Bernhard Dick

DER SATZ VON BAYES UND NATURWISSENSCHAFTLICHE INDUKTION

In den Naturwissenschaften geht es darum, aus Beobachtungen oder experimentellen Daten Schlüsse zu ziehen. Das Ergebnis ist in der Regel ein Satz, etwa „In Wasser gelöste Salze leiten den elektrischen Strom“, oder „alle Körper fallen unabhängig von ihrer Masse im Vakuum gleich schnell“. Meist gibt man sich aber mit solch allgemeinen Sätzen nicht zufrieden sondern möchte zu quantitativen Aussagen kommen. In einem ersten Schritt kann man Eigenschaften wie z.B. den Schmelzpunkt T_S einer Substanz definieren, deren Wert gemessen werden kann. Zur Beschreibung von Zusammenhängen zwischen verschiedenen Größen werden Modelle entwickelt, welche den interessierenden Teil der Wirklichkeit abbilden sollen. Diese Modelle enthalten Parameter, für welche das Modell mathematische Beziehungen postuliert. Für das Problem des freien Falls wären das z.B. die schwere Masse m_S (d.h. das Gewicht), die träge Masse m_T (welche die Beschleunigung mit der Kraft verknüpft), die Fallhöhe x, die Fallzeit t, und schließlich eine Konstante g, welche die schwere Masse mit der Kraft verknüpft, welche der Körper auf eine Waage ausübt. Die mathematische Abbildung des Modells besteht aus den beiden Gleichungen für die Kraft, die einerseits aufgrund der Gravitation auf den Körper wirkt, und die andererseits diesen Körper beschleunigt:

$$F = g\, m_S$$

$$F = m_T\, \frac{d^2x}{dt^2}$$

Das Modell setzt nun beide Kräfte gleich und löst die Differentialgleichung für die Fallhöhe, mit dem Ergebnis

$$x = \frac{1}{2}\, g\, \frac{m_S}{m_T}\, t^2$$

Misst man mit verschiedenen Körpern und Fallhöhen die Fallzeiten, dann könnte die Schlussfolgerung so aussehen:

Satz 1: „Für alle Körper ist die schwere Masse gleich der trägen Masse.“

Satz 2: „Die Konstante g hat den Wert 9.81 m s^{-2}.“

Prof. Dr. Bernhard Dick
Institut für Physikalische und Theoretische Chemie
Universität Regensburg,
Universitätsstraße 31, D-93053 Regensburg
Tel: +49 941 943 4486, Fax: +49 941 943 4488,
E-Mail: bernhard.dick@chemie.uni-regensburg.de

Nun wurden die Messungen aber gar nicht für „alle Körper“ gemacht, sondern nur für eine endliche Zahl von Testkörpern. Ist die Verallgemeinerung im Satz 1 also zulässig? In der Mathematik existiert das Beweisverfahren der vollständigen Induktion, ein entsprechendes Beweisverfahren existiert in den Naturwissenschaften aber nicht. Beim zweiten Satz gibt es noch ein weiteres Problem. Auch wenn eine andere Serie von Experimenten den ersten Satz bestätigen könnte, würde sie im zweiten Satz einen anderen Zahlenwert finden. Beide Zahlenwerte liegen vielleicht nahe beieinander, sind aber nicht völlig identisch. Ein Ausweg, der beide Ergebnisse vereint, wäre, statt eines einzigen Wertes ein Intervall anzugeben, etwa:

Satz 2a: „Der Wert der Konstanten g liegt im Intervall (9.80 – 9.82) m s^{-2}.“

Aber auch hier kann nicht ausgeschlossen werden, dass ein weiteres Experiment einen Wert außerhalb dieses Intervalls findet. (Es sei denn, man wählt das Intervall unendlich groß – aber dann hat der Satz keinen Erkenntniswert.)

Während also in der Mathematik eine Aussage entweder bewiesen oder widerlegt werden kann, ist in der Naturwissenschaft nur das zweite möglich. Der Wissenschaftstheoretiker Popper hat das bekanntlich so formuliert, dass sich Theorien nicht verifizieren sondern nur falsifizieren lassen. Daher sollte die vornehmste Aufgabe eines Experimentes darin bestehen, zu versuchen, eine Theorie zu widerlegen. Gelingt dies, erhält man eine sichere Erkenntnis. Gelingt dies nicht, so ist die Aussage dennoch nicht bewiesen. Nach Popper würde ein Experiment, welches die Theorie bestätigt, überhaupt keinen Erkenntnisgewinn liefern.

In der wissenschaftlichen Praxis liegen die Dinge aber komplizierter. Einerseits kann man ja auch beim Falsifizieren einen Fehler machen – also dürfte man die Theorie nicht auf Grund einer einzigen ihr widersprechenden Beobachtung verwerfen. Dies würde man erst tun, wenn sich die Falsifikation reproduzieren lässt, sich also bewährt. Andererseits sollte es für eine Theorie nicht unerheblich sein, ob ihre Prognosen nur einmal oder tausendmal eingetroffen sind. Jede unabhängige Bestätigung einer Theorie sollte diese irgendwie „besser“ machen. Dies geschieht in der Bayesischen Sichtweise dadurch, dass man jeder Theorie nicht nur die extremen Attribute „wahr“ (1) oder „falsch“ (0) zuordnet, sondern eine „Glaubwürdigkeit“, deren Wert zwischen diesen Extremen liegt. Eine Bestätigung der Theorie würde deren Glaubwürdigkeit erhöhen, ein ihr widersprechendes Experiment ihre Glaubwürdigkeit reduzieren.

Im Folgenden werde ich zuerst den Satz von Bayes vorstellen, der ursprünglich ein Theorem aus der Wahrscheinlichkeitstheorie darstellt. Die Anwendung auf das Verfahren des wissenschaftlichen Schlussfolgerns geschieht dann dadurch, dass der Begriff der Wahrscheinlichkeit umgedeutet wird als ein Maß für die Glaubwürdigkeit einer Aussage (im Gegensatz zu einer relativen Häufigkeit). Anschließend wird an Hand einiger einfacher Beispiele gezeigt, wie dieses Prinzip in der Praxis funktioniert. So erklärt es z.B., unter welchen Bedingungen der Mittelwert mehrerer Messungen die beste Schätzung für den „wahren" Wert einer Messgröße ist, oder wann die Minimierung der Fehlerquadrate das geeignete Verfahren für die Bestimmung eines optimalen Fits ist, und wann nicht.

Bayesische Konzepte werden in verschiedenen Varianten schon einige Jahrzehnte zur Datenanalyse genutzt. Finden sich in der Datenbank von Scifinder in der Dekade 1961 – 1970 erst 53 Artikel mit dem Stichwort „Bayesian", so sind es in den folgenden Jahrzehnten 323 (1971 – 1980), 1234 (1981 – 1990) und 3437 (1991 – 2000). Ab dem Jahr 2000 steigt die Zahl dramatisch an: von 2001 bis 2010 waren es schon 24252 Artikel, und von 2011 bis ins laufende Jahr 2018 weitere 38624. Auch wenn vieles davon auf die allgemeine Inflation an wissenschaftlichen Zeitschriftenbeiträgen zurückzuführen sein dürfte, so legen diese Zahlen doch nahe, dass das Thema inzwischen „angekommen" ist. Dennoch ist die Bayesische Theorie bisher kaum in den physikalisch-chemischen Unterricht oder gar in Lehrbücher der PC vorgedrungen. Dieser Artikel soll daher dem Novizen einen allerersten Eindruck verschaffen, und gleichzeitig (hoffentlich) Appetit auf mehr anregen. Diejenigen, die tiefer eindringen möchten, sollten in folgenden (mehr oder weniger zufällig ausgewählten) Büchern [1-3] fündig werden.

DER SATZ VON BAYES

Thomas Bayes wurde um 1701 in London geboren und starb am 7. April 1761 im Ort Tunbridge Wells in England, in dem er zuvor als presbyterianischer Pfarrer gewirkt hatte. Zum Theologiestudium der damaligen Zeit gehörte auch eine Ausbildung in Mathematik und Logik. Auf diesem Gebiet war er auch später als Privatgelehrter mit solchem Erfolg aktiv, dass er am 4. November 1742 zum Fellow of the Royal Society gewählt wurde.

Im Jahr 1763 veröffentlichte sein Freund Richard Price in den *Philosophical Transactions of the Royal Society* einen Artikel, den er im Nachlass von Thomas Bayes gefunden hatte [4]. Der Artikel trägt den Titel „An essay towards solving a problem in the doctrine of chances". Darin beweist er unter anderem ein Theorem, das heute als „Satz von Bayes" bekannt ist. In der heute üblichen Nomenklatur hat es die Form

$$p(A|B) = p(B|A)\frac{p(A)}{p(B)} \qquad (1)$$

Im Folgenden werde ich diesen Satz zunächst am Beispiel eines Würfels demonstrieren, und dann für beliebige Wahrscheinlichkeiten verallgemeinern.

Betrachten wir eine Menge an Ereignissen, die alle die gleiche a-priori Wahrscheinlichkeit besitzen. Ein Beispiel wäre die Augenzahl, die beim Wurf eines idealen Würfels angezeigt wird. Die Menge soll zudem vollständig sein, das heißt, sie muss alle möglichen dieser elementaren Ereignisse umfassen. Die Zahl dieser Möglichkeiten sei N. Im Fall des Würfels ist das die Menge der Augenzahlen $M = \{1,2,3,4,5,6\}$, und die Zahl der Elemente ist $N = 6$. Das ist in Abb. 1(a) skizziert.

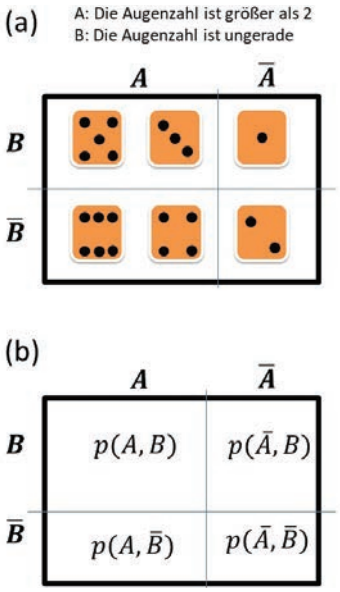

Abb. 1: (a) Die 6 möglichen Ergebnisse eines Würfelwurfs können in 4 disjunkte Teilmengen aufgeteilt werden, für die jeweils die Aussagen A bzw. B wahr oder falsch sind. (b) Diese Aufteilung gilt auch für die Wahrscheinlichkeiten für die 4 möglichen Kombinationen der Wahrheit oder Falschheit der beiden Aussagen, auch wenn diese Wahrscheinlichkeiten keine Verhältnisse ganzer Zahlen sind.

Weiterhin formulieren wir eine Aussage, welche für ein bestimmtes Ereignis entweder wahr oder falsch ist. Ein Beispiel wäre der Satz

(A): Die Augenzahl ist größer als 2

Die Aussage, welche das logische Gegenteil besagt, bezeichnen wir mit einem Querbalken über dem Symbol:

($\bar{A}$): Die Augenzahl ist nicht größer als 2

Die Menge aller möglichen Ereignisse lässt sich dann aufteilen in zwei disjunkte Teilmengen. Für die Elemente der einen Teilmenge ist der Satz wahr, für die in der anderen Teilmenge ist er falsch. Die Zahl der Elemente in der ersten Menge sei $N(A)$, die in der zweiten Menge $N(\bar{A})$. Dann definieren wir die Wahrscheinlichkeit von (A) als

$$p(A) = \frac{N(A)}{N}$$

Da $N(A) + N(\bar{A}) = N$ gilt, folgt $p(A) + p(\bar{A}) = 1$. Für unser Beispiel des idealen Würfels finden wir $p(A) = 2/3$. Nun formulieren wir einen zweiten Satz

(B): Die Augenzahl ist eine ungerade Zahl

In diesem Fall enthalten beide Teilmengen je 3 Elemente (siehe Abb. 1(a)), folglich ist $p(B) = 1/2$.

Nun werden zwei neue Wahrscheinlichkeiten eingeführt, die von beiden Sätzen abhängen. Die kombinierte Wahrscheinlichkeit $p(A, B)$ bezeichnet die Wahrscheinlichkeit, dass beide Sätze wahr sind. Sie ist gegeben durch die Anzahl an Ereignissen, für die das zutrifft, dividiert durch die Gesamtzahl aller möglichen Ereignisse:

$$p(A, B) = \frac{N(A, B)}{N}$$

Die bedingte Wahrscheinlichkeit $p(A|B)$ bezeichnet die Wahrscheinlichkeit, dass (A) wahr ist, unter der Bedingung, dass (B) wahr ist. In diesem Fall besteht der Ereignisraum also nur aus den Ereignissen in der Teilmenge, für die (B) wahr ist. In unserem Beispiel ist dies die obere Reihe in Abb. 1(a). Die Wahrscheinlichkeit ist daher

$$p(A|B) = \frac{N(A, B)}{N(B)}$$

Für unser Beispiel des idealen Würfels finden wir $p(A, B) = 2/6$ und $p(A|B) = 2/3$ (d.h. der Anteil der Augenzahlen größer als 2 unter den ungeraden). Während $p(A, B) = p(B, A)$ symmetrisch in den beiden Argumenten ist, gilt dies nicht für die bedingte Wahrscheinlichkeit, denn $p(B|A) = 1/2$ (d.h. die Anzahl der ungeraden Zahl unter denen, die größer als 2 sind).

Aus den oben angeführten Definitionen folgt nun die Beziehung

$$p(A, B) = \frac{N(A, B)}{N} \frac{N(B)}{N(B)} = p(A|B)\, p(B) \qquad (2)$$

Und da $p(A, B)$ symmetrisch in den beiden Argumenten ist, gilt ebenfalls

$$p(A, B) = \frac{N(A, B)}{N} \frac{N(A)}{N(A)} = p(B|A)\, p(A) \qquad (3)$$

Setzt man die rechten Seiten von (2) und (3) gleich und löst nach $p(A|B)$ auf, erhält man den Satz von Bayes (1). Was ist daran nun so spektakulär? Dieser Satz von Bayes erlaubt es, in bedingten Wahrscheinlichkeiten die Bedingung und die bedingte Größe zu vertauschen! Darauf werden wir im übernächsten Kapitel zurückkommen.

Für die Herleitung des Satzes von Bayes ist die Annahme gleich-wahrscheinlicher Elementarereignisse (wie die Augenzahlen des idealen Würfels) nicht erforderlich. Es genügt, zwei Aussagen (A) und (B) zu nehmen, die jeweils mit den Wahrscheinlichkeiten $p(A)$ und $p(B)$ wahr sind. Die kombinierten Aussagen (A, B), $(\bar{A}, B)$, $(A, \bar{B})$ und $(\bar{A}, \bar{B})$ schließen einander aus und decken zugleich alle Möglichkeiten ab. Dies ist in Abb. 1(b) schematisch dargestellt, in der die Flächen der Rechtecke die jeweiligen Wahrscheinlichkeiten symbolisieren. Für die Summe der vier Wahrscheinlichkeiten gilt

$$p(A, B) + p(A, \bar{B}) + p(\bar{A}, B) + p(\bar{A}, \bar{B}) = 1$$

Die Wahrscheinlichkeiten $p(A)$ und $p(B)$ erhält man, indem man alle kombinierten Wahrscheinlichkeiten addiert, in denen (A) beziehungsweise (B) wahr sind:

$$p(A, B) + p(A, \bar{B}) = p(A)$$

$$p(A, B) + p(\bar{A}, B) = p(B)$$

Für die bedingten Wahrscheinlichkeiten wird der Ereignisraum auf den Teilraum reduziert, für den die Bedingung wahr ist. Die Wahrscheinlichkeit $p(A|B)$ ist also der Anteil, den die Wahrscheinlichkeit $p(A, B)$ zur Wahrscheinlichkeit $p(B)$ beiträgt:

$$p(A|B) = \frac{p(A, B)}{p(A, B) + p(\bar{A}, B)} = \frac{p(A, B)}{p(B)} \qquad (2a)$$

Entsprechend erhalten wir für :

$$p(B|A) = \frac{p(A, B)}{p(A, B) + p(A, \bar{B})} = \frac{p(A, B)}{p(A)} \qquad (3a)$$

Und aus der Kombination den Satz von Bayes:

$$p(A|B) = p(B|A) \frac{p(A)}{p(B)}$$

VON DER WAHRSCHEINLICHKEIT ZUR GLAUBWÜRDIGKEIT

Der Begriff der Wahrscheinlichkeit wird nicht nur in der Umgangssprache sondern auch in der Theorie in mehreren Bedeutungen verwendet, die auf den ersten Blick nicht voneinander abhängen. Die beiden häufigsten Konzepte sind einerseits das einer „relativen Häufigkeit“ (englisch: frequency) andererseits das einer „vernünftigen Erwartung“ (englisch: reasonable expectation), für die ich im Folgenden auch den Begriff „Glaubwürdigkeit“ verwenden möchte. In einem grundlegenden Artikel hat Richard T. Cox im Jahr 1946 gezeigt [5], dass beide Konzepte durchaus Gemeinsamkeiten haben. In diesem Artikel stellt er zunächst Situationen vor, die man mit beiden Konzepten verstehen kann, z.B. die Wahrscheinlichkeit für das Ziehen von weißen und schwarzen Kugeln aus einer Urne, deren Inhalt bekannt ist. Der „Frequentist“ fragt, was er im statistischen Mittel beobachten wird, wenn er das Experiment unendlich oft wiederholt, oder wenn er ein unendlich großes Ensemble solcher Urnen vor sich hat. Eine Glaubwürdigkeit kann man aber bei Kenntnis des Inhalts der Urne auch für eine einzige Urne und einen einzigen Versuch definieren. In diesem Fall führen beide Konzepte zum gleichen Ergebnis.

Es gibt aber Fragestellungen, bei denen die Annahme eines (quasi) unendlich großen Ensembles oder der (quasi) unendlich häufigen Wiederholbarkeit des Versuches absurd erscheint. Cox geht nun nicht von Messungen oder Ereignissen aus, sondern von Aussagen „A“ und „B“, die entweder wahr oder falsch sein können. Auf solche Aussagen lassen sich die Regeln der Boolschen Logik anwenden. Dann definiert er ein Maß $f(A|B)$ für die Glaubwürdigkeit, dass A wahr ist, wenn B als wahr feststeht. Man kann B dann auch eine Hypothese für die Wahrheit der Aussage A nennen. Wenn eine aus zwei Aus-

sagen kombinierte Hypothese $B = (A, X)$ die Aussage A selbst enthält, dann ist $f(A|A, X)$ eine Konstante, die von X nicht abhängt, denn A ist ja aufgrund der Hypothese bereits wahr. Diese Konstante $f(A|A)$ entspricht daher der Glaubwürdigkeit für etwas, das mit Sicherheit wahr ist. Andererseits ist es unmöglich, dass das logische Gegenteil von A, d.h. $\bar{A}$ wahr ist, wenn man die Bedingung vorgibt, dass A wahr sein soll.

Es liegt nahe, der Unmöglichkeit das Maß „0“ und der Gewissheit das Maß „1“ zuzuordnen, also

$$f(A|A) = 1 \;;\; f(\bar{A}|A) = 0$$

Dann kann Cox zeigen, dass alle Glaubwürdigkeiten $f(A|X)$ zwischen diesen beiden Werten liegen müssen. Weiter kann er beweisen, dass für die so definierten Glaubwürdigkeiten gelten muss [6]

$$f(A, B|X) = f(A|B, X)\, f(B|X) \tag{4}$$

Es fällt auf, dass in dieser Betrachtung keine „Glaubwürdigkeiten an sich“ auftreten, sondern nur bedingte Glaubwürdigkeiten. Wenn man aber für eine Hypothese X, die immer wahr ist, $f(B|X) = f(B)$ definiert, dann nimmt Gleichung (4) die Form

$$f(A, B) = f(A|B)\, f(B)$$

an. Dies aber führt wegen $f(B, A) = f(A, B)$ unmittelbar zum Satz von Bayes für die Funktion der Glaubwürdigkeit $f(\cdot)$. Nimmt man die vollständige Form aus Gleichung (4), dann folgt wegen der Vertauschbarkeit von (A, B) in $f(A, B|X)$:

$$f(B|A, X)\, f(A|X) = f(A|B, X)\, f(B|X)$$

Und daraus

$$f(A|B, X) = f(B|A, X)\frac{f(A|X)}{f(B|X)} \tag{5}$$

Dies ist der Bayesische Satz, nun aber erweitert um eine allen Wahrscheinlichkeiten gemeinsame Hypothese X. In der Anwendung auf induktives Schlussfolgern repräsentiert A unsere Theorie, B die Daten, und X unser Vorwissen. Wir suchen die Glaubwürdigkeit unserer Theorie vor dem Hintergrund unseres Vorwissens und unserer Daten. In der Bayesischen Methode heißt $f(A|B, X)$ daher auch Inferenz. Damit wir diese Inferenz aus den Daten finden können, benötigen wir ein Modell, mit dem wir die sogenannte Likelihood $f(B|A, X)$ berechnen können. Diese sagt, wie wahrscheinlich unsere Daten sind, wenn das gewählte Modell richtig ist. $f(B|X)$ nennt man auch die a-priori Wahrscheinlichkeit oder Evidenz der Daten. Diese ist für einen gegebenen Datensatz eine Konstante, die letztlich die Inferenz normiert. Der vierte Term in Gleichung 5, $f(A|X)$, heißt a-priori Wahrscheinlichkeit des Modells. Beim Vergleich verschiedener Modelle hat diese a-priori Wahrscheinlichkeit offensichtlich einen Einfluss auf unsere Inferenz. Im Folgenden werden wir für Glaubwürdigkeit $f(\cdot)$ und Wahrscheinlichkeit $p(\cdot)$ die gleiche Funktion verwenden.
I

IST DIESE MÜNZE FAIR?

Der Satz von Bayes kann genutzt werden, um aus einem Satz von Daten eine Verteilungsfunktion für einen damit verknüpften Modellparameter zu gewinnen. Als Beispiel für das Experiment betrachten wir den Wurf einer Münze. Das Ergebnis kann nur zwei Werte annehmen, nämlich „Kopf“ oder „Zahl“. Wir wiederholen das Experiment N mal, wobei wir K mal das Ergebnis „Kopf“ erhalten. Nennen wir die Wahrscheinlichkeit, dass die Münze auf Kopf fällt, x. Dann ist die Wahrscheinlichkeit, bei N Münzwürfen K mal Kopf zu beobachten, durch die Binomialverteilung gegeben.

$$p(N, K|x) = \binom{N}{K} x^K\,(1-x)^{N-K} \tag{6}$$

Dies ist daher unser theoretisches Modell für den Münzwurf, und die Gleichung definiert die entsprechende Likelihood. Für eine faire Münze erwarten wir $x = 1/2$. Wenn wir wissen, dass die Münze fair ist, können wir die Wahrscheinlichkeit berechnen, dass wir z.B. fünfmal hintereinander Kopf erhalten. Aber wie stellen wir fest, ob die Münze überhaupt fair ist? Können wir aus dem Experiment auf den Wert von x zurückschließen?

Der Satz von Bayes erlaubt es, in der bedingten Wahrscheinlichkeit die beiden Argumente zu vertauschen. Wir erhalten aber nicht einen einzigen Wert für x, sondern eine Wahrscheinlichkeitsverteilung:

$$p(x|N, K) = p(N, K|x)\,\frac{p(x)}{p(N, K)}$$

Die Größe x nimmt, im Gegensatz zu der Zahl der Münzwürfe und der Kopf-Würfe, nicht nur ganzzahlige Werte an, sondern kann alle reellen Werte im Intervall [0,1] annehmen. Dabei entspricht $x = 0$ einer Münze, die immer auf Zahl fällt, $x = 1$ dagegen einer Münze, die immer Kopf zeigt. Die Verteilungsfunktion für x ist folgendermaßen normiert:

$$\int_0^1 p(x|N, K)\, dx = 1$$

Die a-priori Wahrscheinlichkeit der Daten $p(N, K)$ hängt nicht von x ab, daher können wir sie in die Normierungskonstante hineinziehen. Jetzt benötigen wir nur noch die Funktion $p(x)$, d.h. die a-priori Wahrscheinlichkeit von x.

Allerdings sagt uns der Satz von Bayes nicht, wie wir diese a-priori Wahrscheinlichkeit finden können. Die Funktion $p(x)$ soll unseren Wissensstand beschreiben, bevor wir das Experiment gemacht haben. Sie ist quasi unser Vor-Urteil hinsichtlich der Verteilung von x. Das hat der Bayesischen Methode den Vorwurf eingetragen, sie sei in gewisser Weise „subjektiv“. Denn je nach der Wahl dieses Vorurteils erhält man eine andere Deutung der Daten. Allerdings kommen mit dem Verfahren nach Bayes alle, die vom selben Vorwissen ausgehen, auch zum selben Schluss. Insofern ist das Verfahren durchaus „objektiv“. Außerdem entspricht es durchaus unserer Erfahrung, dass verschiedene Personen aufgrund unterschiedlicher Grundkenntnisse dieselben Daten unterschiedlich deuten. Im Folgenden werden wir auch erkennen, dass mit zunehmender Menge an Daten alle Annahmen

über $p(x)$ zum selben Endergebnis konvergieren – es sei denn, die Wahl von $p(x)$ schließt dieses Endergebnis ausdrücklich aus.

Im vorliegenden Fall können wir uns mit der Annahme „maximaler Ignoranz“ auf der sicheren Seite fühlen, d.h., wir nehmen als a-priori Verteilung eine Gleichverteilung $p(x) = 1$ im Intervall [0,1] an. Wenn wir dies in die Definitionsgleichung einsetzen und die Normierung berechnen, erhalten wir als Endergebnis

$$p(x|N,K) = (N+1)\binom{N}{K} x^K (1-x)^{N-K}$$

Abbildung 2 zeigt die Verteilungsfunktionen, die auf diese Weise für verschiedene Wertepaare (N, K) erhalten werden. Bevor die Münze überhaupt geworfen wurde, also für $(N, K) = (0, 0)$, ist jeder Wert von x gleich wahrscheinlich (schwarze Kurve). Nach dem ersten Wurf mit dem Ergebnis „Kopf“ ist unser Kenntnisstand durch die rote Kurve gegeben. Der wahrscheinlichste Wert für x ist $x = 1$, wir können aber mit Sicherheit ausschließen, dass die Münze immer auf Zahl fällt ($p(0) = 0$). Nach 5 Würfen mit zweimal Kopf, oder 10 Würfen mit 6 mal Kopf, erhalten wir die grüne bzw. die blaue Kurve. Man erkennt, dass mit steigender Zahl an Münzwürfen die Verteilung schmaler wird. Ferner kann man das Maximum der Verteilung analytisch leicht bestimmen:

$$x_{max} = \frac{K}{N}$$

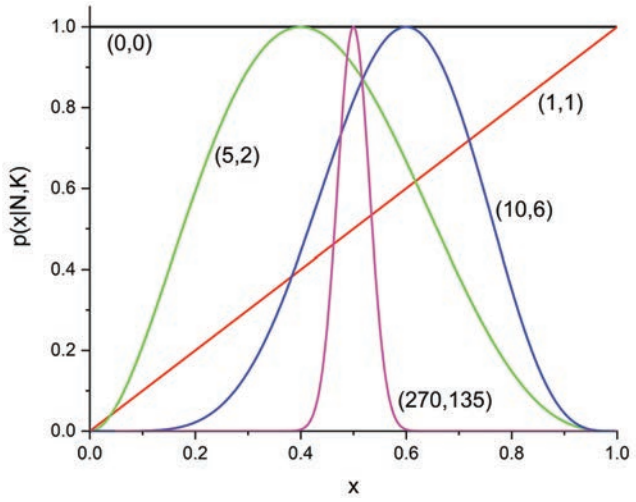

Abb. 2: Verteilungsfunktionen für die Fairness einer Münze, $p(x|N, K)$, nachdem die Münze N-mal geworfen und K-mal Kopf erhalten wurde. Alle Kurven wurden auf denselben Maximalwert skaliert.

Das entspricht unserer intuitiven Erwartung: Die beste Schätzung für die Wahrscheinlichkeit, dass die Münze auf Kopf fällt, ist der Anteil an Kopfwürfen, den wir insgesamt beobachtet haben. Der Satz von Bayes liefert uns nicht nur eine exakte wahrscheinlichkeitstheoretische Begründung für diese Schätzung, er informiert uns auch über die Unsicherheit in dieser Schätzung. Diese ist gegeben durch das Intervall, das einen bestimmten Anteil der Gesamtwahrscheinlichkeit enthält. Wollen wir also z.B. zu 90% sicher gehen, dann finden wir nach 6 Kopfwürfen bei 10 Versuchen, dass x im Intervall [0.35, 0.80] liege sollte, mit Maximum bei $x_m = 0.6$. Der Wert $x = 0.5$ liegt noch in diesem Intervall, die Münze könnte also doch fair sein. Um die Intervallbreite auf $\Delta x = 0.1$ zu reduzieren, braucht man mindestens $N = 270$ Münzwürfe (magenta Kurve).

Zum Schluss stellen wir noch die Frage, wie sich unser Kenntnisstand durch einen weiteren Münzwurf ändert, nachdem wir bereits N mal geworfen und K mal Kopf erhalten haben. Je nachdem, ob die Münze beim nächsten Wurf Kopf anzeigt oder nicht, ist unser neuer Kenntnisstand gegeben durch

$$p(x|N+1,K+1) = \frac{N+2}{K+1} \cdot x \cdot p(x|N,K)$$

Oder

$$p(x|N+1,K) = \frac{N+2}{N+1-K} \cdot (1-x) \cdot p(x|N,K)$$

Die Likelihood für den einzelnen Münzwurf ist x für Kopf bzw. $1 – x$ für Zahl. Vergleichen wir das mit der Formel von Bayes so erkennen wir, dass – abgesehen von der Normierungskonstante – die a-priori Wahrscheinlichkeit durch $p(x|N, K)$ ersetzt wurde. Das ist auch unmittelbar einsichtig: Nachdem wir die Münze bereits N mal geworfen und dabei K mal Kopf erhalten haben, sind wir nicht mehr völlig ignorant! Vielmehr wird unsere aktuelle Kenntnis, vor dem erneuten Münzwurf, genau durch $p(x|N, K)$ beschrieben.

Statt eines Münzwurfes können wir auch ein anderes Ereignis nehmen, bei dem nur zwei Resultate möglich sind, von denen das erste mit der Wahrscheinlichkeit x eintrifft. So würde z.B. die blaue Kurve in Abb. 2 die Glaubwürdigkeit eines Politikers repräsentieren, der in 10 Aussagen 6 mal die Wahrheit gesagt und 4 mal gelogen hat. In den Naturwissenschaften ist es zwar nicht möglich, wie in der Mathematik einen Beweis durch vollständige Induktion zu führen. Das gerade skizzierte Verfahren erlaubt uns aber, die verbleibende Ungewissheit nach N erfolgreichen Tests einer These zu quantifizieren. Abbildung 3 zeigt die entsprechenden Verteilungen.

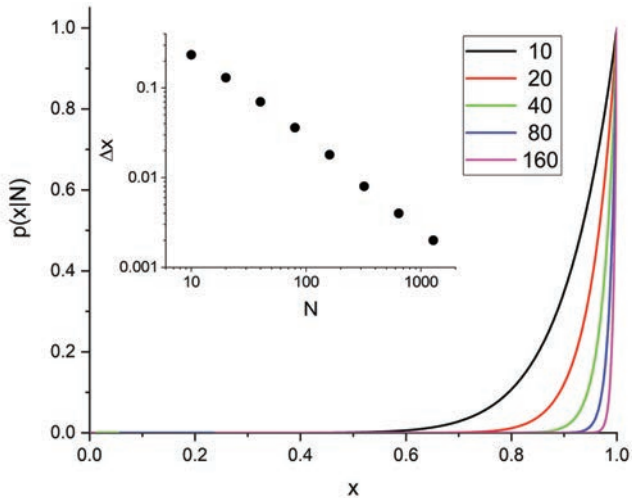

Abb. 3: Verteilungen $p(x|N)$ für die Allgemeingültigkeit einer These, nachdem sie in N Versuchen bestätigt wurde. Der Inset zeigt wie die Breite Δx, die 90% der Wahrscheinlichkeit enthält, mit steigender Zahl erfolgreicher Versuche abnimmt.

Mit zunehmender Zahl erfolgreicher Bestätigungen rückt die Kurve immer näher an den Wert $x = 1$ (d.h. „Die These ist wahr“) heran, die Unsicherheit (d.h. $p(x < 1)$) bleibt aber immer endlich. Das Intervall, in dem sich 90 % der Wahrscheinlichkeit befinden, schrumpft umgekehrt proportional zur Zahl der erfolgreichen Versuche, wie der Inset in Abb. 3 zeigt. Aber ein einziger Fehlversuch führt dazu, dass die Kurve bei $x = 1$ auf Null fällt: Die These ist damit widerlegt. Das Verfahren nach Bayes führt also zum selben Schluss über die Falsifizierbarkeit einer Theorie wie die Wissenschaftstheorie von Popper. Allerdings zeigt es, im Gegensatz zu Popper, wie sich die Glaubwürdigkeit einer Theorie durch ihre Bestätigung im Experiment verbessert.

WANN UND WIE DARF MAN FITTEN?

Bei der Analyse von Daten wird sehr oft eine Modellfunktion durch Variation ihrer Parameter an die Daten angepasst (gefittet). Die Übereinstimmung zwischen den Daten und der Modellfunktion wird durch ein Fitkriterium gemessen, in der Regel die Summe der Fehlerquadrate. Als optimale Parameter werden die Werte betrachtet, für die dieses Fitkriterium minimal wird. Wie sieht dieses Problem nun aus bayesischer Sicht aus?

Betrachten wir als Beispiel Temperaturwerte $\Theta = \{\Theta_1, \Theta_2, \ldots \Theta_N\}$ die an einem Thermometer z.B. von verschiedenen Personen abgelesen werden. Für die Anwendung des Satzes von Bayes benötigen wir nun zweierlei: Erstens eine Theorie für die Likelihood, d.h. eine Formel, die angibt, wie wahrscheinlich diese Messwerte sind, wenn die wahre Temperatur den Wert T hat. Und zweitens eine a-prior Verteilung für diese Temperatur. Ein häufig gewähltes Modell für die Likelihood ist die Gaussverteilung. Für eine Einzelmessung gilt dann

$$p(\Theta|T) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(T-\Theta)^2}{2\sigma^2}\right) \qquad (7)$$

Wenn die einzelnen Messungen statistisch unabhängig sind, ist die Likelihood für den ganzen Datensatz gegeben durch das Produkt der Likelihood-Funktionen für die Einzelmessungen:

$$p(\Theta_1, \Theta_2, \ldots, \Theta_N\,|T) = A \exp\left(-\frac{1}{2}\chi^2\right) \quad ; \; \chi^2 = \sum_{j=1}^{N} \left(\frac{T-\Theta_j}{\sigma}\right)^2$$

Wobei A eine Normierungskonstante ist, die nicht von T abhängt. Der Satz von Bayes nimmt dann die Form

$$p(T|\boldsymbol{\Theta}) = B \exp\left(-\frac{1}{2}\chi^2\right) p(T)$$

an, wobei in der Konstanten B die Normierung und die a-priori Wahrscheinlichkeit der Daten subsumiert sind. Die Wahl von $p(T)$ ist nun nicht mehr so offensichtlich wie im Fall des Münzwurfes. Wenn wir überhaupt keine Vorstellung von der Größenordnung der Temperatur haben, außer dass sie positiv sein muss, dann wäre z.B. $p(T) \sim 1/T$ der theoretisch beste Ausdruck für völlige Ignoranz (d.h. eine Gleichverteilung über einer logarithmischen Temperaturskala). In der Regel haben wir aber eine grobe Vorstellung davon, in welchem Intervall die Temperatur liegen sollte. Nehmen wir der Einfachheit halber an, dass wir ein hinreichend großes Intervall definieren können, und setzen in diesem Intervall eine konstante a-priori Wahrscheinlichkeit an. Abb. 4 zeigt dann, wie sich die Inferenz mit jedem weiteren Messwert entwickelt.

Nach der ersten Messung ($\theta = 220$) ist die Inferenz identisch mit der Likelihood (rote Kurve) [7]. Die zweite Messung ($\theta = 225$) verschiebt das Maximum nach 222.5, und die Verteilung wird schmaler (grüne Kurve). Die dritte Messung ($\theta = 190$) verschiebt das Maximum wieder zu kleineren Temperaturen, und die Verteilung wird weiter schmaler (blaue Kurve). Wir können leicht analytisch das Maximum und die Breite der Verteilung nach einer beliebigen Zahl von Messungen berechnen. Da $p(T|\boldsymbol{\theta})$ positiv ist, können wir statt des Maximums von $p(T|\boldsymbol{\theta})$ auch das Minimum des negativen Logarithmus analysieren. Es gilt

$$L = -\ln p(T|\boldsymbol{\Theta}) = \frac{1}{2}\chi^2$$

$$\frac{dL}{dT} = 0 \quad \rightarrow \quad T_{min} = \frac{1}{N}\sum_{j=1}^{N} \Theta_j$$

Maximierung der Inferenz ist also äquivalent zur Minimierung der Fehlerquadrate, und der optimale Schätzwert für T ist der Mittelwert der Messwerte. Dieses Ergebnis ist unabhängig von der Annahme über die Standardabweichung.

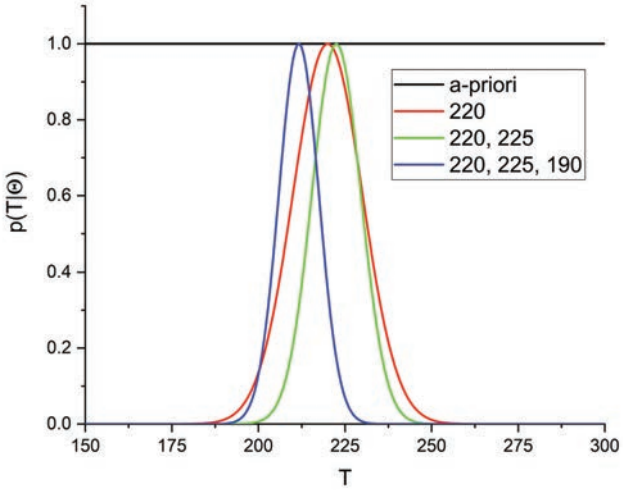

Abb. 4: Inferenz für die Wahrscheinlichkeitsverteilung der Temperatur vor der Messung (schwarz) und nach einer (rot), zwei (grün), und drei (blau) Messungen. Für die Likelihood wurde eine Standardabweichung von σ = 10 angenommen.

Dieses Verfahren lässt sich für den Fall verallgemeinern, dass die Messwerte $\theta(\mathbf{x})$ als Funktion von mehreren experimentell zugänglichen Variablen $\boldsymbol{x} = (x_1, x_2, \ldots, x_V)$ bestimmt werden, und die Theorie die Messung zu diesen Variablen und einer gewissen Zahl von Parametern $\boldsymbol{a} = (a_1, a_2, \ldots, a_P)$ in Beziehung setzt. Der Likelihood-Parameter hat dann die Form

$$L_{Gauss} = \frac{1}{2}\chi^2 = \frac{1}{2}\sum_{j=1}^{N}\left(\frac{\Theta(\mathbf{x}^{(j)}) - \mathrm{T}(\boldsymbol{a}; \mathbf{x}^{(j)})}{\sigma_j}\right)^2$$

Hierbei läuft die Summe über alle Variablen-Tupel $\mathbf{x}^{(j)}$ *für* die eine Messung gemacht wurde, und jede dieser Messungen kann eine eigene Standardabweichung haben. Die optimalen Parameter $\boldsymbol{a}$ werden dann durch Minimieren von χ^2 bestimmt. In der Bayesischen Sichtweise würde man eine P-dimensionale Verteilungsfunktion $p(\boldsymbol{a}|\boldsymbol{\theta})$ bestimmen. Deren Maximum stimmt mit dem durch minimieren der Fehlerquadrate bestimmten Parametern unter zwei Bedingungen überein:

i) Der Fehler ist tatsächlich durch eine Gaussverteilung korrekt beschrieben,
ii) Die Zahl der Parameter in $\boldsymbol{a}$ ist hinreichend klein verglichen mit der Zahl der Datenwerte, so dass jede einigermaßen vernünftige a-priori-Verteilung von der Likelihood überstimmt wird.

Annahme (i) ist z.B. nicht erfüllt, wenn es sich bei den Daten um Zählereignisse handelt. Situationen, in denen die Zahl der Parameter ähnlich groß oder sogar größer als die Zahl der Datenpunkte ist, treten z.B. bei sogenannten inversen Problemen auf. In diesen Fällen wird die Wahl der a-priori Verteilung $p(\boldsymbol{a})$ entscheidend.

ZÄHLEN VON EREIGNISSEN

Manche Experimente liefern als Ergebnis ganze positive Zahlen, etwa die Zahl an Photonen, die in einem bestimmten Zeitraum auf einen Detektor treffen, oder die Zahl der Autos, die in einer Viertelstunde durch eine bestimmte Straße fahren. Die Gaussverteilung beschreibt die Verteilung einer kontinuierlichen Veränderlichen, die auch negative Werte annehmen kann. Sie ist daher als Likelihood für die Beschreibung von Zählereignissen nicht geeignet.

Wenn die Zählereignisse nicht miteinander korreliert sind, dann ist eine Poissonverteilung oft ein gutes theoretisches Modell. Die Likelihood für eine einzelne Messung, die den Zählwert n liefert, ist gegeben durch

$$p(n|\lambda) = \frac{\lambda^n}{n!}\, e^{-\lambda} \qquad (8)$$

Wir fragen nun, wie die Verteilung der Glaubwürdigkeit für den Theorieparameter λ aussieht, nachdem wir eine Sequenz von M Messungen mit den Ergebnissen $\{n_1, n_2, \ldots, n_M\}$ durchgeführt haben. Der Satz von Bayes liefert die Antwort:

$$p(\lambda|n_1, n_2, \ldots, n_M) = \prod_{j=1}^{M}\left(\frac{\lambda^{n_j}}{n_j!}\, e^{-\lambda}\right) \frac{p(\lambda)}{p(n_1, n_2, \ldots, n_M)}$$

Fassen wir alle Terme, die nicht von λ abhängen, zu einer Normierungskonstante A zusammen, dann kann man vereinfachen

$$p(\lambda|n_1, n_2, \ldots, n_M) = A\, \lambda^Z\, e^{-M\lambda}\; p(\lambda)$$

$$Z = \sum_{j=1}^{M} n_j$$

Die Verteilung hängt also nur davon ab, wie viele Ereignisse in wie vielen Versuchen insgesamt gezählt wurden, nicht aber davon, wie sich die Einzelereignisse auf die einzelnen Versuche verteilen. Abb. 5 zeigt einige Beispiele für den Fall, dass wir unsere ursprüngliche Ignoranz bezüglich λ durch einen konstanten Wert für die a-priori Verteilung $p(\lambda)$ ausdrücken. Hat man nur einen Versuch unternommen und dabei kein Ereignis gezählt, liegt der Wert von λ mit 90 % Wahrscheinlichkeit im Intervall [0.0, 3.0] (schwarze Kurve). Wiederholt man das Experiment fünfmal mit demselben Ergebnis, schrumpft das Intervall auf [0.0, 0.57] (schwarze gestrichelte Kurve). Zählt man in einem Versuch 5 Ereignisse, dann liegt λ im Intervall [2.6, 10.4] (rote Kurve) mit Maximum bei $\lambda = 5$. Zählt man 25 Ereignisse in 5 Experimenten, dann liegt das Maximum weiterhin bei $\lambda = 5$, aber das Intervall wird kleiner auf [3.6, 7.0] (blaue Kurve). Zählt man dagegen die 5 Ereignisse in zwei Experimenten, halbiert sich der Wert für λ am Maximum (grüne Kurve).

Der optimale Wert für λ lässt sich wieder leicht analytisch bestimmen. Man findet die Nullstelle der Ableitung des Logarithmus der Inferenz nach λ bei

$$\lambda_{max} = \frac{Z}{M}$$

Der optimale Wert ist also, wie bei der Gaussverteilung, der Mittelwert.

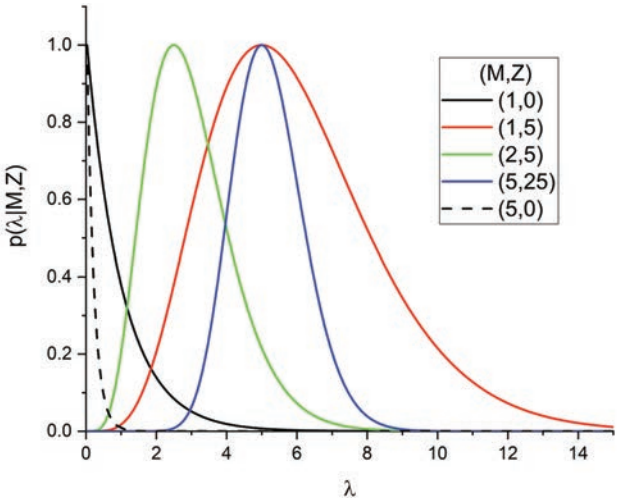

Abb. 5: Glaubwürdigkeitsverteilung (Inferenz) für den Parameter λ einer Poissonverteilung nachdem man in M Versuchen insgesamt Z Ereignisse gezählt hat.

Wie bei der Gaussverteilung lässt sich das Verfahren auch für die Poissonverteilung verallgemeinern, wenn die Messwerte $\mathrm{n}(\mathbf{x})$ als Funktion von mehreren experimentell zugänglichen Variablen $\boldsymbol{x} = (x_1, x_2, \ldots, x_V)$ bestimmt werden, und die Theorie die Messung zu diesen Variablen und einer gewissen Zahl von Parametern $\boldsymbol{a} = (a_1, a_2, \ldots, a_P)$ in Beziehung setzt. Der Likelihood-Parameter, d.h. der negative Logarithmus der Likelihood, hat dann die Form

$$L_{Poisson} = \sum_{j=1}^{N}\left(\lambda_j - n_j \ln(\lambda_j) + \ln(n_j!)\right)$$

Mit

$$\lambda_j = \lambda\left(\boldsymbol{a}; \mathbf{x}^{(j)}\right) \quad ; \quad n_j = \mathrm{n}\left(\mathbf{x}^{(j)}\right)$$

Hierbei läuft die Summe wieder über alle Variablen-Tupel $\mathbf{x}^{(j)}$ *für* welche eine Messung gemacht wurde. Anstelle von χ^2 muss nun L_P durch Variation der Parameter $\boldsymbol{a}$ minimiert werden. Eine solche Situation tritt z.B. bei der Analyse der Daten auf, die bei einem TCSPC-Experiment erhalten werden. TCSPC steht für „time correlated single photon counting" und bezeichnet ein Verfahren, mit dem man Fluoreszenzabklingzeiten bestimmt. Dazu wird eine Probe mit einem kurzen Lichtpuls angeregt und die Zeit t_j bis zum Erscheinen des ersten Fluoreszenzphotons gemessen. Die Zeitskala wird in viele äquidistante Bins eingeteilt und die Zahl mit der Adresse des Bins, in das t_j fällt, wird um eins erhöht. So erhält man ein Histogramm der Verzögerungszeiten der Fluoreszenzphotonen, welches der Fluoreszenzabklingkurve entspricht. Vermuten wir, dass es sich um einen biexponentiellen Zerfall handelt, dann würden wir das durch

$$\lambda_j = A_1 \exp\left(-\frac{t_j}{\tau_1}\right) + A_2 \exp\left(-\frac{t_j}{\tau_2}\right)$$

modellieren. D.h., $\mathbf{x}^{(j)} = t_j$ besteht jeweils aus einem einzelnen Zeitwert, und der Parametersatz $\boldsymbol{a} = \{A_1, A_2, \tau_1, \tau_2\}$ besteht aus den beiden Amplituden und Abklingzeiten. Die Bayesische Betrachtung zeigt, dass es in diesem Fall nicht korrekt wäre, die Anpassung durch Minimieren der Fehlerquadrate durchzuführen. Für große Zählraten erhält man zwar praktisch dasselbe Ergebnis, wenn man gleichzeitig in der Gaussverteilung die Standardabweichung $\sigma_j = \sqrt{n_j}$ setzt. Für kleine Zählraten, also zum Ende der Zerfallskurve, macht man damit aber einen Fehler. Für die Möglichkeiten, die Daten durch einen Fit mit der Poisson-Likelihood zu analysieren, gilt analog zum Fit durch Minimieren der Fehlerquadrate: Die Possionverteilung muss die korrekte Wahrscheinlichkeitsverteilung für das Problem sein, und die Zahl der Daten muss viel größer als die Zahl der zu bestimmenden Parameter sein.

WIE VIELE PHOTONEN BRAUCHT MAN FÜR EINE ABKLINGZEIT?

Wir wollen uns das Problem der Bestimmung einer Fluoreszenzabklingzeit noch einmal genauer anschauen. Eine Anwendung zum Studium der Dynamik von Proteinen besteht darin, an zwei Stellen im Protein zwei Farbstoffe anzubinden, von denen der eine als Donor und der andere als Akzeptor in einem FRET-Prozess fungiert. FRET steht hier für „Förster Resonance Energy Transfer". Kommen sich Donor und Akzeptor hinreichend nahe, dann kann der Donor seine Anregungsenergie auf den Akzeptor übertragen. Dadurch verkleinert sich sowohl die Quantenausbeute als auch die Abklingzeit der Donor-Fluoreszenz. Die Effizient dieses Energietransfers nimmt mit der sechsten Potenz des Abstandes ab. Hat das Protein zwei bevorzugte Konformationen, die sich im Abstand unterscheiden, so zeigt die Abklingzeit der Donorfluoreszenz an, in welcher der beiden Konformationen sich das Protein gerade befindet. Durch wiederholte schnelle Messung der Abklingzeit erhält man daher Information über die Dynamik des Proteins.

Eine schnelle Messung der Abklingzeit der Fluoreszenz ist möglich, wenn man ein Ensemble von Molekülen mit einem kurzen Lichtpuls anregt und dann die Intensität des Fluoreszenzlichtes in Realzeit mit einem schnellen Detektor misst. Die Abklingkurve kann dann mit dem Modell eines exponentiellen Zerfalls gefittet werden. Da man bei der Messung aber über das gesamte Ensemble mittelt, erhält man so keine Information über die Proteindynamik.

Man muss daher ein einzelnes Proteinmolekül beobachten, das aber nach jeder Anregung nur höchstens ein einzelnes Fluoreszenzphoton abgibt. Mit der oben beschriebenen Methode des TCSPC könnte man eine solche Einzelmolekülmessung machen und würde die Abklingzeit durch Maximieren der Poissonschen Likelihood fitten, wie oben beschrieben. Dieses Verfahren hat aber den Nachteil einer langen Beobachtungszeit. Wählt man das Zeitfenster für die Erstellung des Histogramms etwa doppelt so groß wie die erwartete Lebensdauer, und teilt das Fenster in 100 Bins ein, so entspricht ein Zeitschritt etwa 2 % der Lebensdauer. Zählt man so lange, bis der erste Kanal 1000 Photonen enthält, dann würden insgesamt 50000 Photonen emittiert, von denen ca. 43000 gezählt wurden. Da typischerweise auf 100 Anregungen ein Zählereignis kommt, und die Probe typischerweise mit einer Rate von 80 MHz angeregt wird, benötigt diese Messung 62.5 ms. Änderungen der Proteinkonformation, die schneller erfolgen, kann man so also nicht erfassen.

Die Bayesische Methode zeigt uns aber, wie es viel schneller und effizienter gehen kann. Wir benötigen zunächst ein Modell für die Likelihood, dass ein Molekül mit Lebensdauer τ ein Photon zur Zeit t aussendet. Das Modell des unimolekularen Zerfalls liefert hierfür

$$p(t|\tau) = \frac{1}{\tau}\exp\left(-\frac{t}{\tau}\right) \tag{9}$$

Für eine Sequenz von Messungen $\{t_1, t_2, \ldots, t_N\}$ gilt dann

$$p(t_1, t_2, \ldots, t_N|\tau) = \prod_{j=1}^{N}\left(\frac{1}{\tau}\exp\left(-\frac{t_j}{\tau}\right)\right) = \frac{1}{\tau^N}\exp\left(-\frac{T}{\tau}\right)$$

Mit

$$T = \sum_{j=1}^{N} t_j$$

Die Likelihood hängt also nur von der Zahl der beobachteten Photonen und der Summe ihrer Ankunftszeiten ab. Nun ist zwar $p(t|\tau)$ als Funktion von t auf eins normiert, das Integral über die Variable τ aber divergiert. Bei der Anwendung des Satzes von Bayes,

$$p(\tau|t_1, t_2, \ldots, t_N) = A\, p(t_1, t_2, \ldots, t_N|\tau)\, p(\tau)$$

Können wir daher nicht wie bisher $p(\tau) = 1$ setzen, da die so erhaltene Inferenz (zumindest für das erste Photon) nicht normierbar ist. Eine pragmatische Lösung wäre, $p(\tau)$ für τ oberhalb einer maximalen Lebensdauer auf Null zu setzen. Das

erfordert ein gewisses Vorwissen über die Eigenschaften des verwendeten Farbstoffs, das aber in der Regel vorhanden ist. Spaßeshalber versuchen wir es aber mal mit der Annahme vollständiger Ignoranz, welche durch $p(\tau) = 1/\tau$ gegeben ist (was streng genommen nur normierbar ist, wenn man $\tau = 0$ ausschließt). Die erhaltene Inferenz ist jetzt schon für das erste Photon normierbar. Für die Likelihood-Funktion erhalten wir so

$$L_E = -\ln p(\tau|N,T) = (N+1)\ln\tau - \frac{T}{\tau}$$

Diese hat ihr Minimum bei

$$\tau_{min} = \frac{T}{N+1} = \langle t\rangle\left(1 - \frac{1}{N+1}\right)$$

Die optimale Schätzung für die Lebensdauer geht also mit steigender Zahl beobachteter Photonen gegen den Mittelwert $\langle t\rangle$ der Ankunftszeiten.

Abb. 6 zeigt die a-posteriori Verteilung der Lebensdauer für das erste Photon (blaue Kurve), nach 10 Photonen (rote Kurve) und nach 100 Photonen (schwarze Kurve), für eine mittlere Ankunftszeit von 1 (durchgezogene Kurve) bzw. 2 (gestrichelte Kurven). Mit 100 Photonen sind beide Kurven bereits so schmal, dass sich die jeweiligen 90 %-Intervalle nicht mehr überlappen. Damit ist eine Entscheidung über die aktuelle Konformation des Proteins 500 mal schneller möglich als mit der Auswertung eines TCSPC Histogramms.

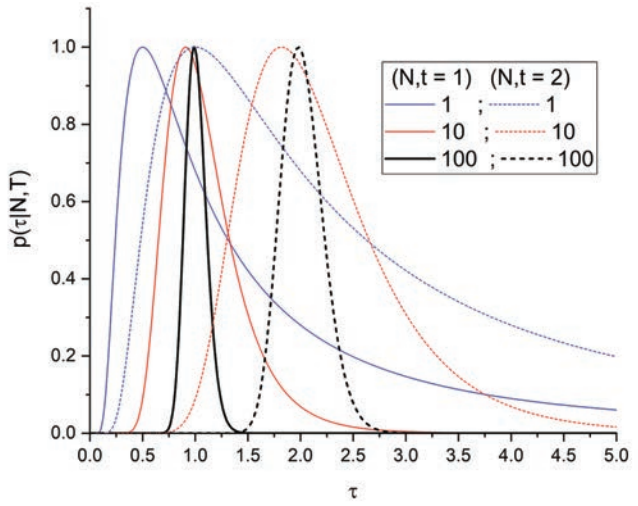

Abb. 6: Wahrscheinlichkeitsverteilung der Fluoreszenzlebensdauer nach der Messung von N = 1 (blau), 10 (rot) und 100 (schwarz) Photonen. Die mittlere Ankunftszeit der Photonen ist t = 1 für die durchgezogenen Kurven und t = 2 für die gestrichelten Kurven.

FRET-EFFIZIENZ EINES FLUKTUIERENDEN MOLEKÜLS

Zum Schluss soll noch ein Beispiel aus der Literatur [8] vorgestellt werden, das mehrere der oben beschriebenen Konzepte vereinigt. Statt die Lebensdauer des Donors in einem FRET Prozess zu messen, kann man einfacher die Photonen zählen, die in den beiden Emissionskanälen des Donors und des Akzeptors in einem bestimmten Zeitintervall auftreten. Im folgenden Modell wird vorausgesetzt, dass nur der Donor angeregt wird, und das Fluoreszenzphoton dann entweder aus dem Donor, oder nach FRET aus dem Akzeptor kommt. Aus diesen beiden Zahlen könnte man eine FRET Effizienz nach

$$E_n = \frac{n_A}{n_A + n_D}$$

berechnen. Korrekter ist es aber, statt mit den Photonenzahlen eines einzelnen Experimentes die FRET Effizienz über die entsprechenden Erwartungswerte zu definieren:

$$E_\mu = \frac{\mu_A}{\mu_A + \mu_D}$$

Die Wahrscheinlichkeit, die Photonenzahlen (n_A, n_D) zu beobachten wenn die Erwartungswerte (μ_A, μ_D) sind, ist die Likelihood $p(n_A, n_D \mid \mu_A, \mu_D)$. Ein Modell für diese Likelihood wäre z.B. das Produkt von zwei Poissonverteilungen. In [8] wird als Alternative auch eine andere Modellfunktion getestet, welche die Tatsache widerspiegelt, dass die Emissionsprozesse antikorreliert sind: Wenn der Donor emittiert, wird es der Akzeptor nicht tun, und umgekehrt.

Gesucht ist die Inferenz, d.h. $p(\mu_A, \mu_D \mid n_A(t), n_D(t))$, die zusätzlich noch von der Zeit abhängen soll. Hier hat t die Bedeutung eines Index, welcher die konsekutiven Messungen nummeriert, die jeweils die Photonen in einem festgelegten Zeitintervall zählen. Die Formel von Bayes liefert dann für die Inferenz nach der aktuellen Messung:

$$p\left(\mu_A, \mu_D \,|n_A(t), n_D(t)\right) =$$

$$p(\mu_A, \mu_D \,|n_A(t-1), n_D(t-1)) \times p(n_{A(t)}, n_D(t) \,|\mu_A, \mu_D) \quad (10)$$

Der erste Term auf der rechten Seite ist die Inferenz, die sich bis zum Zeitpunkt vor der aktuellen Messung akkumuliert hat. Der zweite Term ist die Likelihood für die aktuelle Messung.

Das Ergebnis ist eine zweidimensionale Verteilung in den Variablen μ_A und μ_D, die numerisch als Matrix dargestellt werden kann. Die Verteilung der FRET-Effizienz erhält man dann in jeder Iteration durch Marginalisierung:

$$P(E,t) =$$

$$\iint d\mu_A\, d\mu_D \;\; p\left(\mu_A, \mu_D \,|n_A(t), n_D(t)\right) \times \delta\left(E - \frac{\mu_A}{\mu_A + \mu_D}\right)$$

Solange das Protein in einer Konformation verharrt, wird die Kurve schärfer und zentriert ihr Maximum bei der FRET-Effizienz dieser Konformation. Ändert sich die Konformation, so folgt auch die Kurve für die FRET-Effizienz dieser Änderung. Allerdings benötigt man mehr Photonen aus der neuen Konformation, um die Information aus der ersten Konformation zu überstimmen. Damit der Algorithmus schneller auf eine Konformationsänderung reagieren kann, wenden die Autoren einen Trick an: Sie addieren in jeder Iteration einen Untergrund zur Inferenz (Gleichung 10) hinzu, deren Amplitude von einem Zufallszahlengenerator bestimmt wird. So wird in jeder Iteration etwas „Ignoranz“ in die Inferenz eingeführt, wodurch der Algorithmus „vergesslich“ gemacht wird.

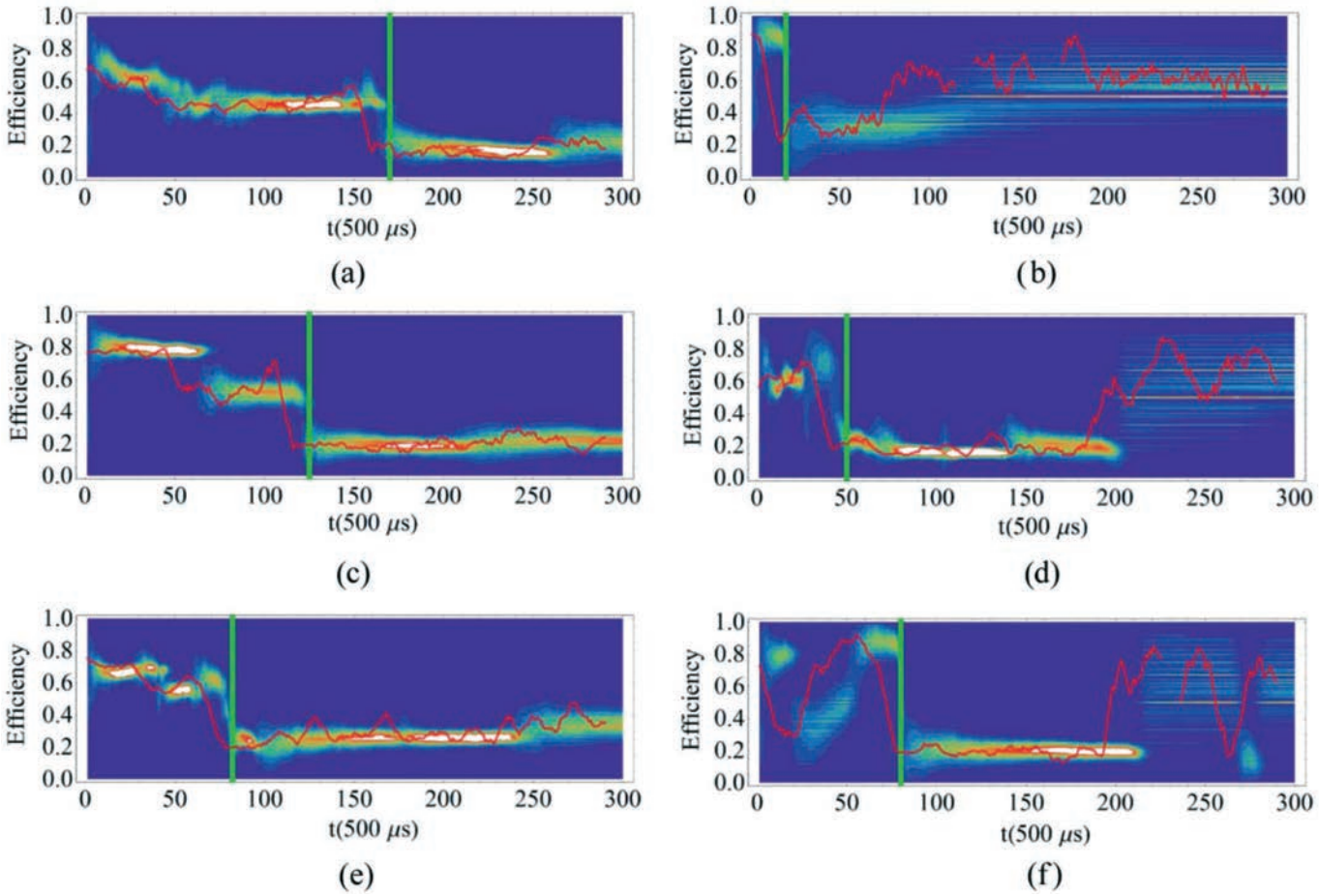

Abb. 7: Beispiele für die zeitabhängige Verteilung der FRET-Effizienz als Funktion der Zeitschritte von jeweils 500 µs Dauer. Die Farbskala geht von blau (Null) über grün, rot bis weiß. Die rote Kurve ist ein gleitender Mittelwert über 10 Zeitschritte. Reproduziert aus [8].

Abb. 7 zeigt Ergebnisse für $P(E, t)$ für mehrere Messungen an dem Protein Calmodulin, das mit zwei FRET Farbstoffen markiert ist. Die Farbcodierung geht von blau über grün, rot bis weiß für steigende Werte von $P(E, t)$. Die rote Kurve zeigt das Ergebnis einer gleitenden Mittelwertbildung über 10 Zeitschritte von jeweils 500 µs an. Letztere fluktuiert deutlich stärker als das Maximum der Bayesischen Verteilung, insbesondere dann, wenn der Donor-Farbstoff ausbleicht.

ZUSAMMENFASSUNG UND AUSBLICK

Wir haben zunächst den Satz von Bayes vorgestellt, den Thomas Bayes vor über 250 Jahren in seiner posthum veröffentlichten Arbeit bewiesen hatte. Dieser Artikel hat die Grundlagen für die exakte Theorie der Wahrscheinlichkeiten gelegt. Vor etwa 70 Jahren hat Richard Cox gezeigt, dass man ein Maß für die Glaubwürdigkeit von Aussagen definieren kann, für das dieselben Rechenregeln gelten wie für Wahrscheinlichkeiten. Durch die Kombination beider Methoden erhält man ein Instrument, um aus experimentellen Daten die optimalen induktiven Schlüsse ziehen zu können. Diese Schlüsse führen nie zu einem exakten Beweis wie etwa in der Mathematik, sie machen aber eine quantitative Aussage darüber, wie wahrscheinlich (d.h. glaubwürdig) die entsprechende Aussage ist. In den vorgestellten Beispielen war die gesuchte Modellgröße jeweils ein einzelner Parameter, und die induktive Methode führte nicht zu einem einzigen Wert für diesen Parameter, sondern zu einer Wahrscheinlichkeitsverteilung.

Die Anwendung dieses Verfahrens setzt voraus, dass man für jeden Fall eine Hypothese für die entsprechende Likelihood in der Form eines mathematischen Zusammenhangs zwischen einem Theorieparameter und der experimentell zugänglichen Größe aufstellen kann. In unseren Beispielen haben wir vier verschiedene solcher Modelle vorgestellt: Die Binomialverteilung (6), die Gaussverteilung (7), die Poissonverteilung (8), und die Exponentialverteilung (9). Allen Beispielen war gemeinsam, dass wir die Entwicklung der Wahrscheinlichkeit mit steigender Zahl von Experimenten zum selben Modellparameter betrachten konnten. In diesem Fall wird die im Satz von Bayes benötigte a-priori-Verteilung der theoretischen Modellgröße zunehmend irrelevant. Als Folge davon kann die Bestimmung des optimalen Wertes des Modellparameters, d.h. des Wertes, für den die Wahrscheinlichkeitsverteilung maximal wird, auf ein Fitproblem zurückgeführt werden. In der Regel maximiert man nicht die Wahrscheinlichkeit selbst, sondern minimiert deren negativen Logarithmus. Dies ist äquivalent zu der Annahme, dass das Ergebnis des konkreten Experimentes auch das theoretisch wahrscheinlichste Ergebnis des Experimentes ist.

Die hier vorgestellten Resultate hätte man vielleicht auch ohne Rückgriff auf den Satz von Bayes begründen können, da die a-priori Wahrscheinlichkeit des Modells letztlich nicht wesentlich in das Ergebnis eingeht. Damit wird Maximieren der Inferenz identisch zum Maximieren der Likelihood. Anders sieht die Sache aber aus, wenn z.B. die Zahl der Modellparameter ähnlich groß oder sogar größer wird als die Zahl der Datenpunkte. In diesem Fall gibt es eine sehr große Zahl von Modellen, welche im Rahmen der Likelihood kompatibel mit dem Experiment sind. Maximieren der Likelihood liefert dann nicht notwendig die glaubwürdigste Lösung. Vielmehr wird nun die Wahl der a-priori Verteilung für das Modell entscheidend. Um ein Beispiel zu skizzieren: Photographien des Himmels mit großen Teleskopen liefern Bilder, auf denen jeder Stern als Beugungsbild mit endlichem Durchmesser zu sehen ist. Theoretisch müsste das Bild bei der enormen Entfernung des Sterns punktförmig sein, aber die endlichen Dimensionen der optischen Elemente im Teleskop erlauben nur eine endliche Auflösung. Andererseits kann man die Verzerrung des Bildes bei Kenntnis des Teleskopaufbaus exakt berechnen. Wäre das ideale Punktmuster der Sternpositionen auf der Photographie bekannt, dann könnte man die Verzerrung durch das Teleskop exakt simulieren. Die Likelihood vergleicht diese Simulation mit den tatsächlichen Daten. Wollte man dieses Problem mit einem Fit lösen, stünde man vor dem Problem, dass es genauso viele Unbekannte (Pixel im idealen Bild) wie Datenpunkte (Pixel im experimentellen Bild) gibt. Hat man aber ein plausibles Modell für die a-priori Wahrscheinlichkeiten der idealen Bilder, dann kann man dieses Problem erstaunlich gut lösen und so z.B. die Frage klären, ob ein etwas „unrunder“ Fleck auf dem Photo vielleicht durch zwei benachbarte Sterne entstanden ist oder ob dies unwahrscheinlich ist.

Diese gerade beschriebene Situation ist das klassische Problem für die Anwendung der „Maximum Entropie“ Methode, wie sie von Skilling und Bryan in ihrem grundlegenden Artikel [9] beschrieben wurde. Diese und weitere Anwendungen der Bayesischen Methode der Datenanalyse wären vielleicht ein Thema für einen späteren Artikel in dieser Zeitschrift.

REFERENZEN

[1] Devinderjit Sivia und John Skilling, Data Analysis: A Bayesian Tutorial, Oxford University Press, 2. Rev. Auflage, ISBN: 978-0198568322.

[2] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, und Donald B. Rubin, Bayesian Data Analysis, 3rd Edition, CRC Press, ISBN: 978-1439840955.

[3] Allen B. Downey, Think Bayes: *Bayesian Statistics Made Simple*, O'Reilly and Associates, ISBN: 978-1449370787; Die PDF-Version gibt es kostenlos bei “Green Tea Press”, http://greenteapress.com/wp/think-bayes/.

[4] Thomas Bayes, *Phil. Trans.* **53**, 370-418 (1763). Einen Scan der Originalseiten findet man unter: doi:10.1098/rstl.1763.0053 2053-9215. Eine deutsche Übersetzung findet sich in dem Taschenbuch: Versuch zur Lösung Eines Problems der Wahrscheinlichkeitsrechnung (Classic Reprint), Forgotten books Verlag, ISBN 978-1334264023

[5] R. T. Cox, Probability, Frequency, and Reasonable Expectation, *Am. J. Phys.* **14** 1 – 13 (1946).

[6] Diese Gleichung, hier übersetzt in unsere Nomenklatur, steht in [5] ohne Nummer in dem Absatz vor Gleichung (14) auf Seite 8.

[7] Streng genommen sind beide Formeln nur identisch über den Bereich, über den die a-priori Wahrscheinlichkeit $p(T)$ den Wert 1 hat. So verschwindet die Inferenz $p(T|\theta)$ für negative Temperaturen, während die Likelihood $p(\theta|T)$ dort positiv ist. Wenn die Standardabweichung σ viel kleiner als θ ist, ist diese Differenz unerheblich. Andererseits verhindert eine korrekt gewählte a-priori Wahrscheinlichkeit physikalisch unsinnige Ergebnisse.

[8] M. Backowic, E. Shane Price, C. K. Johnson, J. P. Ralston, A distribution-based method to resolve singe molecule Förster resonance energy transfer observations, *J. Chem. Phys.* **134**, 145101 (2011).

[9] J. Skilling, R. K. Bryan, Maximum-Entropy Image-Reconstruction - General Algorithm; *Mon. Not. R. Astron. Soc.*, **211**, 111–124 (1984).