Intelligent Agents
Justification Logic

Özgür L. Özçep
Universität zu Lübeck
Institut für Informationssysteme

Todays lecture based on

- Slides of talk „Justification Logic“ of Thomas Studer, 2016
  https://home.inf.unibe.ch/~tstuder/papers/Studer_Iran_16_JL.pdf

- Slides of two talks by N. Kotsani on justification logic availabe at
  - http://corelab.ntua.gr/~nkotsani/slides/JL_session01.pdf
  - http://corelab.ntua.gr/~nkotsani/slides/JL_session02.pdf

- Parts of course CS154, „Polynomial Time with Oracles“ by Omer Reingold
  https://omereingold.files.wordpress.com/2020/10/37p-polynomial-hierarchy.pptx

MOTIVATION

Two Traditions

Modal logic adds a new connective $\Box$ to the language of logic. Two traditions:

- Epistemic logic:
  $\Box A$ means „A is known / believed"
- Proof theory:
  $\Box A$ means „A is provable in system S"

Problem with Epistemic Tradition

- We saw: defining „Knowledge is justified true belief“ according to Plato is problematic
- ⇒ Gettier Paradoxa

- No explicit treatment of justifications in modal logic

Recap: Gettiers two counterexamples

Scenario 1

- Smith and Jones apply for a job
- Smith believes (justifiably):
  (p) Jones will get the Job &
  Jones has ten coins in his pocket
- Smith believes also in the entailed assertion:
  (r) The one who gets the job has ten coins in his pocket.
- Coincidence : Smith gets the job and Smith has ten coins in his pocket.
- Smith „knew“ (r) only by chance

Scenario 2

- Smith justifiably believes
  (p) Jones owns a Ford
- Smith also believes in entailed assertion
- (r) = (p or q): Jones owns a Ford, or Brown lives in Barcelona (Though Smith has no justification for q)
- Coincidence: Jones does not own Ford, but Brown lives in Barcelona
- Smith „knew“ (r) only by chance

General idea: decouple justification and truth conditions of propositional content of belief

Recap: General remarks

- What is a justification at all?
  - „Solutions“ to Gettier‘s problem deal with this problem
  - A formal treatmant of justification similar to provability logic: Justification Logic (Artemov 2008) -> Today

- Gettier‘s problem „formalized“
  - Suppose logic of belief and justification such that
    (*) $\phi \rightarrow \psi \vdash \quad hasJust(a, \phi) \rightarrow hasJust(a, \psi)$
  - Suppose: a wrongly but justifiably believes in p
    $$\neg p \wedge B_a p \wedge hasJust(a, p)$$
  - By M(B_a): $B_a(p \vee q\ ) \wedge B_a(p \vee \neg q)$
  - By (*): $hasJust(a, (p \vee q\ )) \wedge \mathrm{hasJust}(a, (p \vee \neg q))$
  - Hence: $\vDash B_a p \wedge hasJust(a, p) \rightarrow (K_a(p \vee q) \vee K_a(p \vee \neg\, q))$

Problems of proof theoretic tradition

- $\Box \perp \rightarrow \perp$ is an axiom
- So $\neg\Box \perp$ is provable
- By necessication: $\Box\neg\Box \perp$ is provable

- $\Box \perp$ means system S proves $\perp$
- $\neg\Box \perp$ means S does not proves$\perp$, i.e.
  $\neg\Box \perp$ means S is consistent
- $\Box\neg\Box \perp$ means S proves that S is consistent

- Famous result of Gödel: if S has a certain strength, it cannot prove its own consistency

Justification Logic

- Explicitly account for justifications of assertions

Example

A ist justified with r	$r : A$
$A \to B$ is justified with s	$s : (A \to B)$
B is justified by s, r	$s \cdot r : B$

- A book-length treatment of justification logic by one of the founders (Artemov/Fitting 19)

SYNTAX AND CALCULUS

Syntax of the logic of proofs

- The logic of proofs LP_{CS} is the justification counterpart of the modal logic S4
  (this statement will be made precise in the following)
- LP was suggested by Gödel (Gödel 1995) and formalized by Artemov

Definition (Syntax LP_{CS})

- Justification terms Tm
  $t ::= x \mid c \mid (t \cdot t) \mid (t + t) \mid !t$
- Formulas $\mathcal{L}_j$

$$A ::= p \mid \neg A \mid (A \to A) \mid t{:}A$$

Axioms for LP

Definition (Axioms for LP)

- (CL) All propositional tautologies
- (J) $t : (A \to B) \to (s : A \to (t \cdot s) : B)$ („application“)
- (+) $t : A \to (t + s) : A, \quad s : A \to (t + s) : A$ („sum/monotonicity“)
- (jt) $t : A \to A$ („factivity“)
- (j4) $t : A \to\, !t : t : A$ („proof checker“)

Notes:

- Application rule as in (typed) lambda calculus
- Can think of $t + s$ as the whole containing parts t, s
- „!“ is an operator for positive introspection (knowing that one knows) justification $!t$ for $t : A$ can be thought of meta-evidence such as the evidence of a proof checker
- Different relevant (weaker) systems follow by deleting one or other axiom
  - Eg.: Factivity may be dropped when focus is rather on beliefs (not knowledge)

Wake-Up Question

- Q: Consider the logic J0 given as
  (J), (+), propositional axioms + modus ponens. Sometimes this is characterized as the logic of a skeptical agent. In which sense is this true?
- A: (according to SEP entry (Artemov/Fitting 21))
  - „J0 is the logic of general (not necessarily factive) justifications for an absolutely skeptical agent for whom no formula is provably justified, i.e., J0 does not derive $t{:}F$ for any t and F. Such an agent is, however, capable of drawing *relative justification conclusions* of the form
  - If $x{:}A, y{:}B, \ldots, z{:}C$ hold, then $t{:}F$.
  - With this capacity J0 is able to adequately emulate many other Justification Logic systems in its language.

Calculus für LP_{CS}

Definition (Constant specification)

A constant specification CS is any subset

$$CS \subseteq \{\ (c, A)\ |\ c \text{ is a constant and } A \text{ is an axiom}\}$$

Definition (Calculus für $LP_{CS}($

- Axioms for LP (mentioned before)
- Rule of modus ponens: From A and A→ B infer B
- Rule of necessitation: From $(c, A) \in CS$ infer $c: A$

The role of constant specifications

- Principle of Logical Awareness
  „all (logical) axioms are justified"
- This applies only for ideal agents
- Constants specifications weaken this principle:
  „all axioms occuring CS are justified"

Extended ex

Recap

- Necessitation: From $(c, A) \in CS$ infer $c{:}\,A$
- (J) $t{:}\,(A \rightarrow B) \rightarrow (s{:}\,A \rightarrow (t \cdot s){:}\,B)$
- (+) $t{:}\,A \rightarrow (t + s){:}\,A, \quad s{:}\,A \rightarrow (t + s){:}\,A$

Example (Justified version of $\mathrm{A} \vee \mathrm{B} \rightarrow \Box(A \vee B)$)

- Assume LP_{CS} with
  $(a, A \rightarrow (A \vee B)) \in CS$ and $(b, B \rightarrow (A \vee B)) \in CS$
- With necessitation
  $\mathrm{LP_{CS}} \vdash a{:}\,(A \rightarrow (A \vee B))$ and $\mathrm{LP_{CS}} \vdash b{:}\,(B \rightarrow (A \vee B))$
- With (J) and (MP) we obtain
  $\mathrm{LP_{CS}} \vdash x{:}\,A \rightarrow (a \cdot x)(A \vee B)$ and $\mathrm{LP_{CS}} \vdash y{:}\,B \rightarrow (b \cdot y)(A \vee B)$
- With (+) we have
  $\mathrm{LP_{CS}} \vdash (a \cdot x) : (A \vee B) \rightarrow (a \cdot x + b \cdot y) : (A \vee B)$ and
  $\mathrm{LP_{CS}} \vdash (b \cdot y) : (A \vee B) \rightarrow (a \cdot x + b \cdot y) : (A \vee B)$
- Using propositional axioms one obtains
  $\mathrm{LP_{CS}} \vdash (x{:}\,A \vee y{:}\,B) \rightarrow (a \cdot x + b \cdot y) : (A \vee B)$

Internalization

Definition (axiomatical appropriate)

A constant specification CS for LP is called axiomatically appropriate if for each for each axiom of LP there is a constant c such that $(c, F) \in CS$

Lemma (Internalization)

Let CS be an axiomatically appropriate constant specification.

For arbitrary formulas $A, B_1, \ldots, B_n$:
If $B_1, \ldots, B_n \vdash_{LP_{CS}} A$, then there is a term t such that

$$x_1 : B_1, \ldots, x_n : B_n \vdash_{LP_{CS}} t : A$$

for fresh variables $x_1, \ldots, x_n$.

Forgetful Projection

Definition (forgetful projection)

The mapping $\circ$ of forgetful projection from justified formulas to modal formulas is defined as follows:

- $P^{\circ} = P$ for P atomic
- $(\neg A)^{\circ} = \neg A^{\circ}$
- $(A \rightarrow B)^{\circ} = A^{\circ} \rightarrow B^{\circ}$
- $(t{:}A)^{\circ} = \Box A^{\circ}$

Lemma (forgetful projection)

For any constant specification CS and any formula F we have that $LP_{CS} \vdash F$ entails $S4 \vdash F^{\circ}$

Realization

Definition (justifications‵ realization)

A realization is a mapping r from modal to justified formulas such that $\left(r(A)\right)^{\circ} = A$

Definition (justifications‵ realization)

We say a justification logic LP_{CS} realizes S4 if there is a realization r such that for any formula A we have
$S4 \vdash A$ implies $LP_{CS} \vdash r(A)$

Realization Theorem

Definition (Schematic CS)

We say that a constant specification CS is schematic if it satisfies the following:
for each constant c, the set of axioms $\{A \mid (c, A) \in CS\}$ consists of all instances of one or several (possibly zero) axiom schemes of LP.

Theorem (realization)

Let CS be an axiomatically appropriate and schematic constant specification. Then the logic LP_{CS} realizes S4 , i.e., there exists a realization r such that for all formulas A

$S4 \vdash A$ entails $LP_{CS} \vdash r(A)$

SEMANTICS

Arithmetical Semantics

- Originally, LP_{CS} was developed to provide classical provability semantics for intuitionistic logic

- Arithmetical Semantics for LP_{CS}
  - Justification terms interpreted as proofs in Peano arithmetic
  - Operations on terms correspond to computable operations on proofs in Peano Arithmetics (PA)

$$\text{Intuitionistic Logic} \xrightarrow{\text{Gödel}} \text{S4} \xrightarrow{\text{realization}} \text{JL} \xrightarrow{\text{Arithm. semantics}} \text{CL + proofs}$$

Intuitionism

- Intuitionistic logic is an offshoot of mathematical intuitionism
- In classical mathematics: Showing existence of an object does not mean finding a verifier
- Interesting debate in philosophy of mathematics whether non-constructive proofs are acceptable
- Mathematical Intuitionism: field allowing only constructive proofs
  - truth = provable = constructively provable
  - Deviates in many aspects from classical logic
    - Double Negation elimination $\neg\neg A \vdash A$ does not hold
    - Tertium non datur $\vdash \neg A \vee A$ does not hold

Intuitionism

L.E.J. Brouwer (1881 to 1966)

Fun facts

- Guru of intuitionism
- Irony of history: Proved many interesting results in classical (non-constructive) mathematics (Brouwer's fixed point theorem)

Self-referentiality

- Gödel‘s famous incompletess result for PA uses self-references: „I am not provable“
  - See also (Halbach/Visser 14a,b) for an overview of self reference in arithmetics
- In modal logic (reading □ as „is provable“) such self-referentiality is not easy to define
- Justification logic helps

Self-referentiality

Definition (Self-referential CS)

A constant specification CS is called self-referential if $(c, A) \in CS$ for some axiom A that contains at least one occurrence of the constant c.

- S4 and LP_{CS} describe self-referential knowledge.
- That means if LP_{CS} realizes S4 for some constant specification CS, then that constant specification must be self-referential.

Lemma

Consider the S4-theorem $G := \neg\Box((P \to \Box P) \to \bot)$
and let F be any realization of G.
If $LP_{CS} \vdash F$, then CS must be self-referential.

Towards a semantics I

Definition (Basic Evaluation)

A basic evaluation $*$ for LP_{CS} is a function defined on propositions and terms $*: Prop \rightarrow \{0,1\}$ and $*: Tm \rightarrow Pow(\mathcal{L}_j)$ such that

- $F \in (s \cdot t)^*$ if $(G \rightarrow F) \in s^*$ and $G \in t^*$ for some G
- $F \in (s + t)^*$ if $F \in s^*$ or $F \in t^*$
- $F \in t^*$ if $(t, F) \in CS$
- $s{:}F \in (!s)^*$ if $F \in s^*$

Towards a semantics II

Definition (quasimodel)

A quasimodel is a tuple $\mathcal{M} = (W, R, *)$ with
- a domain of possible worlds $W \neq \emptyset$,
- an accessibility relation $R \subseteq W \times W$
- And an evaluation functions $*$ mapping each world $w \in W$ to a basic evaluation $*_w$

Definition (Truth in quasimodel)

- $\mathcal{M}, w \vDash p$ iff $p_w^* = 1$ for $p \in Prop$;
- $\mathcal{M}, w \vDash F \rightarrow G$ iff not $\mathcal{M}, w \vDash F$ or $\mathcal{M}, w \vDash G$
- $\mathcal{M}, w \vDash \neg F$ iff not $\mathcal{M}, w \vDash F$
- $\mathcal{M}, w \vDash t{:}F$ iff $F \in t_w^*$

Towards a semantics III: Model

Given $\mathcal{M} = (W, R, *)$ and $w \in W$, we define

$\square_w := \{F \in \mathcal{L}_j \mid \mathcal{M}, v \vDash F \text{ whenever } R(w, v)\}$

= formulae true at all successors of w

Definition (Modular Model)

A modular model $\mathcal{M} = (W, R, *)$ is a quasimodel with

1. $t^*_w \subseteq \square_w$ for all terms $t \in Tm$ and $w \in W$
2. R is reflexive
3. R is transitive

Theorem (Soundness and Completeness)

For all formulas F $\in \mathcal{L}_j$ and let F be any realization of G.

$LP_{CS} \vdash F$ iff $\mathcal{M} \vDash F$ for all modular models $\mathcal{M}$

ALGORITHMIC PROBLEMS

- In modal logic, decidability is a consequence of the finite model property.
- For LP_{CS} the situation is more complicated since CS usually is infinite.

Theorem

LP_{CS} is decidable for decidable schematic constant specifications CS.

- A decidable CS is not sufficient:

Theorem

There exists a decidable constant specification CS such that LP_{CS} is undecidable.

Complexity

Theorem

Let CS be a schematic constant specification.
The problem whether $LP_{CS} \vdash t{:}B$ is in NP

Definition

A constant specification is called schematically injective if it is schematic and each constant justifies no more that one axiom scheme.

Theorem

Let CS be a schematically injective and axiomatically appropriate constant specification. The derivability problem for LP_{CS} is Π_2^p $- complete$

Reminder: Polynomical Hierarchy

- Two possible definitions, either based on oracle machines (considered here) or quantified boolean formula

- How to think about oracles?
  - Think in terms of Turing Machine pseudocode or a subroutine
  - An oracle Turing machine M with oracle $B \subseteq \Gamma^*$ lets you include the following kind of branching instructions:

    “if $z \in B$ then <do something>
    else <do something else>”

    where z is some string defined earlier in pseudocode.
  - By definition, the oracle TM can always check the condition ($z \in B$) in one step

Some Complexity Classes With Oracles

- $P^B = \{ L \mid L$ can be decided by some *polynomial-time* TM with an oracle for B $\}$

- P^{SAT} = the class of languages decidable in polynomial time with an oracle for SAT

- P^{NP} = the class of languages decidable by *some* polynomial-time oracle TM with an oracle for *some* B in NP

Wake-Up Exercise

- Q: Is $P^{SAT} \subseteq P^{NP}$?
- A: Yes. By definition…
- Q: Is $\mathrm{P}^{\mathrm{NP}} \subseteq \mathrm{P}^{\mathrm{SAT}}$?
- A: Yes! Every NP language can be reduced to SAT!
  - For every poly-time TM M with oracle $B \in NP$, we can simulate every query z to oracle B by reducing z to a formula ϕ in poly-time, then asking an oracle for SAT instead

Polynomial Hierarchy (PH)

Definition

- $\Delta_0^P \;:= \Sigma_0^P \;:= \Pi_0^P \;:= P$

- $\Delta_{i+1}^P \;:= P^{\Sigma_i^P}$
- $\Sigma_{i+1}^P \;:= NP^{\Sigma_i^P}$
- $\Pi_{i+1}^P \;:= coNP^{\Sigma_i^P}$

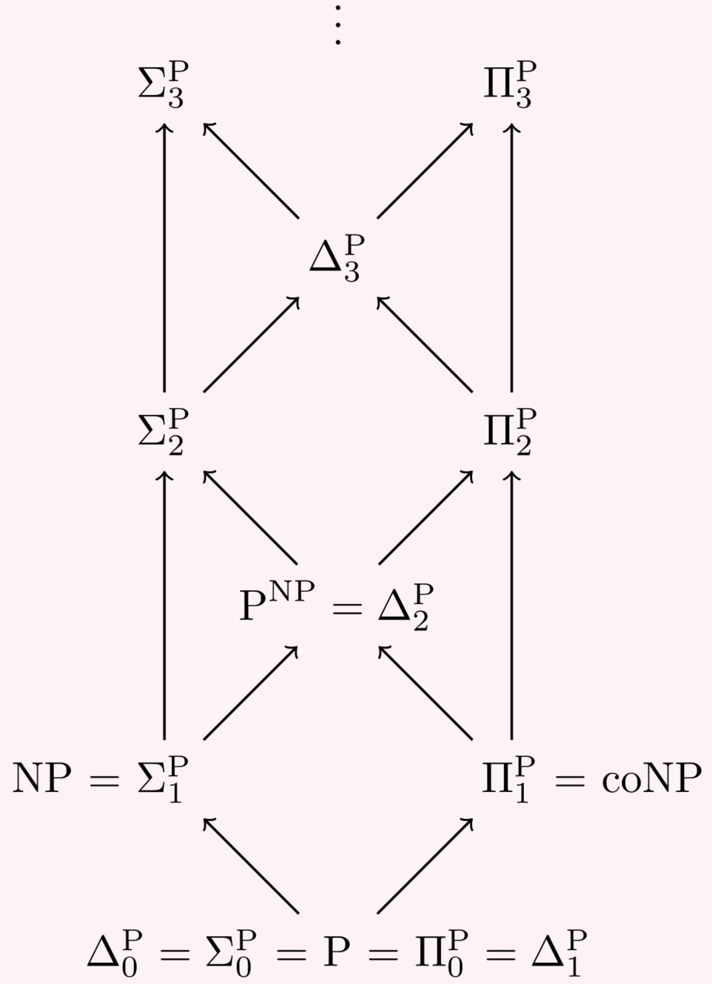
Example: $\Pi_2^p = coNP^{\Sigma_1^P} = coNP^{NP^P}$

Theorem

$$PH := \bigcup \Sigma_i^p \subseteq PSPACE$$

Open: PH = PSPACE

Relations within the heirarchy



RECONSIDERING GETTIER

Recap: Gettiers two counterexamples

Scenario 1

- Smith and Jones apply for a job
- Smith believes (justifiably):
  (p) Jones will get the Job &
  Jones has ten coins in his pocket
- Smith believes also in the entailed assertion:
  (r) The one who gets the job has ten coins in his pocket.
- Coincidence : Smith gets the job and Smith has ten coins in his pocket.
- Smith „knew“ (r) only by chance

Scenario 2

- Smith justifiably believes
  (p) Jones owns a Ford
- Smith also believes in entailed assertion
- (r) = (p or q): Jones owns a Ford, or Brown lives in Barcelona
  (Though Smith has no justification for q)
- Coincidence: Jones does not own Ford, but Brown lives in Barcelona
- Smith „knew“ (r) only by chance

General idea: decouple justification and truth conditions of propositional content of belief

Gettier Examples in Justification Logic

Main intention with justification logic (according to Artemov 08) w.r.t. Gettier paradoxa

- Show that Gettier reasoning is formally correct
- Thereby identify (logical) principles in the reasoning
  - These have lead to the axioms in LP
- Gettier examples inconsistent within Justification Logic systems of factive justifications (factivity axiom)
- Can be used also for analyzing approaches that try to resolve the paradox:

  Justified True Belief + 4th Condition

  ("no-Gettier-problem condition")

Principles Involved in Gettier Examples

- Gettier uses a version of the epistemic closure principle, closure of justification under logical consequence:

If Smith is justified in believing Q	For some t, $t: P$
and Smith deduces Q from P	$P \rightarrow Q$
Then Smith is justified in believing Q	For some t, $t: Q$

- Holds for all justification logic systems due to
  - Internalization: If $\vdash F$, then there is a t such that $\vdash t : F$
  - Application axiom: $t: (A \rightarrow B) \rightarrow (s: A \rightarrow (t \cdot s): B)$
  - Modus ponens

Goldman‘s reliabilism → Factivity

- Goldman (1967) offered the fourth condition to be added to the Justified True Belief definition of knowledge, according to which:
- *"A subjects belief is justified only if the truth of a belief has caused the subject to have that belief, and for a justified true belief to count as knowledge, the subject must also be able to correctly reconstruct (mentally) that causal chain."*
- A situation t justifies F for some t only if F is true, which provides the Factivity Axiom for knowledge-producing justifications:
- Factivity axiom: $t: A \to A$

Lehrer/Paxson's indefeasibility → monotonicity

- Lehrer and Paxson (1969) offered the following 'indefeasibility condition':
  "There is no further truth which, had the subject known it, would have defeated [subjects] present justification for the belief."
- Criticism of this condition: a defeater fact cannot be made precise enough to rule out the Gettier cases without also ruling out a priori cases of knowledge

Lehrer/Paxson's indefeasibility → monotonicity

- *„there is no justification"*
- => „for any further evidence, it is not the case"
- $s{:}\,F$: „present justification for the belief"
  given $s{:}\,F$, for any evidence t, it is not the case that t would have defeated $s{:}\,F$
- $s + t$: the joint evidence of s and t:
- if $s{:}\,F$ holds, then $s\ +\ t$, is also an evidence for F
- $s{:}\,F\ \rightarrow\ (s + t){:}\,F$

Gettier‘s implicit assumptions

- In the first Gettier example we have the following assumptions which cannot hold:
- $J(Smith), C(Smith), C(Jones), \neg J(Jones),$ (*)

$$u: [\big(Jones = \iota x\, J(x)\big) \wedge\ C(Jones)].$$

- Notation
  - $J(x) = x$ gets the job;
  - $C(x) = x$ has coins in his pocket
  - $\iota x\ S(x) =$ the x that has the property S(x) (a so-called definite description)

Gettier's implicit assumptions

- In the firt Gettier example we have the following assumptions which cannot hold:
- $J(Smith), C(Smith), C(Jones), \neg J(Jones),$ (*)

  $$u: [(Jones = \iota x\, J(x)) \;+\; C(Jones)].$$
- With factivity we get a contradiction:
  - $u: [\big(Jones = \iota x\, J(x)\big) \wedge\; C(Jones)]$ from (*)
  - $Jones = \iota x\, J(x),$ Factivity and some propositional logic;
  - $\big(Jones = \iota x\, J(x)\big) \rightarrow\; J(Jones),$ natural property of definite descrs;
  - $J\,(Jones)$ by Modus Ponens.

    This contradicts the condition $\neg\, J\,(Jones)$ from (*).

Uhhh, a lecture with a hopefully useful

APPENDIX

References

- (Artemov/Fitting 21)
  Artemov, Sergei and Melvin Fitting, "Justification Logic", *The Stanford Encyclopedia of Philosophy* (Spring 2021 Edition), Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/spr2021/entries/logic-justification/
- (Artemov/Fitting 19)
  S. Artemov and M. Fitting. Justification Logic: Reasoning with Reasons. Cambridge Tracts in Mathe- matics. Cambridge University Press, 2019.
- (Gödel 1995)
  Vortrag bei Zilsel/Lecture at Zilsel's (*1938a). In Feferman, S., Dawson, J. W., Jr., Goldfarb, W., Parsons, C., and Solovay, R. M., editors. *Unpublished Essays and Lectures, Volume III of Kurt Go¨del Collected Works*. New York: Oxford University Press, pp. 86–113.
- (Halbach/Visser 14a)
  V. Halbach and A. Visser. Self-reference in arithmetic i. The Review of Symbolic Logic, 7:671–691, 2014.
- (Halbach/Visser 14b)
  V. Halbach and A. Visser. Self-reference in arithmetic ii. Review of Symbolic Logic, 7(4):692–712, 2014.
- (Artemov 08)
  S. ARTEMOV. The logic of justification. The Review of Symbolic Logic, 1(4):477–513, 2008.

Color Convention in this course

- Formulae, when occurring inline
- Newly introduced terminology and definitions
- Important results (observations, theorems) as well as emphasizing some aspects
- Examples are given with standard orange with possibly light orange frame
- Comments and notes
- Algorithms