教学交流

概率论与随机过程教程

何毓琦
美国哈佛大学工程与应用科学学院

【编者按】作者何毓琦先生是著名的控制科学家，美国工程院院士，中国科学院和中国工程院外籍院士。2013 年他在中国科学网上发表了五篇关于《概率论与随机过程教程》（Probability and Stochastic Process Tutorial）的精彩博文。经何先生同意，本刊有幸得以转载。为便于广大中文读者阅读，我们特别请中国科学院数学与系统科学研究院的万林副研究员把这五篇博文翻译成了中文。本文的五节分别对应于这五篇博文。

概率论，通常被认为是“一种用以探究事物的未知性或不确定性的精确方法”。“（某件事）发生的机率有多少？”每个人对这样的问题都有一些直观的认识。随机过程，则是研究以时间（或者其它独立的指标变量，如距离）为指标的概率问题。在概率论与随机过程领域，已有大量优秀的经典教科书。这是我喜欢的口试题目之一，常常告诉学生们提前准备[1]。并且，我认为对应用数学家及工程师来说，概率论与随机过程是最有用的工具[2]。

然而，根据我的亲身经历，这也是很多学生在学习中感到最困惑的学科之一。为什么呢？

基于我的学习经验，我期望以自己的方式来阐释概率论与随机过程这一学科。本文是以我在科学网博客上的 5 篇博文为基础发展而来的（见[3, 4, 5, 6, 7]）。我的目的绝不是要取代那些优秀的教科书。**本文的主要目的，是希望大家通过阅读，更加容易接受和理解概率论与随机过程，不再望而生畏，而绝非替代这个学科中大量优秀的教科书**。本文的写作没有采用随机过程教科书所要求的精确的数学语言，而是从教师和学生面对面交流的角度下笔。这种方式虽然不够正式，却比较易于向读者呈现概率论与随机过程的概貌。同时，希望本文有助于大家今后阅读以测度论的语言撰写的教科书和文章。我将严格的从一个使用者的角度落笔，所需知识也仅限于初等微积分以及从 3 维空间到n维空间的推广。让我们开始吧！

1. 概率论基础

首先，让我们做一个简化的假设——对从事实际应用的人们来说，这一假设完全不重要，也不严格。这个假设是：

有限性假设：我们假设不存在无穷大的数，也就是说，不存在无穷大但可以有非常大的数字，例如：10^{100}（估计这个数

字大于宇宙中所有原子数目的总和）。如果只运用数字计算机进行实际运算，这一假设是自动满足的。该假设使我们无需再考虑那些充斥在概率论理论文献中的、令外行人困惑的测度论相关术语了。

基于有限性假设，我下面定义什么是随机变量。

随机变量：随机变量就是一个变量，当对这个变量进行抽样或观测时，变量可以从一个有限数值集合中取任意值。我们通过直方图来描述一个随机变量。直方图指出了采样值在该随机变量可能的取值范围中所占的百分比。图 1 是一个典型的直方图。它描述了过去四年中我的博客文章的阅读量/点击率的随机变量。

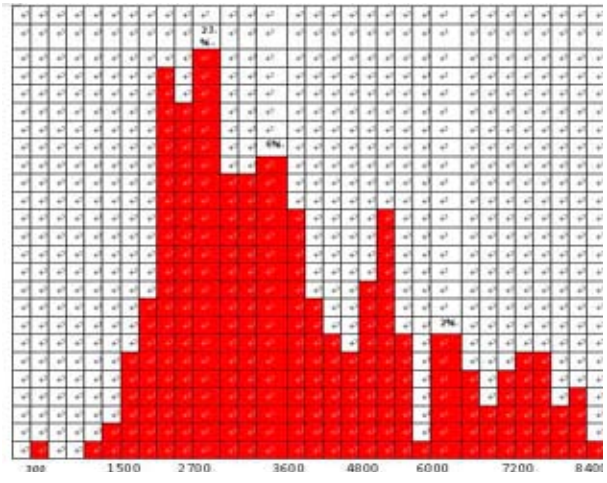

图 1. 2009～2013 年，我的博文阅读量的直方图：x-轴表示的是文章的点击次数，y-轴表示的是相应点击次数范围内的文章数目。

注意，图中每个柱状条的取值表示的是百分比，因而所有柱状条的取值加在一起等于 1 或 100%。也就是说，随机变量必然以概率 1 在整个范围的某一个位置上取值。根据有限性假设，随机变量的取值范围是有限的。不过，完全地描述一个随机变量仍需要大量的数据。（实际上，我花了 3 个小时搜集数据以及制作图 1，因此我没有搜集 5 年及 5 年以上博文的数据。）从计算上讲，这并不简便。出于简化的目的，人们一般运用下面两个常见的粗略特性指标对数据进行描述/刻画。

随机变量的**均值**——直观的解释是，想象把一个纸板剪成直方图的形状，如果将刀刃垂直于x-轴放置在该轴的某点上，以保持纸板的平衡，这个点就是随机变量的均值。从数学上讲，均值就是每篇文章点击率的平均值。实际上，科学网会为每位博主计算其文章点击率的均值，并列出排名前 100 位的博主。目前，我的博文点击率均值为 4130，位列第 26 位。

随机变量的**方差**——这是用来度量直方图的扩展性的指标。粗略的说，小方差意味着直方图在其均值附近扩展范围较小，对于大方差而言则相反。它描述了随机变量取值的可变性。在股票市场术语中，简单而言，股票的 β 值就是股票每日市值的方差，是其波动性的一种度量。数学上，方差称为直方图的**二阶中心矩**。

为了进一步对直方图的特性进行粗略的描述，我们还可以定义所谓的高阶中心矩——例如直方图的**偏度**，即**三阶中心矩**。但是，实践中很少用得到这样的高阶

矩或者缺少这些高阶矩的数据。

对单个随机变量就说到这里。在实际工作中我们常常需要处理多个随机变量。让我们考虑两个随机变量x和y。现在随机变量x-y的直方图变成了一个三维图形。它看起来像是一个多峰的地形图（想象一下中国南部的广西桂林或者是纽约市曼哈顿岛上的摩天大楼）。为此，需要引入一个新的概念，即，随机变量x和y的**“联合概率”**或者在近似描述的情形下用到的**“相关性 / 协方差”**。它抓住了随机变量之间可能有的关联性。比如大家都认为，聪明的父母更加倾向于生育聪明的子女。如果我们把父母的智商作为随机变量x，同时相应的子女的智商作为随机变量y，于是从数学上说，y正相关于x。如果我们从x和y的三维直方图上俯瞰，我们将看到如图 2 所示的那样，峰值沿着东北向西南方向散布。

换句话说，当y值已知时，会改变我们对x的可能值的判断。更一般地，我们称x和y是不独立，而是相关的。作为一般的三维函数，数学上我们记其联合概率为$p(x,y)$（即直方图）。同样，我们定义在给定x值下，y的**条件概率**为

$$p(x|y) \triangleq \frac{p(x,y)}{p(y)} \quad \text{or} \quad p(y|x) \triangleq \frac{p(x,y)}{p(x)},$$

其中$p(y)$和$p(x)$分别称为y和x的**边缘概率**，即将三维的直方图分别沿着y轴或者x轴折叠成二维的直方图。从图形上看，条件概率$p(x|y)$可以理解为，在三维直方图上，沿着特定的y值的横截面图，即二维直方图。从数学上讲，我们需要将$p(x,y)$除以$p(y)$用以归一化$p(x|y)$值，使其满足直方图的定义，即面积等于 1(100%)。

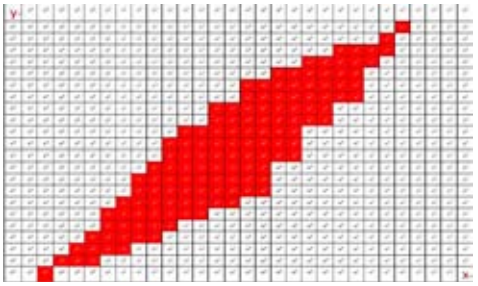

图 2. 相关性，从三维直方图上鸟瞰。

另一方面，三维直方图的鸟瞰图也有可能是一个长方形（与图 2 的视图相比）。换句话说，不论我们对y取什么值，都有$p(x|y) = p(x)$。这种情况下，根据$p(x|y)$的定义，我们有$p(x,y) = p(y)p(x)$。 我们说随机变量x和y**相互独立**。这符合我们关于这一概念的直观印象，即知道y的值不能告诉你关于x的可能取值的任何新信息；对于知道x时考虑y的情形同样如是。从计算上讲，这将 2 元变量的函数简化成为单变量函数的乘积；对于n个随机变量的情形，其对计算的简化相当可观。

为了粗略地刻画两个一般的随机变量，我们使用一个均值向量$[x\ \ y]$以及一个2×2的协方差矩阵，该矩阵的对角元素为x和y的方差，在非对角位置上是对称的协方差，即

$$\begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{yx} & \sigma_y^2 \end{bmatrix}.$$

对本节进行一个小结，我们到目前为止介

绍了以下概念：

1. 随机变量的直方图描述；

2. 使用均值和方差对直方图的粗略刻画；

3. 两个随机变量的联合概率（三维直方图）；

4. 独立性与条件概率；

5. 协方差矩阵。

下面，假设我们有n个随机变量$[x_1, x_2, \cdots, x_n]$，则前面关于两个随机变量的所有内容同样适用。我们仅仅需要把二维和三维相应的变为n维和$n+1$维。将n维随机变量的均值变成n维的向量，协方差阵变成$n \times n$的矩阵。在你的脑海中，可以像图 1 和 2 一样，以同样的方式可视化n维情形下的一切。联合分布（直方图）$p(x_1, x_2, \cdots, x_n)$是一个n元变量的函数。如果n个变量彼此相互独立，我们可以写成

$$p(x_1, x_2, \cdots, x_n) = p(x_1)p(x_2)\cdots p(x_n)$$

这里没有涉及新的概念。

从概念上讲，不管你信或者不信，我认为以上所讲的关于概率和随机过程的内容可以完全满足工程界朋友的需要，甚至对有志于从事学术或理论研究的朋友而言也一样。根据我在随机控制与优化方面 46 年的潜心研究经验以及在工程界的咨询经历，我所用到的知识从来没有超出上述领域。下面的各节我将简要地说明，如何把上述知识应用到实际工作中。

由于计算的指数级增长，处理任意n元函数是不可能的[8]。数据方面，这也涉及到天文数字般的大量数据。为了至少在理论上简化符号，我们对这些离散数据进行连续逼近，同时引入连续变量和函数。必须强调，对于我们的目的而言，这些仅仅是为了方便的近似和简化，不涉及到任何新的思想。这些将是本文下面的内容。除了引入连续型变量外，我们还需要发展各种特殊类型的联合概率结构，以便简化描述与计算，本文下面将会解决这些问题。我要再次强调，从我的角度来看，这些简化和特殊情况是计算的可行性与实用性的需要。不涉及任何新的概念。

2. 测度论背后的基本原理以及作为使用者需要知道的内容

本节，我们使用一个解析函数，**概率密度函数**，来取代第 1 节中较为笨拙的直方图描述。从我们的观点来看，这就是用连续变量函数逼近一个直方图。最著名的例子是下式给出的高斯型密度函数：

$$p(x) \sim \exp\left\{-\frac{(x-m)^2}{\sigma^2}\right\}. \qquad (2.1)$$

式(2.1)表示着大家熟悉的钟型曲线。我们随后将会讨论它所具有的一些性质。在这之前，我们首先要解决一个问题。一旦我们开始使用连续型变量，马上就会遇到数学上的一个问题。式(2.1)当作概率密度函数处理是否是定义明确的对象？一个实变量可假设有无穷多个值。这里有两类无穷性。第一类是**可列无穷**，比如正整数序列$(1, 2, 3, \cdots, \infty)$；另一类是**不可列无穷**，即实变量。流传千年的“Zeno 悖论”所衍生的现代版本可

以很好的解释这一概念：“如想拥抱距离自己 10 英尺远的一位美丽的裸女郎，必须首先靠近她，使自己距离她 5 英尺远。然而，要想实现距离5英尺，必须先移动距离的四分之一，即 2.5 英尺（距离目标 7.5 英尺）……以此类推，你会发现：任何一个距离都可以被无穷多的划分，因此你根本无法移动。[1)] 换句话说，一个实变量的任何子区间似乎含有与原始区间相同多的点。这为数学家提出了一个难题。据说，19 世纪著名数学家、集合论的创始人 G. Cantor 曾饱受这一悖论困扰以至神经衰弱。于是，当说到实变量的概率密度函数时，大家都无法回避一个实际问题：对不可列个点如何分配其概率？**测度论**就是用来解决这个问题的。粗略的说，我们扔掉了实数区间上的一个子点集（该子集是测度为零的集合），使得剩下来的子集是可列无穷集，于是我们就可以在这个集合上定义概率了。这样做的妙处在于，它不仅具有数学上的严格性，同时与应用科学家/工程师们在学习与实践用到的大学课程中不涉及测度论的概率论相一致。更为重要的是，因为其数学上的精确定义，使得我们可以使用我们熟悉的诸如微积分等工具，对概率（比如密度函数）进行运算。总之，方方面面都很满意。

当谈到随机过程时，我们需要引入一个特别的实变量——时间。直观地说，就是一个随机变量随时间演化。于是，我们将在连续时间随机过程上有不可列无穷个随机变量，每一个随机变量对应着一个时间点（于是有不可列无穷多个点）。基于测度论理论的随机过程相应地也被发展起来。标志性的奠基性工作是 J. Doob 于 1953 年出版的著名的教科书[9]。 但是，凡事有利必有弊，严格的随机过程教科书中的测度论术语，使得工程专业的学生学习起来感到挫折和困惑。

但是对我们实际应用者而言，我们不必担心测度论。所有我们需要知道的是，非测度论的概率论是有着严格性和相容性的理论基础的。知道上面提到的测度论的基本原理和它的一些专业术语对我们来说就足够了。可以放心使用如微积分和代数等所有的工具。概率论和随机过程的教科书可以不参考测度论的语言来写。我学习这门课时用到的由 Davenport 和 Root 于 1957 年出版的著名教科书就是以这样的方式撰写的（见[10]）。下面教程的内容就体现了这样的思想。（我以这种非严格的方式阐释测度论可能让纯粹数学家觉得太粗糙并无法接受，我向他们表示歉意。）

下面回到高斯型随机变量。

1. 注意到，完全定义一个高斯型密度函数只需要两个参数，均值μ和方差σ。（对于高斯向量，我们有**均值向量**和**协方差矩阵**，二者都仅有有限个参数。这都将大大简化计算。但是严格的讲，由式(2.1)刻画的高斯随机型变量可以在连续区间$(-\infty,+\infty)$上取任何值。）

2. 经验和理论都说明，为什么高斯型随机变量会在自然界中经常出现。中心极限定理说明，任何由许多复杂相互作用

[1)] 工程师对这个悖论的回答是：“不要担心这个悖论，我只关心我能否足够接近目标。”

的基本随机变量而导致的随机现象倾向于高斯型密度。

3. 如果你仅仅知道均值和方差，假设随机变量是高斯分布，则需要加入的无根据假设最少。

4. 下面是一个最有用，但是非常困难，并且没有得到一般性解决的问题。给定$y = f(x)$,如果我们知道x的概率密度，当函数f明确知道并且不可逆时，那么y作为随机变量的密度函数是什么？换而言之，知道输入以及系统方程，输出是什么？系统理论学家都知道这是业界的一个价值 64,000 美元的问题。但是这却是应用数学家在教科书中从不指出的隐晦的秘密。然而，当函数f是线性的并且x是高斯型随机变量时，则y也同样是高斯的，并且可以很容易计算其均值和方差。Kalman 滤波的成功是建立在这一事实上的。

由于这些原因，除非我们知道具体的信息，否则在涉及实变量的计算中，我们通常假设所有的随机变量都是高斯型的。事实上，著名的 Kalman 滤波已经成功地应用于众多不同情形，即便我们知道所涉及的噪声是非高斯型的。类似地，在一般的应用中，我会毫不犹豫的把如图 1 中的直方图用高斯型密度函数近似。

当然，一些其他概率密度函数也具有些不错的性质，并且在离散事件系统的离散变量中很有用，例如泊松密度函数和指数型密度函数。这些将在我今后的文章中介绍。

以此为背景，我们现在可以继续进行，接着讨论随机过程。

3. 随机序列与过程

以前面两节对基本概念的讨论为基础，我们将在本节谈谈一个非常具有实用性的学科—随机过程。我们先研究**随机序列**$\cdots, X_1, X_2, X_3, \cdots, X_i, \cdots, X_t, X_{t+1}, \cdots$，该序列除了是由相互独立的整数变量$i = \cdots, 1, 2, 3, \cdots, t, t+1, \cdots$作为指标的随机变量外，没有其他特别之处。我们在第 1 节讨论的关于n维随机变量的内容在这里同样适用。如果聚焦到随机变量$1, \cdots, t$，我们需要指定其联合概率密度函数$p(x_1, x_2, x_3, \ldots, x_t)$；对于粗略描述，我们有均值的$t$维向量，$[E(x_i)]$，其中的每一个元素是随机变量$x_i$的均值，同时有$t \times t$的协方差矩阵$[\sigma_{ij}]$,其中$\sigma_{ij}$是$x_i$与$x_j$的协方差。矩阵$[\sigma_{ij}]$的第一行的元素有$\sigma_{11}^2, \sigma_{12}, \sigma_{13}, \cdots, \sigma_{1t}$（如果难以理解以上内容，请复习第 1 节）。这里，与通常的术语略有不同的是，我们不是说x_1与$x_2, x_3, \cdots, x_t$的协方差，而是称为**相关序列**，并且记为$x_{1i}, i = 1, 2, \cdots, t$。使用这种复杂的符号系统是因为：$x_i$一般是随机向量，因此我们需要讨论其元素的协方差。于是，对于随机变量在时间上的依赖性，我们采用了相关性这一术语。为了避免混淆，协方差仅指在同一个时间点上的随机变量的协方差。除了计算上的考虑外，所有需要研究随机序列的内容我们已

经在第 1 节介绍过了。

涉及到时间点的计算处理时，我们需要的时间点的数目往往是在几百或几千的量级。对一千乘以一千的矩阵或者有一千个变量的密度函数计算很麻烦，而且往往不切实际。为此，我们需要研究能够把事情化繁为简的特殊情形。首先，最显著的简化是考虑独立随机序列，即

$$p(x_1,\cdots,x_t)=p(x_1)p(x_2)\cdots p(x_t).$$

这种序列通常称为**白噪声**序列。进一步，如果$p(x_1)=p(x_2)=\cdots=p(x_t)$，我们便得到了**平稳白噪声**序列。这一序列通常也叫做**独立同分布**序列。[2)]

然而，独立同分布对于实际中观察到的随机序列做了太强的假设与限制。为此，我们引入了以下的更为复杂的高阶假设。

Markov 假设：这一假设可以很好地用一句话诠释：“**当前的状态将过去与未来区分开来**”。用数学的语言表述就是：

$$p(x_{t+1}|x_t,x_{t-1},\cdots,x_1)=p(x_{t+1}|x_t)\ \forall t. \quad (3.1)$$

利用式(3.1)可以马上把一个t元变量的函数变成 2 元变量函数的乘积，即

$$p(x_t,x_{t-1},\cdots,x_1)=p(x_t|x_{t-1})\times p(x_{t-1}|x_{t-2})\cdots p(x_2|x_1)p(x_1). \quad (3.2)$$

当然，大家会质疑 Markov 假设不现实，例如，当x_t不仅依赖其最近邻的过去x_{t-1}，而且还依赖其较近的过去，例如x_{t-2}。理论上讲，我们很容易克服这一质疑。通过重新定义变量$y_1=[x_1,x_{t-1}]$，稍加思索就会明白，y_t是 Markov 随机序列。于是，**依赖于有限的过去的任何序列可以被转换成一个 Markov 序列**。这也就说明了有大量的理论文献研究 Markov 过程的原因。当然，从计算上讲，这仅仅是符号变换而已。简化计算必须分开处理。特别地，下一节我们将看到，Gauss-Markov 序列因其一般性、实用性以及计算的简化性等，在随机系统研究中扮演着独一无二的角色。

同时，我特别反复强调的是

1. 随机序列无非就是一串由指标标记的随机变量（随机变量的性质我们在第 1 节已经讲过）；

2. Markov 随机序列使得人们可以更一般地研究那些依赖于有限过去时间的随机现象；

3. “高斯型”、“平稳性”、“Markov”以及“独立同分布”等形容词可以选择性地应用到随机序列上，用以简化这些序列的符号与计算。

下面一道思考题检测你是否掌握了本节的内容：“下面的论述是否有意义：系统输出（可能是向量）是非高斯型 Markov 序列，其输出协方差矩阵的非对角线元素是非零的。”

基于本节内容，我们可以直接推广到

[2)] 我们进一步区分为宽平稳序列和严平稳序列。前者仅需σ_{ij}依赖于指标$i-j$的差，而后者需要$p(x_i,x_j)$仅依赖指标$i-j$的差（对所有i和j）。

连续时间随机过程，除了需要注意第 2 节的说明外，不需要引入新的概念。

4. Gauss-Markov 线性系统实例

为了实际应用的目的，上面三节我列出了概率论与随机过程的基础知识。至少从使用者的角度，我说明了为什么某些事情是以特定的方式发展。当伟大的数学家 Gauss 被问及“为什么你使你的工作如此漂亮，却又让我们凡人难以理解?”时，他回答道：“当天主大教堂的建筑者完成建造时，他必须取下所有的脚手架，这样你会仅仅赞叹建筑的美丽，而不会被脚手架吸引注意力，同时你也无法知道教堂是怎样修建的。”我本文的目的之一，就是要把其中的一些“脚手架”展示给你们，使得大家作为使用者理解这门学科是什么，以及为什么如此。我的主要原则是**“使该学科从计算的角度有用”**。

本节，我将着重强调这一点。基于 Markov 假设，我们已经把随机过程的研究简化到$1 \sim 2$元变量的函数（暂时先不要考虑这些变量是向量的情形，后面我们还会进行这方面的扩展）。但是，即使在低维的情形下，函数的计算也绝不是一件容易的事。我们必须把问题简化为有限个参数的情形。正如第 1、2 节提及的很多原因，高斯型密度函数刚好符合我们的要求。其密度方程完全由均值μ和方差σ刻画。通常，我们用记号$x \sim N(\mu, \sigma)$表示随机变量x是高斯（正态）分布。

于是，我们假设$p(x_1)$是$N(\bar{x}_1, p_1)$以及$p(x_t|x_{t-1}) \sim N(\phi x_{t-1}, q)$，其中$\phi$和$q$是已知参数，这样我们可以用一对均值和方差的公式刻画整个 Gauss-Markov 序列：对所有$t = 1, 2, \cdots$，我们有

$$\bar{x}_{t+1} = \phi \bar{x}_t, \tag{4.1}$$

$$p_{t+1} = \phi p_t \phi + q, \tag{4.2}$$

其中$p(x_{t+1})$是$N(\bar{x}_{t+1}, p_{t+1})$。（请思考上面的结论并令你自己信服。如果需要的话，再读一遍第 1 节。）这样，一个 Gauss-Markov 随机序列可以完全用式(4.1)-(4.2)刻画，其中$\bar{x}_1, p_1, \phi$和q是给定的初始条件和系统参数。式(4.1)-(4.2)可有进一步好的解释。考虑由向量线性差分方程控制的多维线性离散时间系统

$$x_{t+1} = \Phi x_t + D w_t,$$

其中w_t是均值为零，协方差为Q的高斯型独立同分布（白噪声）序列，并且$p(x_1)$为$N(\bar{x}_1, P_1)$。于是得出那些学过随机线性系统课程的人们都熟知的结论：$p(x_t)$为$N(\bar{x}_t, P_t)$，其中

$$\bar{x}_{t+1} = \Phi \bar{x}_t, \tag{4.3}$$

$$P_{t+1} = \Phi P_t \Phi^T + DQD^T. \tag{4.4}$$

式(4.3)-(4.4)是对式(4.1)-(4.2)的推广到向量和矩阵的情形。其推导的步骤很直接，任何标准教材都可以找到。我们的动机是想说明，Gauss-Markov 随机

序列以及由白噪声驱动的线性动态系统，他们看似不同，实则等价。他们堪称天作之合。这是系统理论高年级本科生以及一年级研究生课程中最先讲授的内容，也是系统工程师最有用的工具之一。这也是著名的 Kalman 滤波方程的其中的一半。

巧合的是，$\bar{x}_t$和P_t也是$p(x_t)$的**充分统计量**，也就是说这些有限的数值完全决定了多元密度函数$p(x_t)$。

为了介绍著名的 Kalman 滤波的另一半，我们需要引入在不确定情形下关于估计的一个基本事实。我们会在下一节讨论。

5. Bayes 准则与 Kalman 估计

这里还有一个值得一提的简单公式，即 Bayes 准则。它是由条件概率定义衍生出的结果。考虑随机变量x和y，根据定义

$$p(x|y)=\frac{p(x,y)}{p(y)}=\frac{p(y|x)}{p(y)}p(x) \tag{5.1}$$

在上面的式(5.1)中，$p(x)$记为**先验概率**。这是我们可知的关于随机变量x的信息。另外，$p(x|y)$记为**后验概率**，它表示的是在观测到样本值y后，对x的认识的改进。其中因子$\frac{p(y|x)}{p(y)}$把先验概率转变成后验概率。从式(5.1)看出，由于涉及到对函数的运算，其计算上并不是很有用。于是，我们又一次考虑到高斯型密度函数。令

$$p(x)\sim N(\bar{x},M),$$

并且$p(y|x)\sim N(hx,r)$，其中$\bar{x},M,h,r$为已知参数，于是$p(x|y)\sim N(\hat{x},P)$，其中

$$\hat{x}=\bar{x}+Phr^{-1}(y-h\bar{x}), \tag{5.2}$$

$$P=M+Mh(hMh+r)^{-1}hM. \tag{5.3}$$

式(5.2)-(5.3)可视为一阶段 Kalman 滤波，其利用观测到的y，把先验概率$p(x)$更新为后验概率$p(x|y)$。式(5.2)-(5.3)通过更新参数（均值和方差），把先验高斯型密度更新至后验高斯型密度。我们说高斯型密度函数属于一类**再生密度函数**，因为先验概率密度再生成为具有不同参数的后验概率密度函数。

同时，高斯密度函数的 Bayes 准则估计具有线性系统的解释（和第 4 节中 Gauss-Markov 序列具有线性系统的解释是类似的涵义）。我们有

$$y=hx+v,\quad p(v)\sim N(0,r), \tag{5.4}$$

其中y表示对系统状态x的带有噪声的观测。至于推广到一般的向量和矩阵情形则比较直接，可以在所有的标准教材中找到，我在此不重复。（包括我 44 年前撰写的教科书中的推导[11]。 读者也可以参考我的两篇博文[12,13]，以便了解关于 Kalman 滤波以及扩展非线性滤波的直观的解释。）

最后，总结一下本文讨论的五个方面的内容

1. 概率论基础；
2. 测度论背后的基本原理以及作为

使用者需要知道的内容；

3. 随机序列与过程；

4. Gauss-Markov 线性系统实例；

5. Bayes 准则与 Kalman 估计。

相信我，在过去 46 年的教书和咨询生涯中，以上讨论的关于概率论和随机过程的内容支撑了我工作的前 30 年。他们是我所需要的全部。在后面的 25 年，为了研究离散事件系统，我又使用了泊松与指数分布。然而，后者的实用性和美感都无法与 Gauss-Markov 线性二次型指标控制系统相提并论。未来，我打算写一个关于随机离散事件系统的教程。

正如古谚所说，“不劳无获”。在本教程之外，你们仍需要通过一本常用的教材学习概率论和随机过程。我写的教程仅仅提供了一个概略，希望能够指导你们的学习。天下没有免费的午餐。

参考文献

[1] Yu-Chi Ho. On Oral Examinations. http://blog.sciencenet.cn/blog-1565-13708.html, 2008.

[2] Yu-Chi Ho. The Markov Centennial. http://blog.sciencenet.cn/blog-1565-656455.html, 2013.

[3] Yu-Chi Ho. Probability and Stochastic Process Tutorial (1).http://blog.sciencenet.cn/blog-1565-664051.html, 2013.

[4] Yu-Chi Ho. Probability and Stochastic Process Tutorial (2).http://blog.sciencenet.cn/blog-1565-665359.html, 2013.

[5] Yu-Chi Ho. Probability and Stochastic Process Tutorial (3). http://blog.sciencenet.cn/blog-1565-666599.html, 2013.

[6] Yu-Chi Ho. Probability and Stochastic Process Tutorial (4). http://blog.sciencenet.cn/blog-1565-669532.html, 2013.

[7] Yu-Chi Ho. Probability and Stochastic Process Tutorial (5).http://blog.sciencenet.cn/blog-1565-671318.html, 2013.

[8] Yu-Chi Ho. Some Dirty Secrets of Applied Mathematics. http://blog.sciencenet.cn/blog-1565-26889.html, 2008.

[9] Joseph L. Doob. Stochastic Processes. John Wiley & Sons Inc, 1953.

[10] Wilbur B Davenport, William L Root, et al. An introduction to the theory of random signals and noise. McGraw-Hill New York, 1957.

[11] Jr. Arthur E. Bryson and Yu-Chi Ho. Applied Optimal Control: Optimization, Estimation and Control. Taylor & Francis, 1975.

[12] Yu-Chi Ho. Proximity to Fame (#2) – the 2008 Draper Prize. http://blog.sciencenet.cn/blog-1565-14253.html, 2008.

[13] Yu-Chi Ho. A Short Course on Nonlinear Filtering. http://blog.sciencenet.cn/blog-1565-426323.html, 2011.