Regular Policies in Stochastic Optimal Control and Abstract Dynamic Programming

Dimitri P. Bertsekas

Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology

Conference in honor of Steven Shreve

Carnegie Mellon University
June 2015

Classical Total Cost Stochastic Optimal Control (SOC)

System: $x_{k+1} = f(x_k, u_k, w_k)$

- x_k: State at time k, from some space X
- u_k: Control at time k, from some space U
- w_k: Random "disturbance" at time k, from a countable space W, with $p(w_k \mid x_k, u_k)$ given

Policies: $\pi = \{\mu_0, \mu_1, \ldots\}$

- Each μ_k maps states x_k to controls $u_k = \mu_k(x_k) \in U(x_k)$ (a constraint set)
- Cost of π starting at x_0, with discount factor $\alpha \in (0, 1]$:

  $J_\pi(x_0) = \limsup_{N\to\infty} E\left\{\sum_{k=0}^{N} \alpha^k g(x_k, \mu_k(x_k), w_k)\right\}$
- Optimal cost starting at x_0: $J^*(x_0) = \inf_\pi J_\pi(x_0)$
- Optimal policy π^*: Satisfies $J_{\pi^*}(x) = J^*(x)$ for all $x \in X$

Bellman's (Optimality) Equation:

$$J^*(x) = \inf_{u\in U(x)} E\Big\{g(x,u,w) + \alpha J^*\big(f(x,u,w)\big)\Big\}, \qquad \forall\, x \in X$$

Three Main Classes of Total Cost SOC Problems

Discounted:

- $\alpha < 1$ and bounded g
- Dates to 50s (Bellman, Shapley)
- Nicest results; key fact is contraction property in Bellman's equation

Undiscounted ($g \leq 0$ or $g \geq 0$):

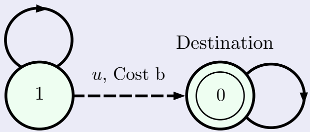
- N-step horizon costs are going $\downarrow$ or $\uparrow$ with N
- Dates to 60s (Blackwell, Strauch); positive and negative DP
- Not nearly as powerful results compared with the discounted case

Stochastic Shortest Path (SSP):

- Dates to 60s (Eaton-Zadeh, Derman, Pallu de la Barriere)
- Also known as first passage or transient programming
- Aim is to reach a special termination state at min expected cost
- Under favorable assumptions (including finite state space), results are almost as strong as for the discounted case (some contraction properties)
- In general, very complex behavior is possible

Complexities of Noncontractive Problems with $g \geq 0$ or $g \leq 0$

A deterministic shortest path problem



Bellman's equation: $J(1) = \min\{b + J(0),\ J(1)\},\quad J(0) = J(0)$
Solutions with $J(0) = 0$: All $J(1) \leq b$

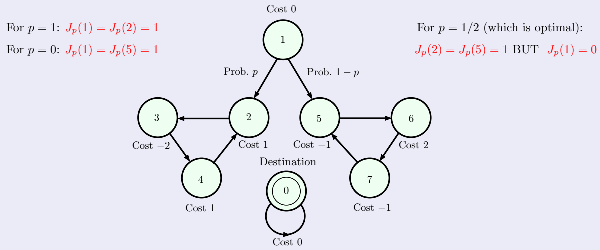
Value iteration (VI) starting from any J_0 with $J_0(0) = 0$

- VI for the terminating policy: $J_{\mu,k}(1) = b$ (works)
- VI for the nonterminating policy: $J_{\mu',k+1}(1) = J_{\mu',k}(1)$ (fails)
- VI for the entire problem: $J_{k+1}(1) = \min\{b, J_k(1)\}$
- If $b < 0$: $J_k(1) \to J^*(1)$ starting with $J_0(1) \geq b$ (works depending on J_0)
- If $b > 0$: $J_k(1) \to J^*(1)$ only if $J_0(1) = 0$; starting from $J_0(1) \geq b$, $J_k(1) \to J_\mu(1)$

Policy iteration (PI) starting from μ

- If $b < 0$: Oscillates between μ and μ'. If $b > 0$: Converges to suboptimal μ

A stochastic shortest path problem (from Bertsekas and Yu, 2015)



- The Bellman Eq. is violated at 1 for $p = 1/2$: $J_p(1) \neq pJ_p(2) + (1-p)J_p(5)$
- Mathematically, the difficulty is that $\limsup E\{\cdot\} \neq E\{\limsup \{\cdot\}\}$

Consider the deterministic problem that chooses either $p = 1$ or $p = 0$

- Belman's equation $J^*(1) = \min\{J^*(2), J^*(5)\}$ is satisfied
- Introducing randomization
  - Lowers the optimal cost and invalidates Bellman's equation
  - VI fails to converge to J^* from any initial condition

What is the Root of the Anomalies?

A (partial) answer

The presence of policies that are not well-behaved in terms of VI (e.g., involve zero length cycles)

We call these policies “irregular" and we investigate

- What problems can they cause?
- Under what assumptions are they “harmless"?

References

- D. P. Bertsekas, Abstract Dynamic Programming, Athena Scientific, 2013. (Regularity introduced in the context of semicontractive models, i.e., models where some policies involve contraction-like properties, and some do not.)
- D. P. Bertsekas, “Regular Policies in Abstract Dynamic Programming," Lab. for Information and Decision Systems Report LIDS-P-3173, MIT, May 2015.
- D. P. Bertsekas, “Value and Policy Iteration in Optimal Control and Adaptive Dynamic Programming," Lab. for Information and Decision Systems Report LIDS-P-3174, MIT, May 2015.
- D. P. Bertsekas and H. Yu, “Stochastic Shortest Path Problems Under Weak Conditions," Lab. for Information and Decision Systems Report LIDS-P-2909, MIT, August 2013 (revised March 2015).
- H. Yu and D. P. Bertsekas, “A Mixed Value and Policy Iteration Method for Stochastic Control with Universally Measurable Policies," Lab. for Information and Decision Systems Report LIDS-P-2905, MIT, July 2013.

Outline

Regularity: A Summary of Ideas

S-Regular stationary policy μ (S is a set of “value" functions on X)

μ is S-regular if it behaves well with respect to VI when started from S, i.e., if VI using μ converges to J_μ starting from all $J \in S$

Extension: S-Regular set of policy-state pairs

A set $\mathcal{C}$ of policy-state pairs (π, x) is S-regular if for all $(\pi, x) \in \mathcal{C}$, VI using π and starting from x converges to $J_\pi(x)$ starting from all $J \in S$

Key idea: Exclude the irregular pairs (i.e., optimize over the S-regular set)

- The (restricted) optimal cost function,

$$J^*_{\mathcal{C}}(x) = \inf_{(\pi,x)\in\mathcal{C}} J_\pi(x),$$

  may be the unique solution of Bellman's equation within S, while J^* may not be!
- This is an interesting and (possibly) better-behaved problem
- Also $J^*_{\mathcal{C}}$ may be obtained by VI starting from within S

Regular Collections of Policy-State Pairs

Definition: For a set of functions $S \subset E(X)$ (the set of extended real-valued functions on X), we say that a collection $\mathcal{C}$ of policy-state pairs (π, x_0) is *S-regular* if

$$J_\pi(x_0) = \limsup_{N\to\infty} E\left\{\alpha^N J(x_N) + \sum_{k=0}^{N-1} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big)\right\}, \qquad \forall\, (\pi, x_0) \in \mathcal{C},\ J \in S$$

Notes:

- Interpretation: Addition of a terminal cost function $J \in S$ does not matter in the definition of $J_\pi(x_0)$
- Example: $\alpha = 1$ and $J \in S$ are s.t. $J(x_k) \to 0$ for generated $\{x_k\}$ under π
- Example: $\alpha < 1$ and $J \in S$ are s.t. $\{J(x_k)\}$: bounded for generated $\{x_k\}$ under π
- For $(\mu, x) \in \mathcal{C}$ with μ stationary: $J_\mu(x)$ is obtained by VI starting with any $J \in S$
- A set $\mathcal{C}$ of policy-state pairs (π, x) may be S-regular for many different sets S

Optimal cost function over regular collections

$$J^*_{\mathcal{C}}(x) = \inf_{\{\pi \mid (\pi, x) \in \mathcal{C}\}} J_\pi(x), \qquad x \in X$$

Abstract Notation - Connection with Abstract DP

- Mapping of a stationary policy μ: For any control function μ, with $\mu(x) \in U(x)$ for all x, and $J \in E(X)$ define the mapping $T_\mu : E(X) \mapsto E(X)$ by

$$(T_\mu J)(x) = E\big\{g(x, \mu(x), w) + \alpha J\big(f(x, \mu(x), w)\big)\big\}, \qquad x \in X$$

- Value Iteration mapping: For any $J \in E(X)$ define the mapping $T : E(X) \mapsto E(X)$

$$(TJ)(x) = \inf_{u \in U(x)} E\big\{g(x, u, w) + \alpha J\big(f(x, u, w)\big)\big\}, \qquad x \in X$$

- Note that Bellman's equation is $J = TJ$ and VI starting from J is $T^k J$, $k = 0, 1, \ldots$

Abstract notation relating to regularity

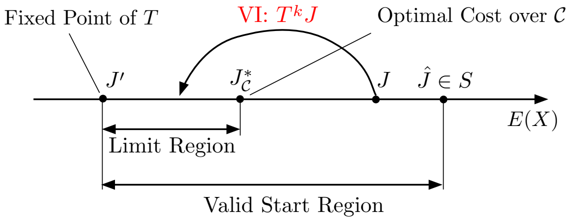
- We have

$$(T_{\mu_0} \cdots T_{\mu_{N-1}} J)(x_0) = E\left\{\alpha^N J(x_N) + \sum_{k=0}^{N-1} \alpha^k g\big(x_k, \mu_k(x_k), w_k\big)\right\}$$

- $\mathcal{C}$ is S-regular if

$$J_\pi(x) = \limsup_{N \to \infty} (T_{\mu_0} \cdots T_{\mu_N} J)(x), \qquad \forall\, (\pi, x) \in \mathcal{C},\ J \in S$$

Upper Bounding the Fixed Points of T



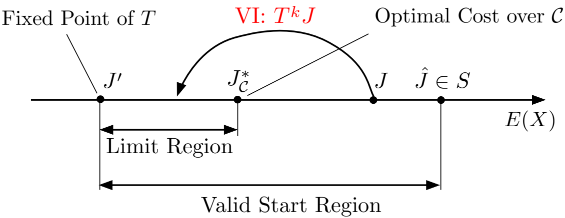
Let $\mathcal{C}$ be an S-Regular Collection

- For all fixed points J' of T, and all $J \in E(X)$ such that $J' \leq J \leq \hat{J}$ for some $\hat{J} \in S$,

$$J' \leq \liminf_{k\to\infty} T^k J \leq \limsup_{k\to\infty} T^k J \leq J^*_{\mathcal{C}}$$

- If in addition $J^*_{\mathcal{C}}$ is a fixed point of T (a common case), then $J^*_{\mathcal{C}}$ is the largest fixed point

Characterizing VI Convergence



VI-Related Properties

- If $J_{\mathcal{C}}^*$ is a fixed point of T, then VI converges to $J_{\mathcal{C}}^*$ starting from any $J \in E(X)$ such that $J_{\mathcal{C}}^* \leq J \leq \hat{J}$ for some $\hat{J} \in S$
- J^* does not enter the picture! It is possible that VI converges to $J_{\mathcal{C}}^*$ and not to J^* (which may not even be a fixed point of T)
- When J^* is a fixed point of T, a useful analytical strategy is to choose $\mathcal{C}$ such that $J_{\mathcal{C}}^* = J^*$. Then a VI convergence result is obtained

Nonnegative Cost Optimal Control

Cost nonnegativity, $g \geq 0$, provides a favorable structure (Strauch 1966)

- J^* is the smallest fixed point of T within $E^+(X)$
- VI converges to J^* starting from 0 under some mild compactness conditions

Regularity-based analytical approach

- Define a collection $\mathcal{C}$ such that $J^*_{\mathcal{C}} = J^*$
- Define a set $S \subset E^+(X)$ such that $\mathcal{C}$ is S-regular
- Use the main result in conjunction with the fixed point property of J^* to show that J^* is the unique fixed point of T within S
- Use the main result to show that the VI algorithm converges to J^* starting from J within the set $\{J \in S \mid J \geq J^*\}$
- Enlarge the set of functions starting from which VI converges to J^* using a compactness condition

We use this approach in three major applications

Application to Nonnegative Cost Deterministic Optimal Control

Classic problem of regulation to a terminal set

- System: $x_{k+1} = f(x_k, u_k)$. Cost per stage: $g(x_k, u_k) \geq 0$
- Cost-free and absorbing terminal set of states X_s that we aim to reach or approach asymptotically at minimum cost

Assumptions

- $J^*(x) > 0$ for all $x \notin X_s$
- Controllability: For all x with $J^*(x) < \infty$ and $\epsilon > 0$, there exists a policy π that reaches (in a finite number of steps) X_s starting from x with cost $J_\pi(x) \leq J^*(x) + \epsilon$

Define

- $\mathcal{C} = \big\{(\pi, x) \mid J^*(x) < \infty,\ \pi \text{ reaches } X_s \text{ starting from } x\big\}$
- $S = \big\{J \in E^+(X) \mid J(x) = 0,\ \forall\, x \in X_s\big\}$

Results

- J^* is the unique solution of Bellman's equation within S
- VI converges to J^* starting from any $J_0 \in S$ with $J_0 \geq J^*$ (and for any $J_0 \in S$ under a compactness condition)

Application to Nonnegative Cost Stochastic Optimal Control

Problem

- System: $x_{k+1} = f(x_k, u_k, w_k)$
- Cost per stage: $g(x_k, u_k, w_k) \geq 0$

Define

- $\mathcal{C} = \big\{(\pi, x) \mid J_\pi(x) < \infty\big\}$; so $J^*_{\mathcal{C}} = J^*$
- $S = \big\{J \in E^+(X) \mid E^\pi_{x_0}\big\{J(x_k)\big\} \to 0,\ \forall\, (\pi, x_0) \in \mathcal{C}\big\}$

Results

- J^* is the unique solution of Bellman's equation within S
- VI converges to J^* starting from any $J_0 \in S$ with $J_0 \geq J^*$ (and for any $J_0 \in S$ under a compactness condition)

An interesting consequence (Yu and Bertsekas, 2013)

If a function $J \in E^+(X)$ satisfies $J^* \leq J \leq cJ^*$ for some $c \geq 1$, VI converges to J^* starting from J

Application to Discounted Nonnegative Cost Stochastic Optimal Control

The problem with discount factor $\alpha < 1$

Terminology and definitions

- $X_f = \{x \in X \mid J^*(x) < \infty\}$
- π is stable from $x_0 \in X_f$ if there is bounded subset of X_f s.t. the sequence $\{x_k\}$ generated starting from x_0 and using π lies with probability 1 within that subset
- $\mathcal{C} = \{(\pi, x) \mid x \in X_f,\ \pi \text{ is stable from } x\}$
- $J \in E^+(X)$ is bounded on bounded subsets of X_f if for every bounded subset $\tilde{X} \subset X_f$ there is a scalar b such that $J(x) \leq b$ for all $x \in \tilde{X}$
- $S = \{J \in E^+(X) \mid J \text{ is bounded on bounded subsets of } X_f\}$

Assumption

$\mathcal{C}$ is nonempty, $J^* \in S$, and for every $x \in X_f$ and $\epsilon > 0$, there exists a policy π that is stable from x and satisfies $J_\pi(x) \leq J^*(x) + \epsilon$

Results

- J^* is the unique solution of Bellman's equation within S
- VI converges to J^* starting from any $J_0 \in S$ with $J_0 \geq J^*$ (and for any $J_0 \in S$ under a compactness condition)

S-Regular Collections Involving Stationary Policies

Definitions: For a nonempty set of functions $S \subset E(X)$

- We say that a stationary policy μ is S-regular if $T_\mu^k J \to J_\mu$ for all $J \in S$
- Equivalently, μ is S-regular if the set $\mathcal{C} = \{(\mu, x) \mid x \in X\}$ is S-regular
- Let $\mathcal{M}_S$ be the set of policies that are S-regular, and define

$$J_S^*(x) = \inf_{\mu \in \mathcal{M}_S} J_\mu(x), \qquad \forall\, x \in X$$

- Equivalently, $J_S^* = J_{\mathcal{C}}^*$ when $\mathcal{C} = \mathcal{M}_S \times X$

VI Convergence Result

Given a set $S \subset E(X)$, assume that

- There exists at least one S-regular policy
- J_S^* is a fixed point of T

Then $T^k J \to J_S^*$ for every $J \in E(X)$ such that $J_S^* \leq J \leq \hat{J}$ for some $\hat{J} \in S$.

Policy Iteration

Definitions:

- Standard PI: $T_{\mu^{k+1}} J_{\mu^k} = T J_{\mu^k}$
- Optimistic PI: $T_{\mu^k} J_k = T J_k$, $J_{k+1} = T_{\mu^k}^{m_k} J_k$ (evaluation of the current policy is approximate, using m_k iterations of VI)

Convergence of standard PI, assuming $J^* \geq 0$

- The sequence $\{\mu^k\}$ satisfies $J_{\mu^k} \downarrow J_\infty$, where J_∞ is a fixed point of T with $J_\infty \geq J^*$
- If for a set $S \subset E(X)$, the policies μ^k generated are S-regular and we have $J_{\mu^k} \in S$ for all k, then $J_{\mu^k} \downarrow J_S^*$ and J_S^* is a fixed point of T

Convergence of optimistic PI

- The sequence $\{J_k\}$ satisfies satisfies $J_k \downarrow J_\infty$, where J_∞ is a fixed point of T
- If for a set $S \subset E(X)$, the policies μ^k generated are S-regular and we have $J_{\mu^k} \in S$ for all k, then $J_k \downarrow J_S^*$ and J_S^* is a fixed point of T

With more analysis and conditions, we can show that $J_\infty = J^*$. This is true for the deterministic and stochastic nonnegative cost problems.

Stochastic Shortest Path Problems

Problem Formulation

- Finite state space $X = \{0, 1, \ldots, n\}$ with 0 being a cost-free and absorbing state
- Transition probabilities $p_{xy}(u)$
- $U(x)$ is finite for all $x \in X$
- No discounting ($\alpha = 1$)

Proper policies

- μ is proper if the terminal state t is reached w.p.1 under μ (is improper otherwise)
- Let $S = \Re^n$. Then μ is S-regular if and only if it is proper. (The idea of an S-regular policy evolved as a generalization of a proper policy.)

Contraction properties

- The mapping T_μ of a policy μ is a weighted sup-norm contraction iff μ proper
- If all stationary policies are proper, then T is a sup-norm contraction, and the problem behaves like a discounted problem
- SSP is a prime example of a semicontractive model (some policies correspond to contractions while others do not)

Stochastic Shortest Path Problems Results

Case where improper policies have infinite cost

If there exists a proper policy and for every improper μ, $J_\mu(x) = \infty$ for some x, then:

- J^* is the unique fixed point of T in $\Re^n$
- VI converges to J^* starting from every $J \in \Re^n$
- PI converges to an optimal proper policy, if started with a proper policy

Case where improper policies have finite cost (due to zero length “cycles")

Let $\hat{J}$ be the optimal cost function over proper stationary policies only, and assume that $\hat{J}$ and J^* are real-valued. Then:

- $\hat{J}$ is the unique fixed point of T in the set $\{J \in \Re^n \mid J \geq \hat{J}\}$
- VI converges to $\hat{J}$ starting from any $J \geq \hat{J}$
- PI need not converge to an optimal policy even if started with a proper policy
- A “perturbed" version of PI (add a $\delta_k > 0$ to g, with $\delta_k \downarrow 0$) converges to an optimal policy within the class of proper policies, if started with a proper policy
- An improper policy may be (overall) optimal, while J^* need not be a fixed point of T

Abstract DP

Main Objective

- Unification of the core theory and algorithms of total cost DP
- Simultaneous treatment of a variety of problems: MDP, sequential games, sequential minimax, multiplicative cost, risk-sensitive, etc

Main Idea

- Define a DP problem by its "mathematical signature": an abstract monotone mapping $H : X \times U \times E(X) \mapsto [-\infty, \infty]$

$$J \le J' \qquad \Longrightarrow \qquad H(x, u, J) \le H(x, u, J'), \qquad \forall\, x, u$$

  where $E(X)$ is the set of functions $J : X \mapsto [-\infty, \infty]$
- Stochastic optimal control example: $H(x, u, J) = E\big\{g(x, u, w) + \alpha J\big(f(x, u, w)\big)\big\}$
- Minimax example: $H(x, u, J) = \sup_{w \in W} \big\{g(x, u, w) + \alpha J\big(f(x, u, w)\big)\big\}$

Abstract DP Mappings

- State and control spaces: X, U
- Control constraint: $u \in U(x)$
- Stationary policies: $\mu : X \mapsto U$, with $\mu(x) \in U(x)$ for all x

Monotone Mappings

- Abstract monotone mapping $H : X \times U \times E(X) \mapsto \Re$

$$J \leq J' \qquad \Longrightarrow \qquad H(x, u, J) \leq H(x, u, J'), \qquad \forall\, x, u$$

where $E(X)$ is the set of functions $J : X \mapsto [-\infty, \infty]$

- For a stationary policy μ

$$(T_\mu J)(x) = H\big(x, \mu(x), J\big), \qquad \forall\, x \in X,\ J \in E(X)$$

and for VI

$$(TJ)(x) = \inf_{u \in U(x)} H(x, u, J), \qquad \forall\, x \in X,\ J \in E(X)$$

Abstract Optimization Problem

- Given an initial function $\bar{J} \in E(X)$ and policy $\pi = \{\mu_0, \mu_1, \ldots\}$, define

$$J_\pi(x) = \limsup_{N \to \infty} (T_{\mu_0} \cdots T_{\mu_N} \bar{J})(x), \qquad x \in X$$

- Find $J^*(x) = \inf_\pi J_\pi(x)$ and an optimal π attaining the infimum

Notes

- Theory revolves around fixed point properties of mappings T_μ and T:

$$J_\mu = T_\mu J_\mu, \qquad J^* = TJ^*$$

These are generalized forms of Bellman's equation

- Algorithms are special cases of fixed point algorithms

Principal Abstract Models

Contractive:

- Patterned after discounted
- The DP mappings T_μ are weighted sup-norm contractions (Denardo 1967)

Monotone Increasing/Decreasing:

- Patterned after positive and negative DP
- No reliance on contraction properties, just monotonicity of T_μ (Bertsekas 1977, Bertsekas and Shreve 1978)

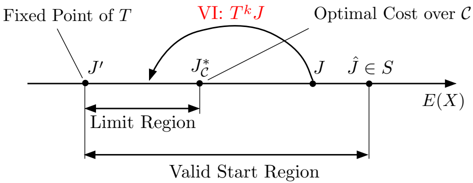
Semicontractive:

- Patterned after stochastic shortest path
- Some policies μ are “regular" (T_μ is contractive-like); others are not, but focus is on optimization over “regular" policies

Verbatim Extension of the Notion of S-Regularity

Let $\mathcal{C}$ be a collection of policy-state pairs (π, x) that is S-regular. For all fixed points J' of T, and all $J \in E(X)$ such that $J' \leq J \leq \hat{J}$ for some $\hat{J} \in S$, we have

$$J' \leq \liminf_{k \to \infty} T^k J \leq \limsup_{k \to \infty} T^k J \leq J^*_{\mathcal{C}}$$



- If $J^*_{\mathcal{C}}$ is a fixed point of T, then VI converges to $J^*_{\mathcal{C}}$ starting from any $J \in E(X)$ such that $J^*_{\mathcal{C}} \leq J \leq \hat{J}$ for some $\hat{J} \in S$
- When J^* is a fixed point of T, a useful analytical strategy is to choose $\mathcal{C}$ such that $J^*_{\mathcal{C}} = J^*$. Then a VI convergence result is obtained

Diverse Applications to Various Types of DP Problems

Bellman equation, VI, and PI analysis

- To minimax problems (also zero sum games); e.g.,

$$H(x, u, J) = \sup_{w \in W} \left\{ g(x, u, w) + \alpha J\big(f(x, u, w)\big) \right\}, \qquad \bar{J}(x) \equiv 0$$

- To robust shortest path planning (minimax with a termination state)
- To multiplicative and risk-sensitive cost functions

$$H(x, u, J) = E\left\{ g(x, u, w) J\big(f(x, u, w)\big) \right\}, \qquad \bar{J}(x) \equiv 1$$

or

$$H(x, u, J) = E\left\{ e^{g(x,u,w)} J\big(f(x, u, w)\big) \right\}, \qquad \bar{J}(x) \equiv 1$$

- More ... see the references

Thank you!