Exercises
**Introduction to Machine Learning**
SS 2018

**Institute for Machine Learning**
Dept. of Computer Science, ETH Zürich
**Prof. Dr. Andreas Krause**
Web: https://las.inf.ethz.ch/teaching/introml-s18
**Email questions to:**
Mohammad Reza Karimi, mkarimi@ethz.ch

# Series 7, May 22, 2018
# (EM Convergence)

*The problems marked with an asterisk * are intended for deeper understanding. One should look at these problems as an opportunity to have more insight into the theory. Note that these problems are not necessarily harder than the other ones.*

**Problem 1 (EM for Censored Linear Regression):**

Suppose you are trying to learn a model that can predict how long a program will take to run for different settings. In some situations, when the program is taking too long, you abort the program and just note down the time at which you aborted. These values are *lower bounds* for the actual running time of the program. We call this type of data **right-censored**. Concretely, all you know is that the running time $y_i \geq c_i$, where $c_i$ is the censoring time. Written in another way, one can say $y_i = \min\{z_i, c_i\}$ where $z_i$ is the true running time. Our goal is to derive an EM algorithm for fitting a linear regression model to right-censored data.

(a) Let $z_i = \mu_i + \sigma\varepsilon_i$, where $\varepsilon_i \sim \mathcal{N}(0,1)$. Suppose that we do not observe $z_i$, but we observe the fact that it is higher than some threshold. Namely, we observe the event $E = \mathbb{I}(z_i \geq c_i)$. Show that

$$\mathbb{E}[z_i \,|\, z_i \geq c_i] = \mu_i + \sigma R\left(\frac{c_i - \mu_i}{\sigma}\right)$$

and

$$\mathbb{E}[z_i^2 \,|\, z_i \geq c_i] = \mu_i^2 + \sigma^2 + \sigma(c_i + \mu_i) R\left(\frac{c_i - \mu_i}{\sigma}\right),$$

where we have defined

$$R(x) := \frac{\phi(x)}{1 - \Phi(x)}.$$

Here, $\phi(x)$ is the pdf of the standard Gaussian, and $\Phi(x)$ is its cdf.

(b) Derive the EM algorithm for fitting a linear regression model to right-censored data. Describe completely the E-step and M-step.

**Solution 1:**

**(a)** First note that $p(\varepsilon_i|E) = \frac{p(\varepsilon_i,E)}{p(E)}$. Also for brevity, define $a_i := \frac{c_i-\mu_i}{\sigma}$. Then we have $E = \mathbb{I}(z_i \geq c_i) = \mathbb{I}(\varepsilon_i \geq a_i)$. So we can write

$$\begin{aligned}
\mathbb{E}[z_i \,|\, z_i \geq c_i] &= \int_{\mathbb{R}} z_i \, p(\varepsilon_i|E) \, d\varepsilon_i \\
&= \int_{\mathbb{R}} z_i \, \frac{p(\varepsilon_i, E)}{p(E)} \, d\varepsilon_i \\
&= \frac{1}{p(E)} \int_{a_i}^{\infty} (\mu_i + \sigma\varepsilon_i) \, p(\varepsilon_i) \, d\varepsilon_i \\
&= \mu_i + \frac{\sigma}{p(E)} \int_{a_i}^{\infty} \varepsilon_i \, p(\varepsilon_i) \, d\varepsilon_i
\end{aligned}$$

The equality follows from the fact that $p(E) = \int_{a_i}^{\infty} p(\varepsilon_i) \, d\varepsilon_i = 1 - \Phi(a_i)$. Now observe that for the Standard Normal distribution density $\phi(x)$, we have

$$\frac{d}{dx}\phi(x) = -x \, \phi(x),$$

implying that

$$\int_{a_i}^{\infty} \varepsilon_i \, p(\varepsilon_i) \, d\varepsilon_i = \phi(a_i) - \phi(+\infty) = \phi(a_i).$$

Putting all together we get

$$\mathbb{E}[z_i \,|\, z_i \geq c_i] = \mu_i + \sigma \frac{\phi(a_i)}{1 - \Phi(a_i)} = \mu_i + \sigma R\left(\frac{c_i - \mu_i}{\sigma}\right).$$

To compute $\mathbb{E}[z_i^2 \,|\, z_i \geq c_i]$, we first note that

$$\frac{d^2}{dx^2}\phi(x) = -\phi(x) + x^2\phi(x),$$

implying that

$$\int_a^b x^2 \phi(x) \, dx = \Phi(b) - \Phi(a) + a\phi(a) - b\phi(b). \tag{1}$$

Now we have

$$\begin{aligned}
\mathbb{E}[z_i^2 \,|\, z_i \geq c_i] &= \frac{1}{p(E)} \int_{a_i}^{\infty} (\mu_i^2 + 2\mu_i\sigma\varepsilon_i + \sigma^2\varepsilon_i^2) \, p(\varepsilon_i) \, d\varepsilon_i \\
&= \mu_i^2 + \frac{2\mu_i\sigma}{p(E)} \int_{a_i}^{\infty} \varepsilon_i \, p(\varepsilon_i) \, d\varepsilon_i + \frac{\sigma^2}{p(E)} \int_{a_i}^{\infty} \varepsilon_i^2 \, p(\varepsilon_i) \, d\varepsilon_i \\
&= \mu_i^2 + \frac{2\mu_i\sigma}{p(E)}\phi(a_i) + \frac{\sigma^2}{p(E)}(1 - \Phi(a_i) + a_i\phi(a_i)) \\
&= \mu_i^2 + \frac{2\mu_i\sigma}{1 - \Phi(a_i)}\phi(a_i) + \frac{\sigma^2}{1 - \Phi(a_i)}(1 - \Phi(a_i) + a_i\phi(a_i)) \\
&= \mu_i^2 + \sigma^2 + (2\mu_i\sigma + a_i\sigma^2)R(a_i) \\
&= \mu_i^2 + \sigma^2 + \sigma(\mu_i + c_i)R(a_i)
\end{aligned}$$

**(b)** The model we have for linear regression is $z_i \sim \mathcal{N}(\mathbf{w}^\top \mathbf{x}_i, \sigma)$ where $z_i$ is missing. Our observed variable is $y_i = \min\{z_i, c_i\}$. For ease of notation, let

$$d_i = \begin{cases} 1 & \text{if } z_i \leq c_i \\ 0 & \text{if } z_i > c_i \end{cases}$$

to be the censoring indicator, i.e. it is 1 if the observation is not censored, and is 0 otherwise. We denote by $\mathbf{z}$ the set of all $z_i$'s, by $\mathbf{X}$ the set of all $\mathbf{x}_i$'s, by $\mathbf{y}$ the set of all $y_i$'s, by $\mathbf{c}$ the set of all $c_i$'s, and by $\mathbf{d}$ the set of all $d_i$'s.

The complete-data log-likelihood would be

$$\log p(z_i \,|\, \mathbf{w}) = -\frac{1}{2\sigma^2}(z_i - \mathbf{w}^\top \mathbf{x}_i)^2 + \text{const.}$$

For the first step, we need to find the posterior of the missing data given the observed data and parameters. We have

$$p(z_i \,|\, \underbrace{\mathbf{x}_i, y_i, c_i, d_i}_{\text{observed}}, \mathbf{w}) = \begin{cases} \delta(z_i - y_i) & \text{if } d_i = 1 \\ \frac{\mathcal{N}(z_i \,|\, \mathbf{w}^\top \mathbf{x}_i, \sigma)}{1 - \Phi\left(\frac{c_i - \mathbf{w}^\top \mathbf{x}_i}{\sigma}\right)} & \text{if } d_i = 0 \end{cases},$$

in which $\delta(\cdot)$ is the dirac delta function, and $1 - \Phi(\frac{c_i - \mu_i}{\sigma})$ is the probability that $z_i > c_i$.

Now we should compute the expected value of the complete-data log-likelihood w.r.t the posterior $p(z_i \,|\, \mathbf{x}_i, y_i, c_i, d_i, \mathbf{w}')$. This can be computed as

$$\int_{\mathbb{R}} \log p(z_i \,|\, \mathbf{w}) \cdot p(z_i \,|\, \mathbf{x}_i, y_i, c_i, d_i, \mathbf{w}')\, dz_i.$$

Note that if $d_i = 1$, the integral is evaluated as $-\frac{1}{2\sigma^2}(y_i - \mathbf{w}^\top \mathbf{x}_i)^2$, and if $d_i = 0$, we can use part (a) to compute the expectation. For ease of notation, we call $\mu_i := \mathbf{w}^\top \mathbf{x}_i$ and $\mu_i' := \mathbf{w}'^\top \mathbf{x}_i$ and $a_i = \frac{c_i - \mu_i'}{\sigma}$. We then have

$$\begin{aligned} \mathbb{E}[\log p(z_i \,|\, \mathbf{w}) \,|\, z_i > c_i] &= -\frac{1}{2\sigma^2}\left\{\mu_i^2 + \mathbb{E}[z_i^2 \,|\, z_i > c_i] - 2\mu_i\, \mathbb{E}[z_i \,|\, z_i > c_i]\right\} \\ &= -\frac{1}{2\sigma^2}\left\{\mu_i^2 + \mu_i' + R(a_i) - 2\mu_i \underbrace{((\mu_i')^2 + \sigma^2 + \sigma(\mu_i' + c_i)R(a_i))}_{:=b_i}\right\}. \end{aligned}$$

Adding the evaluated expectation for all data, and removing the terms that are not dependent on $\mathbf{w}$ we get

$$\begin{aligned} Q(\mathbf{w}, \mathbf{w}') &= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu_i)^2 \cdot d_i + (\mu_i^2 - 2b_i\mu_i) \cdot (1 - d_i) \\ &\doteq -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\mu_i^2 - 2y_i\mu_i) \cdot d_i + (\mu_i^2 - 2b_i\mu_i) \cdot (1 - d_i) \\ &= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\mu_i^2 - 2\mu_i \underbrace{(y_i d_i + b_i(1 - d_i))}_{:=e_i} \\ &= -\frac{1}{2\sigma^2}(\mathbf{w}^\top \mathbf{X}^\top \mathbf{X}\mathbf{w} - 2\mathbf{w}^\top \mathbf{X}^\top \mathbf{e}) \end{aligned}$$

The maximizer for $Q(\mathbf{w}, \mathbf{w}')$ would be

$$\mathbf{w}^\star = -\frac{1}{2}(\mathbf{X}^\top \mathbf{e})(\mathbf{X}^\top \mathbf{X})^{-1},$$

which sums up the M-step.

**Problem 2 (Soft $k$-means, Revisited):**

(a) Consider the following optimization problem:

$$\max_{\mathbf{c}\in\mathbb{R}^k} \sum_{i=1}^{k} v_i \log(c_i) \quad \text{s.t.} \quad c_i > 0, \ \sum_{i=1}^{k} c_i = 1,$$

where $\mathbf{v} \in \mathbb{R}_+^k$ is a vector of non-negative weights. Check that the M-step of soft $k$-means includes solving such an optimization problem.

(b) Let $\mathbf{c}^\star = \frac{1}{\sum_i v_i}\mathbf{v}$. Verify that $\mathbf{c}^\star$ is a probability vector.

(c) Show that the optimization problem is equivalent to the following problem:

$$\min_{\mathbf{c}\in\mathbb{R}^k} \mathrm{D}_{\mathrm{KL}}(\mathbf{c}^\star \| \mathbf{c}) \quad \text{s.t.} \quad c_i > 0, \ \sum_{i=1}^{k} c_i = 1,$$

(d) Using the properties of KL divergence, prove that $\mathbf{c}^\star$ is indeed the solution to the optimization problem.

**Solution 2:**

(a) check the slides of the course.

(b) The components of $\mathbf{c}^\star$ are non-negative (since $\mathbf{v}$ is nonnegative), and add up to 1.

(c) Since the optimization is over $\mathbf{c}$, it makes no difference if we divide it by a positive number or add/subtract terms which are not dependent on $\mathbf{c}$. We first divide the objective by $\sum_i v_i$ and then subtract from the sum $\sum_{i=1}^k c_i^\star \log c_i^\star$. We get

$$\sum_{i=1}^k c_i^\star \log(c_i) - \sum_{i=1}^k c_i^\star \log(c_i^\star) = \sum_{i=1}^k c_i^\star \log \frac{c_i}{c_i^\star} = -\mathrm{D}_{\mathrm{KL}}(\mathbf{c}^\star \| \mathbf{c}).$$

Thus maximizing the objective is equivalent to minimizing $\mathrm{D}_{\mathrm{KL}}(\mathbf{c}^\star \| \mathbf{c})$.

(d) Since KL Divergence is always nonnegative and is zero if and only if the two distributions are equal, we get that the optimal solution to the optimization problem is indeed $\mathbf{c} = \mathbf{c}^\star$.

**Problem 3 (Yet another perspective on EM):**

The EM algorithm is a general technique for finding maximum likelihood solutions for probabilistic models having latent variables. Take a probabilistic model in which we denote all of the *observed* variables as $\mathbf{X}$ and all of the hidden variables as $\mathbf{Z}$ (here we assume $\mathbf{Z}$ is discrete, for the sake of simplicity). Let us assume that the joint distribution is $p(\mathbf{X}, \mathbf{Z} \,|\, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the set of all parameters describing this distribution (*e.g.* for a Gaussian distribution, $\boldsymbol{\theta} = (\mu, \Sigma)$). The goal is to maximize the likelihood function

$$p(\mathbf{X} \,|\, \boldsymbol{\theta}) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z} \,|\, \boldsymbol{\theta}).$$

(a) For an arbitrary distribution $q(\mathbf{Z})$ over the latent variables, show that the following decomposition holds:

$$\ln p(\mathbf{X} \,|\, \boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{D}_{\mathrm{KL}}(q \| p_{\mathrm{post}}), \tag{2}$$

where $p_{\mathrm{post}} = p(\mathbf{Z} \,|\, \mathbf{X}, \boldsymbol{\theta})$ is the posterior distribution. Also find the formulation of $\mathcal{L}(q, \boldsymbol{\theta})$.

(b) Verify that $\mathcal{L}(q, \boldsymbol{\theta}) \leq \ln p(\mathbf{X} \,|\, \boldsymbol{\theta})$, and that equality holds if and only if $q(\mathbf{Z}) = p(\mathbf{Z} \,|\, \mathbf{X}, \boldsymbol{\theta})$.

(c) Suppose that the current value of the parameters is $\boldsymbol{\theta}_{\mathrm{curr}}$. Verify that in the E-step, the lower bound $\mathcal{L}(q, \boldsymbol{\theta}_{\mathrm{curr}})$ is maximized with respect to the distribution $q(\mathbf{Z})$, while keeping $\boldsymbol{\theta}_{\mathrm{curr}}$ fixed. Since the left-hand-side of (2) does not depend on $q(\mathbf{Z})$, maximizing $\mathcal{L}(q, \boldsymbol{\theta}_{\mathrm{curr}})$ will result in minimizing the KL divergence between $q$ and $p_{\mathrm{post}}$, which happens at $q^\star = p_{\mathrm{post}}$.

(d) Verify that in the M-step, the lower bound $\mathcal{L}(q, \boldsymbol{\theta})$ is maximized with respect to $\boldsymbol{\theta}$ while keeping $q(\mathbf{Z})$ fixed, resulting in a new value of parameters $\boldsymbol{\theta}_{\mathrm{new}}$. This step will result in an increase in left-hand-side of (2) (if it is not already in a local maximum).

(e) Substitute $q(\mathbf{Z}) = p(\mathbf{Z} \,|\, \mathbf{X}, \boldsymbol{\theta}_{\mathrm{curr}})$ in (2), and observe that

$$\mathcal{L}(q, \boldsymbol{\theta}) = \mathbb{E}_q[\text{complete-data log likelihood}] - H(q).$$

In other words, in the M-step we are maximizing the expectation of the complete-data log likelihood[1], since the entropy term is independent of $\boldsymbol{\theta}$. Compare this result with the EM for Gaussian mixture models.

(f) Show that the lower bound $\mathcal{L}(q, \boldsymbol{\theta})$, where $q(\mathbf{Z}) = q^\star(\mathbf{Z}) = p(\mathbf{Z} \,|\, \mathbf{X}, \boldsymbol{\theta}_{\mathrm{curr}})$, has the same gradient w.r.t. $\boldsymbol{\theta}$ as the log likelihood function $p(\mathbf{X} \,|\, \boldsymbol{\theta})$ at the point $\boldsymbol{\theta} = \boldsymbol{\theta}_{\mathrm{curr}}$. This shows that the lower bound becomes tangent to the log likelihood function at the end of E-step.

(g) Have you found an argument to prove the convergence of EM algorithm?

[1] $p(\mathbf{X}, \mathbf{Z} \,|\, \boldsymbol{\theta})$

**Solution 3:**

**(a)** By computing the KL Divergence of $q$ to $p_{\text{post}}$ we get

$$\mathrm{D}_{\mathrm{KL}}(q\|p_{\text{post}}) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})}.$$

Knowing that $p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}) = \frac{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})}{p(\mathbf{X}|\boldsymbol{\theta})}$, we get

$$\mathrm{D}_{\mathrm{KL}}(q\|p_{\text{post}}) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{q(\mathbf{Z})\, p(\mathbf{X}|\boldsymbol{\theta})}{p(\mathbf{Z},\mathbf{X}|\boldsymbol{\theta})} = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{q(\mathbf{Z})}{p(\mathbf{Z},\mathbf{X}|\boldsymbol{\theta})} + \log p(\mathbf{X}|\boldsymbol{\theta}).$$

This implies that

$$\log p(\mathbf{X}|\boldsymbol{\theta}) = \mathrm{D}_{\mathrm{KL}}(q\|p_{\text{post}}) + \mathcal{L}(q,\boldsymbol{\theta}),$$

where

$$\mathcal{L}(q,\boldsymbol{\theta}) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{Z},\mathbf{X}|\boldsymbol{\theta})}{q(\mathbf{Z})}. \tag{3}$$

**(b)** Since KL divergence is always nonnegative and is zero only if the distributions are the same, we have

$$\mathcal{L}(q,\boldsymbol{\theta}) \leq \log p(\mathbf{X}|\boldsymbol{\theta}).$$

with equality iff $q = p_{\text{post}}$.

**(c)** This is for you to verify. As seen in the examples in the slides of the course, in some situations, the E-step is just *computing* the posterior, which is equivalent to minimizing the KL divergence of $q$ to the posterior.

**(d)** Another thing to verify by yourself. Look at GMMs as an example and try to relate the variables defined in here and the parameters and variables there.

**(e)** Putting $q = p_{\text{post}}$ in (3) we get

$$\begin{aligned}
\mathcal{L}(p_{\text{post}},\boldsymbol{\theta}) &= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}_{\text{curr}}) \log \frac{p(\mathbf{Z},\mathbf{X}|\boldsymbol{\theta})}{p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})} \\
&= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}) \log p(\mathbf{Z},\mathbf{X}|\boldsymbol{\theta}) - \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}) \log p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}) \\
&= \mathbb{E}_q[\text{complete-data log likelihood}] + H(q)
\end{aligned}$$

**(f)** We have

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}}\, \mathcal{L}(q,\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\text{curr}}} &= \nabla_{\boldsymbol{\theta}} \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{Z},\mathbf{X}|\boldsymbol{\theta})}{q(\mathbf{Z})} \\
&= \sum_{\mathbf{Z}} q(\mathbf{Z}) \frac{\nabla_{\boldsymbol{\theta}}\, p(\mathbf{Z},\mathbf{X}|\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\text{curr}}}}{p(\mathbf{Z},\mathbf{X}|\boldsymbol{\theta}_{\text{curr}})} \\
&= \sum_{\mathbf{Z}} \frac{\nabla_{\boldsymbol{\theta}}\, p(\mathbf{Z},\mathbf{X}|\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\text{curr}}}}{p(\mathbf{X}|\boldsymbol{\theta}_{\text{curr}})} \\
&= \frac{\nabla_{\boldsymbol{\theta}}\, p(\mathbf{X}|\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\text{curr}}}}{p(\mathbf{X}|\boldsymbol{\theta}_{\text{curr}})} = \nabla_{\boldsymbol{\theta}}\, \log p(\mathbf{X}|\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\text{curr}}}.
\end{aligned}$$

**(g)** Compare with Homework 6.

**Problem 4 (On Statistical Distances*):**

In some situations in statistics and machine learning, the objective that we are going to optimize is some distribution (*e.g.* in the E-step of EM algorithm). This motivates us to understand a bit more about the *space* of probability distributions.

For simplicity, let $\mathcal{P}$ be the set of all probability distributions over the set $\{1, \ldots, n\}$, *i.e.*

$$\mathcal{P} = \{(p_1, \ldots, p_n) \,|\, \textstyle\sum p_i = 1, p_i \geq 0\}.$$

Usually $\mathcal{P}$ is called the probability simplex, or simply the $n$-simplex. One can equip $\mathcal{P}$ with a metric, inducing a geometry on $\mathcal{P}$. Recall that a metric is a function $d : \mathcal{P} \times \mathcal{P} \to \mathbb{R}_+$ satisfying the following criteria:

- (Non-negativity) $d(p, q) \geq 0$ for all $p, q \in \mathcal{P}$ and equality holds iff $p = q$,
- (Symmetry) $d(p, q) = d(q, p)$,
- (Triangle inequality) $d(p, q) + d(q, r) \geq d(p, r)$.

A metric on the probability simplex is also called a **statistical distance**. Here we mention a few distances and some of their properties:

(a) **Total Variation Distance.** For $p, q \in \mathcal{P}$, we define their TV distance as

$$\mathrm{D}_{\mathrm{TV}}(p, q) := \frac{1}{2}\|p - q\|_1 = \frac{1}{2}\sum_{i=1}^{n} |p_i - q_i|.$$

Prove that TV distance is indeed a metric, and equals to the largest possible difference between the probabilities that the two probability distributions $p$ and $q$ can assign to the same event, *i.e.*

$$\mathrm{D}_{\mathrm{TV}}(p, q) = \max_{E \subseteq \{1, \ldots, n\}} |p(E) - q(E)|.$$

(b) **Kullback-Leibler Divergence.** For $p, q \in \mathcal{P}$, we define their KL divergence as

$$\mathrm{D}_{\mathrm{KL}}(p\|q) = -\sum_{i=1}^{n} p_i \log \frac{q_i}{p_i}.$$

(b.1) Prove that KL divergence satisfies the first property of a metric: it is non-negative, and it is zero if and only if the distributions are equal.

(b.2) Give an example that $\mathrm{D}_{\mathrm{KL}}(p\|q) \neq \mathrm{D}_{\mathrm{KL}}(q\|p)$.

(b.3) Give a counter-example for the triangle inequality for KL divergence.

(b.4) Prove the Pinsker's Inequality:

$$\mathrm{D}_{\mathrm{TV}}(p, q) \leq \sqrt{\tfrac{1}{2}\mathrm{D}_{\mathrm{KL}}(p\|q)}.$$

(b.5) Although KL divergence fails to be a metric on $\mathcal{P}$, it satisfies some convergence properties. As an example, prove the following theorem: Let $p^{(1)}, p^{(2)}, \ldots$ be a sequence of probability distributions in $\mathcal{P}$, such that

$$\lim_{n \to \infty} \mathrm{D}_{\mathrm{KL}}(p^{(n)}\|q) = 0,$$

*i.e.* the sequence is "converging" to $q$ with respect to KL divergence. Prove that this sequence is actually converging to $q$ in Euclidean sense, *i.e.*

$$\lim_{n \to \infty} \|p^{(n)} - q\|_2 = 0.$$

(b.6) Let $X$ and $Y$ be two random variables with distributions $p_X$ and $p_Y$ and joint distribution $p_{X,Y}$. If $X$ and $Y$ were independent, the we had $p_{X,Y} = p_X p_Y$. Otherwise, if one tries to give a "measure of independence" of $X$ and $Y$, one idea is to consider

$$\mathrm{D}_{\mathrm{KL}}(p_{X,Y} \| p_X p_Y).$$

This value is called the **mutual information** between $X$ and $Y$, denoted by $I(X,Y)$. Prove that

$$I(X,Y) = H(X) - H(X \,|\, Y),$$

where $H(X)$ is the entropy[2] of $X$ and $H(X \,|\, Y)$ is the conditional entropy of $X$ given $Y$. In Bayesian point-of-view, the mutual information shows how much information does knowledge about $Y$ reveal about $X$.

[2]Entropy of a random variable $X$ is defined as $H(X) := \mathbb{E}_X[-\log X] = -\sum_x p_X(x) \log p_X(x)$, and is a measure of "uncertainty" of $X$. For example if $X$ has the uniform distribution, it has the highest entropy. If the base of $\log$ is 2, entropy is measured with the unit "bits", suggesting the idea that one needs $H(X)$ bits to encode the outcome of $X$ with zeros and ones. Convince yourself that this definition makes sense.

**Solution 4:**

**(a)** We first prove that Total Variation is a distance function. Let $p, q, r \in \mathcal{P}$ be three probability distributions over the set $\{1, \ldots, n\}$. Non-negativity follows by definition,

$$\mathrm{D_{TV}}(p, q) = \frac{1}{2}\|p - q\|_1 \geq 0.$$

Symmetry follows from $\|p - q\|_1 = \|q - p\|_1$. Also Triangle inequality follows from the triangle inequality for $\ell_1$ norm.

These three properties show that the Total Variation distance is indeed a distance function.

Now we prove the second argument. Let $E = \{i : p_i \geq q_i\}$ be the event that contains the elements which $p$ gives higher probability than $q$. We claim that $E$ attains the maximum value of $|p(F) - q(F)|$ among all events $F \subseteq \{1, \ldots, n\}$.

By writing $F = (F \cap E) \cup (F \cap E^c)$ we observe that

$$p(F) - q(F) = p(F \cap E) - q(F \cap E) + \underbrace{p(F \cap E^c) - q(F \cap E^c)}_{\leq 0} \leq p(E) - q(E). \tag{4}$$

This is true since for all elements $i \in E^c$ we have $p_i < q_i$, which makes $p(F \cap E^c) - q(F \cap E^c) \leq 0$, and adding elements in $E$ to $F \cap E$ will not decrease the difference in probability, meaning $p(F \cap E) - q(F \cap E) \leq p(E) - q(E)$.

With the same argument, but for $E^c$ this time, we arrive at

$$q(F) - p(F) \leq q(E^c) - p(E^c). \tag{5}$$

Since $p(E) - q(E) = q(E^c) - p(E^c)$, the upper bounds for (4) and (5) become the same and we get

$$p(E) - q(E) = \max_F |p(F) - q(F)|.$$

Also, by definition of $E$ we can write

$$\|p - q\|_1 = \sum_{i \in E}(p_i - q_i) + \sum_{j \notin E}(q_j - p_j) = p(E) - q(E) + q(E^c) - p(E^c) = 2(p(E) - q(E)),$$

where the last equality is because $p(E) - q(E) = q(E^c) - p(E^c)$.

**(b.1)** We prove that for $p, q \in \mathcal{P}$, we have $\mathrm{D_{KL}}(p\|q) \geq 0$. Note that postivity of the KL Divergence is regardless of the basis of the logarithm, since for all $a > 1$ we have $\log_a(x) = \ln(x)/\ln(a)$. Hence we prove that for all $p, q \in \mathcal{P}$ we have

$$-\sum_{i=1}^{n} p_i \ln \frac{q_i}{p_i} \geq 0.$$

A useful inequality about logarithms is $\ln(x) \leq x - 1$ for all $x > 0$, with equality iff $x = 1$. Using this inequality we have

$$\begin{aligned} -\sum_{i=1}^{n} p_i \ln \frac{q_i}{p_i} &\geq -\sum_{i=1}^{n} p_i \left(\frac{q_i}{p_i} - 1\right) \\ &= -\sum_{i=1}^{n}(q_i - p_i) = 0. \end{aligned}$$

The equality only happens if $p_i = q_i$ for all $i$, or equivalently when $p = q$.

**(b.2)** Take $p = (0.1, 0.9)$ and $q = (0.5, 0.5)$. Then we have

$$\mathrm{D}_{\mathrm{KL}}(p\|q) = 0.1 \times \log 0.2 + 0.9 \times \log 1.8 \approx 0.531,$$

while

$$\mathrm{D}_{\mathrm{KL}}(q\|p) = 0.5 \times \log 5 + 0.5 \times \log \tfrac{5}{9} \approx 0.737.$$

**(b.3)** To give an counterexample for Triangle inequality, we should provide three distributions $p, q, r \in \mathcal{P}$, such that

$$\mathrm{D}_{\mathrm{KL}}(p\|q) + \mathrm{D}_{\mathrm{KL}}(q\|r) < \mathrm{D}_{\mathrm{KL}}(p\|r).$$

By moving the first term to the right hand side and expanding the definition of KL divergence we need to have

$$\sum_i q_i \log \frac{q_i}{r_i} < \sum_i p_i \left( \log \frac{p_i}{r_i} - \log \frac{p_i}{q_i} \right) = \sum_i p_i \log \frac{q_i}{r_i}.$$

This suggests that we take $q$ and $r$ two arbitrary distributions and the find a $p$ that makes this inequality possible. Take $q = (0.5, 0.5)$ and $r = (0.1, 0.9)$. Then $\log \frac{q_1}{r_1} \approx 2.322$ and $\log \frac{q_2}{r_2} \approx -0.847$. Now take $p = (1, 0)$. In this way we have

$$\sum_i q_i \log \frac{q_i}{r_i} \approx 0.737,$$

but

$$\sum_i p_i \log \frac{q_i}{r_i} = \log \frac{q_1}{r_1} \approx 2.322.$$

**(b.4)** We first prove this inequality for the case that $p$ and $q$ are two probability distributions over a set of two elements, i.e. $p = (p, 1-p)$ and $q = (q, 1-q)$. In this case we have $\|p - q\|_1 = 2|p - q|$. We shall prove

$$2(p-q)^2 = \frac{1}{2}\|p - q\|_1^2 \overset{?}{\leq} \mathrm{D}_{\mathrm{KL}}(p\|q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}.$$

To prove this inequality, we fix $p$ and look at

$$f(q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q} - 2(p-q)^2.$$

So it suffices to prove $f(q)$ is always nonnegative. First, observe that at $q = p$, we have $f(p) = 0$. Taking the derivative w.r.t. $q$ we have

$$f'(q) = -\frac{p}{q} + \frac{1-p}{1-q} + 4(p-q) = (q-p)\left( \frac{1}{q(1-q)} - 4 \right)$$

Since for $0 < q < 1$ we know $q(1-q) \leq \frac{1}{4}$, it follows immediately that the sign of the derivative is the same as $q - p$. This concludes that the point $p$ is the minimum of $f$, with a value of 0.

For the more general case, we try to reduce the problem to the case we already solved. Before doing this, let us state (without proof; the proof can be carried out by the reader) a useful and important lemma:

**Lemma 1 (Log-Sum inequality)** *Let $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ be nonnegative numbers. Then*

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq a \log \frac{a}{b},$$

*where $a = \sum a_i$ and $b = \sum b_i$.*

Now let $p$ and $q$ be two probability distributions over a set of $n$ elements. Let $A = \{i : p_i \geq q_i\}$. Define $\tilde{p} = \sum_{i \in A} p_i$ and $\tilde{q} = \sum_{i \in A} q_i$ and take the distributions $\tilde{p} = (\tilde{p}, 1 - \tilde{p})$ and $\tilde{q} = (\tilde{q}, 1 - \tilde{q})$. We show that the Pinsker's inequality for $p$ and $q$ is reduced to the Pinsker inequality for $\tilde{p}$ and $\tilde{q}$, which we have already proved.

To show this reduction, we first show that $\mathrm{D}_{\mathrm{TV}}(p, q) = \mathrm{D}_{\mathrm{TV}}(\tilde{p}, \tilde{q})$. This follows from

$$\begin{aligned}
\mathrm{D}_{\mathrm{TV}}(p, q) &= \frac{1}{2} \sum_{i=1}^{n} |p_i - q_i| \\
&= \frac{1}{2} \sum_{i \in A} (p_i - q_i) + \frac{1}{2} \sum_{i \notin A} (q_i - p_i) \\
&= \tilde{p} - \tilde{q} = \mathrm{D}_{\mathrm{TV}}(\tilde{p}, \tilde{q})
\end{aligned}$$

Next, we show that $\mathrm{D}_{\mathrm{KL}}(p \| q) \geq \mathrm{D}_{\mathrm{KL}}(\tilde{p} \| \tilde{q})$:

$$\begin{aligned}
\mathrm{D}_{\mathrm{KL}}(p \| q) &= \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i} \\
&= \sum_{i \in A} p_i \log \frac{p_i}{q_i} + \sum_{i \notin A} p_i \log \frac{p_i}{q_i} \\
&\geq \tilde{p} \log \frac{\tilde{p}}{\tilde{q}} + (1 - \tilde{p}) \log \frac{1 - \tilde{p}}{1 - \tilde{q}} = \mathrm{D}_{\mathrm{KL}}(\tilde{p} \| \tilde{q}),
\end{aligned}$$

where the inequality follows from the Log-Sum inequality. Putting all together we have

$$\mathrm{D}_{\mathrm{KL}}(p \| q) \geq \mathrm{D}_{\mathrm{KL}}(\tilde{p} \| \tilde{q}) \geq 2\mathrm{D}_{\mathrm{TV}}(\tilde{p}, \tilde{q})^2 = 2\mathrm{D}_{\mathrm{TV}}(p, q)^2.$$

**(b.5)** By the Pinsker's inequality, we know that if $\mathrm{D}_{\mathrm{KL}}(p^{(n)} \| q) \to 0$, then $\mathrm{D}_{\mathrm{TV}}(p^{(n)}, q) \to 0$, meaning that $\|p^{(n)} - q\|_1 \to 0$. But since in finite dimensions all norms are equivalent, meaning that there is a constant $C$ such that $\|p - q\|_2 \leq C\|p - q\|_1$ for all $p, q$, then this implies that $\|p^{(n)} - q\|_2 \to 0$.

**(b.6)** By definition we have

$$I(X, Y) = \mathrm{D}_{\mathrm{KL}}(p_{X,Y} \| p_X p_Y) = \sum_{i,j} p_{X,Y}(i, j) \log \frac{p_{X,Y}(i, j)}{p_X(i)\, p_Y(j)}.$$

Using the chain rule for probabilities, we have $p_{X,Y}(i, j) = p_{X|Y}(i|j)\, p_Y(j)$. Hence

$$\begin{aligned}
I(X, Y) &= \sum_{i,j} p_{X,Y}(i, j) \log \frac{p_{X|Y}(i|j)\, p_Y(j)}{p_X(i)\, p_Y(j)} \\
&= -\sum_{i,j} p_{X,Y}(i, j) \log p_X(i) + \sum_{i,j} p_{X,Y}(i, j) \log p_{X|Y}(i|j).
\end{aligned}$$

In the first sum, the summation on $j$ changes $p_{X,Y}(i,j)$ to $p_X(i)$ and the sum becomes $H(X)$. The second sum, is by definition, minus the conditional entropy. So we have

$$I(X,Y) = H(X) - H(X|Y).$$