ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Machine Learning
Session Examination

Prof. Joachim M. Buhmann

Examination Questions

January 26th, 2016

First and Last name: _____

ETH number: _____

Signature: _____

General Remarks

- The examination lasts for two hours. There are five sections.
- Please write your answers directly on the exam sheets. At the end of the exam you will find supplementary sheets, feel free to separate them from the exam. If you submit the supplementary sheets, put your name and ETH number on top of each.
- Please answer the questions in English. Do not use a pencil or red color pen.
- You may provide at most one valid answer per question. Invalid solutions must be clearly canceled out.

We wish you great success for this examination!

	Topic	Max. Points	Points	Signature
1	ML & Bayesian inference	20		
2	Kernels	20		
3	Neural Networks	20		
4	Gaussian processes	20		
5	Unsupervised learning	20		
Total		100		

Grade: .....

This page has been intentionally left blank.

Question 1: Maximum Likelihood and Bayesian Learning (20 pts.)

1. Given are n independently distributed binary vectors, each of dimension d. They are distributed according to:

$$X_{ij} \sim \text{ber}(1-\theta), \qquad \lambda_j = 1,$$
$$X_{ij} \sim \text{ber}(\theta), \qquad \lambda_j = 0,$$

for $1 \leq j \leq d$ and $1 \leq i \leq n$ where λ_j selects the Bernoulli source.

(a) What is the likelihood function $p(\mathcal{X}|\theta, \lambda)$?

$$p(\mathcal{X}|\theta,\lambda) = \prod_i \prod_j p(x_{ij}|\theta,\lambda)$$
$$= \prod_i \prod_j (1-\theta)^{\lambda_j x_{ij} + 1 - \lambda_j - x_{ij}} \theta^{x_{ij} - \lambda_j x_{ij} + \lambda_j}$$
$$= (1-\theta)^{\sum_i \sum_j (\lambda_j x_{ij} + 1 - \lambda_j - x_{ij})} \theta^{\sum_i \sum_j (x_{ij} - \lambda_j x_{ij} + \lambda_j)}$$

3 pts.

(b) Derive the maximum likelihood estimate for θ given λ.

$$\frac{\partial \log p(\mathcal{X}|\theta,\lambda)}{\partial \theta}$$
$$= \frac{\partial \left[\left(\sum_i \sum_j \lambda_j x_{ij} + 1 - \lambda_j - x_{ij} \right) \log(1-\theta) + \left(\sum_i \sum_j x_{ij} - \lambda_j x_{ij} + \lambda_j \right) \log \theta \right]}{\partial \theta}$$
$$= -\frac{\sum_i \sum_j (\lambda_j x_{ij} + 1 - \lambda_j - x_{ij})}{1-\theta} + \frac{\sum_i \sum_j (x_{ij} - \lambda_j x_{ij} + \lambda_j)}{\theta}$$
$$= 0$$
$$\Rightarrow \theta = \frac{\sum_i \sum_j (x_{ij} - \lambda_j x_{ij} + \lambda_j)}{n} = \frac{\sum_{j \in d} (n_j - 2 n_j \lambda_j + n \lambda_j)}{n}$$

4 pts.

Let $\sum_i x_{ij} = n_j$

(c) Derive the posterior distribution $p(\theta|\mathcal{X}, \lambda, \alpha)$ given that the parameter θ is beta distributed: $p(\theta|\alpha) = B(\alpha,\alpha)^{-1}\theta^{(\alpha-1)}(1-\theta)^{(\alpha-1)}$, where $B(\alpha,\alpha)$ normalizes the distribution. Please determine the normalization of the posterior explicitly.

$$p(\theta|X,\lambda,\alpha) \propto \frac{p(X|\theta,\lambda)\,p(\theta|\alpha)}{p(X|\lambda,\alpha)}$$

$$\frac{\theta^{\sum_{j=1}^{d}(x_{ij}+\lambda_j-2\lambda_j x_{ij})+\alpha-1}(1-\theta)^{\sum_{j=1}^{d}(x_{ij}+\lambda_j-2\lambda_j x_{ij})+\alpha-1}}{B(\alpha,\alpha)\,p(X|\lambda,\alpha)}$$

$$= \prod \text{Beta}\left(\sum_{j=1}^{d}2\lambda_j x_j + n - \lambda_j n - n_j,\ \sum_{j=1}^{d} n_j + \lambda_j n - 2\lambda_j n_j + \alpha\right)$$

3 pts.

(d) Derive the evidence $p(\mathcal{X}|\lambda,\alpha)$. Please determine the normalization of the evidence explicitly.

$$p(X|\lambda,\alpha) = \frac{B(\alpha,\alpha)^{-1}(1-\theta)^{\sum_{j=1}^{d}2\lambda_j n_j + n - \lambda_j n - n_j + \alpha}\ \theta^{\left(\sum_{j=1}^{d} n_j + \lambda_j n - 2\lambda_j n_j\right)+\alpha}}{\text{Beta}\left(\sum_{j=1}^{d}2\lambda_j n_j + n - \lambda_j n - n_j,\ \sum_{j=1}^{d} n_j + \lambda_j n - 2\lambda_j n_j + \alpha\right)}$$

3 pts.

(e) Under what condition is the maximum likelihood estimator for θ equal to the maximum a posterior estimator?

$$MAP = \frac{\sum_{j=1}^{d}(n_j - 2n_j\lambda_j + n\lambda_j) + \alpha - 1}{n + 2\alpha - 2} = \frac{\sum_{j=1}^{d}(n_j - 2n_j\lambda_j + n\lambda_j)}{n}$$

$\Rightarrow \quad \alpha = 1$ prior is uniform

or when $\lambda \to \infty$

1 pts.

2. Given is an input vector $X = (X_1, \ldots, X_d)^T$ and model parameters $\beta \in \mathbb{R}^d$. A linear regression model predicts the response Y:

$$Y = X^T\beta + \epsilon, \qquad \epsilon \sim N(\epsilon|0, \sigma^2),$$

(a) What is the likelihood function for n i.i.d. observations of Y?

$$L = \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n} e^{-\frac{\sum_{i=1}^{n}(y^{(i)} - x^{(i)T}\beta)^2}{2\sigma^2}}$$

2 pts.

(b) Recall that ridge regression minimizes the following function:

$$\hat{\beta}_{\text{ridge}} = \arg\min_{\beta} \sum_{i=1}^{n} \left(y_i - x_i^T\beta\right)^2 + \lambda \sum_{j=1}^{d} \beta_j^2$$

Given that $\sigma = 1$, how should one choose the prior on β such that the maximum a posteriori estimator is equal to $\hat{\beta}_{\text{ridge}}$? Please explain your answer.

选取 $\beta \sim N(0, \Lambda^{-1})$

$\Lambda^{-1} = \frac{\lambda}{\sigma^2} I_d$

in this case

$$\max P(\beta | \Lambda, X, Y, \sigma) = \max \alpha \cdot e^{-\frac{\sum_{i=1}^{n}(y^{(i)} - x^{(i)T}\beta)^2}{2\sigma^2} - \frac{\beta^T\Lambda\beta}{2}}$$

when $\Lambda = \frac{\lambda}{\sigma^2} I_d$

4 pts.

Question 2: Maximum Likelihood and Bayesian Learning (20 pts.)

1. Given are n random binary vectors, each of dimension d, $\mathcal{X} = \{X_1, \ldots, X_n\}, X_i \in \{0,1\}^d$. The components of the binary vectors are generated from one of the two sources:

$$\begin{aligned} X_{ij} &\sim \text{ber}(1-\theta), & \lambda_j &= 1, \\ X_{ij} &\sim \text{ber}(\theta), & \lambda_j &= 0, \end{aligned}$$

where λ_j selects the Bernoulli source and $1 \leq j \leq d$.

(a) What is the likelihood function $p(\mathcal{X}|\theta, \lambda)$?

$$p(\mathcal{X} \mid \theta, \lambda) = \prod_{i \leq n} \prod_{j \leq d} \text{ber}(X_{ij} \mid 1-\theta)^{\lambda_j} \text{ber}(X_{ij} \mid \theta)^{1-\lambda_j}$$

2 pts.

(b) Derive the maximum likelihood estimate for θ given λ. Use the following abbreviation $n_j = \sum_{i \leq n} X_{ij}$.

$$\begin{aligned} \frac{d}{d\theta}(-\log p(\mathcal{X} \mid \theta, \lambda)) &= \frac{d}{d\theta}\left(-\sum_{j \leq d} \lambda_j \left[n_j \log(1-\theta) + (n-n_j)\log\theta \right] \right. \\ &\qquad \left. + (1-\lambda_j)\left[n_j \log\theta + (n-n_j)\log(1-\theta) \right] \right) \\ &= -\sum_{j \leq} \lambda_j \left[\frac{n_j}{1-\theta} + \frac{n-n_j}{\theta} \right] + (1-\lambda_j)\left[\frac{n_j}{\theta} + \frac{n-n_j}{1-\theta} \right] \\ &\propto -\sum_{j \leq d} \lambda_j \left[\theta n_j + (n-n_j)(1-\theta) \right] + (1-\lambda_j)\left[n_j(1-\theta) + (n-n_j)\theta \right] \\ &= -\sum_{j \leq d} \left[\theta(2n_j - n) + n - n_j \right] + (1-\lambda_j)\left[\theta(n-2n_j) + n_j \right] = 0 \end{aligned}$$

$$\hat{\theta} = \frac{\sum_{j \leq d} \lambda_j (n-n_j) + (1-\lambda_j) n_j}{\sum_{j \leq d} \lambda_j (2n_j - n) + (1-\lambda_j)(n-2n_j)}$$

4 pts.

(c) Derive the posterior distribution $p(\theta|\mathcal{X}, \lambda, \alpha)$ in closed form given that the prior distribution on θ is beta distributed: $p(\theta|\alpha) = B(\alpha,\alpha)^{-1}\theta^{(\alpha-1)}(1-\theta)^{(\alpha-1)}$, where $B(\alpha,\alpha)$ normalizes the distribution.

$$p(\theta \mid \mathcal{X}\lambda, \alpha) = a \prod_{j \leq d} (1-\theta)^{(n_j\lambda_j + (n-n_j)(1-\lambda_j)+\alpha-1)} \theta^{((n-n_j)\lambda_j + n_j(1-\lambda_j)+\alpha-1)} B(\alpha,\alpha)^{-1}$$
$$= \prod_{j \leq d} \text{beta}\left(\theta \mid n_j\lambda_j(n-n_j)(1-\lambda_j)+\alpha; (n-n_j)\lambda_j + n_j(1-\lambda_j)+\alpha\right)$$

3 pts.

(d) Derive the evidence $p(\mathcal{X}|\lambda, \alpha)$.

$$p(\mathcal{X} \mid \lambda, \alpha) = \prod_{j \leq d} \frac{B\left(n_j\lambda_j + (n-n_j)(1-\lambda_j)+\alpha, (n-n_j)\lambda_j + n_j(1-\lambda_j)+\alpha\right)}{B(\alpha,\alpha)}$$

3 pts.

(e) Under what conditions is the maximum likelihood estimator for θ equal to the maximum a posterior estimator?

When the prior is uniform or when $n \to \infty$.

2 pts.

2. Given is an input vector $X = (X_1, \ldots, X_d)^T$ and model parameters $\beta \in \mathbb{R}^d$. A linear regression model predicts the response Y:

$$Y = X^T\beta + \epsilon, \qquad \epsilon \sim N(\epsilon|0, \sigma^2),$$

(a) Given n i.i.d. observations, what is the likelihood function for β?

$$p(Y \mid \beta, X) = \mathcal{N}\left(Y \mid X^T\beta, \mathbb{I}\sigma^2\right)$$

2 pts.

(b) Recall that ridge regression maximizes the following function:

$$\hat{\beta}_{\text{ridge}} = \arg\min_{\beta} ||\mathbf{y} - \mathbf{X}\beta||^2 + \lambda||\beta||^2$$

Given that $\sigma = 1$, how should one choose the prior on β such that the maximum a posteriori estimator is equal to $\hat{\beta}_{\text{ridge}}$? Please explain your answer.

The MAP estimator is given by:

$$\hat{\beta} = \arg\max_{\beta} \log p(\mathbf{y} \mid \mathbf{X}, \beta, \sigma^2) + \log p(\beta \mid \Lambda)$$
$$\stackrel{(a)}{=} \arg\max_{\beta} -\frac{1}{2\sigma^2}||\mathbf{y} - \mathbf{X}\beta||^2 - \frac{1}{2}\beta^T \Lambda \beta,$$

where in (a) we use a Gaussian prior on β. If we choose $\Lambda = \mathbb{I}\lambda$ we obain the ridge regression.

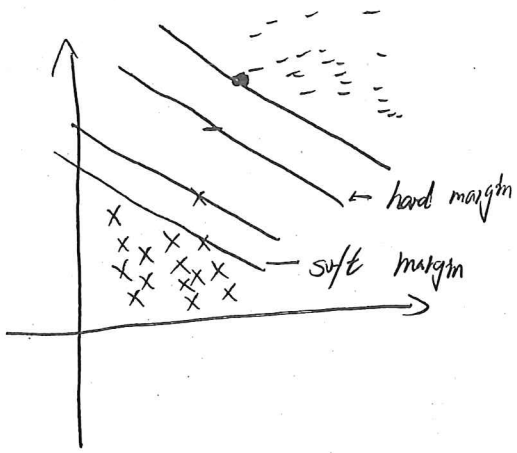
4 pts.

Question 3: **Kernels (20 pts.)**

1. What is the difference between hard-margin and soft-margin SVM for classification? Please illustrate your argument with a figure. What is the meaning of the support vectors in each case?

in the hard margin it is the vectors that lies on the boundary that compose the support vectors

in the latter one, the boundary and the slack vectors lieson the variable.



3 pts.

2. Suppose your training dataset is perfectly linearly separable. Is there any reason (motivation) why to use soft-margin SVM for classification in this case?

avoid over fitting

1 pts.

3. What are the three main (distinct) advantages of using the Kernel trick?

① do not need to explicit express the feature transformation, more expressive

② computational efficiency, since the kernel function.

~~③ can ex~~

③ non-linear transformation, high-dimension representation

3 pts.

4. The Sigmoid Kernel is defined as: $K(x, y) = \tanh(\alpha x^T y + r)$ where α is a scaling parameter and r is a shifting parameter.

   (a) Prove that the Sigmoid Kernel is not always a valid kernel. Which property of a valid kernel is violated by the Sigmoid Kernel? Please give a short description.

semi-positive

$\left(\frac{e^x - e^{-x}}{e^x + e^{-x}}\right)$

Let $\alpha = 1$

$\begin{bmatrix} \tanh(\frac{1}{k^2}) & \tanh(1) \\ \tanh(1) & \tanh(k^2) \end{bmatrix}$

$x_1 = \frac{1}{k}$

$x_2 = k$

1 pts.

(b) Suppose that x and y are feature vectors in $\{-1, 0, 1\}^N$. Is there any circumstance in which the Sigmoid Kernel is a valid kernel? If yes, please provide such an example.

$\alpha > 0$ 足够小 $\gamma \to 0$

4 pts.

5. Support Vector Machines are inherently binary classifiers. How would you adapt the original SVM technique in order to perform multi-class classification? In particular:

(a) Describe the main idea of your technique.

$\min \frac{1}{2} w^T w$

s.t. $(w_{z_i}^T x_i + w_{z_i 0}) \geq \max_{z \neq z_i} (w_z^T x_i + w_z) + 1$

or $\min \frac{1}{2} w^T w$

s.t. $w^T \psi(x_i, y_i) - \max_{z \neq z_i} [w^T \psi(x_i, y) + \delta(z, z_i)] \geq -\xi_i \quad \forall y \in Y$

1 pts.

(b) Provide a formal definition of the optimization problem during training.

3 pts.

(c) Provide the formula that you would use to do classification in test phase.

$z = \max(w_z^T x + w_{z,0})$

2 pts.

(d) Mention one advantage and one disadvantage of your solution.

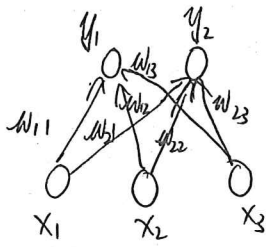
dis: no probability / probalistic explannation; did not discriminate the difference ~~between~~ of the class differences, may not be feasible

ad: simple data model, easy to train. In linear, seprable case

2 pts.

Question 4: **Neural Networks (20 pts.)**

1. Sketch a neural-network-diagram of a linear discriminant model $y_k = \sum_{j=0}^{2} w_{kj} x_j$ with two outputs $k = 1, 2$ and label all inputs, outputs and weights.



1 pts.

2. Why are activation functions in multilayer networks in general non-linear?

in order to generalization generalization of linear model
allow more non-linear for both regression
and classification

1 pts.

3. Explain why weights are generally initialized to small values close to zero?

no ~zero break the symmetry

after normalization
(1) Expectation is 0 (2) should not set to 0 since the backpropagation is sort of gradient decent, when the initial weight is large, the error will be large, so the step will be huge.

(2) since the expectation for weight is around 0 Vanish gradient

1 pts.

(3) logistic function is close to flat for large positive or negative inputs.

(4) the same values of 0 will decrease -- output and does not break the symmetry of the net and

4. What is the computational advantage of activation functions like the sigmoid or the hyperbolic tangent?

$f'(x) = f(x)(1-f(x))$

① Since the derivative of them are simple and can be reused.

② they are differentiable.

~~③ yield a zero gradient~~

1 pts.

5. What difficulty is caused by step activation functions for the learning algorithm `backpropagation`?

non-differentiable

Vanishing gradient problem especially when the neural networks are deep, $f'(x) \to 0$ when $x \to \infty$

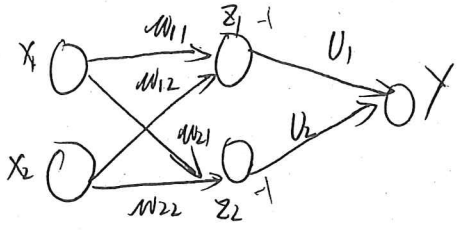
1 pts.

6. Consider a two layer feed-forward neural network with a 2 dimensional input (x_1, x_2). The hidden layer (z_1, z_2) is also two dimensional. We will use the sigmoid $\sigma(x)$ as activation function in the hidden layer. In addition, both layer have biases, defined as b_1 and b_2. The output y is one dimensional. Thus we can define the network by the following equations:

$$z_j = \sigma(\sum_{k=1}^{2} w_{jk} x_k + b_1)$$

$$y = \sum_{k=1}^{2} U_k z_k + b_2$$

(a) Sketch the neural-network-diagram of the above model.



1 pts.

(b) Initialize all w_{ij} and U_i to -1, and set all biases b_i to 1. The loss function is $L(t,y) = \frac{1}{2}(t-y)^2$ and you are given one data point $(1, \frac{1}{2}, \frac{1}{2})$ with label $t = 1$. With backpropagation, calculate the gradient of the loss with respect to the weight w_{11}, and use it, together with a learning rate $\eta = 4$, to update w_{11}.

$$\frac{\partial L}{\partial w_{ij}} = \frac{\partial L}{\partial z_1}\frac{\partial z_1}{\partial w_{11}}$$

$$= (t-y)\frac{\partial y}{\partial z_1}\frac{\partial z_1}{\partial w_{11}}$$

$$= (t-y)\,U_1 \cdot \sigma'\left(\sum_{k=1}^{2} w_{1k}x_k + b_1\right)x_1 = (1-0)\times(-1)\times \sigma(1)(1-\sigma(1))\times 1$$

$$= -1 \times \frac{1}{2}\times\frac{1}{2}\times 1 = -\frac{1}{8}$$

$$z_1 = \frac{1}{1+e^{-(-1\times\frac{1}{2}-1\times\frac{1}{2}+1)}} = \frac{1}{2}$$

$$z_2 = \frac{1}{1+e^{-(-1\times\frac{1}{2}-1\times\frac{1}{2}+1)}} = \frac{1}{2}$$

$$y = -\frac{1}{2}-\frac{1}{2}+1 = 0$$

thus

$$w_{11}' = w_{11} + \eta\left(-\frac{\partial L}{\partial w_{ij}}\right)$$

4 pts.

$$= -1 + 4\times\frac{1}{8} = -\frac{1}{2}$$

7. Multilayer Perceptrons

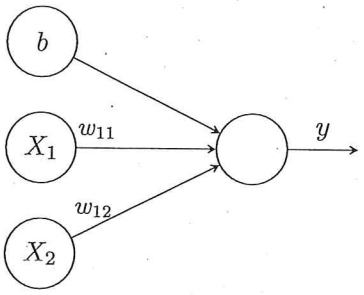


Figure 1: Single layer network.

(a) Formally prove, that the single layer perceptron shown in Figure ?? can not correctly classify the following training data set $\mathcal{Z}$:

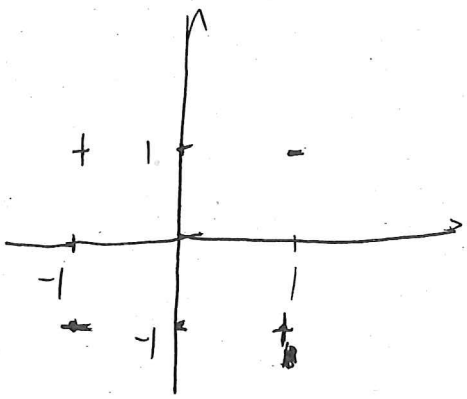
$$\mathcal{Z} = \{((1, 1), A), (-1, -1), A), ((1, -1), B), ((-1, 1), B)\}$$

Each element $((x_1, x_2), t)$ denotes a pattern and its associated label. A pattern is classified as class A if the network output y is less than zero and as class B else. The third unit b denotes the bias. Note that the activation function of the output is the identity.

$y = w_{11} x_1 + w_{12} x_2 + b$

↖ it is an linear combination of the input

these classification samples are not linear seperable



from add (0,0) $b < 0$

from add (1,0) $b \geq 0$ → contradict

suppose we have

① $w_{11} + w_{12} + b < 0 \cdots ①$

② $-w_{11} + (-w_{12}) + b \leq 0 \cdots ②$

③ $w_{11} - w_{12} + b > 0 \cdots ③$

④ $-w_{11} + w_{12} + b > 0 \cdots ④$

4 pts.

from

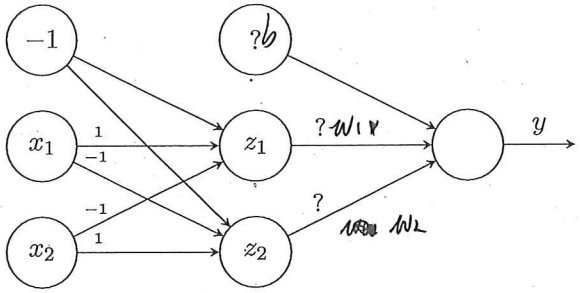


Figure 2: Two layer network.

(b) Now consider the two-layer network in Figure ??. Assume that the activation function in the hidden layer is a step functions, which is 0 if their input is less than zero and 1 otherwise. Show that it can correctly classify the training data set $\mathcal{Z}$ by selecting an appropriate set of second layer weights and bias.

~~$O(z_1)$ $h(z_1)$~~

$input(z_1) = x_1 - x_2 - 1$

$input(z_2) = -x_1 + x_2 - 1$

for label A, output(z_1) = 0
... B, z_1 (1,-1), 1
(-1,1), 0

for z with A output (z) = 0

B, (1,-1) 0

(-1,1) 1

in this case (-1,-1) → (0,0)
(1,1) → (0,0)

(1,-1) → (0,0)

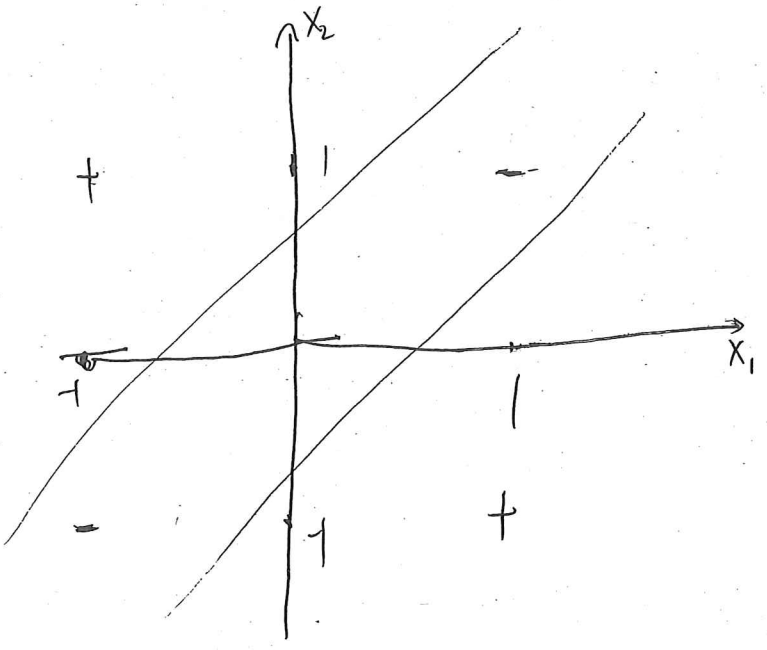
(-1,1) → (0,1)

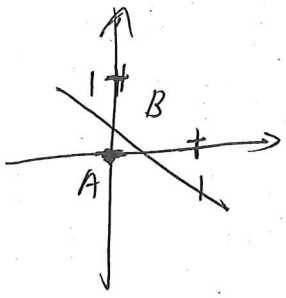
$y = w_1 z_1 + w_2 z_2 + b$

$b < 0$

$\begin{cases} w_1 + b > 0 \\ w_2 + b > 0 \end{cases}$ which can be met by selection **2 pts.**

(c) Plot the decision regions in the (x_1, x_2) plane, i.e. the areas classified as A and B, respectively.





2 pts.

(d) Plot how the original input space (x_1, x_2) is mapped to a new feature space (z_1, z_2). Is the training data set linearly separable in the new feature space?

Yes

2 pts.

Question 5: **Gaussian process regression (20 pts.)**

Consider the Gaussian process regression model, $y = f(\boldsymbol{x}) + \epsilon$, where $\boldsymbol{x} \in \mathbb{R}^d$ is the input vector, y is the response, ϵ is the Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$, $f(\boldsymbol{x}) \sim \mathcal{GP}(\mathbf{0}, k(\boldsymbol{x}, \boldsymbol{x}'))$ with mean $\mathbf{0}$ and covariance function $k(\boldsymbol{x}, \boldsymbol{x}')$.

1. There are n training vectors associated with the responses: $X \in \mathbb{R}^{n \times d}, \boldsymbol{y} \in \mathbb{R}^n$. A single test vector is $\boldsymbol{x}_*$, assume the predictive target of it is y_*. For notation simplicity, let

$$K(X, X) := \begin{bmatrix} k(\boldsymbol{x}_1, \boldsymbol{x}_1) & \dots & k(\boldsymbol{x}_1, \boldsymbol{x}_n) \\ \vdots & \ddots & \vdots \\ k(\boldsymbol{x}_n, \boldsymbol{x}_1) & \dots & k(\boldsymbol{x}_n, \boldsymbol{x}_n) \end{bmatrix}, K(\boldsymbol{x}_*, X) := [k(\boldsymbol{x}_*, \boldsymbol{x}_1), \dots, k(\boldsymbol{x}_*, \boldsymbol{x}_n)]$$

   (a) Derive the joint distribution of $\begin{bmatrix} \boldsymbol{y} \\ y_* \end{bmatrix}$;

$$\begin{bmatrix} y \\ y_* \end{bmatrix} = \mathcal{N}\left[\begin{bmatrix} 0 \\ 0 \end{bmatrix} \middle| \begin{bmatrix} C_n & k \\ k^T & c \end{bmatrix} \right]$$

with $C_n = K(X, X) + \sigma^2 \mathbb{I}$

$$c = k(X_*, X_*) + \sigma^2$$

$$k = K(X_*, X)^T$$

$$k^T = K(X_*, X)$$

2 pts.

   (b) According to the joint distribution, determine the predictive distribution $y_* | X, \boldsymbol{y}, \boldsymbol{x}_* \sim \mathcal{N}(\bar{y}_*, \mathbb{V}[y_*])$. Hint: using the conditioning rule: Let $\boldsymbol{a}$ and $\boldsymbol{b}$ be jointly Gaussian random vectors

$$\begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{b} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix}, \begin{bmatrix} A & C \\ C^{\mathrm{T}} & B \end{bmatrix} \right)$$

the conditional distribution of $\boldsymbol{b}$ given $\boldsymbol{a}$ is: $\boldsymbol{b}|\boldsymbol{a} \sim \mathcal{N}(\boldsymbol{\mu}_b + C^{\mathrm{T}}A^{-1}(\boldsymbol{a} - \boldsymbol{\mu}_a), B - C^{\mathrm{T}}A^{-1}C)$.

$\bar{y}_* = k(X_*, X)\,[k(X,X) + \sigma^2 I]^{-1} y$

$V(y_*) = k(X_*, X_*) - k(X_*, X)\,[k(X,X) + \sigma^2 I]^{-1} k(X_*, X)^{\mathrm{T}}$

2 pts.

2. Consider the special situation: Given one scalar data point (x, y), and the squared exponential kernel with length-scale l: $k(x, x') = \exp(-\frac{(x - x')^2}{2l^2})$, for the test point x_*:

   (a) Write down the specific predictive mean and variance.

$\bar{y}_* = \exp\left(-\frac{(x_* - x)^2}{2l^2}\right)\,[1 + \sigma^2]^{-1} y$

$= \frac{y}{1+\sigma^2} \cdot e^{-\frac{(x_* - x)^2}{2l^2}}$

$V(y_*) = 1 + \sigma^2 - \frac{\exp\left(-\frac{(x_* - x)^2}{l^2}\right)}{1+\sigma^2}$

2 pts.

(b) Fixing the noise level $\sigma = 0.1$, in Figure ?? you can see four different plots showing predictions made using length-scales of $l = \{0.2, 0.5, 1.2, 2\}$, with ten scalar training data points (crosses in the Figure). Specify which l was used for each of these four predictions. Please indicate your choice by writing the value of l below the corresponding plot.

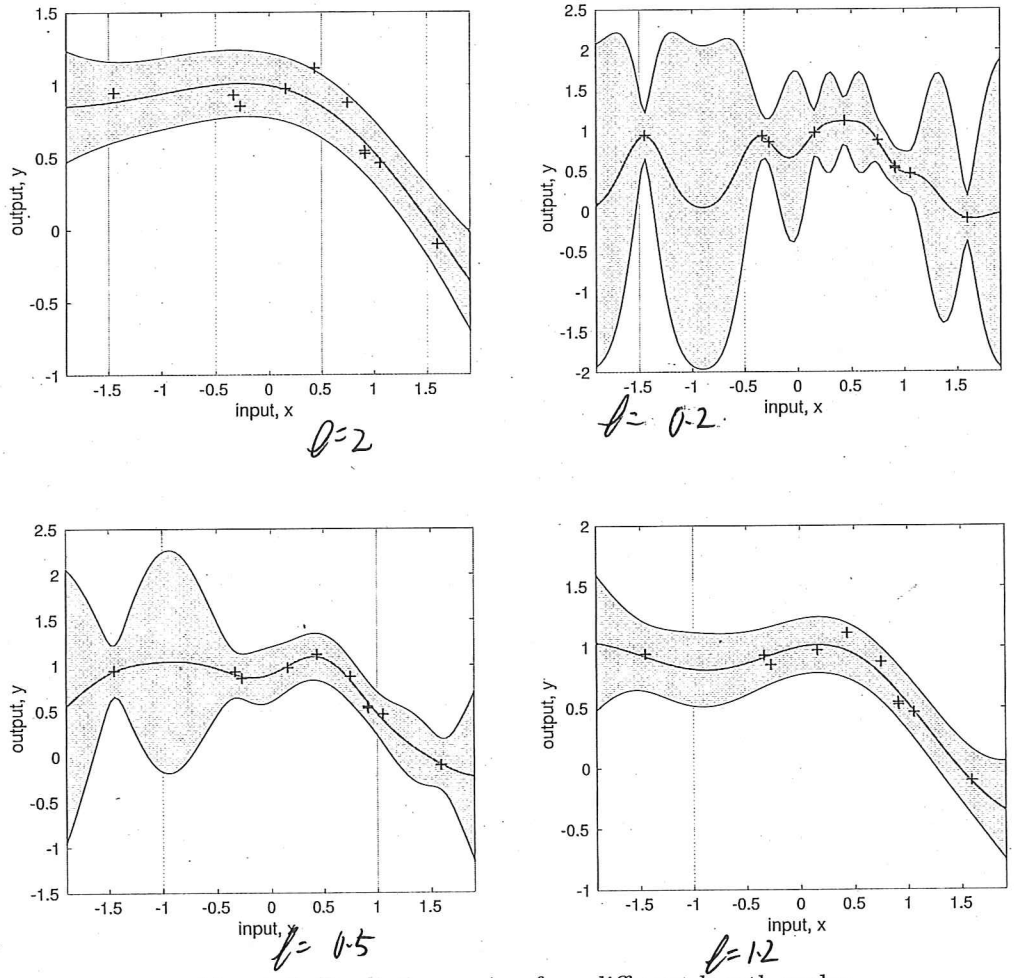


Figure 3: Predictions using four different length-scales.

$$\frac{\gamma}{1+0.01} e^{\frac{-(x^x-x)^2}{2\ell^2}}$$

$$\frac{\exp^{\frac{-(x^x-x)^2}{2\ell^2}}}{1.01}$$

3 pts.

(c) Briefly describe what role the length-scale l plays.

the scale of the dissimilarity, how to discriminate

↓ Sensitivity ↓ (regularization)

$\ell \uparrow$, more smooth

$\ell \downarrow$, more adpt to the data change

3 pts.

3. Let $\boldsymbol{\theta}$ represent all the hyperparameters of the $\mathcal{GP}$ (including hyperparameters in the kernel, e.g. l, and noise level σ). Usually the optimal hyperparameters are found by maximizing the log marginal likelihood ($\log \mathbb{P}(\boldsymbol{y}|X, \boldsymbol{\theta})$) with respect to $\boldsymbol{\theta}$.

(a) Prove that the log marginal likelihood is

$$\log \mathbb{P}(\boldsymbol{y}|X, \boldsymbol{\theta}) = -\frac{1}{2}\boldsymbol{y}^{\mathrm{T}}[K(X,X)+\sigma^2 I]^{-1}\boldsymbol{y} - \frac{1}{2}\log|K(X,X)+\sigma^2 I| - \frac{n}{2}\log 2\pi$$

where I is the identity matrix, $|\cdot|$ is the determinant of a matrix.

$P(y|x,\theta) = \frac{1}{\sqrt{(2\pi)^n |V(x)|}} \left[-\frac{1}{2}(y-\bar{y})^T V^{-1}(x)(y-\bar{y})\right]$

$\log P(y|x,\theta) = -\frac{n}{2}\log 2\pi - \frac{1}{2}\log[K(X,X)+\sigma^2 I] - \frac{1}{2} y^T [K(X,X)+\sigma^2 I]^{-1} y$

4 pts.

(b) Generally, can you find an analytic expression for the optimal hyperparameter θ^* that maximize the log marginal likelihood? If not, suggest an approach to compute θ^*.

Or hard to find the close-form solution since the inversion of the matrix can be too huge

use Newton update

$$\theta^* = \theta_0 - \frac{\partial \log p(y|X,\theta)}{\partial \theta} \Big/ \frac{\partial^2 \log p(y|X,\theta)}{\partial \theta \partial \theta'}$$

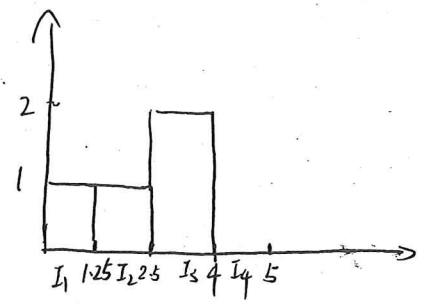
4 pts.

Question 6: Unsupervised learning (20 pts.)

1. *Non-parametric density estimation.*

   Consider the following datasets of 1-dimensional data points: $D^+ = \{d_1, d_2, d_3, d_4\}$ where $d_1 = 1$, $d_2 = 1.5$, $d_3 = 3$ and $d_4 = 3.1$ and $D^- = \{d_5, d_6, d_7\}$ where $d_5 = 1.3$, $d_6 = 3.9$, $d_7 = 4$.

   (a) Show the histogram of the dataset D^+ with intervals $I_1 = [0, 1.25)$, $I_2 = [1.25, 4)$ and $I_3 = [4, 5)$.



1 pts.

   (b) Kernel density estimators and nearest neighbor estimates use the following formula to approximate the probability density for any point $\mathbf{x}$:

$$p(\mathbf{x}) = \frac{K(\mathbf{x})}{nV}$$

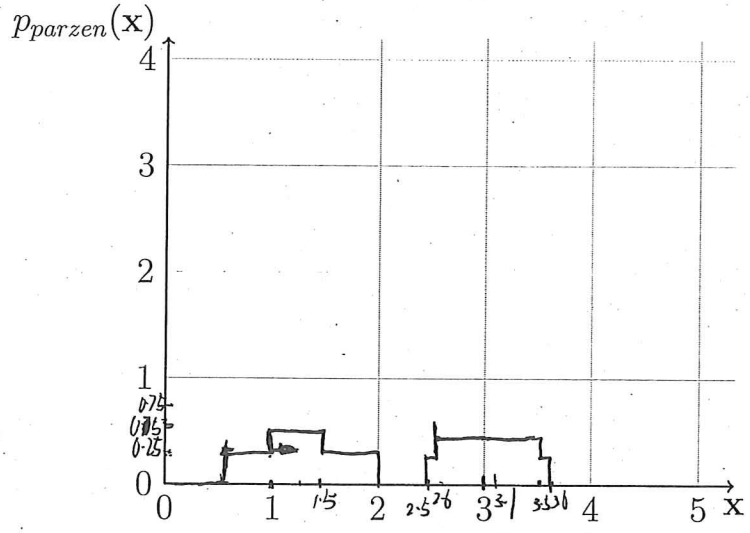
   where $K(\mathbf{x})$ is the number of samples falling within a window around $\mathbf{x}$, n is the total number of training samples and V is the area of the window around $\mathbf{x}$.

   Let $k(\frac{\mathbf{x}-\mathbf{x}_j}{h})$ be a Parzen window of the unit cube function scaled by the bandwith h. The unit cube function is defined as follows:

$$k(\mathbf{u}) = \begin{cases} 1 & \textbf{if } |u_i| \leq 1/2 \textbf{ for all } i \in 1, ..., d \\ 0 & \textbf{otherwise} \end{cases}$$

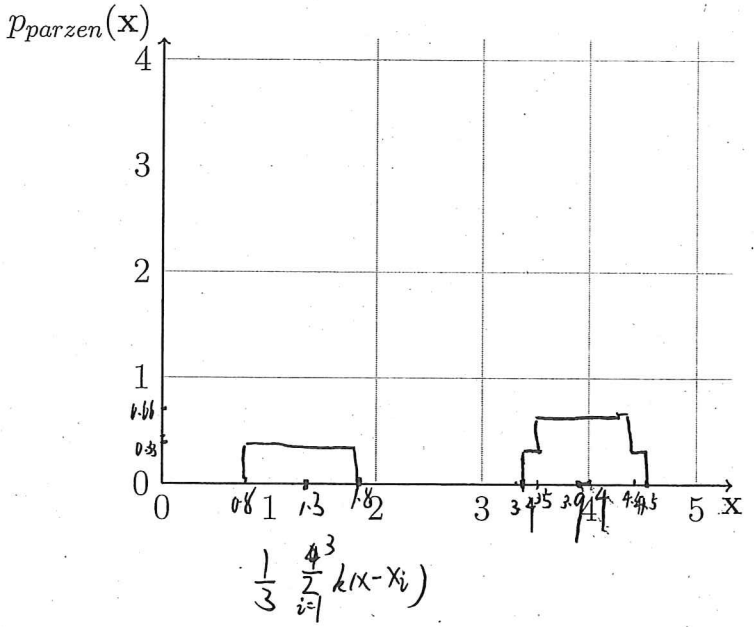
   i. Plot (draw) the resulting probability density function $p_{parzen}(\mathbf{x})$ for the 1-dimensional dataset D^+ and for $h = 1$. Use the provided plotting space.

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{V}k\left(\frac{x-x_i}{h}\right) = \frac{1}{4}\sum_{i=1}^{4}k(x-x_i)$$



2 pts.

ii. Plot (draw) the resulting probability density function $p_{parzen}(\mathbf{x})$ for the 1-dimensional dataset D^- and for $h = 1$. Use the provided plotting space.



$\frac{1}{3}\sum_{i=1}^{3} k(x-x_i)$

2 pts.

iii. We are now given the point $\mathbf{x} = 1.2$, which we want to classify in either one of the following two classes $C = \{c^+, c^-\}$. The training data is $D = D^+ \cup D^-$ where D^+ and D^- are defined as before and contain the samples of the class c^+ and c^-. In order to do classification based on the previously introduced density we need to estimate the following table of probabilities. Fill in the table using

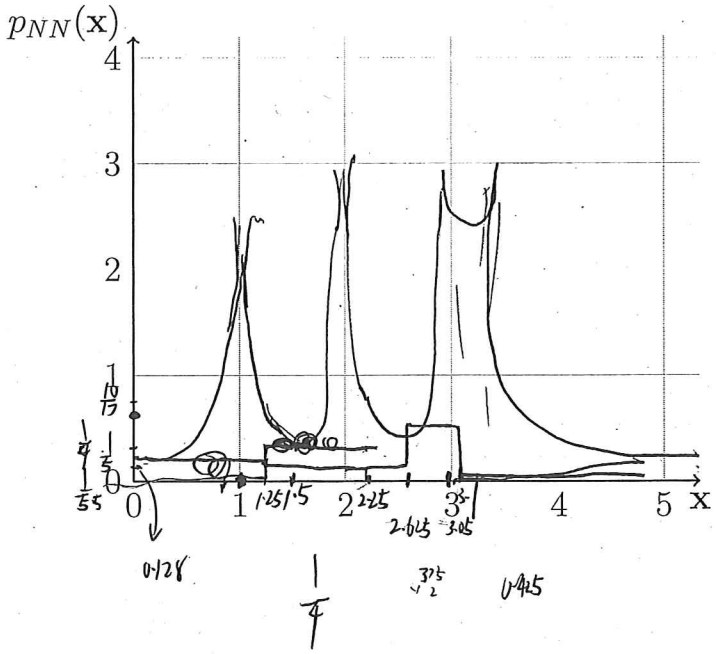
your results from (i) and (ii) and applying Bayes rule:

For $\mathbf{x} = 1.2$ and using a Parzen window estimate from (b)				
	$p(c\|\mathbf{x})$	$p(\mathbf{x}\|c)$	$p(c)$	$p(\mathbf{x})$
$c = c^+$	$\frac{2}{3}$	$\frac{1}{2}$	$\frac{4}{7}$	$\frac{3}{7}$
$c = c^-$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{3}{7}$	$\frac{3}{7}$

$\sum_c p(x|c)P(c) = \frac{1}{2}\times\frac{4}{7} + \frac{1}{3}\times\frac{3}{7} = \frac{3}{7}$

1 pts.

iv. Now plot the related $p_{NN}(\mathbf{x})$ for D^+ using a nearest neighbor density estimator. In nearest neighbor density estimation the size of the window in $p(\mathbf{x}) = \frac{K(\mathbf{x})}{nV}$ is chosen such that exactly one sample falls into the hyper-sphere around x (i.e. $K(\mathbf{x}) = 1$). You can resolve cases where it is not possible to set $K(\mathbf{x}) = 1$ (i.e. tie-breaks) arbitrarily. Use the provided space below.



4 pts.

v. Compute the values of $\int_{-\infty}^{\infty} p_{parzen}(x)dx$ and $\int_{-\infty}^{\infty} p_{NN}(x)dx$ obtained in (i) and (iv) and compare them.

$$\int_{-\infty}^{\infty} p_{parzen}(x) = \frac{1}{3} + \frac{1}{3}\times(0.2) + \frac{2}{3}\times(0.9)$$

$$= 1$$

$$\int_{-\infty}^{\infty} p_{NN}(x)\,dx = 1$$

2 pts.

$$= \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

$$\int_0^1 \frac{1}{8(1-x)}dx + \int_1^{1.25} \frac{1}{8(x-1)}dx + \int_{1.25}^{1.5} \frac{1}{8(1.5-x)}dx$$

$$+ \int_{1.5}^{2.25} \frac{1}{8(x-1.5)}dx + \int_{2.25}^{3} \frac{1}{8(3-x)}dx + \int_{3}^{3.05} \frac{1}{8(x-3)}dx$$

$$+ \int_{3.05}^{3.5} \frac{1}{8(3.1-x)}dx + \int_{3.1}^{\infty} \frac{1}{8(x-3.1)}dx$$

$$= \infty$$

2. *Expectation-Maximization algorithm.*

Assume the following mixture of d-dimensional densities, consisting of K components:

$$p(\mathbf{x}|\mu_1, \ldots, \mu_K) = \frac{1}{K}\sum_{c=1}^{K} p(\mathbf{x}|\mu_c),$$

where

$$p(\mathbf{x}|\mu_c) = (2\pi)^{-d/2} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu_c)^T(\mathbf{x}-\mu_c)\right)$$

Here, the parameters μ_c, $c = 1, ..., K$ are d-dimensional vectors.

Let $\{\mathbf{x}_i\}_{i=1}^N$ be a dataset generated from the mixture above. For the EM-algorithm we introduce latent variables $\gamma_{i,c}, i = 1, \ldots, N; c = 1, \ldots, K$.

$\gamma_{i,c}$ indicates the probability of the point $\mathbf{x}_i$ belonging to class c.

(a) What is the difference between this model and the traditional Gaussian Mixture Model?

the prior are the uniform distribution and same σ^2 for each gaussian.

2 pts.

(b) Derive the update step for $\gamma_{i,c}$ given the model parameters μ_c.
Hint: recall that $\gamma_{i,c}$ is defined to be $\gamma_{i,c} = p(c|\mathbf{x}_i)$

$$\gamma_{ic} = \frac{p(c|x_i, \theta)\, p(c|\theta)}{p(x_i)}$$

$$= \frac{\frac{1}{K}\exp\left(-\frac{1}{2}(x_i-\mu_c)^T(x-\mu_c)\right)}{\frac{1}{K}\sum_{j=1}^{K}\exp\left(-\frac{1}{2}(x_i-\mu_j)^T(x-\mu_j)\right)}$$

3 pts.

(c) Derive the update step for the model parameters $\mu_c, c = 1, \ldots, K$ given the latent variables $\gamma_{i,c}$.

$$\frac{\sum_i^N \gamma_{ic} x_i}{\sum_i^N \gamma_{i,c}} = \mu_c'$$

$$Q(\theta; \theta^{(t)}) = \sum_{i=1}^{N} \sum_{c=1}^{k} \gamma_{ic} \log \frac{1}{K} \cdot (2\pi)^{-\frac{d}{2}} \exp\left[-\frac{1}{2}(x_i - \mu_c)^T (x_i - \mu_c)\right]$$ **3 pts.**

$$= \sum_{i=1}^{N} \sum_{c=1}^{k} \gamma_{ic} \left(-\log K - \frac{d}{2}\log 2\pi - \frac{1}{2}(x_i - \mu_c)^T (x_i - \mu_c)\right)$$

$$\frac{\partial Q}{\partial \mu_c} = \sum_{x \in X} \gamma_{xc} \Sigma_c^{-1} (x - \mu_c) = 0$$

Supplementary Sheet

Supplementary Sheet

Supplementary Sheet