STANFORD UNIVERSITY
CS 229, Autumn 2016
Midterm Examination

Wednesday, November 9, 6:00pm-9:00pm

Question	Points
1 Short answers	/24
2 Linear regression	/12
3 Generative models	/12
4 Generalized linear models	/22
5 Kernels	/16
6 Learning theory	/10
Total	/96

Name of Student: ______________________________

SUNetID: ______________________________@stanford.edu

The Stanford University Honor Code:

I attest that I have not given or received aid in this examination, and that I have done my share and taken an active part in seeing to it that others as well as myself uphold the spirit and letter of the Honor Code.

Signed: ______________________________

1. **[24 points] Short answers**

The following questions require a reasonably short answer (usually at most 2-3 sentences or a figure for each question part, though some may require longer or shorter explanations).

To discourage random guessing, one point will be deducted for a wrong answer on true/false or multiple choice questions. Also, no credit will be given for answers without a correct explanation.

(a) [5 points] Given a cost function $J(\theta)$ that we seek to minimize and $\alpha \in \mathbb{R} > 0$, consider the following update rule:

$$\theta^{(t+1)} = \arg\min_{\theta} \left\{ J(\theta^{(t)}) + \nabla_{\theta^{(t)}} J(\theta^{(t)})^T (\theta - \theta^{(t)}) + \frac{1}{2\alpha} \|\theta - \theta^{(t)}\|_2^2 \right\}.$$

i. [3 points] Show that this yields the same $\theta^{(t+1)}$ as the gradient descent update with step size α.

ii. [2 points] Provide a sketch (i.e. draw a picture) of the above update for the simplified case where $\theta \in \mathbb{R}$, $J(\theta) = \theta$, and $\theta^{(t)} = 1$. Make sure to clearly label $\theta^{(t)}$, $\theta^{(t+1)}$ and α.

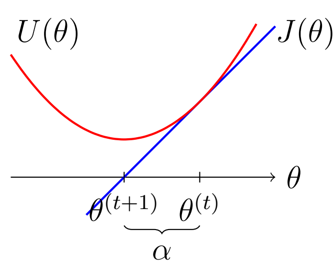
Answer:

i. Denote $U(\theta) = J(\theta^{(t)}) + \nabla_{\theta^{(t)}} J(\theta^{(t)})^T (\theta - \theta^{(t)}) + \frac{1}{2\alpha} \|\theta - \theta^{(t)}\|_2^2$. To find the minimum over θ, we compute the gradient of $U(\theta)$ w.r.t. θ and set it to 0:

$$\begin{aligned} \nabla_\theta U(\theta) &= 0 \\ \nabla_{\theta^{(t)}} J(\theta^{(t)}) + \frac{1}{2\alpha}(-2\theta^{(t)} + 2\theta) &= 0 \\ \alpha \nabla_{\theta^{(t)}} J(\theta^{(t)}) - \theta^{(t)} + \theta &= 0 \\ \implies \theta = \theta^{(t)} - \alpha \nabla_{\theta^{(t)}} J(\theta^{(t)}), & \end{aligned}$$

which is the gradient descent update, as desired. To confirm this is a minimum, we compute the Hessian $\nabla_\theta^2 U = \frac{1}{\alpha} I$ which is positive definite as expected.

ii. We provide an example sketch for $\alpha = 1$. Note that $\alpha = \theta^{(t)} - \theta^{(t+1)}$ since $\nabla_\theta J(\theta) = 1$.



(b) [4 points] In the binary classification setting where $y \in \{-1, +1\}$, we define the margin as $z = y\theta^T x$ where θ and x lie in $\mathbb{R}^n$. Consider each of the following three loss functions:

i. zero-one loss: $\varphi_{\text{zo}}(z) = 1\{z \leq 0\}$

ii. exponential loss: $\varphi_{\text{exp}}(z) = e^{-z}$

iii. hinge loss: $\varphi_{\text{hinge}}(z) = \max\{1 - z, 0\}$

Suppose we have margin $z < 0$ for our current parameters θ. Give the expression for $\frac{\partial}{\partial \theta_k} \varphi(y\theta^T x)$ for each of the given loss functions. Which loss would we fail to minimize with gradient descent, no matter the step size we choose?

Answer:

i. $\frac{\partial}{\partial \theta_k} \varphi_{\text{zo}}(y\theta^T x) = 0$

ii. $\frac{\partial}{\partial \theta_k} \varphi_{\text{exp}}(y\theta^T x) = -yx_k e^{-z}$

iii. $\frac{\partial}{\partial \theta_k} \varphi_{\text{hinge}}(y\theta^T x) = -yx_k$

Since the zero-one loss is 0 for margin $z < 0$, no matter the step size our parameter values would remain unchanged, and hence we fail to minimize the loss with gradient descent.

(c) [5 points] Consider performing spam classification where each e-mail is represented as a vector x of the same size as the number of words in the vocabulary $|V|$, where x_i is 1 if the the e-mail contains word i and 0 otherwise. We saw in class that Naive Bayes with Laplace smoothing is one simple method for performing classification in this setting. For this question, to simplify we set $p(y = 1) = p(y = -1) = 0.5$.

Consider classifying x by instead using the boosting algorithm with $2|V|$ decision stumps as the weak learners. In this setting, which of the two methods, Naive Bayes or boosting with decision stumps, would you expect to yield lower bias? Explain your reasoning.

Answer: First, note that since x is a vector of only 0s and 1s, the decision stump thresholds can all be set to any value strictly between 0 and 1 and have the same effect. One possible output of the boosting algorithm would simply be of the form $\mathrm{sign}(\theta^T[x; x])$ for $\theta \in \mathbb{R}^{2|V|}$ (where we replace the 0s in x with -1s).

For each possible word, Naive Bayes learns two parameters, $p(x_j|y = 1)$ and $p(x_j|y = -1)$, and hence also has $2|V|$ parameters (this is crucial for comparing the two classifiers!). The decision rule in log space is also linear: output $\mathrm{sign}(\sum_j \log p(x_j|y = 1) - \sum_j \log p(x_j|y = -1))$. However, Naive Bayes makes the generative modeling assumption that $p(x|y)$ is modeled by independent word counts. On the other hand, as a discriminative model boosting allows for more possible values of θ, and hence has a larger hypothesis space and should achieve lower bias.

(d) [4 points] Suppose we trained a linear SVM classifier to perform binary classification using the hinge loss $L(\theta^T x, y) = \max\{0, 1 - y\theta^T x\}$. For each of the following scenarios, does the optimal decision boundary necessarily remain the same? Explain your reasoning, perhaps by sketching a picture. Assume that after we perform the action described in each scenario we still have at least one training example in the positive class as well as in the negative class.

i. Remove all examples $(x^{(i)}, y^{(i)})$ with margin > 1.
ii. Remove all examples $(x^{(i)}, y^{(i)})$ with margin < 1.
iii. Add an ℓ_2-regularization term $\frac{\lambda}{2}\theta^T\theta = \frac{\lambda}{2}\|\theta\|_2^2$ to the training loss.
iv. Scale all $x^{(i)}$ by a constant factor α.

Answer:

i. Yes; the loss is not affected by examples with margin > 1.
ii. No; the loss is affected by these examples and hence we may have different optimal θ.
iii. No; the regularization term directly encourages θ with smaller ℓ_2-norm, hence changing the decision boundary.
iv. No; consider 1-D counter-example with $\alpha = 2$, $x^{(1)}$ at the origin, and $x^{(2)}$ at 1; the decision boundary moves from 0.5 to 1.

(e) [6 points] We consider a binary classification task where we have m training examples and our hypothesis $h_\theta(x)$ is parameterized by θ. For each of the following scenarios, select whether we should expect bias and variance to increase or decrease. Explain your reasoning.

i. Project the values of θ to lie between -1 and 1 after each training update, that is $\theta_j = \min\{1, \max\{-1, \theta_j\}\}$.

bias: increase decrease **variance:** increase decrease

ii. Smooth the estimates of our hypotheses by outputting

$$h(x) = (1/3) \sum_{x^{(i)} \in N_3(x)} h_\theta(x^{(i)}),$$

where $N_3(x)$ are the 3 points in the training set closest to x.

bias: increase decrease **variance:** increase decrease

iii. Remove one of the feature dimensions of x.

bias: increase decrease **variance:** increase decrease

Answer:

i. Bias should increase and variance should decrease since we're reducing the hypothesis space of the model.

ii. Bias should increase and variance should decrease since smoothing encourages more similar outputs for different examples. For example, consider the extreme case where we smooth by outputting the mean over all m examples; we then have very high bias and 0 variance since we make the same prediction for every input.

iii. Bias should increase and variance should decrease since for the same reason as in (i); the hypothesis space is now a strict subset of the previous space.

2. **[12 points] Linear regression: First order convergence for least squares**

   Consider the least squares problem, where we pick θ to minimize the objective $J(\theta) = \frac{1}{2}(X^T\theta - y)^T(X^T\theta - y)$. The solution to this problem is given by the normal equation, where $\theta = (XX^T)^{-1}Xy$. In Problem Set 1, we showed that a single Newton step will converge to the correct solution. Now we will examine how gradient descent performs on the same problem.

   (a) [4 points] Find the gradient of J with respect to θ, and write the gradient descent update step for $\theta^{(t+1)}$, given $\theta^{(t)}$ and step size α.

   Answer: $\nabla_\theta J = XX^T\theta - Xy$; $\theta^{(t+1)} = \theta^{(t)} - \alpha(XX^T\theta^{(t)} - Xy)$

(b) [8 points] Show that as $t \to \infty$, $\theta^{(t+1)} \to (XX^T)^{-1}Xy$, for gradient descent with step size α and $\theta^{(0)} = 0$. You may use the fact that $(\alpha A)^{-1} = \sum_{i=0}^{\infty}(I - \alpha A)^i$ for small $\alpha > 0$, and you may assume that your choice of α is small enough.

Answer: From (a), we have the gradient descent update

$$\theta^{(t+1)} = \theta^{(t)} - \alpha XX^T\theta^{(t)} + \alpha Xy$$
$$\theta^{(t+1)} = (I - \alpha XX^T)\theta^{(t)} + \alpha Xy$$
$$\theta^{(t+1)} = (I - \alpha XX^T)^{t+1}\theta_0 + \sum_{i=1}^{t+1}(I - \alpha XX^T)^{t+1-i}\alpha Xy$$
$$\theta^{(t+1)} = 0 + \alpha\sum_{i=0}^{t}(I - \alpha XX^T)^i Xy$$

As $t \to \infty$, $\sum_{i=0}^{t}(I - \alpha XX^T)^i = (\alpha XX^T)^{-1}$. Using this, we now have,

$$\theta^{(t+1)} = \alpha\alpha^{-1}(XX^T)^{-1}Xy$$
$$\theta^{(t+1)} = (XX^T)^{-1}Xy$$

3. **[12 points] Generative models: Gaussian discriminant analysis, continued**

Consider the 1-dimensional Gaussian discriminant analysis model where $x \in \mathbb{R}$ and we assume

$$
\begin{aligned}
p(y) &= \phi^{1\{y=1\}}(1-\phi)^{1\{y=-1\}} \\
p(x|y=-1) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_{-1})^2\right) \\
p(x|y=1) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_{1})^2\right)
\end{aligned}
$$

In this problem we will assume that σ *is a fixed quantity* that we have been given and is therefore not a parameter of the model.

Recall from Problem Set 1 that we can express $p(y|x;\phi,\mu_{-1},\mu_1)$ in the form

$$p(y|x;\theta) = \frac{1}{1+\exp(-y(\theta_1 x + \theta_0))}$$

where for the model described above we have,

$$
\begin{aligned}
\theta_0 &= \frac{1}{2\sigma^2}(\mu_{-1}^2 - \mu_1^2) - \log\frac{1-\phi}{\phi} \\
\theta_1 &= \frac{1}{\sigma^2}(\mu_1 - \mu_{-1}).
\end{aligned}
$$

(a) [2 points] Write the joint log-likelihood $\ell(\phi,\mu_{-1},\mu_1) = \log p(x,y;\phi,\mu_{-1},\mu_1)$ for a single example (x,y).

Answer:

$$
\begin{aligned}
p(x,y;\phi,\mu_{-1},\mu_1) &= p(y;\phi)p(x|y;\mu_{-1},\mu_1) \\
\log p(x,y;\phi,\mu_{-1},\mu_1) &= \log p(y|\phi) + \log p(x|y;\mu_{-1},\mu_1) \\
&= \log(1-\phi)^{1\{y=-1\}}\log(\phi)^{1\{y=1\}} + \log\frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}(x-\mu_y)^2
\end{aligned}
$$

(b) [7 points] Show that the log-likelihood of all training examples $\{(x^{(i)}, y^{(i)})\}_{i=1}^m$ is concave (and hence any maximum we find must be the global maximum) by first computing $\frac{\partial^2 \ell}{\partial \phi^2}$, $\frac{\partial^2 \ell}{\partial \mu_{-1}^2}$, and $\frac{\partial^2 \ell}{\partial \mu_1^2}$ for a single example (x, y). Then make an argument that the total log-likelihood is concave. Hint: Recall a function is concave if its Hessian is negative semidefinite. A one-dimensional function f is concave if $f''(x) \leq 0$ for all x.

Answer: First we show that the log-likelihood is concave for a single (x, y).

$$\frac{\partial \ell}{\partial \phi} = -1\{y = -1\}\frac{1}{1-\phi} + 1\{y = 1\}\frac{1}{\phi}$$
$$\frac{\partial^2 \ell}{\partial \phi^2} = \begin{cases} -\phi^{-2} & y = 1 \\ -(1-\phi)^{-2} & y = -1 \end{cases}$$

which is negative for both cases.

$$\frac{\partial \ell}{\partial \mu_y} = \frac{1}{\sigma^2}(x - \mu_y)$$
$$\frac{\partial^2 \ell}{\partial \mu_y^2} = -\frac{1}{\sigma^2}$$

and negative as well.

Since ϕ and μ_y are in separate terms, the Hessian H must be diagonal and negative along the diagonal. Hence H is negative semidefinite, and ℓ is concave in both ϕ and μ_y.

Due to linearity of differentiation, the sum of concave functions is concave, and thus log-likelihood over all training m examples must be concave as well.

(c) [3 points] Derive an expression for the decision boundary for classifying x as either $y = -1$ or 1.

Answer: We want $p(y = -1|x;\theta) = p(y = 1|x;\theta) = 0.5$ and hence set $\theta_1 x + \theta_0 = 0$ where θ_1 and θ_0 are given in the problem statement.

Solving, we find

$$x = \frac{2\sigma^2 \log \frac{1-\phi}{\phi} + (u_1^2 - u_{-1}^2)}{2(\mu_1 - \mu_{-1})}.$$

Note that setting $p(x|y = -1) = p(x|y = 1)$ does *not* work, since this does not take into account $p(y)$.

4. **[22 points] Generalized linear models: Gaussian distribution**

Assume we are given $x_1, x_2, \dots, x_n$ drawn i.i.d. $\sim \mathcal{N}(\mu, \sigma^2)$, that is,

$$p(x_i; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2\right)$$

Define $s^2 = \sum_{i=1}^n (x_i - \bar{x})^2$ where $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$.

(a) [3 points] Prove $g(x) = \frac{s^2}{n-1}$ is an unbiased estimator of σ^2, that is

$$\mathbb{E}[g(x)] = \sigma^2$$

Hint: $\mathbb{E}[x_i] = \mu, \mathrm{Var}(x_i) = \sigma^2, \mathrm{Cov}(x_i, x_j) = 0$.

Answer:

$$\begin{aligned}
\mathbb{E}[g(x)] &= \frac{1}{n-1}\mathbb{E}\left[\sum_{i=1}^n x_i^2 - n\bar{x}^2\right] \\
&= \frac{1}{n-1}\left(n(\sigma^2 + \mu^2) - \frac{1}{n}\left(n(\sigma^2 + \mu^2) + \mu^2 n(n-1)\right)\right) \\
&= \frac{1}{n-1}\left((n-1)(\sigma^2 + \mu^2) - (n-1)\mu^2\right) \\
&= \sigma^2
\end{aligned}$$

(b) [5 points] Find the maximum-likelihood estimate of μ and σ^2. Hint: You should be able to express your final expression for σ^2 in terms of s^2.

Answer:

$$\begin{aligned}
L &= \prod_{i=1}^{n} p(x_i; \mu, \sigma^2) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) \\
l &= -\frac{1}{2}\sum_{i=1}^{n}\left(\log 2\pi\sigma^2 + \frac{(x_i-\mu)^2}{\sigma^2}\right) \\
\nabla_{\sigma^2} l &= -\frac{1}{2}\sum_{i=1}^{n}\left(\frac{1}{\sigma^2} - \frac{(x_i-\mu)^2}{\sigma^4}\right) \\
\nabla_{\mu} l &= \frac{1}{2}\sum_{i=1}^{n}\left(\frac{2(x_i-\mu)}{\sigma^2}\right)
\end{aligned}$$

Setting $\nabla_{\sigma^2} l = 0$ and $\nabla_{\mu} l = 0$, we have

$$\begin{aligned}
\mu &= \frac{\sum_{i=1}^{n} x_i}{n} \\
\sigma^2 &= \sum_{i=1}^{n} \frac{1}{n}(x_i-\mu)^2 = \frac{s^2}{n}
\end{aligned}$$

(c) [6 points] Show that the general form of the Gaussian distribution is a member of the exponential family by finding $b(x)$, η, $T(x)$, and $a(\eta)$. Hint: Since both μ and σ^2 are parameters, η and $T(x)$ will now be two dimensional vectors. Denote $\eta = [\eta_1, \eta_2]^T$ and try to express $a(\eta)$ in terms of η_1 and η_2.

Answer:

$$\begin{aligned}
b(x) &= \frac{1}{\sqrt{2\pi}} \\
\eta &= \left[\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right]^T \\
T(x) &= \left[x, x^2\right]^T \\
a(\eta) &= \frac{\mu^2}{2\sigma^2} + \log\sigma = -\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\log(-2\eta_2)
\end{aligned}$$

(d) [4 points] Verify that $\nabla_\eta a(\eta) = \mathbb{E}[T(x);\eta]$ for the Gaussian distribution. Hint: You can prove this either by using the general form of exponential families, or by computing $\nabla_\eta a(\eta)$ directly from part (c).

Answer: In general for an exponential family,

$$\int h(x) \exp\left(\eta^T T(x) - a(\eta)\right) dx = 1$$

Thus we have

$$\begin{aligned}
a(\eta) &= \log \int h(x) \exp\left(\eta^T T(x)\right) dx \\
\nabla_\eta a(\eta) &= \frac{\int h(x) \exp\left(\eta^T T(x)\right) T(x) dx}{\int h(x) \exp\left(\eta^T T(x)\right) dx} \\
&= \frac{\int h(x) \exp\left(\eta^T T(x) - a(\eta)\right) T(x) dx}{\int h(x) \exp\left(\eta^T T(x) - a(\eta)\right) dx} \\
&= \mathbb{E}[T(x);\eta]
\end{aligned}$$

We consider the two components of η separately for the case of the Gaussian distribution:

$$\begin{aligned}
\nabla_{\eta_1} a(\eta) &= \nabla_{\eta_1}\left(-\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\log(-2\eta_2)\right) \\
&= \frac{-2\eta_1}{4\eta_2} \\
&= \mu \\
&= \mathbb{E}[x]
\end{aligned}$$

$$\begin{aligned}
\nabla_{\eta_2} a(\eta) &= \nabla_{\eta_2}\left(-\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\log(-2\eta_2)\right) \\
&= \frac{\eta_1^2}{4\eta_2^2} - \frac{1}{2\eta_2} \\
&= \mu^2 + \sigma^2 \\
&= \mathbb{E}[x^2]
\end{aligned}$$

(e) [4 points] Show that $\nabla_\eta^2 a(\eta)$ is positive semidefinite. Hint: You can compute $\nabla_\eta^2 a(\eta)$ using the results from part (c) and (d). Or instead you may use the following fact: In general for exponential families,

$$\nabla_\eta^2 a(\eta) = \mathbb{E}\left[T(x)T(x)^T\right] - \mathbb{E}[T(x)]\mathbb{E}[T(x)]^T$$

.

Answer:

Applying the above formula, we have

$$\nabla_\eta^2 a(\eta) = \begin{bmatrix} \sigma^2 & 2\mu\sigma^2 \\ 2\mu\sigma^2 & 4\mu^2\sigma^2 + 2\sigma^4 \end{bmatrix}$$

We can then confirm the Hessian (covariance of $T(x)$) is positive semidefinite:

$$\begin{aligned} z^T[\nabla_\eta^2 a(\eta)]z &= \sigma^2 z_1^2 + 4\mu\sigma^2 z_1 z_2 + 4\mu^2\sigma^2 z_2^2 + 2\sigma^4 z_2^2 \\ &= (\sigma z_1 + 2\mu\sigma z_2)^2 + 2\sigma^4 z_2^2 \\ &\geq 0. \end{aligned}$$

5. **[16 points] Shift Invariant Kernels**

   A kernel K on $\mathbb{R}^n$ is said to be shift invariant if:

$$\forall \delta \in \mathbb{R}^n, \forall x, z \in \mathbb{R}^n, K(x, z) = K(x + \delta, z + \delta)$$

   (a) [4 points] Give an example of a shift invariant and a non-shift invariant kernels seen in lectures (no need to prove they are kernels). For the rest of this problem, we will simplify a bit and consider the case where $n = 1$.

   Answer: $K(x, z) = \alpha \exp -\frac{\|x-z\|^2}{2\tau^2}$ with $\alpha \geq 0$; $K(x, z) = \beta(x^T z + 1)^d$ with $\beta \geq 0$.

(b) [6 points] Let $p(\omega)$ be a probability density over $\mathbb{R}$ and ϕ be a function mapping $\mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^d$. We define $F : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ as:

$$F(x, z) = \int_{-\infty}^{\infty} \phi(x, \omega)^T \phi(z, \omega) p(\omega) \mathrm{d}\omega$$

for all $x, z \in \mathbb{R}^n$. Show that F is a kernel.

Answer: We show that F is a kernel by showing that the Gram matrix $K_{ij} = F(x^{(i)}, x^{(j)})$ is positive semidefinite.

$$\begin{aligned}
\sum_{1 \leq i,j \leq m} z_i z_j F\left(x^{(i)}, x^{(j)}\right) &= \sum_{1 \leq i,j \leq m} z_i z_j \int_{-\infty}^{\infty} \phi(x^{(i)}, \omega)^T \phi(x^{(j)}, \omega) p(\omega) \mathrm{d}\omega \\
&= \int_{-\infty}^{\infty} p(\omega) \mathrm{d}\omega \sum_{1 \leq i,j \leq m} z_i z_j \phi(x^{(i)}, \omega)^T \phi(x^{(j)}, \omega) \\
&= \int_{-\infty}^{\infty} p(\omega) \mathrm{d}\omega \left(\sum_{i=1}^{m} z_i \phi(x^{(i)}, \omega) \right)^2 \geq 0
\end{aligned}$$

(c) [4 points] Let's suppose $n = 1$. Let $h : \mathbb{R} \to \mathbb{R}$ be a function such that

$$\forall z \in \mathbb{R}, h(z) = \int_{-\infty}^{\infty} \cos(\omega z) p(\omega) \mathrm{d}\omega$$

Show that there exists ϕ such that $h(x - z) = \int_{-\infty}^{\infty} \phi(x, \omega)^T \phi(z, \omega) p(\omega) \mathrm{d}\omega$. Provide an explicit definition of ϕ. Hint: Use the trigonometric identity $\cos(a-b) = \cos(a)\cos(b) + \sin(a)\sin(b)$, valid for all $a, b \in \mathbb{R}$.

Answer: We have $\cos(\omega(x - z)) = \cos(\omega x)\cos(\omega z) + \sin(\omega x)\sin(\omega z)$ thus we choose $\phi(x, \omega) = (\cos(\omega x), \sin(\omega x))$.

(d) [2 points] Show that $K(x, z) = h(x - z)$ is indeed a kernel.

Answer: Derives directly from applying the general result we proved in (b) to the specific case of of $\phi(x, \omega)$ for $h(x - z)$ that we gave in (c).

6. **[10 points] Learning theory: Relaxed generalization bounds**

Let $Z_1, Z_2, \ldots, Z_m$ be independent and identically distributed random variables drawn from a Bernoulli(ϕ) distribution where $P(Z_i = 1) = \phi$ and $P(Z_i = 0) = 1 - \phi$. Let $\hat{\phi} = (1/m)\sum_{i=1}^m Z_i$, and let any $\gamma > 0$ be fixed. Hoeffding's inequality, as we saw in class, states

$$\mathbb{P}(|\phi - \hat{\phi}| > \gamma) \leq 2\exp(-2\gamma^2 m)$$

However, this relies on the assumption that the random variables $Z_1, \ldots, Z_m$ are all *jointly* independent. In this problem we will relax this assumption by only assuming *pairwise* independence among the Z_i. In this case we cannot apply Hoeffding's inequality, but the following inequality (Chebyshev's inequality) holds:

$$P(|\phi - \hat{\phi}| > \gamma) \leq \frac{\mathrm{Var}(Z_i)}{m\gamma^2}$$

where $\mathrm{Var}(Z_i)$ denotes the variance of the random variable Z_i and for $Z_i \sim \mathrm{Bernoulli}(\phi)$ we have $\mathrm{Var}(Z_i) = \phi(1 - \phi)$.

Given our hypothesis set $\mathcal{H} = \{h_1, \ldots, h_k\}$ and m pairwise but not necessarily jointly independent data samples $(x, y) \sim \mathcal{D}$, we now derive guarantees on the generalization error of our best hypothesis

$$\hat{h} = \operatorname*{argmin}_{h \in \mathcal{H}} \hat{\varepsilon}(h)$$

where as usual we define $\hat{\varepsilon}(h) = \frac{1}{m}\sum_{i=1}^m 1\{h(x^{(i)}) \neq y^{(i)}\}$, where $(x^{(i)}, y^{(i)})$ are examples from the training set.

(a) [2 points] What is the maximum possible value of $\mathrm{Var}(Z_i) = \phi(1 - \phi)$? From now on we will instead use this maximal value such that the bounds we derive hold for all possible ϕ.

Answer: We find the maximum value by using the first and second order conditions. Differentiating and setting to 0 gives $\phi = 1/2$. By finding the second derivative (-2), we confirm that this point is a maximum. Hence we substitute $\mathrm{Var}(Z_i)$ with $1/4$ for the remainder of the question.

(b) [4 points] Let $\gamma > 0$.

i. [2 points] Give a non-trivial (i.e. not the constant 1) upper bound on the probability that $|\hat{\varepsilon}(\hat{h}) - \varepsilon(\hat{h})| > \gamma$.

ii. [1 points] Fix $\delta \in (0, 1)$. Using your upper bound, how large must the sample size m be before you can guarantee that

$$\mathbb{P}(|\hat{\varepsilon}(\hat{h}) - \varepsilon(\hat{h})| > \gamma) \leq \delta,$$

that is, that the training error and generalization error are within γ of one another with probability at least $1 - \delta$?

iii. [1 points] How does this sample size compare to what is achievable using Hoeffding's inequality?

Answer: We first use the Union bound to find

$$\begin{aligned} P(\exists h \in \mathcal{H}, |\varepsilon(h) - \hat{\varepsilon}(h)| > \gamma) &\leq \sum_{i=1}^{k} P(|\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma) \\ &\leq \sum_{i=1}^{k} \frac{1}{4m\gamma^2} \text{ (using Chebyshev's inequality)} \\ &= \frac{k}{4m\gamma^2}. \end{aligned}$$

(Note that applying Chebyshev's inequality to $\hat{h}$ does *not* work.) Setting this equal to δ and solving for m, we find the solution:

$$m = \frac{k}{4\delta\gamma^2}$$

Hence the number of training examples required to make this guarantee is linear in k instead of logarithmic as when we used Hoeffding's inequality.

(c) [4 points] Show that with probability at least $1-\delta$, the difference between the generalization error of $\hat{h}$ and the generalization error of the best hypothesis in $\mathcal{H}$ (i.e. the hypothesis $h^\star = \text{argmin}_{h\in\mathcal{H}}\, \varepsilon(h)$) is bounded by $\sqrt{k/(m\delta)}$.

Answer: First we solve for γ in the bound we found in (b):

$$\gamma = \sqrt{\frac{k}{4m\delta}}$$

Let $h^* = \arg\min_{h\in\mathcal{H}} \varepsilon(h)$. By uniform convergence and the definition of $\hat{h}$ (see Lecture Notes 4, page 7),

$$\varepsilon(\hat{h}) \leq \varepsilon(h^*) + 2\gamma$$

Hence $|\varepsilon(\hat{h}) - \varepsilon(h^*)| \leq 2\gamma = \sqrt{k/(m\delta)}$ as desired.