^T \vec{x}_i < 1$, $\quad y_i \vec{w}(S^{(i)}(z'))^T \vec{x}_i < 1$,

$$l(\vec{w}(S^{(i)}(z'), z_i) - l(\vec{w}, z_i) = 1 - y_i \vec{w}(S^{(i)}(z'))\vec{x}_i - (1 - y_i \vec{w}(S)\vec{x}_i) \le \rho \, \| \vec{w}(S^{(i)}(z') - \vec{w} \|$$

So I prove that $l(\vec{w}, z_i)$ is ρ-Lipschitz

(g)

from (e) we know $l(\vec{w}(S^{(i)}(z'), z_i) - l(\vec{w}(S), z_i) \le \frac{2\rho^2}{\lambda n}$

So $E_{S \sim D^n, z' \sim D, i \sim U(n)} [l(\vec{w}(S^{(i)}(z'), z_i) - l(\vec{w}(S), z_i)] \le \frac{2\rho^2}{\lambda n}$

So $E_{S \sim D^n} [L_D(\vec{w}(S))] - E_{S \sim D} [L_S(\vec{w}(S))] \le \frac{2\rho^2}{\lambda n}$

5. Q: Find the minimum of $f(x) = x^2$ for $1 \le x \le 2$

A: We want to find the minimum of $f(x) = x^2$

Constrained by $x - 1 \geq 0, \quad 2 - x \geq 0$

Let $g_1(x, s, t) = x - 1 - s^2 = 0$

$g_2(x, s, t) = 2 - x - s^2 = 0$

Now the optimization problem becomes $\min f(x) = x$

$\text{s.t. } g_1(x, s, t) = 0$

$g_2(x, s, t) = 0$

Let $L(x, s, t, \lambda_1, \lambda_2) = x^2 + \lambda_1(x - 1 - s^2) + \lambda_2(2 - x - t^2)$

$\frac{\partial L}{\partial x} = 0 \Rightarrow 2x + \lambda_1 - \lambda_2 = 0$ ①

$\frac{\partial L}{\partial s} = 0 \Rightarrow -2\lambda_1 s = 0$ ②

$\frac{\partial L}{\partial t} = 0 \Rightarrow -2\lambda_2 t = 0$ ③

$\frac{\partial L}{\partial \lambda_1} = 0 \Rightarrow x - 1 - s^2 = 0$ ④

$\frac{\partial L}{\partial \lambda_2} = 0 \Rightarrow 2 - x - t^2 = 0$ ⑤

② $\Rightarrow \begin{cases} s \neq 0, \ \lambda_1 = 0, \ x = 1 + s^2 > 1 \\ s = 0, \ \lambda_1 \neq 0, \ x = 1 \end{cases}$

③ $\Rightarrow \begin{cases} t \neq 0, \ \lambda_2 = 0, \ x = 2 - t^2 < 2 \\ t = 0, \ \lambda_2 \neq 0, \ x = 2 \end{cases}$

1. $x = 1$
2. $x = 2$
3. $1 < x < 2, \ \lambda_1 = \lambda_2 = 0, \Rightarrow 2x + \lambda_1 - \lambda_2 = 0 \Rightarrow x = 0$, impossible

So $x = 1 \rightarrow f(x) = 1$

$x = 2 \Rightarrow f(x) = 4$

$\Rightarrow f(x=1) = 1$ is the minimum