ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
School of Computer and Communication Sciences

Foundations of Data Science Assignment date: Wednesday, November 17th, 2021, 10:15
Fall 2021 Due date: Wednesday, November 17th, 2021, 12:00

Midterm Exam – CM 1 2

This exam is closed book, closed notes. You are allowed to bring one A4 sheet (both sides of it) of hand-written and not photocopied notes (”cheat sheet”). No electronic devices of any kind are allowed. There are four problems. Choose the ones you find easiest and collect as many points as possible. We do not necessarily expect you to finish all of them. Good luck!

Name: ____________________________

Problem 1	/ 10
Problem 2	/ 10
Problem 3	/ 10
Problem 4	/ 10
Total	/40

Problem 1 (Even Moments of Subgaussian RV)**.** [10pts]

(i) [5pts] Let Z be a non-negative random variable. Show that

$$\mathbb{E}[Z] = \int_0^\infty \text{Prob}(Z > z)dz. \tag{1}$$

Solution: Note that the expectation is either finite or it diverges to infinity. Let us first assume that $\mathbb{E}[Z] < \infty$. There are various ways to develop the proof for this case. Here are three:

- Observe that $z = \int_0^\infty \mathbb{1}_{\{t\le z\}}dt$, where $\mathbb{1}_{\{t\le z\}}$ is the indicator function of the condition $\{t \le z\}$. Then

$$\begin{aligned}\mathbb{E}[Z] &= \int_0^\infty zf(z)dz \\ &= \int_0^\infty \int_0^\infty \mathbb{1}_{\{t\le z\}}dtf(z)dz \\ &= \int_0^\infty \underbrace{\int_0^\infty \mathbb{1}_{\{t\le z\}}f(z)dz}_{=\text{Prob}(Z>t)}\,dt. \qquad \text{(Because the expectation is finite)}\end{aligned}$$

- Let $f(z)$ be the density of the random variable and $F(z) = \int_0^z f(x)dx$ be its cumulative distribution function. Then, using integration by parts,

$$\begin{aligned}\mathbb{E}[Z] &= \int_0^\infty zf(z)dz \\ &= z(F(z)-1)|_0^\infty - \int_0^\infty (F(z)-1)dz\end{aligned}$$

  Note that,$z(F(z)-1) = z\int_z^\infty f(t)dt \le \int_z^\infty tf(t)dt \to 0$ as $z \to \infty$, therefore

$$\begin{aligned}&= \int_0^\infty (1-F(z))dz \\ &= \int_0^\infty \text{Prob}(Z > z)dz.\end{aligned}$$

- Let $f(z)$ be the density of the random variable. Letting $G(z) = \text{Prob}(Z > z)$, note that $G'(z) = -f(z)$. Then, using integration by parts,

$$\begin{aligned}\int_0^\infty 1\cdot\text{Prob}(Z > z)dz &= \underbrace{[z\text{Prob}(Z > z)]_0^\infty}_{=0} - \int_0^\infty z(-f(z))dz \\ &= \mathbb{E}[Z].\end{aligned}$$

For the case where $\mathbb{E}[Z] = \infty$ (you were not expected to do this part!) here is the proof that the integral, $\int_0^\infty \text{Prob}(Z > z)dz$ also diverges to ∞. Let $Z_K = K\mathbb{1}_{X>K} + X\mathbb{1}_{X\leq K}$, where $K \in \mathbb{R}^+$. Then, since Z_K is bounded it's mean is finite. Therefore, we can use the previous result to say,

$$\begin{aligned}
\mathbb{E}[Z_K] &= \int_0^\infty \text{Prob}(Z_K > z)dz && (2)\\
&= \int_0^K \text{Prob}(Z_K > z)dz && (Z_K \text{ is bounded between 0 and } K)\\
&= \int_0^K \text{Prob}(Z > z)dz && (\text{for } z \in [0,K),\ Z_K > z \text{ iff } Z > z)
\end{aligned}$$

By definition,

$$\begin{aligned}
\int_0^\infty \text{Prob}(Z > z)dz &= \lim_{K\to\infty} \int_0^K \text{Prob}(Z > z)dz && (3)\\
&= \lim_{K\to\infty} \mathbb{E}[Z_K] && (4)\\
&= \mathbb{E}[Z] && (\text{Monotone Convergence Theorem})
\end{aligned}$$

(ii) [5pts] Let X be a σ^2-subgaussian random variable. Show that for even integers $k = 2m$,

$$\mathbb{E}[X^k] \leq C(k)\sigma^k, \tag{5}$$

and find the expression for $C(k)$. (HINT: Use the formulation of the mean above. HINT: $\int_0^\infty x^{2m-1}e^{-x^2/2}dx = 2^{m-1}(m-1)!$)

Solution: When X is a σ^2-subgaussian random variable, we derived together in class that for any $\eta > 0$,

$$\text{Prob}(X > \eta) \leq e^{-\frac{\eta^2}{2\sigma^2}}, \quad \text{Prob}(X < -\eta) \leq e^{-\frac{\eta^2}{2\sigma^2}}. \tag{6}$$

We were assuming that you'd have this formula on your cheat sheets.

Now, to continue, using the result from the first part of this problem,

$$\begin{aligned}
\mathbb{E}[X^k] &= \int_0^\infty \text{Prob}(X^k > \alpha)d\alpha && (7)\\
&= \int_0^\infty \text{Prob}(X > \alpha^{1/k})d\alpha + \int_0^\infty \text{Prob}(X < -\alpha^{1/k})d\alpha && (8)
\end{aligned}$$

Therefore,

$$\mathbb{E}[X^k] \leq 2\int_0^\infty e^{-\frac{\alpha^{2/k}}{2\sigma^2}}d\alpha. \tag{9}$$

Setting $u^k\sigma^k = \alpha$ we obtain

$$\mathbb{E}[X^k] \leq 2\int_0^\infty e^{-\frac{\alpha^{2/k}}{2\sigma^2}}d\alpha = 2\sigma^k k\int_0^\infty u^{k-1}e^{-u^2/2}du = \sigma^k k 2^{k/2}(k/2-1)! \tag{10}$$

Remark: We gave you as a hint the formula for the moments of the "half-Gaussian" (not normalized). But note that it is not difficult to solve this problem just with elementary integral operations. Specifically, setting $u = \alpha^{\frac{2}{k}}/(2\sigma^2)$, we obtain

$$\int_0^\infty e^{-\frac{\alpha^{2/k}}{2\sigma^2}} d\alpha = (2\sigma^2)^{k/2} k \int_0^\infty e^{-u} u^{m-1} du, \tag{11}$$

and the last integral can be easily solved recursively via integration by parts:

$$\int_0^\infty \underbrace{e^{-u}}_{f'} \underbrace{u^{m-1}}_{g} du = \left[-e^{-u} u^{m-1}\right]_0^\infty - \int_0^\infty \left(-e^{-u}\right) \left((m-1)u^{m-2}\right) du \tag{12}$$

$$= (m-1) \int_0^\infty e^{-u} u^{m-2} du, \tag{13}$$

and hence, by recursion,

$$\int_0^\infty e^{-u} u^{m-1} du = (m-1) \cdot (m-2) \cdot \ldots \cdot 2 \cdot 1, \tag{14}$$

which thus gives the same formula.

Problem 2 (Generating fair coin flips from rolling the dice)**.** [10pts] Suppose $X_1, X_2, \ldots$ are the outcomes of rolling a possibly loaded die multiple times. The outcomes are assumed to be iid. Let $\mathbb{P}(X_i = m) = p_m$, for $m = 1, 2, \ldots, 6$, with p_m unknown (but non-negative and summing to one, clearly). By processing this sequence we would like to obtain a sequence $Z_1, Z_2, \ldots$ of *fair* coin flips.

Consider the following method: We process the X sequence in successive pairs, (X_1X_2), (X_3X_4), (X_5X_6), mapping $(3,4)$ to 0, $(4,3)$ to 1, and all the other outcomes to the empty string λ. After processing X_1, X_2, we will obtain either nothing, or a bit Z_1.

(a) [3pts] Show that, if a bit is obtained, it is fair, i.e., $\mathbb{P}(Z_1 = 0|Z_1 \neq \lambda) = \mathbb{P}(Z_1 = 1|Z_1 \neq \lambda) = 1/2$.

Solution: *(a)* $P(Z_1 = 0|Z_1 \neq \lambda) = P(Z_1 = 0, Z_1 \neq \lambda)/P(Z_1 \neq \lambda) = P(Z_1 = 0)/P(Z_1 \neq \lambda)$. Similarly, $P(Z_1 = 1|Z_1 \neq \lambda) = P(Z_1 = 1)/P(Z_1 \neq \lambda)$. Let us now show that $P(Z_1 = 0) = P(Z_1 = 1)$ and this will complete the proof. Note that $P(Z_1 = 1) = P(X_1 = 3, X_2 = 4) = P(X_1 = 3)P(X_2 = 4) = p_3p_4$ and $P(Z_1 = 0) = P(X_1 = 4, X_2 = 3) = P(X_1 = 4)P(X_2 = 3) = p_4p_3$. Therefore $P(Z_1 = 1) = P(Z_1 = 0)$.

In general we can process the X sequence in successive n-tuples via a function $f : \{1,2,3,4,5,6\}^n \to \{0,1\}^*$ where $\{0,1\}^*$ denotes the set of all finite length binary sequences (including the empty string λ). [The case in (a) is the function where $f(3,4) = 0$, $f(4,3) = 1$, and $f(j,m) = \lambda$ for all other choices of j and m.] The function f is chosen such that $(Z_1, \ldots, Z_K) = f(X_1, \ldots, X_n)$ are i.i.d., and fair (here K may depend on $(X_1, \ldots, X_n)$).

(b) [3pts] Letting $H(X)$ denote the entropy of the (unknown) distribution $(p_1, p_2, \ldots, p_6)$, prove the following chain of (in)equalities.

$$\begin{aligned} nH(X) &= H(X_1, \ldots, X_n) \\ &\geq H(Z_1, \ldots, Z_K, K) \\ &= H(K) + H(Z_1 \ldots, Z_K|K) \\ &= H(K) + \mathbb{E}[K] \\ &\geq \mathbb{E}[K]. \end{aligned}$$

Consequently, on the average no more than $nH(X)$ fair bits can be obtained from $(X_1, \ldots, X_n)$.

Solution:

(b)

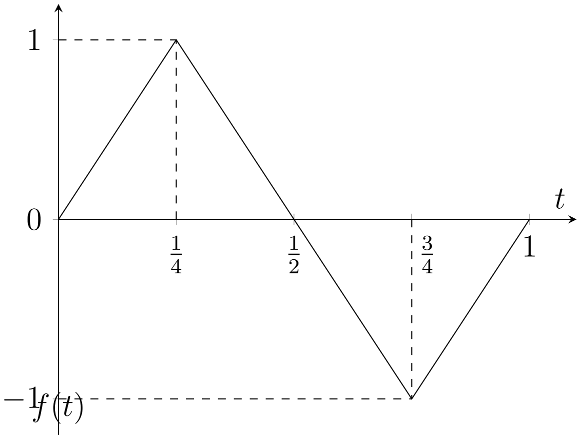
$$
\begin{aligned}
nH(X) &= nH(X_i) \qquad \qquad (15)\\
&= H(X_1, \ldots, X_n) \quad \text{[Independence of } X_i\text{]} \qquad (16)\\
&\geq H(f(X_1, \ldots, X_n)) \quad \text{[Data Processing Inequality]} \qquad (17)\\
&= H(Z_1, \ldots, Z_K, K) \qquad (18)\\
&= H(K) + H(Z_1, \ldots, Z_K|K) \quad \text{[Chain Rule]} \qquad (19)\\
&= H(K) + \sum_k p(K = k) H(Z_1, \ldots, Z_K | K = k) \qquad (20)\\
&= H(K) + \sum_k p(K = k) k \quad [Z_1, \ldots, Z_k \text{ are i.i.d and fair when } K = k] \qquad (21)\\
&= H(K) + \mathbb{E}[K] \qquad (22)\\
&\geq \mathbb{E}[K] \quad \text{[Non-negativity of entropy]} \qquad (23)
\end{aligned}
$$

(c) [4pts] Describe how you would find a good f (with high $\mathbb{E}[K]$) for $n = 4$ which would work for any distribution $(p_1, p_2, ..., p_6)$.

Solution: *(c)*
We have in total 6^4 many possible outcomes. We can only produce fair bits, regardless of the distribution, if we have permutations of the same sequence. e.g., $1555 \rightarrow 00, 5155 \rightarrow 01, 5515 \rightarrow 10, 5551 \rightarrow 11$. Let us do the counting. A sequence can have $1, 2, 3$ or 4 kinds of different symbols. An example to a sequence of 3 different symbols is 1232.

1: We cannot produce bits with 1 kind of different symbols because you cannot permute the sequence and get another sequence. Therefore we map sequences of kind *aaaa* to the null string λ.
2: For 2 different symbols it will be either 3 of the same kind and 1 of another kind which gives 4 different permutations or 2 of the same kind and 2 of another kind, which gives 6 different permutations. From the 4 different permutations of a "3 by 1" (aaab) sequence we can generate 2 fair bits, because there are 4 permutations. From the the first 4 of the 6 different permutations of a "2 by 2" sequence (aabb) we can generate 2 fair bits, and from the remaining 2 permutations we can generate 1 fair bit.
3: For 3 different symbols it has to be 2 of the same symbol, 1 of another symbol and 1 of another symbol (*aabc*). There are $4!/2! = 12$ different ways to permute these sequence of type *aabc*. From the first 8 we can generate 3 bits, and from the remaining 4 we can generate 2 bits.
4: There are $4! = 24$ ways to permute a sequence of kind (a, b, c, d). From the first 16 we can generate 4 bits, and from the remaining 8 we can generate 3 bits.

Problem 3 (Haar Wavelet)**.** [10pts] Consider the following function $f(t)$.



Let $\psi(t)$ be the Haar wavelet. (1 for $t \in [0, 1/2]$ and -1 for $t \in [1/2, 1]$).
As in the class we define $\psi_{m,n}(t) = 2^{-m/2}\psi(2^{-m}t - n)$, for $m, n \in \mathbb{Z}$.

(i) [5pts] Note that $f(t) = \sum_{m,n} a_{m,n}\psi_{m,n}(t)$. Find the scales $m \in \mathbb{Z}$ such that $\forall n, a_{m,n} = 0$.

Solution:
Note that $a_{m,n} = \langle \psi_{m,n}, f \rangle \triangleq \int_{-\infty}^{\infty} \psi_{m,n}(t) f(t) dt$, because $\psi_{m,n}$'s are orthonormal.
We will consider several cases:

- For all $m \geq 1$ and $n \in \mathbb{Z}$, $a_{m,n} = 0$, because the function is supported only on $[0, 1]$ and integrates to 0.
- For $m = 0$, $a_{0,0} \neq 0$ because the inner product $\langle \psi_{m,n}, f \rangle$ gives $1/2$.
- For $m = -1$, the support of $\psi_{m,n}$ and the support of f are not disjoint only on $n \in \{0, 1\}$. For each of these values of n, we can see that $a_{m,n} = 0$. Hence, we can claim that for all value of n, $a_{-1,n} = 0$.
- For $m \leq -2$, $a_{m,0}$ is not 0.

Therefore we see that for $m \in \{1, 2, ...\} \cup \{-1\}$, $\forall n \in \mathbb{Z}$, $a_{m,n} = 0$.

(ii) [5pts] Let $f_{m^*}(t)$ be the projection of $f(t)$ to the space spanned by $\{\psi_{m,n} : m, n \in \mathbb{Z},\ m \geq m^*\}$ w.r.t. the standard L_2 norm. Find

$$\max_{t \in \mathbb{R}} |f_{m^*}(t) - f(t)|$$

as a function of $m^* \in \mathbb{Z}$.

[Hint : Try m^* equals to 0 and sketch $f_0(t)$.]

Solution:
Here we will use the fact that $f_{m^*} = \sum_{m \geq m^*} \sum_{n \in \mathbb{Z}} a_{m,n}\psi_{m,n}$. Note that for $m^* \geq 1$,

$f^*_m(t) = 0$ everywhere. This is due to the fact that for $m \geq 1$, $a_{m,n} = 0$ for all n as we have shown on (i).

Now consider $m^* = 0$, $f_0(t) = \sum_{m\geq 0}\sum_n a_{m,n}\psi_{m,n}(t) = a_{0,0}\psi_{0,0}(t) = 1/2\ \psi(t)$. Therefore the absolute error is at most $1/2$ (i.e., at $t \in \{0, 1/4, 3/4, 1\}$

Note that for $m = -1$, for all n, $a_{m,n} = 0$, so $f_{-1} = f_0$.

To get f_{-2}, we look at the error of f_{-1} and project it to the space spanned by $\{\psi_{-2,n} : n \in \mathbb{Z}\}$. One finds that the error $e_{-1}(t) = f_{-1}(t) - f(t)$ is of the form,

$$e_{-1}(t) = \begin{cases} -1/2 + 4t & t \in [0, 1/4) \\ 1/2 - 4(t - 1/4) & t \in [1/4, 1/2) \\ 1/2 - 4(t - 1/2) & t \in [1/2, 3/4) \\ -1/2 + 4(t - 3/4) & t \in [3/4, 1/4) \\ 0 & \text{otherwise} \end{cases}$$

Let us focus our attention on the segments on $[0, 1/4)$. This segment coincides with the support of $\psi_{-2,0}$ and we have $a_{-2,0} = 1/8$. Since $\psi_{-2,0}$ has height 2. The height of $a_{-2,0}\psi_{-2,0}$ is $1/4$. This results in the maximum absolute error of $1/4$. Similar arguments also works for the other segments.

But more remarkably, for $t \in [0, 1/4)$, the error $e_{-2}(t) = f_{-2}(t) - f(t)$ is of the form $-1/4 + 8t, t \in [0, 1/8)$ and $1/4 - 8(t - 1/8), t \in [1/8, 1/4)$. Hence we can use the same argument to deduce that $a_{-3.0} = 1/(16\sqrt{2})$, therefore the height of $a_{m,n}\psi_{-3,0}$ is $1/(16\sqrt{2}) \times 2\sqrt{2} = 1/8$. and the maximum absolute error is $1/8$. This pattern holds for all $m^* \leq -4$.

Hence we have the error is given by

$$\max_{t\in\mathbb{R}} |f^*_m(t) - f(t)| = \begin{cases} 1 & m^* \in \{1, 2, ...\} \\ 1/2 & m^* = 0 \\ 1/2 & m^* = -1 \\ 2^{m^*} & m^* \leq -2. \end{cases}$$

Problem 4 (UCB With Geometric Intervals)**.** [10pts] Consider the following slight variant of the UCB algorithm. We have K arms. As in the lecture notes, assume that each of these K arms corresponds to a random variable which is 1-subgaussian. For the first K steps we sample each of these arms once. After these K first steps we have an interval of length 1, then an interval of length 2, then one of length 4, and so on. At the beginning of each such interval we choose the arm in the same manner as the UCB algorithm. More precisely, if t marks the beginning of a new interval then

$$A_t = \text{argmax}_k \hat{\mu}_k(t-1) + \sqrt{\frac{2\ln f(t)}{T_k(t-1)}},$$

where $f(g) = 1 + t\ln^2(t)$ as for the case we discussed in the course and where $T_k(t-1)$ denotes the number of times we have chosen arm k in the last $t-1$ steps. But unlike the standard UCB algorithm, for all other steps in this interval we keep the same arm. Why might we be interested in such an algorithm? One motivation is complexity. Computing which arm is best takes some effort. In this way we only have to compute the best arm a logarithmic (in the time horizon) number of times.

Recall that in the analysis of the original algorithm the key to the analysis was to find a good upper bound on $T_k(n)$ for $k > 1$, assuming that arm 1 is the optimum arm. In turn, we upper bounded the probability that we choose arm k at a particular point in time t by the probability that arm 1 had an empirical mean at least an ϵ below its true mean μ_1 and that the empirical mean of arm k was above $\mu_1 - \epsilon$. In formulae we had

$$T_k(n) = \sum_{t=1}^{n} \mathbb{1}_{\{A_t=k\}} \leq \sum_{t=1}^{n} \mathbb{1}_{\{\hat{\mu}_1(t-1)+\sqrt{\frac{2\ln f(t)}{T_1(t-1)}} \leq \mu_1-\epsilon\}} + \sum_{t=1}^{n} \mathbb{1}_{\{\hat{\mu}_k(t-1)+\sqrt{\frac{2\ln f(t)}{T_k(t-1)}} \geq \mu_1-\epsilon \,\wedge\, A_t=k\}} \tag{24}$$

Let us proceed in the same fashion. Let $n = K + 2^L - 1$. In words, we are at the end of the L-th interval, where $L \in \mathbb{N}$.

(i) [5pts] What is the expression equivalent to (**??**) for our case?

Solution: Let $t(l) = K + 2^{l-1}$. This is the time at the beginning of the l-th interval. In the sequel, to lighten the notation, we will sometimes just write t. Then we have

$$T_k(n) = \sum_{l=1}^{L} \mathbb{1}_{\{A_{t(l)}=k\}} 2^{l-1} \leq \sum_{l=1}^{L} 2^{l-1} \mathbb{1}_{\{\hat{\mu}_1(t(l)-1)+\sqrt{\frac{2\ln f(t(l))}{T_1(t(l)-1)}} \leq \mu_1-\epsilon\}} + \sum_{l=1}^{L} 2^{l-1} \mathbb{1}_{\{\hat{\mu}_k(t(l)-1)+\sqrt{\frac{2\ln f(t(l))}{T_k(t(l)-1)}} \geq \mu_1-\epsilon \,\wedge\, A_{t(l)}=k\}} \tag{25}$$

(ii) [5pts] Look at the first of the two terms on the right of (**??**) in your equivalent expression. Derive a suitable upper bound for this first term. If you do not have time for the whole derivation just write down the first few steps. These are the most crucial ones.

Solution:

For the first term the derivation is almost the same as what we saw in the course. to ease the notation burden, we will write $t(l)$ as t in the following (in)equalities. We have

$$
\begin{aligned}
\mathbb{E}[\sum_{l=1}^{L} 2^{l-1}\mathbb{1}_{\{\hat{\mu}_1(t-1)+\sqrt{\frac{2\ln f(t)}{T_1(t-1)}}\leq\mu_1-\epsilon\}}] &= \sum_{l=1}^{L} 2^{l-1}\mathbb{P}\left(\hat{\mu}_1(t-1)+\sqrt{\frac{2\ln f(t)}{T_1(t-1)}}\leq \mu_1-\epsilon\right)\\
&\leq \sum_{l=1}^{L} 2^{l-1}\sum_{s=1}^{t}\mathbb{P}\left(\hat{\mu}_{1,s}+\sqrt{\frac{2\ln f(t)}{s}}\leq \mu_1-\epsilon\right)\\
&\qquad\text{(condition over } T_1 \text{ and use average is less than the sum)}\\
&\leq \sum_{l=1}^{L} 2^{l-1}\sum_{s=1}^{t} e^{-\frac{s}{2}(\sqrt{\frac{2\ln f(t)}{s}}+\epsilon)^2}\\
&= \sum_{l=1}^{L} 2^{l-1}\sum_{s=1}^{t} e^{-\ln(f(t))-\sqrt{2s\ln f(t)}-\frac{s}{2}\epsilon^2}\\
&\leq \sum_{l=1}^{L} 2^{l-1}\frac{1}{f(t)}\sum_{s=1}^{t} e^{-\frac{s}{2}\epsilon^2}\\
&= \sum_{l=1}^{L} 2^{l-1}\frac{1}{f(t)}\frac{e^{-\frac{\epsilon^2}{2}}}{1-e^{-\frac{\epsilon^2}{2}}}\\
&= \sum_{l=1}^{L} 2^{l-1}\frac{1}{f(t)}\qquad\qquad \underbrace{\frac{1}{e^{\frac{\epsilon^2}{2}}-1}}_{\text{take Taylor series}}\\
&\qquad\qquad\text{all terms are positive; keep only first two}\\
&\leq \sum_{l=1}^{L}\frac{1}{f(t)}2^{l-1}\frac{2}{\epsilon^2}\\
&\leq \frac{2}{\epsilon^2}\sum_{l=1}^{L} 2^{l-1}\frac{1}{1+t\ln(t)^2}\\
&\leq \frac{2}{\epsilon^2}(1+\sum_{l=2}^{L}\frac{1}{(l-1)^2})\qquad\qquad (t\ln(t)^2\geq 2^{l-1}(l-1)^2)\\
&\leq \frac{2}{\epsilon^2}(2+\int_{1}^{L-1}\frac{1}{x^2})dx)\\
&\qquad\qquad\text{(upper bound discrete sum with integral)}\\
&\leq \frac{6}{\epsilon^2}
\end{aligned}
$$