**The Australian National University** Semester 2, 2023
**School of Computing** Theory Assignment 2 of 4

# COMP3670/6670: Introduction to Machine Learning

**Release Date.** Aug 16th, 2023
**Due Date.** 11:59pm, Sept 17th, 2023
**Maximum credit.** 100

**Exercise 1** **Orthogonal Projections** (3 + 3 + 3 + 4 + 6 + 3 credits)

Consider the Euclidean vector space $\mathbb{R}^3$ with the dot product. A subspace $U \subset \mathbb{R}^3$ and vector $\mathbf{x} \in \mathbb{R}^3$ are given by:

$$U = \text{span}\left\{ \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ -1 \\ -2 \end{bmatrix} \right\}, \mathbf{x} = \begin{bmatrix} 8 \\ 4 \\ 16 \end{bmatrix}$$

1. Show that $\mathbf{x} \notin U$.

   **Solution.** We can show that $\mathbf{x} \notin U$ by showing that the set

$$\left\{ \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ -1 \\ -2 \end{bmatrix}, \begin{bmatrix} 8 \\ 4 \\ 16 \end{bmatrix} \right\}$$

   Is linearly independent. We can demonstrate this using the fact that this set is linearly independent iff the matrix

$$\begin{bmatrix} -1 & 2 & 8 \\ 1 & -1 & 4 \\ 1 & -2 & 16 \end{bmatrix}$$

   has full rank. We can do this by row reducing the matrix to

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

   which has full rank. Therefore, the set is linearly independent and $\mathbf{x} \notin U$.

   **Solution.** Alternative solution: We can show that $\mathbf{x} \notin U$ by showing that $\mathbf{x}$ is not a linear combination of the vectors in $U$. We can do this by solving the system of equations

$$\begin{bmatrix} -1 & 2 \\ 1 & -1 \\ 1 & -2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 8 \\ 4 \\ 16 \end{bmatrix}$$

Which is equivalent to solving the equations,

$$
\begin{aligned}
-a + 2b &= 8 \\
a - b &= 4 \\
a - 2b &= 16
\end{aligned}
$$

Multiplying the first equation by $-1$ we get

$$
\begin{aligned}
a - 2b &= -8 \\
a - b &= 4 \\
a - 2b &= 16
\end{aligned}
$$

We can see $a - 2b = 16$ and $a - 2b = -8$ are not consistent, therefore the system has no solution and $\mathbf{x} \notin U$.

2. Determine the orthogonal projection of $\mathbf{x}$ onto $U$, denoted $\pi_U(\mathbf{x})$.

   **Solution.** Let $\mathbf{B} = \begin{bmatrix} -1 & 2 \\ 1 & -1 \\ 1 & -2 \end{bmatrix}$. The projection matrix is defined as

$$\mathbf{P} = \mathbf{B}\left(\mathbf{B}^T\mathbf{B}\right)^{-1}\mathbf{B}^T = \begin{bmatrix} 0.5 & 0 & -0.5 \\ 0 & 1 & 0 \\ -0.5 & 0 & 0.5 \end{bmatrix}$$

   The projection is hence

$$\pi_U(\mathbf{x}) = \mathbf{P}\mathbf{x} = \begin{bmatrix} -4 \\ 4 \\ 4 \end{bmatrix}$$

3. Determine the distance $d(\mathbf{x}, U) := \min_{\mathbf{y} \in U} \|\mathbf{x} - \mathbf{y}\|$, where $\|\cdot\|$ denotes the Euclidean norm.

   **Solution.**

$$d(\mathbf{x}, U) = \|\mathbf{x} - \pi_U(\mathbf{x})\|_2 = \left\| \begin{bmatrix} 12 & 0 & 12 \end{bmatrix}^T \right\|_2 = 12\sqrt{2}$$

4. Use Gram-Schmidt orthogonalization to transform the matrix $\mathbf{A} = \begin{bmatrix} -1 & 2 \\ 1 & -1 \\ 1 & -2 \end{bmatrix}$ into a matrix $\mathbf{B}$ with orthonormal columns.

   **Solution.** We can use Gram-Schmidt orthogonalization to transform the matrix $\mathbf{A}$ into a matrix $\mathbf{B}$ with orthonormal columns. We can do this by first finding the orthogonal basis of the column space of $\mathbf{A}$. We calculate:

$$\begin{aligned}
\mathbf{v}_1 &= \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix} \\
\mathbf{v}_2 &= \begin{bmatrix} 2 \\ -1 \\ -2 \end{bmatrix} - \frac{\langle \mathbf{v_2}, \mathbf{v_1} \rangle}{\langle \mathbf{v_1}, \mathbf{v_1} \rangle} \mathbf{v}_1 \\
&= \begin{bmatrix} 2 \\ -1 \\ -2 \end{bmatrix} - \frac{-5}{3} \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix} \\
&= \begin{bmatrix} 2 \\ -1 \\ -2 \end{bmatrix} + \begin{bmatrix} -\frac{5}{3} \\ \frac{5}{3} \\ \frac{5}{3} \end{bmatrix} \\
&= \begin{bmatrix} \frac{6}{3} \\ -\frac{3}{3} \\ -\frac{6}{3} \end{bmatrix} + \begin{bmatrix} -\frac{5}{3} \\ \frac{5}{3} \\ \frac{5}{3} \end{bmatrix} \\
&= \begin{bmatrix} \frac{1}{3} \\ \frac{2}{3} \\ -\frac{1}{3} \end{bmatrix}
\end{aligned}$$

Therefore, the orthogonal basis of $U$ is

$$\left\{ \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} \frac{1}{3} \\ \frac{2}{3} \\ -\frac{1}{3} \end{bmatrix} \right\}$$

We can then normalize the vectors to get the orthonormal basis of $U$

$$\left\{ \begin{bmatrix} -\frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \end{bmatrix}, \begin{bmatrix} \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{6}} \\ -\frac{1}{\sqrt{6}} \end{bmatrix} \right\}$$

We can then construct the matrix $\mathbf{B}$ by placing the vectors in the orthonormal basis of $U$ as columns of $\mathbf{B}$

$$\mathbf{B} = \begin{bmatrix} -\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & \frac{2}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{6}} \end{bmatrix}$$

5. Let $\mathbf{Q} \in \mathbb{R}^{m \times n}$ be a matrix with orthonormal columns and $\mathbf{x} \in \mathbb{R}^m$ be an m-dimensional vector. Find the vector $\boldsymbol{\theta}$ that minimizes $||\mathbf{x} - \mathbf{Q}\boldsymbol{\theta}||^2 + \lambda||\boldsymbol{\theta}||^2$, where $\lambda$ is a positive real number.

**Solution.** We can find the expression

$$||\mathbf{x} - \mathbf{Q}\boldsymbol{\theta}||^2 + \lambda||\boldsymbol{\theta}||^2$$

is equivalent to

$$(\mathbf{x} - \mathbf{Q}\boldsymbol{\theta})^T (\mathbf{x} - \mathbf{Q}\boldsymbol{\theta}) + \lambda\boldsymbol{\theta}^T\boldsymbol{\theta}$$

Taking the gradient of this expression with respect to $\boldsymbol{\theta}$ gives

$$\begin{aligned}
\frac{\partial}{\partial\boldsymbol{\theta}}\left((\mathbf{x} - \mathbf{Q}\boldsymbol{\theta})^T (\mathbf{x} - \mathbf{Q}\boldsymbol{\theta}) + \lambda\boldsymbol{\theta}^T\boldsymbol{\theta}\right) &= \frac{\partial}{\partial\boldsymbol{\theta}}\left(\mathbf{x}^T\mathbf{x} - \mathbf{x}^T\mathbf{Q}\boldsymbol{\theta} - \boldsymbol{\theta}^T\mathbf{Q}^T\mathbf{x} + \boldsymbol{\theta}^T\mathbf{Q}^T\mathbf{Q}\boldsymbol{\theta} + \lambda\boldsymbol{\theta}^T\boldsymbol{\theta}\right) \\
&= \frac{\partial}{\partial\boldsymbol{\theta}}\left(-2\mathbf{x}^T\mathbf{Q}\boldsymbol{\theta} + \boldsymbol{\theta}^T\mathbf{Q}^T\mathbf{Q}\boldsymbol{\theta} + \lambda\boldsymbol{\theta}^T\boldsymbol{\theta}\right) \\
&= \frac{\partial}{\partial\boldsymbol{\theta}}\left(-2\mathbf{x}^T\mathbf{Q}\boldsymbol{\theta} + \boldsymbol{\theta}^T\boldsymbol{\theta} + \lambda\boldsymbol{\theta}^T\boldsymbol{\theta}\right) \\
&= \frac{\partial}{\partial\boldsymbol{\theta}}\left(-2\mathbf{x}^T\mathbf{Q}\boldsymbol{\theta} + (1+\lambda)\,\boldsymbol{\theta}^T\boldsymbol{\theta}\right) \\
&= -2\mathbf{x}^T\mathbf{Q} + 2\,(1+\lambda)\,\boldsymbol{\theta}^T
\end{aligned}$$

Now setting this equal to $\mathbf{0}$ and solving for $\boldsymbol{\theta}$ gives

$$\boldsymbol{\theta} = \frac{1}{1+\lambda}\mathbf{Q}^T x$$

6. Compute the vector $\boldsymbol{\theta}$ for the matrix $\mathbf{B}$ and $\lambda = 10$.

   **Solution.**

$$\theta = \frac{1}{11}\begin{bmatrix} -\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & -\frac{1}{\sqrt{6}} \end{bmatrix}\begin{bmatrix} 8 \\ 4 \\ 16 \end{bmatrix} = \begin{bmatrix} \frac{4\sqrt{3}}{11} \\ 0 \end{bmatrix}$$

**Exercise 2** **Vector calculus practices** (6 + 8 + 8 credits)

Compute the following gradients over $\mathbf{x}$ or $\mathbf{X}$. Represent the result in numerator layout. **Note that you are only allowed to use the rules demonstrated in the lecture**. Show each step clearly.

1. $\dfrac{\partial\mathbf{x}^T\mathbf{ABC}\;\mathbf{x}}{\partial\mathbf{x}}$
2. $\dfrac{\partial(\mathbf{Bx}+\mathbf{b})^T\mathbf{C}(\mathbf{Dx}+\mathbf{d})}{\partial\mathbf{x}}$
3. $\dfrac{\partial\;\mathbf{tr}(\mathbf{X}^2)}{\partial\mathbf{X}}$

**Solution.**

1. Let $\mathbf{D} = \mathbf{ABC}$. Then applying the rule $\frac{\partial\mathbf{x}^T\mathbf{Ax}}{\partial\mathbf{x}} = \mathbf{x}^T(\mathbf{A}+\mathbf{A}^T)$, we get

$$\frac{\partial\mathbf{x}^T\mathbf{ABC}\;\mathbf{x}}{\partial\mathbf{x}} = \mathbf{x}^T(\mathbf{ABC} + (\mathbf{ABC})^T)$$

2. Opening the brackets we have

$$\frac{\partial(\mathbf{Bx}+\mathbf{b})^T\mathbf{C}(\mathbf{Dx}+\mathbf{d})}{\partial\mathbf{x}} = \frac{\partial\mathbf{x}^T\mathbf{B}^T\mathbf{CDx}+\mathbf{b}^T\mathbf{CDx}+\mathbf{x}^T\mathbf{B}^T\mathbf{Cd}+\mathbf{b}^T\mathbf{Cd}}{\partial\mathbf{x}}$$

Note that, $\mathbf{x}^T\mathbf{B}^T\mathbf{Cd}$ is a scalar. For scalar $a$, we have $a = a^T$. Hence, $\mathbf{x}^T\mathbf{B}^T\mathbf{Cd} = \mathbf{d}^T\mathbf{C}^T\mathbf{Bx}$. Thus,

$$\frac{\partial(\mathbf{Bx}+\mathbf{b})^T\mathbf{C}(\mathbf{Dx}+\mathbf{d})}{\partial\mathbf{x}} = \frac{\partial\mathbf{x}^T\mathbf{B}^T\mathbf{CDx}+(\mathbf{b}^T\mathbf{CD}+\mathbf{d}^T\mathbf{C}^T\mathbf{B})\mathbf{x}+\mathbf{b}^T\mathbf{Cd}}{\partial\mathbf{x}}$$

Applying the rule $\frac{\partial\mathbf{a}^T\mathbf{x}}{\partial\mathbf{x}} = \mathbf{a}^T$, we have

$$\frac{\partial(\mathbf{Bx}+\mathbf{b})^T\mathbf{C}(\mathbf{Dx}+\mathbf{d})}{\partial\mathbf{x}} = \mathbf{x}^T(\mathbf{D}^T\mathbf{C}^T\mathbf{B}+\mathbf{B}^T\mathbf{CD})+\mathbf{b}^T\mathbf{CD}+\mathbf{d}^T\mathbf{C}^T\mathbf{B}$$

3. We want to evaluate the elementwise gradient $[\frac{\partial\ \mathbf{tr}(\mathbf{X}^2)}{\partial\mathbf{X}}]_{ij} = \frac{\partial\ \mathbf{tr}(\mathbf{X}^2)}{\partial X_{ij}}$. The numerator can be expressed as

$$\mathbf{tr}(\mathbf{X}^2) = \mathbf{tr}(\mathbf{X}\cdot\mathbf{X}) = \sum_k^n\sum_p^n X_{kp}X_{pk}$$

From this summation, we need to pick terms that are relevant to $X_{ij}$. For fixed $k, p$, there are four cases in total: **(1)** $X_{kp} \neq X_{ij}, X_{pk} \neq X_{ij}$, **(2)** $X_{kp} = X_{ij}, X_{pk} \neq X_{ij}$, **(3)** $X_{kp} \neq X_{ij}, X_{pk} = X_{ij}$ and **(4)** $X_{kp} = X_{ij}, X_{pk} = X_{ij}$. Note that, case (2, 3) and case (4) won't exist at the same time because (2, 3) implies $i \neq j$ while (4) implies $i = j$. We break down the summation accordingly:

$$\mathbf{tr}(\mathbf{X}\cdot\mathbf{X}) = \sum_k^n\sum_p^n X_{kp}X_{pk} = \underbrace{X_{ij}X_{ji}}_{\text{Case 2}} + \underbrace{X_{ji}X_{ij}}_{\text{Case 3}} + \underbrace{X_{ij}^2}_{\text{Case 4}} + \{\textbf{CASE ONE TERMS}\}$$

Case (1) terms won't affect the derivative. Now suppose $i \neq j$, case (4) diminishes. Finding the derivative gives us

$$\frac{\partial\ \mathbf{tr}(\mathbf{X}^2)}{\partial X_{ij}} = 2X_{ji}$$

Suppose $i = j$, case (2, 3) diminish. Finding the derivative gives us

$$\frac{\partial\ \mathbf{tr}(\mathbf{X}^2)}{\partial X_{ij}} = 2X_{ij} = 2X_{ji}$$

So in general,

$$\frac{\partial\ \mathbf{tr}(\mathbf{X}^2)}{\partial X_{ij}} = 2X_{ji}$$

This means

$$[\frac{\partial\ \mathbf{tr}(\mathbf{X}^2)}{\partial\mathbf{X}}]_{ij} = 2X_{ji}, \frac{\partial\ \mathbf{tr}(\mathbf{X}^2)}{\partial\mathbf{X}} = 2\mathbf{X}^T$$

**Exercise 3** **Concavity of a function** (8 + 10 + 10 credits)

A function $f : \mathbb{R}^n \to \mathbb{R}$ with a convex domain is called a **concave** function if and only if its Hessian $\mathbf{H} = \frac{\partial^2 f}{\partial\mathbf{x}^2}$ is negative semidefinite. Consider the following function:

$$f(\mathbf{x}) = \left(\sum_{i=1}^n x_i^p\right)^{1/p}$$

with convex domain $\textbf{dom}(f) = \mathbb{R}^n_{++}$ (n-dim strictly elementwise positive vectors), and $p < 1, p \neq 0$.

1. Evaluate the elementwise second order derivatives $\frac{\partial^2 f}{\partial x_i x_j}$ for arbitrary integer $i, j \in [1, n]$.

   **Solution.** First consider the elementwise first order derivatives $\frac{\partial f}{\partial x_i}$. Obviously

$$\frac{\partial f}{\partial x_i} = \frac{1}{p}\left(\sum_{i=1}^{n} x_i^p\right)^{1/p-1} \cdot p \cdot x_i^{p-1} = x_i^{p-1} \cdot f(\mathbf{x})^{1-p}$$

   Now consider the second order derivative $\frac{\partial^2 f}{\partial x_i x_j} = \frac{\partial x_i^{p-1} \cdot f(\mathbf{x})^{1-p}}{\partial x_j}$.

$$\begin{aligned}\frac{\partial x_i^{p-1} \cdot f(\mathbf{x})^{1-p}}{\partial x_j} &= \frac{\partial x_i^{p-1}}{\partial x_j} \cdot f(\mathbf{x})^{1-p} + x_i^{p-1} \cdot \frac{\partial f(\mathbf{x})^{1-p}}{\partial x_j} \\ &= \frac{\partial x_i^{p-1}}{\partial x_j} \cdot f(\mathbf{x})^{1-p} + x_i^{p-1} \cdot (1-p) f(\mathbf{x})^{1-2p} \cdot x_j^{p-1}\end{aligned}$$

   Obviously, if $i = j$, meaning the elementwise gradient is on the diagonal:

$$\frac{\partial^2 f}{\partial x_i x_j} = (p-1) x_i^{p-2} f(\mathbf{x})^{1-p} + x_i^{p-1} \cdot (1-p) f(\mathbf{x})^{1-2p} \cdot x_j^{p-1}$$

   Otherwise

$$\frac{\partial^2 f}{\partial x_i x_j} = x_i^{p-1} \cdot (1-p) f(\mathbf{x})^{1-2p} \cdot x_j^{p-1}$$

2. Denote the elementwise power of a vector $\mathbf{a} \in \mathbb{R}^n_{++}$ to a real number $t$ as $\mathbf{a}^t = \begin{bmatrix} {a_1}^t & {a_2}^t & \cdots & {a_n}^t \end{bmatrix}^T$. Also, the $\textbf{diag}(\cdot)$ function returns the diagonal matrix with diagonal values input as a vector. Prove that

$$\mathbf{H} = (1-p) f(\mathbf{x})^{1-2p} \cdot \left(\mathbf{x}^{p-1} \cdot \mathbf{x}^{p-1^T} - f(\mathbf{x})^p \cdot \textbf{diag}\left(\mathbf{x}^{p-2}\right)\right)$$

   **Solution.** An additional $(p-1)x_i^{p-2}$ will be applied to the diagonal elements. We leave it for later, as we can easily manipulate the result by adding a $\textbf{diag}(\cdot)$. We consider the general case

$$\frac{\partial^2 f}{\partial x_i x_j} = x_i^{p-1} \cdot (1-p) f(\mathbf{x})^{1-2p} \cdot x_j^{p-1}$$

   This corresponds to the vector form

$$\frac{\partial^2 f}{\partial \mathbf{x}^2} = \mathbf{x}^{p-1} \cdot \mathbf{x}^{p-1^T} \cdot (1-p) f(\mathbf{x})^{1-2p} + \{\textbf{DIAGONAL TERMS}\}$$

   Given the additional term $(p-1)x_i^{p-2} f(\mathbf{x})^{1-p}$, the diagonal term is

$$(p-1)\ \textbf{diag}(\mathbf{x}^{p-2}) f(\mathbf{x})^{1-p}$$

   Merge the terms we have

$$\mathbf{H} = (1-p) f(\mathbf{x})^{1-2p} \cdot \left(\mathbf{x}^{p-1} \cdot \mathbf{x}^{p-1^T} - f(\mathbf{x})^p \cdot \textbf{diag}\left(\mathbf{x}^{p-2}\right)\right)$$

3. Prove $\mathbf{H}$ is negative semidefinite, hence $f$ is concave since it has a convex domain.

**Solution.** Consider the quadratic form $\mathbf{v}^T\mathbf{A}\mathbf{v}$ for arbitrary vector $\mathbf{v}$ where

$$\mathbf{A} = \mathbf{x}^{p-1} \cdot \mathbf{x}^{p-1^T} - f(\mathbf{x})^p \cdot \mathbf{diag}\left(\mathbf{x}^{p-2}\right)$$

Thus,

$$\begin{aligned}
\mathbf{v}^T\mathbf{A}\mathbf{v} &= \mathbf{v}^T\mathbf{x}^{p-1} \cdot \mathbf{x}^{p-1^T}\mathbf{v} - f(\mathbf{x})^p \cdot \mathbf{v}^T\mathbf{diag}\left(\mathbf{x}^{p-2}\right)\mathbf{v} \\
&= \left(\mathbf{v}^T\mathbf{x}^{p-1}\right)^2 - \left(\sum_{i=1}^n x_i^p\right)\left(\sum_{i=1}^n x_i^{p-2}v_i^2\right) \\
&= \sum_{i=1}^n (x_i^{p-1}v_i)^2 - \left(\sum_{i=1}^n x_i^p\right)\left(\sum_{i=1}^n x_i^{p-2}v_i^2\right)
\end{aligned}$$

The Cauchy-Schwarz inequality can be useful here, as for arbitrary real numbers $a_i, b_i$,

$$\left(\sum_{i=1}^n a_i b_i\right)^2 \le \left(\sum_{i=1}^n a_i^2\right)\left(\sum_{i=1}^n b_i^2\right)$$

Here, by setting $a_i = x_i^{p/2}, b_i = x_i^{p/2-1}v_i$,

$$\mathbf{v}^T\mathbf{A}\mathbf{v} = \left(\sum_{i=1}^n a_i b_i\right)^2 - \left(\sum_{i=1}^n a_i^2\right)\left(\sum_{i=1}^n b_i^2\right) \le 0$$

Since $(1-p)f(\mathbf{x})^{1-2p}$ is always positive,

$$(1-p)f(\mathbf{x})^{1-2p} \cdot \mathbf{v}^T\mathbf{A}\mathbf{v} = \mathbf{v}^T\mathbf{H}\mathbf{v} \le 0$$

Thus, we conclude $\mathbf{H}$ is negative semidefinite.

**Exercise 4** **Expectations with respect to a Gaussian distribution** (10+10+8 credits)

A common objective function in modern machine learning is the variational free-energy,

$$\mathcal{F}(q(\theta)) = \int \mathrm{d}\theta q(\theta) \log \left\{ \frac{q(\theta)}{p(\theta)p(y|\theta,x)} \right\} = \int \mathrm{d}\theta q(\theta)[\log\{q(\theta)\} - \log\{p(\theta)\} - \log\{p(y|\theta,x)\}]. \quad (1)$$

Consider a simplified setting in which

$$p(\theta) = \mathcal{N}(\theta; 0, 1), \quad (2)$$

$$p(y|\theta,x) = \mathcal{N}(y; \theta x, \sigma_n^2), \quad (3)$$

$$q(\theta) = \mathcal{N}(\theta; \mu, \sigma^2), \quad (4)$$

where $\mathcal{N}(x; m, v)$ means $x$ is a univariate Gaussian random variable with mean $m$ and variance $v$.

1. Compute $\mathcal{F}$.

**Solution.** We calculate some results beforehand. Consider antiderivative of the following function

$$\int \theta \exp\{-\theta^2\} d\theta \overset{x=\theta^2}{=\!=\!=} \frac{1}{2}\int \exp\{-x\}dx = -\frac{1}{2}\exp\{-\theta^2\} + C$$

Now consider

$$\int_{\mathbb{R}} x^2 \exp\{-x^2\}dx$$

Using integration by parts, we know

$$\int_{\mathbb{R}} x^2 \exp\{-x^2\}dx = -\frac{1}{2}\left[x\exp\{-x^2\}\right]_{\mathbb{R}} + \int_{\mathbb{R}} \frac{1}{2}\exp\{-x^2\}dx = \frac{1}{2}\int_{\mathbb{R}} \exp\{-x^2\}dx$$

Note that for a standard Gaussian distribution, we have the following property:

$$\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{\theta^2}{2}\right\}d\theta = 1$$

$$\int_{\mathbb{R}} \exp\left\{-\frac{\theta^2}{2}\right\}d\theta = \sqrt{2\pi}$$

$$\int_{\mathbb{R}} \exp\{-x^2\}dx = \sqrt{\pi}$$

Thus,

$$\int_{\mathbb{R}} x^2 \exp\{-x^2\}dx = \frac{\sqrt{\pi}}{2}$$

Let's forward to the problem.

$$\mathcal{F} = \underbrace{\int_{\mathbb{R}} q(\theta)\log\{q(\theta)\}\, d\theta}_{(1)} - \underbrace{\int_{\mathbb{R}} q(\theta)\log\{p(\theta)\}\, d\theta}_{(2)} - \underbrace{\int_{\mathbb{R}} q(\theta)\log\{p(y|\theta,x)\}\, d\theta}_{(3)}$$

We evaluate the terms separately.

$$
\begin{aligned}
(1) &= \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(\theta-\mu)^2}{2\sigma^2}\right\}\left(\log\left\{\frac{1}{\sqrt{2\pi}\sigma}\right\} - \frac{(\theta-\mu)^2}{2\sigma^2}\right) d\theta \\
&= -\log\left\{\sqrt{2\pi}\sigma\right\} - \frac{1}{\sqrt{2\pi}\sigma}\int_{\mathbb{R}} \exp\left\{-\frac{(\theta-\mu)^2}{2\sigma^2}\right\}\frac{(\theta-\mu)^2}{2\sigma^2} d\theta \\
&\overset{x=\frac{\theta-\mu}{\sqrt{2}\sigma}}{=\!=\!=} -\log\left\{\sqrt{2\pi}\sigma\right\} - \frac{1}{\sqrt{\pi}}\int_{\mathbb{R}} x^2 \exp\left\{-x^2\right\} dx \\
&= -\log\left\{\sqrt{2\pi}\sigma\right\} - \frac{1}{2}
\end{aligned}
$$

$$
\begin{aligned}
(2) &= \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(\theta-\mu)^2}{2\sigma^2}\right\}\left(\log\left\{\frac{1}{\sqrt{2\pi}}\right\} - \frac{\theta^2}{2}\right) d\theta \\
&= -\log\left\{\sqrt{2\pi}\right\} - \frac{1}{2\sqrt{2\pi}\sigma}\int_{\mathbb{R}} \exp\left\{-\frac{(\theta-\mu)^2}{2\sigma^2}\right\}\left(\theta^2\right) d\theta \\
&\overset{x=\frac{\theta-\mu}{\sqrt{2}\sigma}}{=\!=\!=} -\log\left\{\sqrt{2\pi}\right\} - \frac{1}{2\sqrt{\pi}}\int_{\mathbb{R}} (\sqrt{2}\sigma x+\mu)^2 \exp\left\{-x^2\right\} dx \\
&= -\log\left\{\sqrt{2\pi}\right\} - \frac{1}{2\sqrt{\pi}}(\sigma^2\sqrt{\pi}+\mu^2\sqrt{\pi}) \\
&= -\log\left\{\sqrt{2\pi}\right\} - \frac{1}{2}\cdot(\sigma^2+\mu^2)
\end{aligned}
$$

$$
\begin{aligned}
(3) &= \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(\theta-\mu)^2}{2\sigma^2}\right\}\left(\log\left\{\frac{1}{\sqrt{2\pi}\sigma_n}\right\} + \frac{(y-\theta x)^2}{2\sigma_n^2}\right) d\theta \\
&= -\log\left\{\sqrt{2\pi}\sigma_n\right\} - \frac{1}{2\sqrt{2\pi}\sigma\sigma_n^2}\int_{\mathbb{R}} \exp\left\{-\frac{(\theta-\mu)^2}{2\sigma^2}\right\}(y-\theta x)^2 d\theta \\
&\overset{z=\frac{\theta-\mu}{\sqrt{2}\sigma}}{=\!=\!=} -\log\left\{\sqrt{2\pi}\sigma_n\right\} - \frac{1}{2\sqrt{\pi}\sigma_n^2}\int_{\mathbb{R}} \exp\left\{-z^2\right\}(y-\sqrt{2}\sigma z x-\mu x)^2 dz \\
&= -\log\left\{\sqrt{2\pi}\sigma_n\right\} - \frac{1}{2\sqrt{\pi}\sigma_n^2}\left(\int_{\mathbb{R}} (y-\mu x)^2 \exp\left\{-z^2\right\} dz + \int_{\mathbb{R}} 2\sigma^2 x^2 z^2 \exp\left\{-z^2\right\} dz\right) \\
&= -\log\left\{\sqrt{2\pi}\sigma_n\right\} - \frac{1}{2\sqrt{\pi}\sigma_n^2}\left((y-\mu x)^2\sqrt{\pi}+\sigma^2 x^2\sqrt{\pi}\right) \\
&= -\log\left\{\sqrt{2\pi}\sigma_n\right\} - \frac{1}{2\sigma_n^2}\left((y-\mu x)^2+\sigma^2 x^2\right)
\end{aligned}
$$

In conclusion,

$$\mathcal{F} = -\log\left\{\sqrt{2\pi}\sigma\right\} - \frac{1}{2} + \log\left\{\sqrt{2\pi}\right\} + \frac{1}{2}\cdot(\sigma^2+\mu^2) + \log\left\{\sqrt{2\pi}\sigma_n\right\} + \frac{1}{2\sigma_n^2}\left((y-\mu x)^2 + \sigma^2 x^2\right)$$

**Version 2.** Note that:

$$\begin{aligned}
\log\{p(\theta)\} &= \log\{\mathcal{N}(\theta;0,1)\} = -\frac{1}{2}\log\{2\pi\} - \frac{1}{2}\theta^2,\\
\log\{p(y|\theta,x)\} &= \log\left\{\mathcal{N}(y;\theta x,\sigma_n^2)\right\} = -\frac{1}{2}\log\left\{2\pi\sigma_n^2\right\} - \frac{1}{2\sigma_n^2}(y^2 - 2x\theta y + x^2\theta^2),\\
\log\{q(\theta)\} &= \log\left\{\mathcal{N}(\theta;\mu,\sigma^2)\right\} = -\frac{1}{2}\log\left\{2\pi\sigma^2\right\} - \frac{1}{2\sigma^2}(\theta^2 - 2\mu\theta + \mu^2),
\end{aligned}$$

and

$$\begin{aligned}
\int \theta\mathcal{N}(\theta;\mu,\sigma^2)\mathrm{d}\theta &= \mu\\
\int \theta^2\mathcal{N}(\theta;\mu,\sigma^2)\mathrm{d}\theta &= \mu^2+\sigma^2.
\end{aligned}$$

Thus,

$$\begin{aligned}
&\mathcal{F}_2 = \langle \log\{p(\theta)\}\rangle_{q(\theta)} = -\frac{1}{2}\log\{2\pi\} - \frac{1}{2}(\mu^2+\sigma^2),\\
&\mathcal{F}_3 = \langle \log\{p(y|\theta,x)\}\rangle_{q(\theta)} = -\frac{1}{2}\log\left\{2\pi\sigma_n^2\right\} - \frac{1}{2\sigma_n^2}(y^2 - 2x\mu y + x^2\mu^2 + x^2\sigma^2),\\
&\mathcal{F}_1 = \langle \log\{q(\theta)\}\rangle_{q(\theta)} = -\frac{1}{2}\log\left\{2\pi\sigma^2\right\} - \frac{1}{2},\\
&\mathcal{F} = \mathcal{F}_1 - \mathcal{F}_2 - \mathcal{F}_3.
\end{aligned}$$

2. Find the gradients $\frac{\partial}{\partial\mu}\mathcal{F}$ and $\frac{\partial}{\partial\sigma}\mathcal{F}$.

   **Solution.**

$$\begin{aligned}
\frac{\partial\mathcal{F}}{\partial\mu} &= \mu + \frac{1}{\sigma_n^2}(x^2\mu - xy)\\
\frac{\partial\mathcal{F}}{\partial\sigma} &= -\frac{1}{\sigma} + \sigma + \frac{1}{\sigma_n^2}x^2\sigma
\end{aligned}$$

3. Set these gradients to zero and solve for $\mu$ and $\sigma$ in terms of $y, x$ and $\sigma_n$.

   **Solution.**

$$\mu = \frac{xy}{\sigma_n^2 + x^2}$$

$$\sigma = \frac{\sigma_n}{\sqrt{\sigma_n^2 + x^2}}$$