Koordinator
Florian Krach

Stochastik
Musterlösung 9

1. Auf einem Weihnachtsmarkt findet Laura einen Stand mit einer Maschine, die das mehrfache Abwerfen einer Münze durchführt und das Ergebnis ausliest. So können in kurzer Zeit viele Münzwürfe durchgeführt werden. Beantworte für jede der nachfolgenden Teilaufgaben:
Für welche der je zwei Optionen sollte sich Laura entscheiden: (1), (2), oder sind beide Optionen etwa gleich gut? Bedenke, dass sie keinen Taschenrechner dabei hat, also die Gewinnwahrscheinlichkeiten nicht genau ausrechnen kann.
Tipp: Überlege wie sich die Standardabweichung für $\bar{X}_n$ und S_n mit n ändern.

a) (1) Es wird 100 mal eine Münze geworfen und Laura bekommt 10 CHF, wenn die Münze mehr als 52% Kopf zeigt.
(2) Es wird 1000 mal eine Münze geworfen und Laura bekommt 10 CHF, wenn die Münze mehr als 52% Kopf zeigt.

b) (1) Es wird 100 mal eine Münze geworfen und Laura bekommt 10 CHF, wenn die Münze weniger als 55 mal Kopf zeigt.
(2) Es wird 200 mal eine Münze geworfen und Laura bekommt 10 CHF, wenn die Münze weniger als 105 mal Kopf zeigt.

c) (1) Es wird 100 mal eine Münze geworfen und Laura bekommt 10 CHF, wenn die Münze weniger als 40 mal Kopf zeigt.
(2) Es wird 400 mal eine Münze geworfen und Laura bekommt 10 CHF, wenn die Münze weniger als 180 mal Kopf zeigt.

Lösung: Wir betrachten die unabhängigen Zufallsvariablen X_i wobei $X_i = 1$, wenn der i-te Münzwurf Kopf zeigt, und $X_i = 0$, wenn der i-te Münzwurf zahl zeigt und betrachten für $n \in \mathbb{N}$

$$S_n := \sum_{i=1}^{n} X_i, \qquad \bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i.$$

a) Laura gewinnt, wenn $\bar{X}_n > 0.52$. Wir wissen, dass

$$E[\bar{X}_n] = E[X_i] = 0.5, \quad \mathrm{Var}(\bar{X}_n) = \frac{1}{n}\mathrm{Var}(X_i)$$

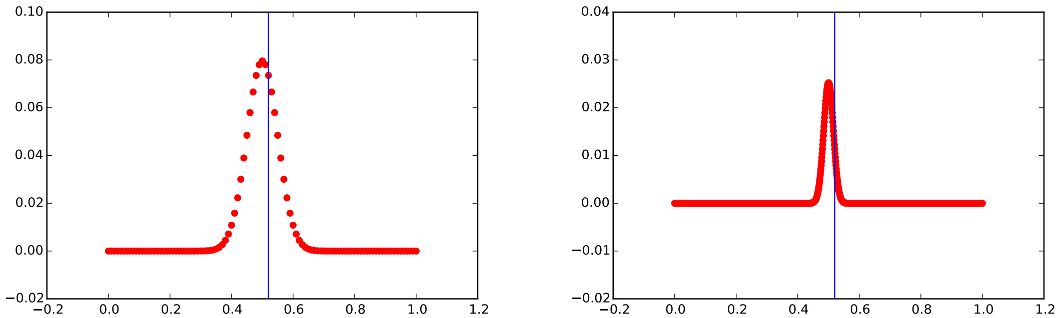

Abbildung 1: Skizzierung der (eigentlich diskreten) Verteilungen von $\bar{X}_{100}$ (links) und $\bar{X}_{1000}$ (rechts) in Aufgabe 1.**a)**. Die vertikalen Linien sind bei $x = 0.52$.

und

$$\sigma_{\bar{X}_n} = \frac{\sigma_{X_i}}{\sqrt{n}}$$

$\bar{X}_n$ ist also zentriert um 0.5 und die Streuung wird kleiner, wenn n grösser ist. Bei kleinerem n hat man also bessere Chancen über 0.52 zu kommen. Option 1 ist daher besser.

Zur Überprüfung: Die Gewinnwahrscheinlichkeiten würde man erhalten, da S_n Bin$(n, 0.5)$ verteilt ist und mit $n = 100$

$$P[S_{100} > 52] = 1 - P[S_{100} \leq 52] \approx 1 - 0.691 = 0.209 = 20.9\%$$

und mit $n = 1000$

$$P[S_{1000} > 520] = 1 - P[S_{1000} \leq 520] \approx 1 - 0.903 = 0.097 = 9.7\%$$

b) Laura gewinnt, wenn $S_{100} < 55$ oder $S_{200} < 105$. Da $E[S_{100}] = 50$ und $E[S_{200} = 100$, erlaubt man in beiden Fällen bis zu 4 Kopfwürfe mehr als der Erwartungswert. Da $\mathrm{Var}(S_n) = n\,\mathrm{Var}(X_i)$ und $\sigma_{S_n} = \sqrt{n}\sigma_{X_i}$ wissen wir, dass die Streuung von S_n zunimmt mit wachsendem n. Bei grossen n hat man also eine grössere Chance mehr als 4 Kopfwürfe mehr als der Ewartungswert zu bekommen - in welchem Fall Laura verlieren würde. Also ist Option 1 besser.

Zur Überprüfung: Die Gewinnwahrscheinlichkeiten sind:

$$P[S_{100} < 55] = P[S_{100} \leq 54] \approx 0.816 = 81.6\%$$

und mit $n = 200$

$$P[S_{200} < 105] = P[S_{200} \leq 104] \approx 0.782 = 78.2\%$$

Siehe nächstes Blatt!

c) Der Fall ist ähnlich zu b), allerdings hat Option 2 nun 4 mal mehr Münzwürfe als Option 1 - die Standardabweichung ist also 2 mal höher. Da wir auch die doppelte Distanz zum Erwartungswert haben können, sind beide Optionen in etwa gleich gut.

Zur Überprüfung: Mit $n = 100$

$$P[S_{100} < 40] = P[S_{100} \leq 39] \approx 0.018 = 1.8\%$$

und mit $n = 400$

$$P[S_{400} < 180] = P[S_{400} \leq 179] \approx 0.02 = 2\%.$$

2. Für eine Studie werden über eine Telefonumfrage zufällig ausgewählte Personen über sensitive Bereiche in ihrem Leben, wie zum Beispiel Drogenkonsum befragt. Bei einer direkten Frage, wird solch eine Frage möglicherweise mit "nein" beantwortet, obwohl die Person Drogen konsumiert hat. Oft wird einfach die sozial besser akzeptierte Anwort gewählt. Um dies zu vermeiden kann man statt direkter Nachfrage wie folgt vorgehen:
Bitte die Person eine Münze zu werfen und sich das Resultat zu merken - dies aber für sich zu behalten. Die Person soll nun, wenn die Münze Kopf gezeigt hat mit "ja" antworten, und wenn die Münze Zahl gezeigt hat, die Frage beantworten.

a) Warum ist es wahrscheinlicher, dass Leute auf diesem Weg die Wahrheit sagen?

b) Nehmen wir eine zufällige Auswahl von 100 Personen aus der Bevölkerung und befragen diese auf dem oben vorgeschlagenen Weg nach dem Drogenkonsum im letzten Jahr. Angenommen 65 Personen antworten mit "ja". Berechne den Momentenschätzer für den Anteil der Bevölkerung, der im vergangenen Jahr Drogen konsumiert hat.

c) Wir führen diese Umfrage in der Vorlesung durch. Benutze die Resultate um den Anteil der Personen, die im letzten Jahr Drogen konsumiert haben, zu schätzen.

Lösung: Betrachte die Zufallsvariablen X_i mit $X_i = 1$, wenn die i-te Person mit "ja" geantwortet hat, und $X_i = 0$ wenn die i-te Person mit "nein" geanwortet hat. Der gesuchte Anteil der Presonen, die Drogen konsumiert haben sei q.

Zum Beispiel mittels Baumdiagram sieht man:

$$E[X_i] = P[X_i = 1] = 0.5 + 0.5 \cdot q.$$

Bitte wenden!

Setzen wir das empirische erste Moment ein und lösen nach q, erhalten wir

$$\bar{x} = 0.5 + 0.5 \cdot \hat{q}$$

und damit

$$\hat{q} = 2 \cdot \bar{x} - 1.$$

In **b)** haben wir $\bar{x} = 0.65$ und somit $\hat{q} = 0.3$. Das ist der Momentenmethodenschätzer für den wahren Anteil q.

3. Um genauere Prognosen bezüglich der Flugzeiten ihrer Flugzeuge machen zu können, möchte die Fluggesellschaft SwitzAir zur Modellierung der absoluten Windgeschwindigkeiten neuerdings eine sogenannte Rayleigh-Verteilung mit Dichte

$$f(x) = \begin{cases} 0 & x < 0 \\ \frac{x}{\sigma^2} \exp\left(-\frac{x^2}{2\sigma^2}\right) & x \geq 0 \end{cases}$$

und kumulativer Verteilungsfunktion

$$F(x) = \begin{cases} 0 & x < 0 \\ 1 - \exp\left(-\frac{x^2}{2\sigma^2}\right) & x \geq 0 \end{cases}$$

benutzen, wobei $\sigma > 0$ ein unbekannter Parameter ist. Sie wurden von der Forschungsabteilung beauftragt, den unbekannten Parameter σ der Verteilung anhand von n beobachteten Datenpunkten zu schätzen.

Tipp: Für eine Rayleigh(σ)-verteilte Zufallsvariable X gilt $\mathbb{E}[X] = \sqrt{\frac{\pi}{2}}\sigma$ und $\mathrm{Var}(X) = \frac{4-\pi}{2}\sigma^2$.

a) Bestimme die Likelihood- und die log-Likelihoodfunktion basierend auf den n unabhängigen, identisch verteilten Beobachtungen $x_1, \ldots, x_n$ einer Zufallsvariable mit obiger Dichte. Berechne daraus den Maximum-Likelihood Schätzer für den unbekannten Parameter σ.

Da die SwitzAir selbst über keine Messinstrumente zur Messung der absoluten Windgeschwindigkeiten verfügt, beauftragt sie ein meteorologisches Institut, 100 Windgeschwindigkeitsmessungen durchzuführen. Das meteorologische Institut liefert folgende Kennzahlen: Mittelwert $\bar{x} = 2.01$, Standardabweichung $s_x = 0.51$, Median $m = 3.12$ und Quartilsdifferenz $q_{0.75} - q_{0.25} = 0.89$.

Siehe nächstes Blatt!

b) Berechne den Momentenschätzer basierend auf dem ersten Moment. Welchen Wert nimmt der Schätzer für die gesammelten Daten des meteorologischen Instituts an?

Lösung:

a) Die Likelihoodfunktion ist gegeben durch

$$\ell(\sigma; x_1, \ldots x_n) = \prod_{i=1}^{n} \left\{ \frac{x_i}{\sigma^2} \cdot \exp\left(-\frac{x_i^2}{2\sigma^2}\right) \right\}.$$

Die log-Likelihoodfunktion ist dann

$$\begin{aligned}
\log \ell(\sigma; x_1, \ldots x_n) &= \sum_{i=1}^{n} \log\left(\frac{x_i}{\sigma^2} \exp\left(-\frac{x_i}{2\sigma^2}\right)\right) \\
&= \sum_{i=1}^{n} \left\{ \log(x_i) - 2\log(\sigma) - \frac{x_i^2}{2\sigma^2} \right\} \\
&= \sum_{i=1}^{n} \log(x_i) - 2n\log(\sigma) - \sum_{i=1}^{n} \frac{x_i^2}{2\sigma^2}.
\end{aligned}$$

Um nun den Maximum Likelihood Schätzer für σ zu erhalten, leiten wir die log-Likelihoodfunktion nach σ ab,

$$\frac{\mathrm{d}}{\mathrm{d}\sigma} \log \ell(\sigma; x_1, \ldots x_n) = 0 - \frac{2n}{\sigma} + \sum_{i=1}^{n} \frac{x_i^2}{\sigma^3}.$$

Null setzen und nach σ auflösen, ergibt

$$\begin{aligned}
& & \frac{2n}{\hat{\sigma}} &= \sum_{i=1}^{n} \frac{x_i^2}{\hat{\sigma}^3} \\
&\Longrightarrow & 2n\hat{\sigma}^2 &= \sum_{i=1}^{n} x_i^2, \\
&\Longrightarrow & \hat{\sigma}_{MLE} &= \sqrt{\frac{1}{2n} \sum_{i=1}^{n} x_i^2}.
\end{aligned}$$

b) Es gilt $E[X] = \sqrt{\frac{\pi}{2}}\sigma$. Einsetzen des empirischen Moments gibt den Momen-

Bitte wenden!

tenschätzer:

$$
\begin{aligned}
& & m_1 &= \sqrt{\frac{\pi}{2}}\hat{\sigma} \\
&\Longrightarrow & \frac{1}{n}\sum_{i=1}^{n} x_i &= \sqrt{\frac{\pi}{2}}\hat{\sigma} \\
&\Longrightarrow & \hat{\sigma}_{MoM} &= \sqrt{\frac{2}{\pi}}\frac{1}{n}\sum_{i=1}^{n} x_i.
\end{aligned}
$$

Um den konkreten Wert des Schätzers anzugeben, benötigt man nur den Mittelwert der Daten. Für die gegebenen 100 Datenpunkte ist der Schätzer gegeben durch:

$$\hat{\sigma}_{MoM} = \sqrt{\frac{2}{\pi}}\,\bar{x} = \sqrt{\frac{2}{\pi}} \cdot 2.01 = 1.60.$$

4. Wir nehmen an, dass der Bleigehalt X in einer Bodenprobe normalverteilt ist,

$$X \sim \mathcal{N}(\mu, \sigma^2).$$

a) Die Standardabweichung ist bekannt ($\sigma = 10$ ppb) und man interessiert sich für den Erwartungswert μ. Man misst den Bleigehalt in 10 Bodenproben und erhält einen Mittelwert von 31 ppb. Gib ein 99% Vertrauensintervall für μ an.

Tipp: Die kumulative Verteilungsfunktion der Standardnormalverteilung erfüllt
$\Phi(2.58) = 0.995$.

b) Wieviel Beobachtungen sind nötig, um die Breite des Vertrauensintervalles auf die Hälfte zu reduzieren?

Wieviele (unabhängige) Bestimmungen des Bleigehaltes müssen geplant werden, wenn der Bleigehalt mit einer Stichprobe "auf 1 ppb genau" bestimmt werden soll, d.h. wenn die Breite des 99 % Konfidenzintervalls nicht grösser als 1 ppb sein soll?

c) Normalerweise ist die Standardabweichung σ unbekannt. Um welchen Faktor verändert sich die Breite des Vertrauensintervalles in a), wenn man die Standardabweichung aus den Daten geschätzt hat?

Lösung:

a) X bezeichne den Schwermetallgehalt in einer Bodenprobe. Es gilt:

$$X \sim \mathcal{N}(\mu, \sigma^2), \text{ mit } \sigma = 10 \text{ und } \mu \text{ unbekannt.}$$

Siehe nächstes Blatt!

Der Mittelwert $\overline{X}$ von $n = 10$ Stichproben ist auch normalverteilt mit Standardabweichung $\sigma/\sqrt{n}$,

$$\overline{X} \sim \mathcal{N}(\mu, \sigma^2/n).$$

Da $\Phi(2.58) = 0.995$ (gegebener Hinweis), liegen 99% aller Beobachtungen von der standardisierten Zufallsvariable

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

in dem Intervall $[-2.58, 2.58]$. Also liegen 99% aller Beobachtungen von $\overline{X} - \mu$ im Intervall $[-2.58 \cdot \sigma/\sqrt{n}, 2.58 \cdot \sigma/\sqrt{n}]$. Ein 99% Vertrauensintervall für μ ist demnach gegeben durch

$$\left[\overline{X} - 2.58 \cdot \frac{\sigma}{\sqrt{n}}, \overline{X} + 2.58 \cdot \frac{\sigma}{\sqrt{n}}\right],$$

Mit dem bekannten Wert von $\sigma = 10$, $\overline{x} = 31$ und $n = 10$ erhält man also ein 99% Vertrauensintervall für μ von $[22.84, 39.16]$.

b) Aus a) sieht man, dass die Breite des Vertrauensintervalles wie $1/\sqrt{n}$ abfällt mit der Anzahl n von Beobachtungen. Also sind viermal so viele Beobachtungen ($4 \cdot 10 = 40$) nötig, um die Breite des Vertrauensintervalles zu halbieren.

Die Breite des 99% Vertrausensintervalles ist, siehe a),

$$2 \cdot 2.58 \cdot \frac{\sigma}{\sqrt{n}}.$$

Um die Breite des Vertrauensintervalles kleiner als 1ppb zu erhalten, muss die Anzahl n der Beobachtungen entsprechend gross werden:

$$\begin{aligned} 2 \cdot 2.58 \cdot \frac{\sigma}{\sqrt{n}} &\leq 1 \\ \rightarrow \qquad 51.6 &\leq \sqrt{n} \\ \rightarrow \qquad n &\geq 2663. \end{aligned}$$

Es müssen mindestens 2663 Beobachtungen vorliegen, um ein 99% Vertrauensintervall von weniger als 1ppb Breite zu erhalten.

c) Die standardisierte Variable

$$\frac{\overline{X} - \mu}{\widehat{\sigma}/\sqrt{n}}$$

mit Schätzwert $\widehat{\sigma}$ für σ ist nicht mehr normalverteilt (wie für bekanntes, festes σ), sondern folgt einer t-Verteilung mit 9 Freiheitsgraden, wobei

$$\hat{\sigma} = S = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2}.$$

Bitte wenden!

Das 99.5% Quantil dieser Verteilung ist bei 3.25 (siehe Tabelle). Insofern fallen 99% der Beobachtungen von

$$\frac{\overline{X} - \mu}{\widehat{\sigma}/\sqrt{n}}$$

in das Intervall $[-3.25, 3.25]$. Ein Vertrauensintervall ist daher gegeben durch

$$\left[\overline{X} - 3.25 \cdot \frac{\widehat{\sigma}}{\sqrt{n}}, \overline{X} + 3.25 \cdot \frac{\widehat{\sigma}}{\sqrt{n}}\right].$$

Für $\hat{\sigma} = 10$ und $n = 10$ ergibt sich das Vertrauensintervall $[20.72, 41.28]$.

Durch Vergleich mit a) findet man, dass das Vertrauensintervall um einen Faktor 3.25/2.58, also um rund 26% grösser geworden ist.

5. (Alte Prüfungsaufgabe - Winter 2019) Eine Firma verkauft Mehl und hat eine Großbestellung von einer Konditorei für 10'000 kg Mehl erhalten. Ein Sack enthält etwa 10 kg Mehl. Genauer gesagt ist die Menge an Mehl in einem Sack gleichverteilt zwischen 9 kg und 11 kg. Die Firma hat nun 1003 Säcke Mehl an den Besteller versendet.

a) Treffen Sie vernünftige Annahmen, sodass Sie die approximative Verteilung für die Gesamtmenge an Mehl in der Sendung bestimmen können. Bestimmen Sie diese Verteilung inklusive der Parameter.

b) Bestimmen Sie approximativ die Wahrscheinlichkeit, dass mit den 1003 versendeten Säcken Mehl die Bestellung von 10'000 kg Mehl auch wirklich abgedeckt ist. Benutzen Sie dazu Ihre Antwort aus (a) und die Standardnormaltabelle.

Die Konditorei blickt auf die Mehlbestellungen im Jahr 2018 zurück. Insgesamt wurden 52 Bestellungen von je 10'000 kg Mehl getätigt. Wir bezeichnen mit X_i die tatsächlich gelieferte Menge an Mehl von Bestellung i für $i \in \{1, ..., 52\}$. Aufgrund der Angaben der liefernden Firmen können wir davon ausgehen, dass die X_i unabgängig und normalverteilt mit Erwartungswert μ und Varianz σ^2 sind.

Die Gesamtmenge des geliefterten Mehls in 2018 sei S_n und die durchschnittlich geliefterte Menge $\bar{X}_n$, wobei $n = 52$.

c) Wie sind $\bar{X}_n$ und S_n verteilt? Bestimmen Sie die Parameter als Funktionen von μ und σ.

Siehe nächstes Blatt!

d) In den ersten 3 Monaten des Jahres hat die Konditorei 12 Lieferungen mit den folgenden Mengen an Mehl erhalten:

Lieferung	1	2	3	4	5	6
Menge an Mehl in kg	10'041	10'019	10'033	10'028	10'005	10'002
Lieferung	7	8	9	10	11	12
Menge an Mehl in kg	10'034	9'990	10'032	10'008	10'016	10'026

(i) Leiten Sie den Maximum-Likelihood Schätzer für μ in allgemeiner Form her. Nehmen Sie $\sigma = 20$ als gegeben an. Berechnen Sie den Wert auf Basis der Daten aus der Tabelle mit den ersten 12 Lieferungen.

(ii) Bestimmen Sie das 95% Vertrauensintervall für μ. Nehmen Sie $\sigma = 20$ wiederum als gegeben an.

Lösung:

a) Für $i = 1, ..., 1003$ sei Y_i die Menge an tatsächlich versendetem Mehl. Es gilt $Y_i \sim U([9, 11])$ und damit $\mathbb{E}[Y_i] = 10$ und $Var(Y_i) = 4/12 = 1/3$. Wir nehmen an, dass alle Y_i unabhängig und identisch $U([9, 11])$-verteilt sind. Dann ist $X := \sum_{i=1}^{1003} Y_i$ nach dem zentralen Grenzwertsatz $\mathcal{N}(\mu_X, \sigma_X^2)$-verteilt, wobei

$$\mu_X = 1003 \cdot 10 = 10'030, \quad \sigma_X = \sqrt{1003/3} \approx 18.28$$

bzw. $\sigma_X^2 = 1003/3 = 334.\overline{3}$.

b)

$$\begin{aligned}\mathbb{P}[X \geq 10'000] &= \mathbb{P}\left[\frac{X - 10'030}{\sigma_X} \geq \frac{-30}{\sigma_X}\right] \\ &= \mathbb{P}\left[\frac{X - 10'030}{\sigma_X} \leq \frac{30}{\sigma_X}\right] \\ &\approx \Phi(1.64) = 94.95\%,\end{aligned}$$

wobei Φ die Verteilungsfunktion der Standardnormalverteilung ist.

c) Es gilt, dass $\bar{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$ also mit $n = 52$: $\bar{X}_{52} \sim \mathcal{N}(\mu, \sigma^2/52)$ und $S_n \sim \mathcal{N}(n\mu, n\sigma^2)$, also mit $n = 52$: $S_{52} \sim \mathcal{N}(52\mu, 52\sigma^2)$

d) (i) Für m Beobachtungen x_1,... x_m und Varianz σ^2 ist der Maximum-Liklihood-Schätzer für μ, der Wert $\hat{\mu}$, welcher

$$L(\mu) = \prod_{i=1}^{m} f_{\mu,\sigma}(x_i)$$

Bitte wenden!

maximiert, wobei $f_{\mu,\sigma}$ die Dichte der $N(\mu, \sigma^2)$-Verteilung ist. Äquivalent können wir hier die log-Likelihood-Funktion

$$l(\mu) = \log((L(\mu)) = \frac{m}{2}\log(2\pi\sigma) - \sum_{i=1}^{m}(x_i - \mu)^2/(2\sigma^2)$$

verwenden. Leiten wir beide Seiten nach μ ab, ergibt das

$$\frac{d}{d\mu}l(\mu) = 2\sum_{i=1}^{m}(x_i - \mu)/(2\sigma^2).$$

Durch 0-setzen des Terms erhalten wir nun den Maximum-Likelihood-Schätzer

$$\hat{\mu} = \frac{1}{m}\sum_{i=1}^{m}x_i = \bar{x}_m.$$

Mit $m = 12$ und den Daten aus der Tabelle erhalten wir

$$\hat{\mu} = \bar{x}_m = 10'019.5$$

(ii) Hier gilt $\sigma = 20$, $\alpha = 0.05$. Das Vertrauensintervall ist nun gegeben als

$$I = [\bar{X}_m - z_{0.975}\sigma/\sqrt{m}, \bar{X}_m + z_{0.975}\sigma/\sqrt{m}].$$

Mit den Daten aus den ersten 12 Lieferungen und $\bar{X}_m = \bar{x}_m$ erhalten wir das Vertrauensintervall

$$[10'008.2, 10'030.8].$$