Advanced Machine Learning

Spring 2021

Prof. Joachim M. Buhmann

Final Exam

August 27th, 2021

First and last name: ______________________

Student ID number: ______________________

Signature: ______________________

General Remarks

- Please check that you have all 28 pages of this exam.
- There are 100 points in total.
- The duration of the exam is 180 minutes.
- Do not spend too much time on a single question. You do not need 100 points to get the top grade.
- Remove all material which is not permitted by the examination regulations from your desk.
- Write your answers directly on the exam sheets. If you need more space, make sure you put your student ID number on top of each supplementary sheet.
- Immediately inform an assistant in case that you are not able to take the exam under regular conditions. Later complaints are not accepted.
- Attempts to cheat/defraud lead to immediate notification to the rector's office with a possible exclusion from the examination and it might entail judicial consequences.
- Use a **black** or a **blue** pen to answer the questions. Pencils or red/green colored pens are not allowed.
- Provide only one solution to each exercise. Invalid solutions have to be clearly and unambiguously cancelled.

	Topic	Points	Points achieved	Checked
1	Regression	10		
2	Gauss-Markov	12		
3	Kernels	12		
4	Linear discriminators	10		
5	Maximum likelihood	12		
6	SVMs	10		
7	Ensembles	12		
8	ELBO	12		
9	PAC learning	10		
Total		100		

Question 1: Regression (10 pts)

Let $\{(x_1, y_1), ..., (x_n, y_n)\} \subseteq \mathbb{R}^d \times \mathbb{R}$. Denote by $X \in \mathbb{R}^{n \times d}$ the matrix where the row i represents x_i and by $Y \in \mathbb{R}^n$ the vector where entry i is y_i.

- For $\lambda > 0$, demonstrate that:

$$\beta^r := \arg\min_{\beta \in \mathbb{R}^d} \sum_{i \leq n} \left(y_i - \beta^\top x_i\right)^2 + \lambda \|\beta\|^2 = \left(X^\top X + \lambda I\right)^{-1} X^\top Y. \quad (1)$$

4 pts ☐

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

Let $L = \sum_{i \leq n} \left(y_i - \beta^\top x_i\right)^2$
Rewriting in matrix notation:
$L = (Y - X\beta)^\top (Y - X\beta) = Y^\top Y - 2Y^\top X\beta + \beta^\top X^\top X\beta$.
Form the derivative:
$\frac{\partial}{\partial \beta}(L + \lambda \beta^\top \beta) = -2X^\top Y + 2X^\top X\beta + 2\lambda\beta$.
Set the derivative to zero and solve for β: .
$\frac{\partial}{\partial \beta}(L + \lambda \beta^\top \beta) = 0 \Rightarrow (X^\top X + \lambda I)\beta = X^\top Y$
$\beta = (X^\top X + \lambda I)^{-1} X^\top Y$.

- Consider the SVD for $X = UDV^\top$ with $U \in \mathbb{R}^{n \times d}$, $D \in \mathbb{R}^{d \times d}$, and $V \in \mathbb{R}^{d \times d}$. Demonstrate that

$$X\beta^r = UD(D^2 + \lambda I)^{-1} DU^\top Y. \quad (2)$$

4 pts ☐

. . . .

. . . .

$$X\beta^r = X(X^\top X + \lambda I)^{-1} X^\top Y$$

$$= UDV^\top(\underbrace{V \underbrace{D^\top \underbrace{U^\top U}_{=I} D}_{= D^2 \text{ (as D is diag.)}} V^\top}_{= D^2 \underbrace{V V^\top}_{=I} = D^2} + \lambda I)^{-1} V D^\top U^\top Y$$

$$= UDV^\top(D^2 + \lambda I)^{-I} V D^\top U^\top Y$$

As D is diagonal, it follows that:

$$= UD(D^2 + \lambda I)^{-1} \underbrace{V^\top V}_{=I} D^\top U^\top Y$$

$$= UD(D^2 + \lambda I)^{-1} D U^\top Y$$

- Briefly explain what the equation above demonstrates about ridge regression and multi-collinearity.

2 pts ☐

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

In case of multicolinearity, the rank of X is less than full and D^2 cannot be inverted. The equation above shows that by adding λI to D^2, ridge regression ensures the equation stays solvable even if D^2 is not invertible on its own. (choosing a relatively large λ can also prevent the estimates from having a large variance.)

Question 2: Gauss-Markov **(12 pts)**

Let X be as defined in the previous question. Let $\mathbf{Y}$ be a random variable such that $\mathbf{Y} = X\beta + \xi$, with $\xi \sim \mathcal{N}(0, \sigma^2 I)$, for fixed $\beta \in \mathbb{R}^d$ and $\sigma \in \mathbb{R}$.

- Let $\hat{\beta} = \left(X^\top X\right)^{-1} X^\top \mathbf{Y}$ be an estimator of β. Mention which function of X and $\mathbf{Y}$ is minimized by $\hat{\beta}$, in the context of linear regression.

2 pts ☐

. . . .

. . . .

. . . .

$||\mathbf{Y} - X\beta||^2$

- Let $\tilde{\beta}$ be another unbiased estimator of β which has the form $C\mathbf{Y}$, for some matrix $C \in \mathbb{R}^{d\times n}$. State and demonstrate the Gauss-Markov theorem.
  Hint: You may use without proof the following fact. For any $a \in \mathbb{R}^d$,

$$Var\left(a^\top \tilde{\beta}\right) = a^\top Cov(\tilde{\beta})a \geq a^\top \sigma^2 \left(X^\top X\right)^{-1} a. \tag{3}$$

10 pts ☐

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

Gauss-Markov Theorem:
For any $a \in \mathbb{R}^d$, it holds that
$\text{Var}(a^\top \tilde{\beta}) \geq \text{Var}(a^\top \hat{\beta})$
Proof:

1. $\text{Var}(\alpha^\top \tilde{\beta}) \geq a^\top \sigma^2 \left(X^\top X\right)^{-1} a$ (inferrable from above)
2. $\text{Var}(\alpha^\top \hat{\beta}) = a^\top \text{Cov}(\hat{\beta}) a$ (inferrable from above)
3. $$\begin{aligned}\hat{\beta} &= (X^\top X)^{-1} X^\top Y \\ &= (X^\top X)^{-1} X^\top (X\beta + \xi) \\ &= \underbrace{(X^\top X)^{-1} (X^\top X \beta)}_{=\beta} + (X^\top X)^{-1} X^\top \xi \\ &= \beta + (X^\top X)^{-1} X^\top \xi\end{aligned}$$
4. $$\begin{aligned}\text{Cov}(\hat{\beta}) &= \underbrace{\text{Cov}(\beta)}_{=0} + \text{Cov}((X^\top X)^{-1} X^\top \xi) \\ &= \mathbb{E}[(X^\top X)^{-1} X^\top \xi \xi^\top X (X^\top X)^{-1}] \\ &= (X^\top X)^{-1} X^\top \underbrace{\mathbb{E}[\xi \xi^\top]}_{=\sigma^2 I} X (X^\top X)^{-1} \\ &= (X^\top X)^{-1} X^\top \sigma^2 I X (X^\top X)^{-1} \\ &= \sigma^2 I (X^\top X)^{-1} X^\top X (X^\top X)^{-1} \\ &= \sigma^2 (X^\top X)^{-1}\end{aligned}$$

Hence:
$\text{Var}(\alpha^\top \tilde{\beta}) \geq a^\top \underbrace{\sigma^2 \left(X^\top X\right)^{-1}}_{\text{Cov}(\hat{\beta})} a = a^\top \text{Cov}(\hat{\beta}) a = \text{Var}(a^\top \hat{\beta})$

Question 3: Kernels (12 pts)

Let $k_1(x, x')$ and $k_2(x, x')$ be valid kernels. Demonstrate that the following functions are also kernels. You are not allowed to use the method of composition from the lecture.

1. $ck_1(x, x')$, for $c > 0$.

3 pts ☐

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

$k(x, x') = ck_1(x, x') = ck_1(x', x) = k(x', x)$ as $k_1(x, x')$ is a valid kernel and hence symmetric.

$$\int_\Omega \int_\Omega f(x)k(x, x')f(x')dxdx' = \int_\Omega \int_\Omega f(x)ck_1(x, x')f(x')dxdx' =$$

$$= c\int_\Omega \int_\Omega f(x)k_1(x, x')f(x')dxdx' \geq 0$$

as $c > 0$, $\int_\Omega \int_\Omega f(x)k_1(x, x')f(x')dxdx' \geq 0$ since $k_1(x, x')$ is a valid kernel.

2. $k_1(x, x') + k_2(x, x')$.

3 pts ☐

. . . .

. . . .

. . . .

. . . .

. . . .

....

....

$k\left(x,x'\right)=k_1\left(x,x'\right)+k_2\left(x,x'\right)=k_1\left(x',x\right)+k_2\left(x',x\right)=k\left(x',x\right)$ as $k_1\left(x,x'\right),k_2\left(x,x'\right)$ are both valid kernels and hence symmetric.

$$\int_\Omega\int_\Omega f(x)k\left(x,x'\right)f(x')dxdx'=\int_\Omega\int_\Omega f(x)\left(k_1\left(x,x'\right)+k_2\left(x,x'\right)\right)f(x')dxdx'=$$

$$=\int_\Omega\int_\Omega f(x)k_1\left(x,x'\right)f(x')dxdx'+\int_\Omega\int_\Omega f(x)k_2\left(x,x'\right)f(x')dxdx'\geq 0$$

as $\int_\Omega\int_\Omega f(x)k_1\left(x,x'\right)f(x')dxdx'\geq 0,\ \int_\Omega\int_\Omega f(x)k_1\left(x,x'\right)f(x')dxdx'\geq 0$

since $k_1\left(x,x'\right),k_2\left(x,x'\right)$ are valid kernels.

3. $k_1(x,x')\,k_2(x,x')$.

3 pts

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

$k(x,x') = k_1(x,x')\,k_2(x,x') = k_1(x',x)\,k_2(x',x) = k(x',x)$ as $k_1(x,x'), k_2(x,x')$ are both valid kernels and hence symmetric.
Since $k_1(x,x'), k_2(x,x')$ they admit an inner product representation. Let $phi(x)$ be a feature map for $k_1(x,x')$ and let $\psi(x)$ be a feature map for $k_2(x,x')$.

$$k(x,x') = k_1(x,x')\,k_2(x,x') = \left(\phi(x)^\top\phi(x')\right)\left(\psi(x)^\top\psi(x')\right) =$$
$$= \left(\sum_i^\infty \phi_i(x)\phi_i(x')\right)\left(\sum_j^\infty \psi_j(x)\psi_j(x')\right) = \left(\sum_i^\infty\sum_j^\infty \phi_i(x)\phi_i(x')\psi_j(x)\psi_j(x')\right) =$$
$$= \left(\sum_i^\infty\sum_j^\infty \phi_i(x)\psi_j(x)\phi_i(x')\psi_j(x')\right) = \left(\sum_{ij}^\infty \eta_{ij}(x)\eta_{ij}(x')\right) = \eta(x)^\top\eta(x')$$
$$\text{where } \eta_{ij}(x) = \phi_i(x)\psi_j(x)$$

So $k(x,x')$ can be represented as an inner product with a feature map $\eta(x)$ and, hence, is a valid kernel (Mercer Theorem).

4. Show that if A is symmetric and positive definite, then $k(x, x') = x^\top A x'$ is a valid kernel.

3 pts ☐

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

$k(x, x') = x^\top A x' = (x^\top A x')^\top = x'^\top A^\top x = x'^\top A x = k(x', x)$ as $x^\top A x' \in \mathbb{R}$ and A is positive definite and, therefore, symmetric.
Since A is PS, it allows Cholesky decomposition $A = L^\top L$. Therefore, $k(x, x') = x^\top A x' = x^\top L^\top L x' = (Lx)^\top (Lx) = \phi(x)^\top \phi(x')$, where $\phi(x) = Lx$. According to Mercer Theorem, $k(x, x')$ is a valid kernel.

Question 4: Linear discriminators (10 pts)

Consider two random variables (X, Y) taking values in $\mathbb{R}^d \times \{-1, +1\}$, such that

- $P(Y = +1) = P(Y = -1)$,
- $X \mid Y = +1 \ \sim \ \mathcal{N}(\mu_+, \Sigma)$, and
- $X \mid Y = -1 \ \sim \ \mathcal{N}(\mu_-, \Sigma)$, for some $\mu_+, \mu_- \in \mathbb{R}^d$ and some symmetric positive definite $\Sigma \in \mathbb{R}^{d\times d}$.

Formally demonstrate that

$$P(Y = +1 \mid X = x) = \sigma\left(\beta^\top x + \beta_0\right), \tag{4}$$

for some $\beta \in \mathbb{R}^d$ and $\beta_0 \in \mathbb{R}$. Here, σ is the sigmoid function $\sigma(t) = \frac{1}{1+e^{-t}}$.
Derive formulas for β and β_0.

10 pts ☐

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

$P(Y = +1 \mid x) = p(x \mid +1)P(+1)/p(x) = p(x \mid +1)P(+1)/(p(x| + 1)P(+1) + p(x| - 1)P(-1)) = \frac{1}{1+\frac{p(x|-1)P(-1)}{p(x|+1)P(+1)}}$

$\frac{p(x|-1)P(-1)}{p(x|+1)P(+1)} = \frac{p(x|-1)}{p(x|+1)} = e^{-\frac{1}{2}(x-\mu_-)^T\Sigma(x-\mu_-)+\frac{1}{2}(x-\mu_+)^T\Sigma(x-\mu_+)}$

$\beta^T x+\beta_0 = \frac{1}{2}(x-\mu_-)^T\Sigma(x-\mu_-)-\frac{1}{2}(x-\mu_+)^T\Sigma(x-\mu_+) = \mu_-^T\Sigma x-\mu_+^T\Sigma x+\frac{1}{2}\mu_-^T\Sigma\mu_- -\frac{1}{2}\mu_+^T\Sigma\mu_+$

$\beta = \Sigma(\mu_- - \mu_+)$, $\beta_0 = \frac{1}{2}\mu_-^T\Sigma\mu_- - \frac{1}{2}\mu_+^T\Sigma\mu_+$.

Question 5: Constrained optimization and maximum likelihood (12 pts)

Assume given a partition of $[0, 1]$ into M disjoint intervals $B_1, \ldots, B_M$. A *piecewise constant pdf* is a pdf $p : [0, 1] \to \mathbb{R}^+$ that is constant on each B_j, for $j \leq M$. That is, for all $x \in B_j$, $p(x) = c_j$, for some $c_j \in \mathbb{R}^+$.

Assume that the intervals $B_1, \ldots, B_M$ are fixed and that you are given a sample $\{x_1, \ldots, x_n\} \subseteq [0, 1]$. Your task is to estimate $c_1, \ldots, c_M$ using maximum likelihood.

a) Let $b : [0, 1] \to \{1, \ldots, M\}$ be the function that maps points in $[0, 1]$ to their corresponding interval in the partition, i.e. $b(x) = j$ iff $x \in B_j$, for $j \leq M$. Write the log likelihood of the sample with respect to $c_1, \ldots, c_M$ and b.

2 pts ☐

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

$\sum_{i=1}^{n} \log c_{b(x_i)}$

b) For $j \leq M$, let V_j be the length of the interval B_j. Formulate the problem of computing $c_1, \ldots, c_M$ that maximize the sample's log likelihood as a constrained optimization problem.

4 pts ☐

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

p is a pdf so we need to impose the constraint that $\sum_{j=1}^{M} c_j V_j = 1$. The constrained optimization problem for MLE is:

$$\max_{c_1,\dots c_M} \sum_{i=0}^{n} \log c_{b(x_i)} \text{ s.t. } \sum_{j=0}^{M} c_j V_j = 1, \forall j \in [M]$$

c) Use Lagrange multipliers to show that the solution for the problem above is given by

$$c_j = \frac{K_j}{nV_j}, \text{ for } j \leq M, \tag{5}$$

where K_j is the number of elements in the sample that are in B_j.

6 pts ☐

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

Using Lagrange multipliers, we obtain the objective function:

$$J(c_i; \lambda) := \sum_{i=1}^{n} \log c_{b(x_i)} + \lambda \left(\sum_{j=1}^{M} c_j V_j - 1 \right)$$

Writing the first order condition wrt c_j we obtain:

$$\frac{K_j}{c_j} + \lambda V_j \stackrel{!}{=} 0$$

It follows that:

$$c_j = \frac{-K_j}{\lambda V_j},$$

and from the constraint we get that:

$$-\sum_{j=1}^{M} \frac{K_j}{\lambda} = 1 \Rightarrow \lambda = -\sum_{j=1}^{M} K_j \Rightarrow \lambda = -n$$

Introducing λ in the first order condition gives the final result:

$$c_j = \frac{K_j}{nV_j}, \forall j \in \{1, ..., M\}$$

Question 6: SVMs (10 pts)

Let $\{(x_1, y_1), \dots, (x_n, y_n)\} \subseteq \mathbb{R}^d \times \{+1, -1\}$. Consider the standard problem of training an SVM:

$$\min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \quad \frac{1}{2} \|w\|^2 \tag{6}$$

$$\text{s.t.} \quad y_i \left(w^\top x_i + b \right) \geq 1, \text{ for } i \leq n. \tag{7}$$

1. Formulate the Lagrangian for this constrained optimization problem.

2 pts ☐

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

2. Formulate the dual for this optimization problem and rewrite it so that w never appears.

3 pts ☐

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

3. Prove that if there exists $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ that fulfil the constraints of this problem, then this problem satisfies Slater's condition.

3 pts ☐

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

4. Suppose that Slater's condition is not fulfilled. What is the consequence of this for strong duality? Briefly explain.

2 pts ☐

. . . .

. . . .

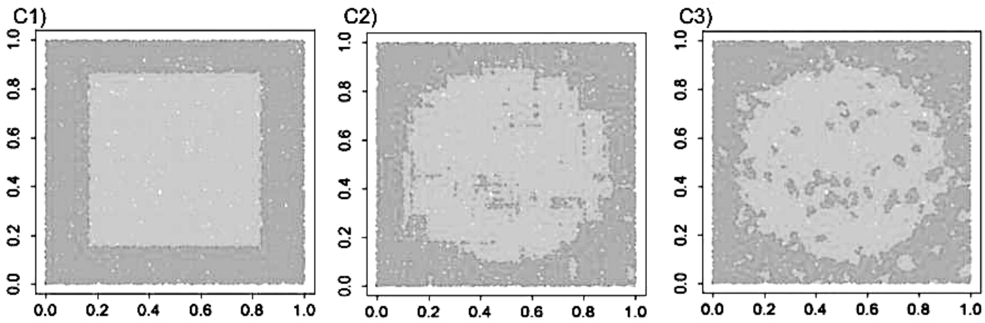
. . . .

. . . .

Question 7: Ensembles **(12 pts)**

1. The following figure, from the paper "Explaining the Success of AdaBoost and Random Forests as Interpolating Classifiers", by Wyler et al, depicts a dataset for binary classification and three ensembles trained on that set. Match the classifiers to their corresponding plots $(C1, C2, C3)$.

3 pts ☐



- AdaBoost with deep decision trees ...
- A shallow decision tree ...
- A 1NN classifier ...

- AdaBoost with deep decision trees C2
- A shallow decision tree C1
- A 1NN classifier C3

2. Let $f : \mathbb{R}^d \to \mathbb{R}$ and p be a distribution over $\mathbb{R}^d$. Suppose that we draw a sample $Z = \{X_1, \ldots, X_n\}$ from p and train M regression functions $f_1, \ldots, f_M$ by repeatedly running some randomized algorithm on $\{(X_1, f(X_1)), \ldots, (X_n, f(X_n))\}$. Assume that $\mathbb{E}_Z[f_i(x)] = f(x)$, for all $x \in \mathbb{R}^d$ and $i \le M$.
Let

$$\bar{f}(x) := \frac{1}{M} \sum_{i \le M} f_i(x). \tag{8}$$

Demonstrate that the ensemble $\bar{f}$ is a better estimator of f than any f_i. That is, for any $i \in 1, \ldots, M$:

$$\mathbb{E}_Z\left[\left(f(x) - \bar{f}(x)\right)^2\right] \le \mathbb{E}_Z\left[(f(x) - f_i(x))^2\right], \text{ for any } x \in \mathbb{R}^d. \tag{9}$$

9 pts ☐

. . . .

Proof 1.

$$
\begin{aligned}
\mathbb{E}_Z\left[\left(f(x)-\bar{f}(x)\right)^2\right] &= \mathbb{E}_Z\left[\left(\mathbb{E}_Z\left[\bar{f}(x)\right]-\bar{f}(x)\right)^2\right] = \mathsf{Var}_Z\left(\bar{f}(x)\right) \\
&= \frac{1}{M^2}\sum_i \mathsf{Var}_Z(f_i(x)) = \frac{1}{M}\mathsf{Var}_Z(f_i(x)) \leq \mathsf{Var}_Z(f_i(x)) \\
&= \mathbb{E}_Z\left[\left(\mathbb{E}_Z\left[f_i(x)\right]-f_i(x)\right)^2\right] = \mathbb{E}_Z\left[\left(f(x)-f_i(x)\right)^2\right]
\end{aligned}
$$

Proof 2.

$$
\begin{aligned}
\mathbb{E}_Z\left[\left(f(x)-\bar{f}(x)\right)^2\right] &= \mathbb{E}_Z\left[\left(\frac{1}{M}\sum_i \left(f(x)-f_i(x)\right)\right)^2\right] \\
&= \frac{1}{M^2}\mathbb{E}_Z\left[\sum_{i,j}\left(f(x)-f_i(x)\right)\left(f(x)-f_j(x)\right)\right] \\
&= \underbrace{\frac{1}{M^2}\sum_{i\neq j}\mathbb{E}_Z\left[\left(f(x)-f_i(x)\right)\left(f(x)-f_j(x)\right)\right]}_{0} + \frac{1}{M^2}\sum_i \mathbb{E}_Z\left[\left(f(x)-f_i(x)\right)^2\right] \\
&= \frac{1}{M}\mathsf{Var}(f_i(x)) \leq Var(f_i(x))
\end{aligned}
$$

Question 8: Evidence lower bound **(12 pts)**

Recall the notation from the lecture:

- $p_{\theta'}(\cdot)$ is a distribution over a representation space $\mathcal{Z}$, parametrized by $\theta' \in \Theta'$.
- For $z \in \mathcal{Z}$, $p_\theta(\cdot|z)$ is a conditional distribution over a measurement space $\mathcal{X}$, parametrized by $\theta \in \Theta$.
- For $x \in \mathcal{X}$, $q_\phi(\cdot|x)$ is a tractable distribution over $\mathcal{Z}$, parametrized by $\phi \in \Phi$. It is intended to approximate $p_{\theta,\theta'}(\cdot|x) \propto p_{\theta'}(\cdot)p_\theta(x|\cdot)$.

Let $\{x_1, \ldots, x_m\} \subseteq \mathcal{X}$. Demonstrate that:

$$\sum_{i \leq n} \log p_{\theta,\theta'}(x_i) \geq \sum_{i \leq n} \mathsf{elbo}_{\theta',\theta,\phi}(x_i), \tag{10}$$

where

$$\mathsf{elbo}_{\theta',\theta,\phi}(x_i) := \mathbb{E}_{z \sim q_\phi(\cdot|x_i)}[\log(p_\theta(x_i|z))] + \mathbb{E}_{z \sim q_\phi(\cdot|x_i)}\left[\log\left(\frac{p_{\theta'}(z)}{q_\phi(z|x_i)}\right)\right] \tag{11}$$

Hints:

- Prove that $\mathbb{E}_{z \sim q_\phi(\cdot|x_i)}[\log p_{\theta,\theta'}(x_i)] = \mathbb{E}_{z \sim q_\phi(\cdot|x_i)}\left[\log\left(\frac{p_{\theta,\theta'}(x_i,z)}{p_{\theta',\theta}(z|x_i)} \frac{q_\phi(z|x_i)}{q_\phi(z|x_i)}\right)\right]$
- Recall that $D_{KL}(P_1||P_2) = \int P_1(z) \log \frac{P_1(z)}{P_2(z)} dz \geq 0$

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

. . . .

Solution:

$$\log p_{\theta,\theta'}(x_i) = \mathbb{E}_{z\sim q_\phi(\cdot|x_i)}[\log p_{\theta,\theta'}(x_i)] = \mathbb{E}_{z\sim q_\phi(\cdot|x_i)}\left[\log\left(\frac{p_{\theta,\theta'}(x_i,z)}{p_{\theta',\theta}(z|x_i)}\frac{q_\phi(z|x_i)}{q_\phi(z|x_i)}\right)\right] = \quad (12)$$

$$= \mathbb{E}_{z\sim q_\phi(\cdot|x_i)}\left[\log\left(\frac{p_{\theta,\theta'}(x_i,z)}{q_\phi(z|x_i)}\right)\right] + \mathbb{E}_{z\sim q_\phi(\cdot|x_i)}\left[\log\left(\frac{q_\phi(z|x_i)}{p_{\theta',\theta}(z|x_i)}\right)\right] = \quad (13)$$

$$= \mathsf{elbo}_{\theta',\theta,\phi}(x_i) + D_{KL}(q_\phi(z|x_i)||p_{\theta',\theta}(z|x_i)) \geq \quad (14)$$

$$\geq \mathsf{elbo}_{\theta',\theta,\phi}(x_i) \quad (15)$$

Q.E.D

Question 9: PAC Learning (10 pts)

1. Consider the definition of PAC learnability:

   A concept class $\mathcal{C}$ is PAC learnable from a hypothesis class $\mathcal{H}$ if there is an algorithm $\mathcal{A}$ s.t. for all $0 < \epsilon, \delta < \frac{1}{2}$, for any distribution on the instance space $\mathcal{X}$, and for any $c \in \mathcal{C}$, there is a polynomial function $poly(\cdot, \cdot, \cdot)$ s.t. if $\mathcal{A}$ receives a sample $\mathcal{Z}$ of size $n \geq poly\left(\frac{1}{\epsilon}, \frac{1}{\delta}, size(c)\right)$, then $\mathcal{A}$ outputs a hypothesis $\hat{c}$ such that $\mathbf{P}\left(\mathcal{R}\left(\hat{c}\right) \leq \epsilon\right) \geq 1 - \delta$.

   Fill in the gaps to complete the definition of *efficient PAC learnability*:
   We say that $\mathcal{A}$ is an efficient PAC algorithm if it runs in time __________ in ______ and ______.

   3 pts ☐

   We say that $\mathcal{A}$ is an efficient PAC algorithm if it runs in time *polynomial* in $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$.

   Schema:

   - ? points for each correctly filled in gap.

2. Let $\mathbf{P}$ be a distribution over $\mathcal{X} \subseteq \mathbb{R}^2$. Let $\mathcal{C}$ be a set of functions mapping $\mathbb{R}^2 \to \{-1, +1\}$.

   - For $c \in \mathcal{C}$, a c*-dataset of size* n is a set of the form $\mathcal{Z} = \{(X_1, c(X_1)), \ldots, (X_n, c(X_n))\}$ with $X_1, \ldots, X_n$ a sample drawn from $\mathbf{P}$.
   - Assume that you are given an algorithm with the following specification:
     - It takes as input any c-dataset of size $n \in \mathbb{N}$, for an arbitrary function $c \in \mathcal{C}$.
     - The output is a function $\hat{c}^* \in \mathcal{C}$ such that

$$\mathbf{P}\left(\mathcal{R}(\hat{c}^*) \leq \epsilon\right) \geq 1 - 4\exp\left(-\frac{n\epsilon}{4}\right), \text{ for any } 0 < \epsilon < \frac{1}{2}. \quad (16)$$

   Show that the class $\mathcal{C}$ above is PAC-learnable from itself. You may assume that $size(c) = 1$, for any $c \in \mathcal{C}$.

   7 pts ☐

   ....

   ....

   ....

   ....

   ....

   ....

. . . .

. . . .

. . . .

. . . .

Fix $\delta = 4\exp\left(-\frac{n\epsilon}{4}\right)$. Choose $n = \frac{4}{\epsilon}\log\left(\frac{4}{\delta}\right)$, then $\mathbf{P}\left(\mathcal{R}(\hat{c}^*) \leq \epsilon\right) \geq 1 - \delta$. Observe that $\frac{4}{\epsilon}\log\left(\frac{4}{\delta}\right)$ is polynomial in $\frac{1}{\epsilon}$, $\frac{1}{\delta}$, and $size(c) = 1$. Thus, we have an algorithm $\mathcal{A}$ which returns $\hat{c}^*(\cdot)$ s. t. $\mathbf{P}\left(\mathcal{R}(\hat{c}^*) \leq \epsilon\right) \geq 1 - \delta$, for all $0 < \epsilon, \delta < \frac{1}{2}$, for $n \geq poly(\frac{1}{\epsilon}, \frac{1}{\delta}, size(c))$. Hence, $\mathcal{C}$ is PAC-learnable from itself.

Schema:

- ? point for defining δ;
- ? points for deriving $n = \frac{4}{\epsilon}\log\left(\frac{4}{\delta}\right)$;
- ? point for stating that $\frac{4}{\epsilon}\log\left(\frac{4}{\delta}\right)$ is polynomial in $\frac{1}{\epsilon}$, $\frac{1}{\delta}$, and $size(c)$.

Supplementary Sheet