DISCOUNTED PROBLEMS/BOUNDED COST

started okay so this is the entire theory of discounted problems in shorthand without groups of course

下面整理一下简洁符号表示的折扣问题的整个理论

just let me remind you the problems we the the infant correlation discounted problem involves a stationary system a cost per stage the stationary and in the involves a discount factor less than one of course is bounded

我来重复一下这个问题，我们有一个包括平稳系统，每阶段的成本也是平稳的，折扣因子小于1并且有界（这个有界指的是总成本和每期成本）的无限期折扣问题

so this is not this is a real function that takes real values for every initial state

所以对于每一个初始状态，成本函数都是一个实函数

and what we want to do is find a policy that minimizes this this cost function

我想要做的就是找到一个策略最小化这个成本函数

and here's our shorthand notation this is the cost per stage the one stage cost and this is for any J this mapping produces another map in T so J by taking the minimum this expression which is really the dynamic programming type of minimization

这是我们的简洁符号的表示形式，这个是每阶段的成本，使用T可以把任意一个J映射到另一个函数上去，通过最小化这个表达式就可以求解，这就是典型的动态规划最小化过程

and for any policy you remember stationary policies the one that applies the same u for every stage we have another mapping where there is no minimization but instead of u will plug in with like mu of X the the control dictated by this policy

对于任意一个平稳策略(每个阶段都使用同一个控制u的策略)，我们有另一种映射，他不是最优的，但是我们用它（mu(x)）代替u进行决策

“SHORTHAND” THEORY – A SUMMARY

so we have these two mappings and the entire theory is described by is given in this slide here there

我们有了这两种映射，整个理论就在这个slide上被展示出来

is this bellman equation J star satisfies a certain equation for every X is an operator equation okay

这是bellman方程，对于任意的x，J\*满足这个方程，这也被叫做operator equation(个人理解是算子方程)

the entire function is J star is obtained by passing J star through teeth fixed point of T and similarly for each policy the cost of the policy is a fixed point of this mapping optimality condition value duration policy duration

这个方程表示，J\*是一个算子T的不动点，相似地，对于每一个策略来说，策略的最优成本也是一个满足最优条件的不动点，这可以通过值迭代和策略迭代计算出来

TWO KEY PROPERTIES

so now to use effectively this mappings T and Tim you we have to recognize some of the major properties and there are two major properties of these mappings

想要使用这个关于T和T\_mu的映射，你必须知道一些主要的性质，这下面就是这些映射的主要性质

the first one is monotonicity for any J and J prime such that J prime is bigger than J okay if you apply the map in T to them inequality is preserved so if J prime is like this and J is like that apply T to them they both change but their order is preserved for every X this is called monotonicity and it applies for T\_mu and for T and this is extremely simple to see the reason is that J is multiplied with a positive number in this J in this TNT new mappings

第一个是单调性，对于任何J和J’，如果有J’大于J，你在使用T进行映射的时候这个不等式关系依然能够被保存下来，所以如果J’在这里，J在这里，对他们使用T的时候他们都被改变，但是他们的大小关系没有变，对任意一个x这种关系都存在，这就是单调性，这种性质对于T\_mu也是成立的，这很简单，想象J乘以一个正整数你就懂了

going back to okay notice the J's multiplied by alpha which is positive and is the plus sign here so if I increase J take it up to a J prime then the right hand side is going to go up the left hand side is also going to go up and similarly for Team u monotonicity is a universal property in dynamic programming problems it holds for all problems not just discounted by any kind of problem dynamic programming problem turns out to have such a monotonicity property it's very fundamental

回到上面的slide，这个J乘以alpha是一个正数(因为他前面是一个加号)，如果我让J的值变大到J’，等号右边的值就会变大，等号左边的值同样也会变大，同样的原因，这个性质对T\_mu也是成立的，单调性是一个动态规划中很普遍的性质，不仅是折扣问题，他对于所有动态规划问题都成立，

here's another property that sort of specific to to discount problems the constant shift property okay if I take J and I increase it uniformly by a constant then TJ is also going to increase uniformly by a constant except that it will be multiplied by alpha and similarly for teen you again if you I ball this team you mapping and T if I if I add a constant R to J then you're going to get a constant alpha R that can be taken out of the braces and the minimization so again a trivial property this constant shift there are variations of the constant shift property for other dynamic programming problems this is shorthand he is the unit function okay again shorthand notation that you may need to eyeball a little bit more closely to get familiar with

这个是第二个性质，他比较特殊，只针对折扣问题，叫做恒定位移属性(直译)，如果我使用一个常数均匀增加J的值，那么TJ的值也会均匀的增加一个常数，这个数等于alpha乘以J增加的常数，同样对于T\_mu也是成立的，如果我给J增加一个常数r，并且把他移出括号，你会看到alpha乘以r而不是r，然后就可以正常最小化了。这是一个比较小的性质，对于其他类型的动态规划问题，有不同种类的恒定位移性质。这里是shorthand的写法，这里有一个单位函数，这种形式你需要多看几次熟悉熟悉

so monotonicity is present in all dynamic programming problems constant shift is special to discounted problems models and it turns out that discounted problems have another property of major importance not all models have this property but that's why they start the problems are so is so much easier because the T mapping and the T new mapping are contraction mappings in a mathematical sense given two functions when you apply T to them their distance shrinks okay that's what contraction means and we will show this later

单调性在所有动态规划问题中都成立，恒定位移性质只对折扣问题成立，实际上折扣问题还有其他很重要的而其他类型的问题不具有的性质，这就是为什么折扣问题比其它问题简单。比如收缩性质，T映射和T\_mu映射在数学角度上是一种收缩映射，给定两个函数，你使用T映射作用在他们两个上，他们的距离会收缩（变短），这就是收缩定理，我们一会会继续讲

CONVERGENCE OF VALUE ITERATION

so now we start proving the basic theory it's not going to be hard

现在我们开始证明这几个基本的理论，一点也不难

first of all let's show that value iteration converges

首先我们证明值迭代的收敛性

starting with a zero function and applying DP to it many many times gives you the optimal cost function at the end

从零函数开始，使用DP计算很多次，最后就可以获得最优成本函数

and here's the proof

下面是证明

for any initial state x and policy PI infinite horizon policy the cost is by definition given like so okay

有一个初始状态x和一个无限期策略pi，成本这样定义

now I break this cost into the cost of the first K stages and the remainder the tail portion the tail portion is from K to infinity

我先把函数J分解成两部分，前k的阶段和剩下的从k到无穷阶段

now the interesting thing is that this trade portion because of the discount factor here becomes negligible as time goes to infinity

一个很有意思的现象是，后面这个部分由于时间趋于无穷，对总成本的影响变得几乎没有了

because it's multiplied by alpha to the K here which goes down to zero geometrically

因为第一项乘以alpha的k次方，后面的项呈几何级数减少到0

so the tail portion satisfies because this G is bounded by M and the remainder of the of the series adds up to alpha K over 1 minus alpha it satisfies this inequality so the state portion goes to 0 as K goes to infinity

因为尾问题由于g的上界是M，剩下的部分累加到一起满足这个不等式，所以后面这项随着k趋于无穷减少到0

and take the minimum here I bound this by this expression

在最小化总成本的时候，第二部分被不等式限制住了上界

take the minimum here over PI and the minimum here over PI the minimum here over PI gives me simply J times T times J times T times the zero function

对等号两边求关于pi的最小值，等号右边等于J0不停地乘以T

so and after I take the minimum ok so I get the j star equals less or equal to P to the K of J 1 plus something that's negligible

这样获得的最小值小于等于T^k J\_0加上那个几乎不造成影响的项

and now take the limit as K goes to infinity and I get this result okay

k趋于无穷的时候计算右式的极限就可以得到J的数值

I mean here mean here take the limit as this is negligible

第二项的极限对整个成本没有影响

this side gives you this this side gives you that

这两项是这么对应的

it's really a very simple proof after you look at it for a little while it becomes very simple

看完这些之后，你就知道证明非常简单

so that's it

证明就这样，讲完了

BELLMAN’S EQUATION

now Bellman's equation the optimal cost function satisfies this expression

bellman方程，成本函数满足这个条件

in other words J stars of X equals min expected value G Plus alpha g star J sir

换种说法，J\*(x)=min E[g+alpha\*J\*]

and here's the proof okay

这是bellman方程的证明

we shown in the previous slide that the case stages cost is obtained is bounded by the infant horizon cost and there's also a tail portion that has this form

之前的slide展示的，成本被无限期成本和尾问题(子问题)提供上下界

apply now T to this relationship now T is monotone applied T to this side is going to give me something that was still bigger

现在把T放到这个不等式里，由于T的单调性，中间那一项还是比左边的项更大

by applying T to this side so what we get is applying T I get this applying T by monotonicity inequalities preserve and by the constant shift property alpha to the K becomes alpha K plus 1 here okay

把T放到右边的项里，由于单调性，不等关系依然成立，然后根据恒量平移性质,把它展开，alpha^k变成了alpha^{k+1}

take now the limit as K goes to infinity this thing goes away this thing converges to J star so I have J star bounded between T sub J star on one side and P sub J star on the other side

在k趋于无穷的时候这几项都在变，中间那一项收敛于J\*，左右两边的界也就被确定了

and that's that using the convergence of value direction this goes to that this goes away and we have this expression that's it again very simple just so big of a base of fundamental properties okay

值迭代收敛之后，中间的一项变成了J\*，右边那项的第二个元素没有了，这就是我们的证明过程很简单但是但是非常基本的性质

THE CONTRACTION PROPERTY

I'm going to go to the third property and I'll give you a chance to ask questions contraction property

下面我要讲第三个性质，收缩性，在这里会给你们机会提问题

given any two bounded functions J and J Prime I have that this expression here which is the soup norm the maximum norm of TJ - piece of J piece of J prime is less than alpha the soup norm of J minus J prime okay so look at what this says view this as a measure of proximity of J and J prime apply t to both of them then the distance between them shrinks this measure of prescribing the strings by a factor alpha similarly for T\_mu

给定两个有界函数J和J\*，我写出这样的表达式，max|TJ-TJ‘|小于alpha max|J-J‘|，他想表达的就是两个接近的度量J和J\*，在使用T映射之后新的函数距离会比alpha折扣后的函数距离更近对于T\_mu也有同样的性质

and the proof of this again two lines denote by C the maximum over all states the supreme north here the maximum X

我要从两个方向证明这个东西，定义c是所有状态下J-J’的最大值

then J prime is bounded between J plus C the upper bound of the distance between them and J prime J - minus C

然后J’就被限制在了J-c和J+c之间

apply T to both sides and apply and use monotonicity use also the constant shift property then apply T gives you this applying T to this side gives you this because of the constant shift property if I change this expression by a constant then after applying T it changes by alpha times the constant and similarly here and now bring this in here and we have that the distance between these two is less than alpha C and that's that

对这个不等式使用算子T，采用单调性和压缩性，可以得到下面的不等式，也就是说，给中间的那项套上绝对值之后这个值小于alpha乘以c，证明就这么结束了

NEC. AND SUFFICIENT OPT. CONDITION

okay so we have that in now the last property

下面我要讲最后一个性质了

a stationary policy is optimal if and only if mu of X attains the minimum in balance equation

如果一个平稳策略是最优策略，只有mu(x)满足最优性方程的时候才成立

in other words in abstract notation this holds or equivalently in more in more expanded notation when you minimize the right-hand side of balance equation involving change star and you take the minimum over u then you get mu of X

换句话说，在抽象记法中，你最小化等号右边的bellman方程获得了一个解集，如果对于所有的状态x，mu(x)都在都在这个解集中那么这个策略也是最优策略

so if you know J star by minimizing here you get an optimal policy conversely if you have a policy that satisfies this minimization then it is optimal it's a guarantee that's optimal it's an if and only if relationship

所以如果你知道J\*的值，你最小化等号右边的表达式也可以获得最优策略，如果一个策略满足这个关系，也可以说他是最优的

to write down analytically the proof using this XT the expanded notation takes perhaps a couple of pages here the proof is a couple of lines couple of short paragraphs askew well

如果使用扩展符号写这个证明过程，需要好多页，但是这里只用了几行就完成了

here's the proof let's assume first that mu attains a minimum here so this relationship holds then I want to show that new is optimal okay

这就是证明过程，我们假设是优化后的策略，也就是mu满足这个关系，下面我要说明这个mu是最优的

now using Bellman's equation we have that this is equal to j star i have this expression but this says that J star is a tricks point of the mapping T\_mu and T\_mu has as its unique fixed point J\_mu because of well it's it's part of the bellman equation here so new is optimal

使用bellman方程，我们可以写出这个等式，等号左边等于J\*，我们知道J\*是这个映射的不动点，J\_mu是T\_mu的不动点，他们都是bellman方程的一部分，所以这个mu是最优策略

conversely suppose that you have an optimal policy I want to show that that takes the minimum here

相反地，假设你知道一个策略mu是最优策略，我要说明他满足上面的表达式

so if it's optimal then J star is equal to J u therefore I have because J nu is a fixed point of t new I have J star equals to T mu J start and combining this with Belmonts equation because J star is equal to this I have that TJ star is equal to mu J star which is what I wanted to prove

如果mu是最优策略，那么J\*就等于J\_mu，同时，J\_mu是T\_mu的不动点，所以J\*等于T\_mu J\*，他们构成了bellman方程，因为J\*等于TJ\*，所以我可以知道TJ\*=T\_muJ\*，证明就结束了

Q&A

again a few lines may seem like they carry but it's not it's correct okay you need to get used a little bit to the abstract notation so now we have the entire theory at least we'll accept for algorithms for infinite varieties and discounted problems and you may want to ask questions about this this about this mapping century and in the proofs that I have given is it clear what these theorems say perhaps you can digest the proofs a later but it's important to understand what the theorem say I have no more slides time for questions there are different equations this equation perhaps I can write actually on the blackboard you can turn on the light please okay Delmas equation is this okay now this is equivalent to if I write it in expanded form it's equivalent to okay minimize over you expected value over W and this holds true for all X so it's a functional equation okay if X takes a finite number of values then this becomes a nonlinear system of a final number ability if X takes n values okay state of an end both of its an N state Markov chain that you have then this becomes a system of n nonlinear equations with n unknowns okay and the unknowns are this J star values J star then becomes a vector of n components so this is a long hand expression of this if you have an infinite number of states then this becomes a functional equation like a partial differential equation you know up an equation with an infinite number of variables okay so this is Bellman's equation it's the equation it has a unique solution for the infinite horizon discounted problem within the class of all bounded functions and now if you write we if you write for a given policy if you write this equation now this has a similar character but it gives us unknowns the components of J nu nought of J star and equivalently this is written as J mu sub X equal to expected value over W G of X muse of X W plus alpha J mu f of X news of X so these two equations are have a similar form but they are different and they have different unknowns this has as unknowns the optimal costs for the corresponding starting States and this has B as cost attached a solution the cost of the policy new starting from all possible states the malady condition I don't know what you mean by optimal you mean this is the optimality condition when you say optimality conditions you mean optimality condition of what optimality of a policy you know what what's optimal okay so that's what this says here it says that this minimum here is attained for new okay so if you combine these two and they are really identical then you have an optimal policy so for an optimal policy these two become the same yeah perhaps you that's what you meant but for an optimal policy they become the same but generally they are not the same okay so this is the theory here and and it's important for algorithms now in in the next lecture on Wednesday we are going to talk about algorithms particularly this policy iteration algorithm and also discuss the related models we're going to focus also a little bit on value iteration and then start talking about approximations to these algorithms which solve approximately the problem-solver cups approximately this bellman equation or find approximate approximately optimal policies so that's where we're headed exhaust the subject of exact over the next covered including next time and then we go into approximation bounded operator okay perhaps a rounded operator but you know the term bounded is used in a different context that's more specialized so you have to have a and a normed vector space perhaps and I don't think that such a connection has been useful in I think what's important here is that the T operator is a contraction yeah yeah I see what you mean what you say is here is look at this you take a look at this iteration and you take it to infinity and eventually you get J stark which is a fixed point you have okay you have the finite horizon dynamic programming equation if you if you take the limit of that when you get balanced equation so it made one may be able to derive it or understand this bellman equation the way you describe but the way that I like to understand it better is that p is a is a contraction mapping and it has a home unique fixed point this is a classical theorem that that sort of relatively simple to prove before it sells it escaped optimality condition that maybe show the existence of the finality point so they show the existence of the six points yes yeah it's useful to know that there exists a unique solution within the class of bounded function so it gives you existence as well as uniqueness that's right the contraction property its existence as well as uniqueness although we do know that the optimal cost function is bounded from elementary arguments by the fact that the the cost of every policy J PI is bounded therefore be how you take the minimum it's also you get also a bounded function yes I don't understand the difference because I don't understand the other technique very well okay yeah there are many schools of several schools I would say for example the people artificial intelligence have a different mentality than people that are in operations research in control theory if I could make a superficial distinction between different schools one is the artificial intelligence schools the other one is the operations research school and the other one is the control engineering or control fielding school and I think there's also another line of research called ADP but not approximate dynamic programming but adaptive dynamic program it's actually quite popular here in China because because because because it applies to control systems in the context of automation robotics and in manufacturing and process control and people like that and it comes from a tradition started by Weber's Paul warbles and it's connected to nonlinear control theory adaptive control and the Frank Lewis for example has done a lot of work his own ping he's earned whom I met actually last week is also another major player in this area and so there are several different lines of research now what's the difference between control theory line of operations research I think that work is basically I think what major part of it has to do with the type of applications that that various various schools of thought consider like Warren Powell for example there's a lot of work with logistics scheduling fleets of trucks to do deliveries and to do things and and there are people also who are more in the traditional decision and control framework like yourself and other people that go to the CDC which have more of a mixture between operations research and and and control theory mentality I think this is good because it provides greater cross fertilization of ideas but if you put together some people from our professional intelligence and people operations research you'll find that the commonality of thinking is minimal and they don't quite understand each other's problems and techniques I think it may be true also of control theory people and operations research in operations research took its people in control people and in field and efficient intelligent people in learning talk about talk about not so much about algorithms but but about learning toe fury look at a lot any other questions okay so see you on Wednesday and I'm told that we're going to have a new room likely so look out for announcements and the check the webpage and Sam will may have more to say about that