

## Lecture 13

We can use the value function to derive in a loose way both the maximum principle and the Hamilton Jacobi Bellman equation. Let the functions  $\bar{x}(t)$  and  $\bar{u}(t)$  solve of the problem

$$\begin{aligned} & \max_{u \in \mathcal{A}} \int_0^T f(x(t), u(t)) dt \\ \text{s.t. } & \frac{dx(t)}{dt} = g(x(t), u(t)), \text{ for all } t, \\ & x(0) = x_0, \text{ and } x(T) = x_1, \end{aligned}$$

where I assume that  $N = K = 1$ , so that  $x(t)$  and  $u(t)$  are numbers. Let the value function  $V(\underline{x}, t)$  be defined as

$$\begin{aligned} V(\underline{x}, t) &= \max_{u \in \mathcal{A}} \int_t^T f(x(s), u(s)) ds \\ \text{s.t. } & \frac{dx(s)}{ds} = g(x(s), u(s)), \text{ for all } s \in [t, T], \\ & \text{and } x(t) = \underline{x} \text{ and } x(T) = x_1. \end{aligned}$$

Assume that the function  $V(x, t)$  is differentiable. If we think of  $dt$  as a small positive number, then by Bellman's equation,

$$\begin{aligned} V(\bar{x}(t), t) &= \max_{u \in U} [f(\bar{x}(t), u)dt + V(\bar{x}(t+dt), t+dt)] \\ &= \max_{u \in U} \left[ f(\bar{x}(t), u)dt + V(\bar{x}(t), t) + \frac{\partial V(\bar{x}(t), t)}{\partial x} \frac{d\bar{x}(t)}{dt} dt + \frac{\partial V(\bar{x}(t), t)}{\partial t} dt \right] \\ &= \max_{u \in U} \left[ f(\bar{x}(t), u)dt + V(\bar{x}(t), t) + \frac{\partial V(\bar{x}(t), t)}{\partial x} g(\bar{x}(t), u)dt + \frac{\partial V(\bar{x}(t), t)}{\partial t} dt \right] \\ &= V(\bar{x}(t), t) + \frac{\partial V(\bar{x}(t), t)}{\partial t} dt + \max_{u \in U} \left[ f(\bar{x}(t), u) + \frac{\partial V(\bar{x}(t), t)}{\partial x} g(\bar{x}(t), u) \right] dt. \end{aligned}$$

If we change  $\bar{x}(t)$  to  $x(t)$ , we obtain

$$\frac{\partial V(x(t), t)}{\partial t} = -\max_u \left[ f(x(t), u) + \frac{\partial V(x(t), t)}{\partial x} g(x(t), u) \right], \quad (13.1)$$

which is the Hamilton Jacobi Bellman equation. This is a partial differential equation in  $V$ ,

which may sometimes be solved for  $V$ . If we substitute  $\lambda(t)$  for  $\frac{\partial V(x(t), t)}{\partial x}$ , we see that the maximum principle applies, that is,  $u(t)$  solves the problem

$$\max_{u \in U} [f(x(t), u) + \lambda(t)g(x(t), u)],$$

for all  $t$ .

So far, the functions  $f$  and  $g$  have not depended directly on  $t$ . I now turn briefly to problems in which the flow of gain is discounted and so does depend directly on time. Consider the problem

$$\begin{aligned} & \max_{u \in \mathcal{A}} \int_0^{\infty} e^{-rt} f(x(t), u(t)) dt \\ \text{s.t. } & \frac{dx(t)}{dt} = g(x(t), u(t)), \text{ for all } t, \\ & x(0) = \underline{x}, \end{aligned}$$

where  $x(t)$  and  $u(t)$  are numbers. The analysis that follows is informal. The value function is defined by the equation

$$\begin{aligned} V(\underline{x}) &= \max_{u \in \mathcal{Q}} \int_0^{\infty} e^{-rt} f(x(t), u(t)) dt \\ \text{s.t. } & \frac{dx(t)}{dt} = g(x(t), u(t)), \text{ for all } t, \\ & \text{and } x(0) = \underline{x}. \end{aligned} \tag{13.2}$$

Because the flow of gain depends on time only through discounting, we could write the value function as

$$\begin{aligned} V(\underline{x}) &= \max_{u \in \mathcal{Q}} \int_0^{\infty} e^{-r(t-T)} f(x(t), u(t)) dt \\ \text{s.t. } & \frac{dx(t)}{dt} = g(x(t), u(t)), \text{ for all } t, \\ & \text{and } x(T) = \underline{x}, \end{aligned}$$

for any time  $T > 0$ . Hence Bellman's principle implies that for any  $t$  and  $T$  such that  $0 < t < T$ ,

$$\begin{aligned}
V(\underline{x}) &= \max_{u \in \mathcal{Q}} \left[ \int_t^T e^{-r(s-t)} f(x(s), u(s)) ds + e^{-r(T-t)} V(x(T)) \right] \\
\text{s.t. } \frac{dx(s)}{ds} &= g(x(s), u(s)), \text{ for all } s \in [t, T] \\
&\text{and } x(t) = \underline{x}
\end{aligned} \tag{13.3}$$

Think of  $dt$  as a small positive number. By Taylor's theorem,

$$e^{-rdt} = 1 - rdt + O(dt^2) = \frac{1}{1 + rdt} + O(dt^2), \tag{13.4}$$

where  $O(x)$  is such that for some positive numbers  $\varepsilon$  and  $b$ ,  $|O(x)| < b|x|$ , for all  $x$  such that

$|x| \leq \varepsilon$ . Assume that the function  $V$  is differentiable and let  $\bar{x}(t)$  and  $\bar{u}(t)$  solve problem 13.2. By equations 13.3 and 13.4,

$$\begin{aligned}
V(\bar{x}(t)) &= \max_{u \in \mathcal{Q}} \left[ \int_t^{t+dt} e^{-r(s-t)} f(x(s), u(s)) ds + e^{-rdt} V(x(t+dt)) \right] \\
\text{s.t. } \frac{dx(s)}{ds} &= g(x(s), u(s)), \text{ for all } s \in [t, t+dt] \\
&\text{and } x(t) = \bar{x}(t) \\
&= \max_{u \in U} \left[ f(\bar{x}(t), u) dt + \frac{1}{1 + rdt} V(\bar{x}(t + dt)) \right] + O(dt^2) \\
&= \max_{u \in U} \left[ f(\bar{x}(t), u) dt + \frac{1}{1 + rdt} \left( V(\bar{x}(t)) + g(\bar{x}(t), u) dt \right) \right] + O(dt^2) \\
&= \max_{u \in U} \left\{ f(\bar{x}(t), u) dt + \frac{1}{1 + rdt} \left[ V(\bar{x}(t)) + \frac{dV(\bar{x}(t))}{dx} g(\bar{x}(t), u) dt \right] \right\} + O(dt^2),
\end{aligned}$$

where the last equation follows from the differentiability of  $V$ . Multiplying this equation by  $1 + rdt$ , we obtain

$$\begin{aligned}
&V(\bar{x}(t)) + rV(\bar{x}(t)) dt \\
&= \max_{u \in U} \left[ f(\bar{x}(t), u) dt + rf(\bar{x}(t), u) dt^2 + V(\bar{x}(t)) + \frac{dV(\bar{x}(t))}{dx} g(\bar{x}(t), u) dt \right] \\
&+ O(dt^2)
\end{aligned}$$

$$= V(\bar{x}(t)) + \max_{u \in U} \left[ f(\bar{x}(t), u)dt + \frac{dV(\bar{x}(t))}{dx} g(\bar{x}(t), u)dt \right] + O(dt^2).$$

If we cancel  $V(\bar{x}(t))$  from both sides of this equation and then divide by  $dt$ , we obtain

$$rV(\bar{x}(t)) = \max_{u \in U} \left[ f(\bar{x}(t), u) + \frac{dV(\bar{x}(t))}{dx} g(\bar{x}(t), u) \right] + O(dt).$$

If we let  $dt$  go to zero and change  $\bar{x}(t)$  to  $x(t)$ , we find that

$$rV(x(t)) = \max_{u \in U} \left[ f(x(t), u) + \frac{dV(x(t))}{dx} g(x(t), u) \right].$$

This also is known as the Hamilton Jacobi Bellman equation. It again is what corresponds to the Bellman equation of discrete time problems. The solution of this partial differential equation over the whole state space is again a necessary and sufficient condition for the existence of an optimum, provided the value function is differentiable.

Example 13.1: In order to illustrate the use of this Hamilton Jacobi Bellman equation, discount the gain in example 12.1 over an infinite horizon. The problem then is

$$\begin{aligned} \max_{u \in \mathcal{A}} \int_0^{\infty} e^{-rt} [-x^2(t) - u^2(t)] dt \\ \text{s.t. } \frac{dx(t)}{dt} = u(t), \text{ for all } t \\ x(0) = \bar{x}, \end{aligned}$$

where  $\mathcal{A}$  is the set of piecewise continuous functions from  $[0, \infty)$  to  $\mathbb{R}$ . In terms of the notation we have been using,  $U = \mathbb{R}$ ,  $f(x, u) = -x^2 - u^2$ , and  $g(x, u) = u$ . Let  $V(x)$  be the value function for this problem. The Hamilton Jacobi Bellman equation in this example is

$$rV(x(t)) = \max_u \left[ -x^2(t) - u^2 + \frac{dV(x(t))}{dx} u \right].$$

Let us guess that  $V(x) = ax^2$ , for some negative number  $a$ . Then  $dV(x)/dx = 2ax$ , so that the Hamilton Jacobi Bellman equation becomes

$$rax^2 = \max_{u \in \mathbb{R}} [-x^2 - u^2 + 2axu].$$

The maximizing value of  $u$  is  $u = ax$ . The above equation therefore becomes

$$rax^2 = -x^2 + a^2x^2,$$

so that

$$ra = -1 + a^2,$$

that is,

$$a^2 - ra - 1 = 0.$$

Solving this quadratic equation, we find that

$$a = \frac{r \pm \sqrt{r^2 + 4}}{2}.$$

Since  $a < 0$ , the appropriate solution is

$$a = \frac{r - \sqrt{r^2 + 4}}{2}.$$

The optimal path for  $x$  then solves the differential equation

$$\frac{dx(t)}{dt} = ax(t),$$

and the solution to the problem is

$$x(t) = x(0) e^{at}$$

with

$$u(t) = \frac{dx(t)}{dt} = ax(0) e^{at},$$

for all  $t$ .

As a check, let's calculate the value function for the path just calculated and make sure that the value equals  $ax^2$ . If  $x(t)$  and  $dx(t)/dt$  are as above, the value of the path is

$$\begin{aligned} V(x) &= -\int_0^{\infty} e^{-rt} (x^2 e^{2at} + a^2 x^2 e^{2at}) dt \\ &= -x^2 (1 + a^2) \int_0^{\infty} e^{(2a-r)t} dt = \frac{x^2 (1 + a^2)}{2a - r}. \end{aligned}$$

If  $V(x) = ax^2$ , then

$$\frac{x^2(1+a^2)}{2a-r} = ax^2,$$

so that

$$1 + a^2 = 2a^2 - ra.$$

That is

$$a^2 - ra - 1 = 0.$$

Since this is precisely the quadratic equation satisfied by  $a = \frac{r - \sqrt{r^2 + 4}}{2}$ , so that we have solved the Hamilton Jacobi Bellman equation and have computed a solution to the problem.

### Transversality Conditions

Recall from the end of the previous lecture the Hahn problem in growth theory is that growth paths can satisfy the first order conditions for optimality and yet not be optimal. The problem was mentioned in the context of the Euler condition for discrete time growth models, but the phenomenon occurs in continuous time optimal growth models as well. Infinite horizon paths that satisfy the maximum principle and the Hamiltonian system may be suboptimal, because consumption converges to zero over time and there is an overaccumulation of capital. A number of conditions labeled transversality conditions have been devised to exclude such paths.

Just as in the case of the discrete time growth models, there is a value function,

$$V(y_0) = \max_{C: [0, \infty) \rightarrow [0, \infty)} \int_0^{\infty} e^{-rt} u(C(t)) dt$$

s.t.  $C(t) + dK(t)/dt = f(K(t))$ , for all  $t$ , and  
 $C(0) + dK(0)/dt = y_0$ ,

where  $y_0$  is given and positive and  $r > 0$ . If  $\lambda(t)$  is the dual or conjugate vector at time  $t$ , then  $\lambda(t)$  is a gradient of  $e^{-rt}V(y)$  at  $y = \bar{y}(t)$ , where  $\bar{y}(t) = f(\bar{K}(t))$  and  $(\bar{C}(t), \bar{K}(t))$  is an optimal path. In this model,  $C(t)$ ,  $K(t)$ , and  $y(t)$  are consumption, capital, and output, respectively, at time  $t$ .

Because  $V$  is concave, the gradient  $\lambda(t)$  of  $e^{-rt}V(\bar{y}(t))$  is a subgradient of  $e^{-rt}V$  at  $\bar{y}(t)$  and so

$$e^{-rt}V(0) \leq e^{-rt}V(\bar{y}(t)) + \lambda(t) \cdot (0 - \bar{y}(t)).$$

Hence

$$\lambda(t) \cdot \bar{y}(t) \leq e^{-rt} [V(\bar{y}(t)) - V(0)]. \quad (13.5)$$

Because  $e^{-rt}V(\bar{y}(t)) = \int_t^\infty e^{-rs}u(\bar{C}(s)) ds$  and  $\int_0^\infty e^{-rs}u(\bar{C}(s)) ds < \infty$ , it follows that

$$\lim_{t \rightarrow \infty} e^{-rt}V(\bar{y}(t)) = \lim_{t \rightarrow \infty} \int_t^\infty e^{-rs}u(\bar{C}(s)) ds = 0 \text{ and hence } \lim_{t \rightarrow \infty} e^{-rt}[V(\bar{y}(t)) - V(0)] = 0. \text{ Hence,}$$

since the vectors  $\lambda(t)$  and  $\bar{y}(t)$  are non-negative, inequality 13.5 implies that  $\lim_{t \rightarrow \infty} \lambda(t) \cdot \bar{y}(t) = 0$ . Thus a necessary condition for optimality is that  $\lim_{t \rightarrow \infty} \lambda(t) \cdot \bar{y}(t) = 0$ . A similar condition is that  $\lim_{t \rightarrow \infty} \lambda(t) \cdot \bar{K}(t) = 0$ . These are typical transversality conditions.

Origin of the term "Transversality": In order to have a feel for the origin of the term, recall the all-terrain vehicle problem described earlier when introducing optimal control theory. Suppose that the goal is to go from point  $x_0$  to a river while consuming as little fuel as possible. Let the river be described by a differentiable function  $h: R \rightarrow R^2$ , such that  $Dh(s) \neq 0$ , for all  $s$ . Let  $\lambda(t)$  in  $R^2$  be the dual vector of the optimal control problem. Let  $T$  be the time the vehicle arrives at the river. The transversality condition is that  $\lambda(T)$  be orthogonal to the river, that is, that  $\lambda(T) \cdot Dh(s) = 0$ , where  $s$  is such that  $h(s) = x(T)$ . Orthogonality suggests cutting across and hence transversality. This endpoint condition is necessary for optimality. In infinite horizon optimal growth models, there is no end point and no boundary to which  $\lambda(T)$  should be transverse. Nevertheless there is a need for a condition on the behavior of a path  $x(t)$  as  $t$  goes to infinity, and the term "transversality" has been extended to such conditions.

That  $\lambda(T)$  should be orthogonal to  $Dh(s)$  makes good sense, when you remember that  $\lambda(t)$  is the gradient of the value function  $V$  at  $x(t)$ , for all  $t$ , so that  $\lambda(t)$  is orthogonal to the level curve of  $V$  through  $x(t)$ . Since the endpoint of the optimization problem is the river, the value there of  $V$  is zero and hence constant along the river. It follows that the level curve of  $V$  through  $x$  is nearly parallel to the river when  $x$  is close to the river. Because the control,  $u(t)$ , has only finitely many points of discontinuity, it is continuous for  $t$  close enough to  $T$ . Therefore  $\lambda(t)$  is continuous for  $t$  close enough to  $T$ . Because  $x(t)$  converges to  $x(T)$  as  $t$  approaches  $T$  and  $\lambda(t)$  is a continuous function of  $t$ , for  $t$  near  $T$ , and  $\lambda(t)$  is orthogonal to level curves that asymptotically become parallel to the river as  $t$  approaches  $T$ ,  $\lambda(T)$  must be orthogonal to the river at  $x(T)$ . That is,  $\lambda(T) \cdot Dh(s) = 0$ , where  $s$  is such that  $h(s) = x(T)$ .

## Probability

A first question is how to represent probability. One approach, the state space approach, is as follows. You define a context and imagine there is a mechanism generating everything you could observe in that context. The set of states or outcomes of the mechanism is a set  $S$ . You observe subsets,  $E$ , of  $S$  called events. The set  $S$  is called the state space or the set of states of the world. If  $S$  is a finite set, a probability on  $S$  is a function  $P: S \rightarrow [0, 1]$  such that

$$\sum_{s \in S} P(s) = 1. \text{ The probability of an event } E \text{ is } P(E) = \sum_{s \in E} P(s). \text{ Notice that } P(S) = 1 \text{ and } P(\phi) = 1, \text{ where } \phi \text{ is the empty set.}$$

Example: (Binomial distribution) There are  $N$  independent trials on each of which a success (A) or a failure (F) occurs. The probability of a success on any one trial is  $p$ , where  $0 \leq p \leq 1$ , and the probability of a failure is  $q = 1 - p$ . The set of states of the world  $S$  is the set of  $N$ -tuples of A's and F's. If  $N = 2$ , then  $S = \{(A, A), (A, F), (F, A), (F, F)\}$ . If  $s \in S$  and A occurs  $n$  times in  $s$ , then  $P(s) = p^n(1-p)^{N-n}$ . If  $E$  is the event that there are  $n$  successes, then  $P(E) = \binom{N}{n} p^n(1-p)^{N-n}$ , where  $\binom{N}{n} = \frac{N!}{n!(N-n)!}$ .

The definition of probability becomes more complicated if the set  $S$  is infinite. Logical difficulties can arise if you attempt to define a probability on the set of all subsets of  $S$ . These difficulties are avoided by defining probability on what is called a  $\sigma$ -field of subsets of  $S$ , call it  $\mathfrak{S}$ . A  $\sigma$ -field  $\mathfrak{S}$  is a set of subsets of  $S$  such that  $S \in \mathfrak{S}$ ,  $\phi \in \mathfrak{S}$ ,  $S \setminus A \in \mathfrak{S}$  whenever  $A \in \mathfrak{S}$ , and

$\bigcup_{n=1}^{\infty} A_n \in \mathfrak{S}$  whenever  $A_n$  is a sequence of sets belong to  $\mathfrak{S}$ . It follows that  $\bigcap_{n=1}^{\infty} A_n \in \mathfrak{S}$  whenever  $A_n$

is a sequence of sets belong to  $\mathfrak{S}$ , since  $\bigcap_{n=1}^{\infty} A_n = S \setminus \left( \bigcup_{n=1}^{\infty} (S \setminus A_n) \right)$ . A probability or probability

distribution on  $S$  is a function  $P : \mathfrak{S} \rightarrow [0, 1]$  such that  $P(S) = 1$ ,  $P(\phi) = 0$ , and

$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n)$  whenever  $A_n$  is a sequence of mutually disjoint sets belong to  $\mathfrak{S}$ . The sets

$A_n$  are mutually disjoint if  $A_n \cap A_m = \phi$  whenever  $n \neq m$ . If there is a probability  $P : \mathfrak{S} \rightarrow [0, 1]$ ,

then the sets that are members of  $\mathfrak{S}$  are said to be measurable. The triple  $(S, \mathfrak{S}, P)$  is called a probability space.

Another approach to probability is to focus on the distribution of observations rather than the probability of events in a state space. The stance is that you observe a function  $x : S \rightarrow \mathbb{R}$  or  $x : S \rightarrow \mathbb{R}^N$ , called a random variable, and what you care about is the probability of the possible observations. Suppose first of all that  $S$  is a finite set with probability  $P$ . Then the image of  $x$ , which is  $x(S) = \{x(s) \mid s \in S\}$ , is also a finite set, and the probability of a point  $y \in x(S)$  is  $Q(y) = P\{s \mid x(s) = y\} = \sum_{s: x(s) = y} P(s)$ . The probability  $Q$  is called the distribution of  $x$ . This use of the word distribution can be confusing, because the word distribution is used also to refer to a probability on the set of states  $S$ .

Now suppose that  $S$  is infinite and let  $(S, \mathfrak{S}, P)$  be a probability space. In this context, it is required that a random variable be measurable or measurable with respect to  $\mathfrak{S}$ , which means that if

$$I_{a,b} = \{y \in \mathbb{R}^N \mid a_n \leq y_n \leq b_n, n = 1, \dots, N\},$$

where  $a$  and  $b$  are vectors in  $\mathbb{R}^N$ , then  $x^{-1}(I_{a,b}) = \{s \in S \mid x(s) \in I_{a,b}\}$  belongs to  $\mathfrak{S}$ .

The set of Borel subsets of  $R$  or  $R^N$  is the smallest  $\sigma$ -field,  $\mathfrak{B}$ , in  $R$  or  $R^N$  containing all the intervals of the form  $I_{a,b}$ . Since  $\mathfrak{S}$  is a  $\sigma$ -field,  $x$  is measurable if and only if  $x^{-1}(B) \in \mathfrak{S}$ , for every  $B \in \mathfrak{B}$ . The distribution of the random variable  $x$  is the probability  $Q$  on  $\mathfrak{B}$  defined by  $Q(B) = P(x^{-1}(B))$ , for every  $B \in \mathfrak{B}$ . In order to know the probability distribution  $Q$ , it is enough to know  $Q\{y \in R^N \mid y \leq a\}$ , for every  $a \in R^N$ . If  $N = 1$ , the function  $F : R \rightarrow [0, 1]$  defined by  $F(r) = Q\{y \in R \mid y \leq r\} = P\{s \in S \mid x(s) \leq r\}$  is called the cumulative distribution function of the random variable  $x$ . Notice that  $F$  is non-decreasing,  $\lim_{r \rightarrow -\infty} F(r) = 0$  and

$\lim_{r \rightarrow \infty} F(r) = 1$ . Also  $F$  may have points of discontinuity where it jumps upward, but it is continuous from the right in the sense that  $\lim_{r \rightarrow a^+} F(r) = F(a)$ , for all  $a$ . If  $F$  is continuous, then

it has a density, where a density is a function  $f : R \rightarrow [0, \infty)$  such that  $F(a) = \int_{-\infty}^a f(r) dr$ , for all  $a$ . Similarly a density function on  $R^N$  for a probability  $Q$  on the Borel subsets of  $R^N$  is a function  $f : R^N \rightarrow [0, \infty)$  such that

$$Q(I_{a,b}) = \int_{a_1}^{b_1} \int_{a_2}^{b_2} \dots \int_{a_N}^{b_N} f(r_1, r_2, \dots, r_N) dr_1 dr_2 \dots dr_N,$$

for all intervals  $I_{a,b}$ . If  $Q$  is the distribution of a random variable  $x : S \rightarrow R^N$ , then  $f$  is said to be the density of  $x$ .

When one considers one random variable in isolation, it may be convenient to take the set of states  $S$  to be the same as the range space of the random variable, whether it be  $R$  or  $R^N$ . When one compares more than one random variable, it may be convenient to suppose that they all are functions on a state space  $S$  that is different from the range space.

Example: The normal or Gaussian distribution on  $R$  has density

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

where  $x$  varies over  $R$ . This distribution is called  $N(\mu, \sigma^2)$  for short. The multivariable normal or Gaussian distribution over  $R^N$  has density

$$f(x_1, \dots, x_N) = f(x) = \frac{1}{(2\pi)^{N/2} \sqrt{\det V}} e^{-\frac{(x-\mu)^T V^{-1} (x-\mu)}{2}},$$

where  $V$  is a positive definite  $N \times N$  matrix,  $\mu$  is an  $N$ -vector, and  $x$  varies over  $R^N$ .

I next define the expected value and variance of a random variable. If  $S$  is a finite set with probability distribution  $P: S \rightarrow [0, 1]$  and  $x: S \rightarrow \mathbb{R}$  is a random variable, the expected value or mean of  $x$  is

$$E_x = \sum_{s \in S} P(s) x(s).$$

The variance of  $x$  is

$$\text{var}(x) = E(x - E_x)^2 = \sum_{s \in S} P(s) (x(s) - E_x)^2.$$

The mean of  $x$  is a central point of its distribution and the variance of  $x$  measures how much  $x$  varies around this central point.

Example: In the example of one binomial trial, let  $x$  be the random variable that is 1 in the case of a success and  $-1$  in the case of a failure. That is, the state space is  $S = \{A, F\}$ , and  $x(A) = 1$  and  $x(F) = -1$ . Then

$$E_x = p - (1 - p) = 2p - 1,$$

and

$$\text{var } x = p(1 - 2p + 1)^2 + (1 - p)(-1 - 2p + 1)^2 = 4p(1 - p).$$

In order to define the mean and variance when the state space has infinitely many points, I need to define a new notion of integral more general than the Riemann integral. Suppose that we have a probability space  $(S, \mathfrak{S}, P)$ . If  $A \in \mathfrak{S}$ , then the indicator function of  $A$  is  $\mathcal{X}_A: S \rightarrow \mathbb{R}$ , where

$$\mathcal{X}_A(s) = \begin{cases} 1, & \text{if } s \in A \text{ and} \\ 0, & \text{otherwise.} \end{cases}$$

Define the integral of  $\mathcal{X}_A$  to be  $\int_S \mathcal{X}_A(s) P(ds) = P(A)$ . A random variable  $x: S \rightarrow \mathbb{R}$  is said to be

simple if  $x(s) = \sum_{n=1}^N a_n \mathcal{X}_{A_n}(s)$ , where  $a_1, \dots, a_N$  are numbers and  $A_1, \dots, A_N$  belong to  $\mathfrak{S}$ . The

integral of a simple function  $x(s) = \sum_{n=1}^N a_n \mathcal{X}_{A_n}(s)$  is defined to be

$$\int_S x(s) P(ds) = \int_S \sum_{n=1}^N a_n \mathcal{X}_{A_n}(s) P(ds) = \sum_{n=1}^N a_n P(A_n).$$

A sequence of simple functions  $x_M$  converges to  $x$  if  $\lim_{M \rightarrow \infty} x_M(s) = x(s)$ , for  $s$  belonging to a set in

$\mathfrak{S}$  of probability 1. Define  $x$  to be integrable if  $\lim_{M \rightarrow \infty} \int_S x_M(s) P(ds)$  converges and converges to the same limit for every sequence of simple functions  $x_M$  that converges to  $x$ . This limit is denoted  $\int_S x(s) P(ds)$  and is called the Lebesgue integral of  $x$ . If  $A \in \mathfrak{S}$  and  $x$  is an integrable random variable, the integral of  $x$  over  $A$  is defined to be  $\int_A x(s) P(ds) = \int_S \chi_A(s) x(s) P(ds)$ . The expected value or mean of a random variable  $x : S \rightarrow R$  is defined to be  $Ex = \int_S x(s) P(ds)$ , and the variance of  $x$  is  $\text{var}(x) = E(x - Ex)^2 = \int_S (x(s) - Ex)^2 P(ds)$ .

It follows from the definition of the expected value that if  $a$  and  $b$  are numbers and  $x$  and  $y$  are random variables from  $S$  to  $R$ , then

$$E(ax + by) = aEx + bEy.$$

Notice that

$$\begin{aligned} \text{var}(x) &= E(x - Ex)^2 = E[x^2 - 2(Ex)x + (Ex)^2] \\ &= E(x^2) - 2(Ex)(Ex) + (Ex)(Ex) = E(x^2) - (Ex)^2. \end{aligned}$$

Also if  $a$  is a number, then

$$\text{var}(ax) = a^2 \text{var}(x).$$

If  $S$  is an infinite set and  $x : S \rightarrow R$  is a random variable,  $Ex$  may not exist or may be infinite and  $\text{var}(x)$  may be infinite even if  $Ex$  is finite.

Example: If  $x : S \rightarrow R$  is a random variable with normal distribution over its range space  $R$  with density

$$f(r) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(r-\mu)^2}{2\sigma^2}},$$

then

$$Ex = \int_{-\infty}^{\infty} \frac{r}{\sigma\sqrt{2\pi}} e^{-\frac{(r-\mu)^2}{2\sigma^2}} dr = \mu$$

and  $\text{var}(x) = \sigma^2$ .

If  $x$  and  $y$  are random variables from  $S$  to  $R$ , then the covariance of  $x$  and  $y$ , written  $\text{cov}(x, y)$ , is defined as

$$\text{cov}(x, y) = E[(x - Ex)(y - Ey)].$$

Hence  $\text{cov}(x, x) = \text{var}(x)$ .

Example: Suppose that the random variables  $x_n : S \rightarrow \mathbb{R}$ , for  $n = 1, \dots, N$ , have a multivariate normal distribution with density

$$f(x_1, \dots, x_N) = f(\mathbf{x}) = \frac{1}{(2\pi)^{N/2} \sqrt{\det V}} e^{-\frac{(\mathbf{x} - \boldsymbol{\mu})^T V^{-1} (\mathbf{x} - \boldsymbol{\mu})}{2}},$$

where  $V$  is a positive definite  $N \times N$  matrix. Then  $\text{cov}(x_n, x_m) = v_{nm}$ , where  $v_{nm}$  is the  $(n, m)$ th entry of the matrix  $V$ . Therefore  $v_{nn} = \text{var}(x_n)$ , for all  $n$ .

I now define stochastic independence of events and random variables. The intuition is that two events are independent if the occurrence of one has no bearing on the likelihood of the other. For instance, if you flip a coin twice in succession, the probability of heads on the second toss does not depend on whether heads occurred on the first toss. The probability should be one half. If the coin is truly fair and the successive tosses are not linked in any way, then the probability of heads on the 101st toss should be one half, even if all the first 100 tosses came up heads.

Let  $(S, \mathfrak{S}, P)$  be a probability space. Two events  $A$  and  $B$  in  $\mathfrak{S}$  are said to be stochastically independent if  $P(A \cap B) = P(A)P(B)$ . For instance, suppose that  $A$  is the occurrence of heads on the first of two tosses of a fair coin and let  $B$  be the occurrence of heads on the second toss. If the tosses are truly independent, then the probability of heads on both tosses, the event  $A \cap B$ , is  $1/4$ , which is the product of the probabilities of  $1/2$  of having heads on each toss.

If  $A$  and  $B$  are events in  $\mathfrak{S}$  such that  $P(B) > 0$ , the probability of  $A$  conditional on  $B$  is defined as  $P[A|B] = \frac{P(A \cap B)}{P(B)}$ . If  $A$  and  $B$  are stochastically independent, then  $P[A|B] = P(A)$ , so that knowledge that  $B$  occurs has no bearing on the probability of  $A$ .