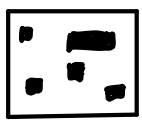
B. Schölkopf (MPI): 'Symbolic, Statistical & Causal AI'

FEYNMAN APPROACH TO INTELLIGENCE — CREATE ⇄ UNDERSTAND

- N. WIENER → CYBERNETICS ⇒ FROM ENERGY TO INFO-PROCESSING
- F. ROSENBLATT → PERCEPTRON - CONNECTIONISM ⇒ LEARNING!
  - → XOR problem: COULD NOT BE LEARNED AT THE TIME!
  - → NOVIKOFF (1962): PERCEPTRON CONVERGE THEOREM
- McCARTHY ET. AL → INTELLIGENCE = PROCESS OF MANIPULATING DISCRETE SYMBOLS
- MINSKY & PAPERT (1969): PARITY PROBLEM



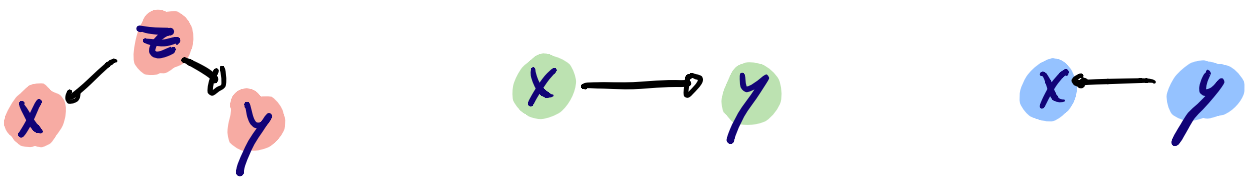
→ COUNT ↓ MEANINGLESS!

  - ↳ PERCEPTRONS FELL OUT OF FAVOR
  - ↳ SYMBOLIC METHODS GAINED MOMENTUM [FROM CONNECTIONIST FAILURE] ↳ WELL SUITED FOR COMPUTERS!
- MORAVEC'S PARADOX (1980s):
  - → PROBLEMS THAT APPEAR HARD ⇒ SIMPLE
    PROBLEMS THAT APPEAR SIMPLE ⇒ HARD
  - → SYMBOLIC METHODS COULDN'T SCALE
  - → BOLTZMANN MACHINES, BACKPROP
  - → PROBABILISTIC EXPERT SYSTEMS

BIRTH OF THE FIELD OF MACHINE LEARNING

- DEEPMIND DQN — PROBLEM OF IID SETTING ⇒ RE-TRAIN + TRICKS
  - → WE ARE INTERESTED IN MORE THAN STATISTICAL REGULARITIES
  - → CAUSAL RELATIONSHIPS VS. DEPENDENCE
    - ↳ REICHENBACH'S COMMON CAUSE PRINCIPLE

SYMMETRY ⇄ ASYMMETRY



→ ALL THREE OPTIONS HAVE SAME OBSERVED DATA

"RUCKSACK PROBLEM" // CARGO CULT → NOT PROBLEM FOR AI

INDUSTRIAL REVOLUTION → PROCESS ENERGY
DIGITAL REVOLUTION → PROCESS INFORMATION
} GENERATE / PROCESS / CONVERT!

B. Schölkopf + S. Bauer (MPI): Causality

→ INDEPENDENCE OF RVs: $X \perp Y : p(x,y) = p(x) \cdot p(y) \; \forall x,y$

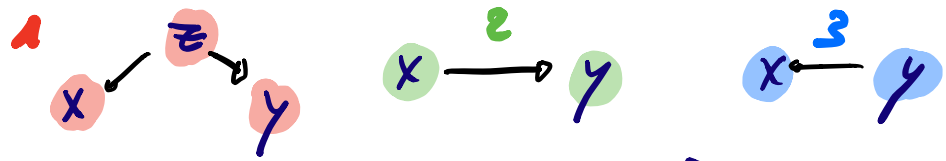
↳ $X \perp Y \Rightarrow \mathbb{E}[X \cdot Y] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$, $Cov[X,Y] = 0$

→ COND. INDEP. OF RVs: $X \perp Y \mid Z : p(x,y|z) \; p(x|z) \cdot p(y|z), \; p(z) > 0$

INDEPENDENCE OF MECHANISMS ASSUMPTION: GENERATIVE PROCESS OF SYSTEM'S VARS. = COMPOSITION OF AUTONOMOUS MODULES THAT DO NOT INFLUENCE EACH OTHER

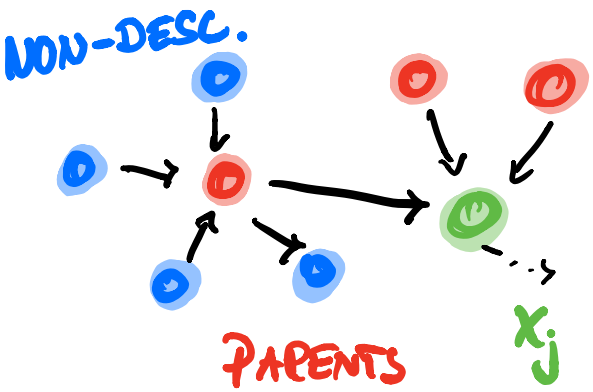
→ STRUCTURAL/FUNCTIONAL CAUSAL MODEL:

- DAG G w. VERTICES $X_1, \dots, X_n$ ⇒ V: OBS., E: DIRECT CAUSATION
- $X_i := f_i(PA_i, U_i)$ → NOISE VARS. / UNMODELLED ⇒ INDEPENDENT
  - ↳ WHY INDEP. U_i? ⇒ REICHENBACH! → CAUSAL SUFFICIENCY
  - ↳ REICHENBACH'S COMMON CAUSE PRINCIPLE → ALL THREE OPTIONS HAVE SAME OBS. DATA!



- ENTAILED DISTRIBUTION → $P(X_1, \dots, X_n)$ ⇒ OBSERV. + NOISE

→ MARKOV CONDITIONS ⇒ EQUIVALENCE:



EXISTENCE OF SCM
⇔ LOCAL CAUSAL MARKOV
⇔ GLOBAL CAUSAL MARKOV
⇔ $p(x_1, \dots, x_n) = \prod_i p(x_i | pa_i)$ → causal Markov kernels

- INTERVENTION: REPLACING $X_i := f_i(PA_i, U_i)$ W. ANOTHER ASSIGNM.
  - ↳ RESULTING ENTAILED DISTR. = INTERVENTIONAL DISTRIB.
- MECHANISM INDEPENDENCE: $p(X_i | PA_i)$ CHANGE DOESN'T CHANGE $p(X_j | PA_j)$ for $j \neq i$. → INVARIANT ⇒ DISENTANGLED FACTORIZATION

☐ COUNTERFACTUALS ⇒ (WOULD BE) → CAUSAL INF. = MISSING DATA P.

→ DO-CALCULUS: $p(y \mid do\, x)$ → DISTR. IF WE INTERVENE ON X AND ASSIGN IT TO x

"DOING VS. SEEING"

CAN BE DIFFERENT FROM $p(y|x)$

→ NEED G & p!

$$p(x_1, \dots, x_n \mid do\, x_i) = \prod_{j \neq i} p(x_j \mid pa_j)\, \delta_{x_i x_i}$$

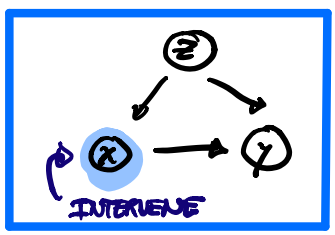
[CAUSAL FACT.] $p(x, y, z) = p(z)\, p(x|z)\, p(y|x, z)$

[INTERVENE] $p(x, y, z \mid do\, x) = p(z)\, \delta_{xx}\, p(y|x, z)$

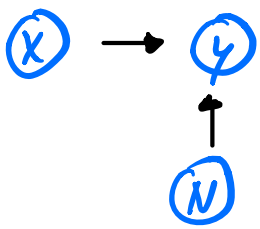
[INT. OUT x] $p(y, z \mid do\, x) = p(z)\, p(y|x, z)$

[MARG. z] $p(y \mid do\, x) = \sum_z p(z)\, p(y|x, z)$ ⇒ ADJUSTMENT FORMULA

[SIMPSON'S P.] $\neq p(y|x)$



☐ ADDITIVE NOISE MODEL FOR 2 VARIABLE-CASE

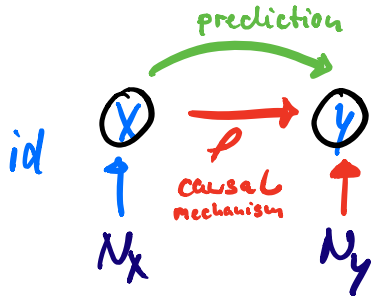


ASSUME TO BE ADDITIVE: $f(X) = Y + N$ w. $X \perp\!\!\!\perp N$

↳ NON-LINEAR REG. + CHECK IF RESIDUALS ARE CORREL. W. X?!

☐ CAUSAL LEARNING:

$p(X)$, $p(Y|X)$ INDEP.



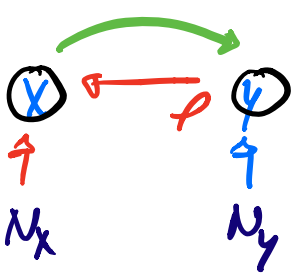
⇒ SEMI-SUPERV. L. IMPOSSIBLE

⇒ $p(y|x)$ INVARIANT UNDER CHANGE IN $p(x)$

↳ COVARIATE SHIFT

ANTIC. LEARNING:

$p(Y)$, $p(X|Y)$ INDEP.



↳ $p(x)$, $p(y|x)$ dependent!

⇒ SEMI-SUPERV. L. POSSIBLE

⇒ $p(y|x)$ CHANGES W. $p(x)$

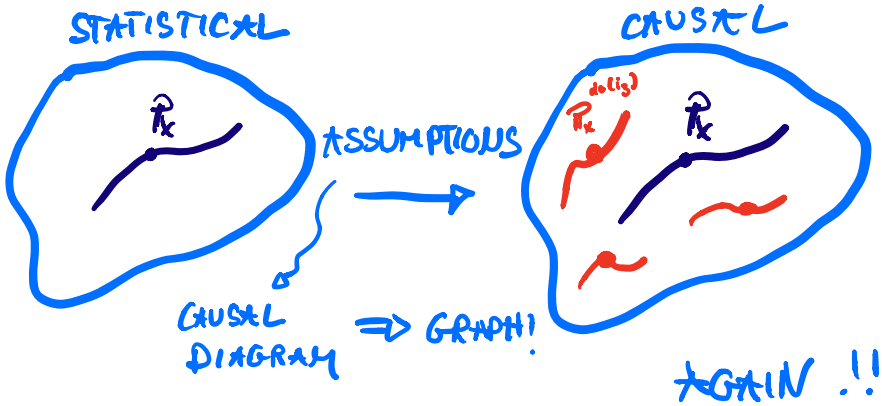
MODELING-TAXONOMY

	STAT	CAUSAL	DE
IID	✓	✓	✓
SHIFT	x	✓	✓
PHYSICS	x	(✓)	✓
REASON	x	?	?
LEARN	✓	(✓)	x

☐ ALGORITHMIC CAUSAL MODELS → KOLMOGOROV COMPLEXITY

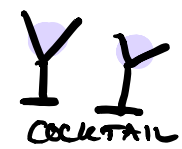
↳ THERMODYNAMIC ARROW OF TIME / ENTROPY PRESERVATION

- CAUSAL MODEL = POSETS OF DISTRIB.
  - CLEARLY STATE ASSUMPTIONS! CAN ALLOW FOR CAUSAL CONCLUSION



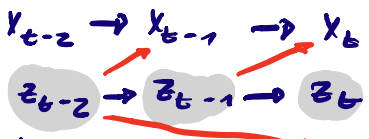
⇒ DATA IS NOT ENOUGH! → NEED INDUCTIVE BIASES! ⇒ AGAIN !!! ASSUMPTIONS

- CAUSAL STRUCTURE LEARNING [HEINZE-DEML ET AL. 17']
  - CONSTRAINT BASED → INDEP. TESTS
  - SCORE BASED → LH-BASED DAG SEARCH
  - RESTRICTED SEM → LiNGAM $X = (I-B)^{-1}\varepsilon$
  - THEOREM: IDENTIFIABILITY OF NON-GAUSSIAN LINEAR MODELS
    - RELATION TO ICA (SOURCE SEPARATION)
    - ⇒ INDEP. COMPONENTS!



⇒ OFTEN REQUIRES LARGE SEARCHES OVER DIFF. GRAPHS!

- TIME SERIES → GRANGER "CAUSALITY" ⇒ PREDICTIVE POWER
  - IF $Y_{present} \not\perp X_{past} \mid Y_{past}$ THERE MUST BE ARROW FROM X TO Y



$Y_{present} \not\perp X_{past} \mid Y_{past}$
$X_{present} \perp\!\!\!\perp Y_{past} \mid X_{past}$
Problem: GRANGER INFERS X→Y

⇒ CONFOUND PROBLEM!

- ODE INVARIANCE ACROSS ENVIRONMENTS → FORM OF CROSS-VAL.
  - PFISTER ET AL. ⇒ STABILIZED REGRESSION → MODEL AVERAGING HELPS WHEN FUNCTIONAL FORM OF ODE NOT KNOWN/REJECTED

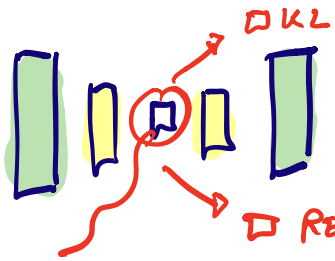
⇒ KEY PROBLEM: CAUSAL VARS ARE ASSUMED TO BE FULLY OBSERVED
- 'WHAT ARE CAUSAL VARIABLES OF AN IMAGE?'
- DEEP REPRESENTATION LEARNING OF INDEP. COMPONENTS!
- GOAL: TRANSFER OF MECHANISMS ⇒ OUT-OF-SAMPLE!

CAUSAL PERSPECTIVE ON DEEP REPRESENTATION LEARNING

→ BENGIO, COURVILLE, VINCENT 12' → MULTI-TASK MOTIVATION

→ DISENTANGLEMENT: SINGLE CHANGE FACTOR ⇒ SINGLE CHANGE REPRES. → ENCOURAGE FACTORIZATION

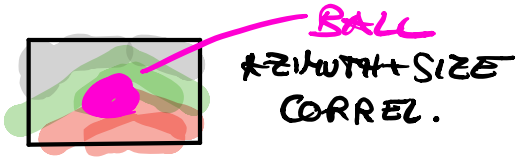
→ DISENTANGLEMENT VIA VAE + REGULARIZER



REPRES. ⇒ SPLIT SOURCES OF VAR.!

RESULT: FOR ARBITRARY UNSUPERVISED LEARNING OF DISENTANGLED REPRESENTATIONS IMPOSSIBLE!
↳ NEED FOR INDUCTIVE BIASES! [OR WEAK SUPERVISION]

→ SCALING TO REAL-WORLD DATA?

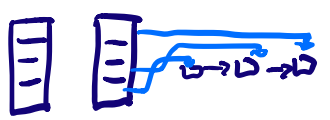


↳ TRAUBLE ET. AL 20' → INDEP. OF MECHANISMS VS. VARS

↳ DISENTANGLEMENT ~ DOWNSTREAM PERFORMANCE

⇒ ENCODE CAUSAL STRUCTURE INTO AUTOENC. → LEEB ET AL. 20'

ATTENTION ENCODER →



STRUCTURAL DECODER

⇒ LEARN TO DISTINGUISH AND ORDER MECHANISMS
↳ STRUCTURE ARCHITECTURE VS. GAUSSIAN REGULARIZER

TOWARDS CAUSAL WORLD MODEL

→ LEARNING INDEP. MECHANISMS → RECURRENT IMS
↳ GOYAL ET AL. 19'

EXPERTS 1 2 3

LEARN DIFF. DE-NOISING MECHANISM

CANDU. → DISCRIMINATOR

→ INDEP. COMPONENTS ⇒ SOCIAL INTELLIGENCE

ATTENTION MODULE ON TOP OF LSTMs

Peter Dayan (MPI): Comp. Neuro in ML

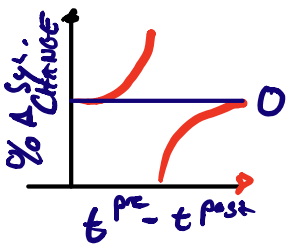
LEVELS OF ANALYSIS

MARRIAN COGNITION

COMPUTATION	⇄ NORMATIVE / BOTTOM-UP	ALGORITHM	⇄ TOP-DOWN / CONSTRAINTS	IMPLEMENTATION
ETHOLOGY		PSYCHOLOGY		NEUROSCIENCE
GOALS		PROCEDURES		NEURAL REALISATIONS

SPIKE-TIME DEP. PLASTICITY

[STDP] + H. MARKRAM



↳ HEBBIAN INTUITION

STRENGTHEN/WEAKEN BASED ON TIMING!

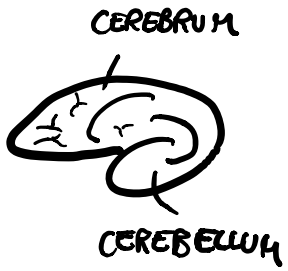
↳ "COINCIDENCE DETECTION"

⇒ MODULATION: DOPAMINE + ACETYLCHOLINE

↳ LOCAL VS. GLOBAL

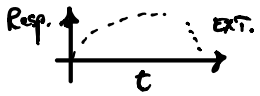
LEARNING / MEMORY

- PROCEDURAL
  - NON-ASSOC.
    - HABITUATION
    - SENSITIZATION
  - ASSOC.
    - CONDITIONING
      - EMOTIONAL
      - MOTORSKILL
- DECLARATIVE
  - EPISODIC
  - SEMANTIC



⇒ MISSING:
* LEARNING OF PREDICTIONS
* FORWARD/INVERSE MODELS
* REPRESENTATION LEARNING

PAVLOVIAN CONDITIONING

UNCOND. STIM.	COND. STIM.	UNC. RESP. & COND. RESP.
r	s_j	

↳ BLOCKING: LEARNING OF ADD. STIMULUS CAN BE BLOCKED

⇒ CONTINGENCY IS NOT ENOUGH

↳ NEED SURPRISE TO LEARN!

↳ 2ND ORDER: TIMING IS CRUCIAL

⇒ LEARN PREDICTOR OF PREDICTOR

$$V_t = \mathbb{E}_\pi \left[\sum_{\tau=t}^{T} r_\tau \right]$$

↳ EXPEC. FUTURE REWARD SEQ.

RESCORLA-WAGNER LEARNING

$$V_{n+1} = \alpha \sum_{v=1}^{n} (1-\alpha)^{n-v} r_v + (1-\alpha)^n V_1$$

CASE FOR: $s_j \in$ Eq 13

↳ LRate → BUT ALSO: FORGET RATE!

⇒ SPEED OF ENVIRONMENT CHANGE

↳ CHOICE VIA PROPENSITIES:

1. $m(s,a) = Q(s,a)$ + SOFTMAX POLICY w. β TEMP.
2. POLICY GRAD. / "DIRECT ACTOR"
   ↳ REINFORCE (Williams)

$$\frac{\partial \ln \pi(s,a)}{\partial \theta} \cdot [r(s,a) - V(s)]$$

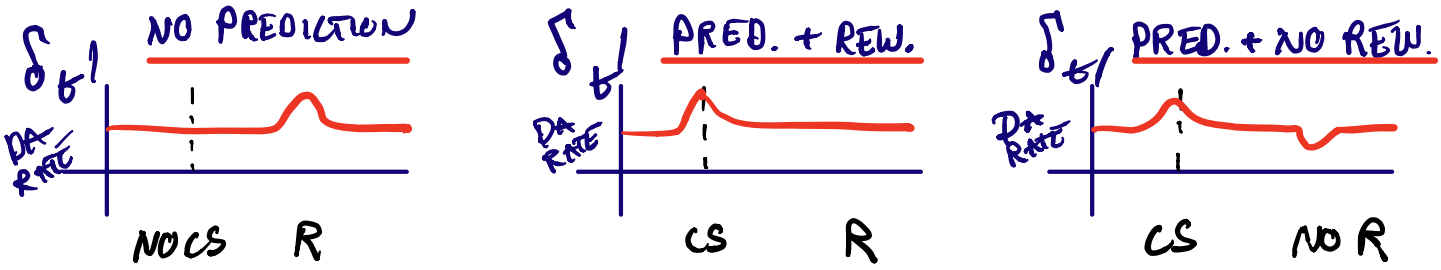
→ KEY: BELLMAN CONSISTENCY ⇒ USE INCONSISTENCY

ERROR FOR LEARNING! → TEMP. DIFFERENCE ERROR

$$V_t \leftarrow V_t + \alpha(r_t + V_{t+1} - V_t) \Rightarrow \text{TD(0)}$$

(δ_t)

⇒ DOPAMINE REWARD PRED. ERROR HYPOTHESIS



↳ MODEL-BASED (PLANNING) VS. MODEL-FREE RL

ACTOR-CRITIC METHODS:

LEARN PROPENSITIES + VALUES SIMULTANEOUSLY!

↳ $\delta > 0$: - V IS TOO PESSIMISTIC
- Q IS BETTER THAN AVG.

↳ SAME SIGNAL ⇒ VENTRAL → DORSAL STRIATUM!

↳ NEURO-WORK BY YAEL NIV

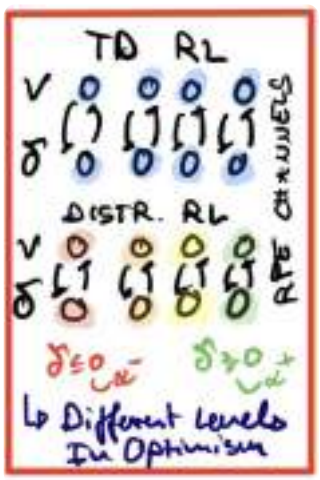
DABNEY ET AL. 20' - DISTR. TD

⇒ DISTRIBUTIONAL

→ NEURON-SPEC α_i!

↳ ENCODE RISK SENSITIVITY/ OPTIMISM!

→ CAPTURES DA NEURON HETEROG.



DIFFERENT SORTS OF LEARNING

SUPERV. REINFORL. SELF-SUPERV.

DIFFERENT ALGORITHMS

BACKPROP HEBBIAN 3-TERM

DIFFERENT NEURAL SUBSTRATES

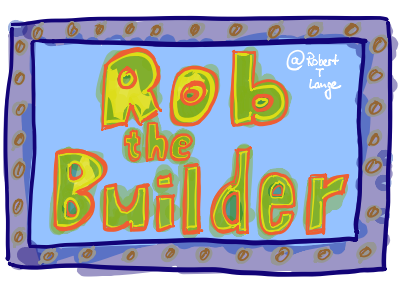
HIPPOCAMPUS CORTEX STRIATUM CEREBELLUM

LECUN CAKE

Q&t:

TABULAR VS. FCT. APPROX.

↳ Q OF STATE REPRES.
Q GENERALISATION



Costis Daskalakis (MIT): Game Theory & ML

UTILITIES ⟷ LEARN + DECIDE

⇒ STRATEGIC REASONING

- SYMMETRY VS. ASYMMETRY
  ⇒ OPTIMIZATION NOT ENOUGH!

- PLATFORM DESIGN
  - ↳ OPTIM. NOT ENOUGH → NEED BEHAV. MODEL → COUNTERFACTUAL
- PHYSICAL RECOMMENDER SYSTEM
  - ↳ COMBINE LEARNING + MARKET DESIGN

MULTI-AGENT EQ. LEARNING

MINIMIZATION → CURRENT AI ⇒ STATIONARY WORLD

MIN-MAX OPTIM. → FUTURE AI ⇒ NON-STAT. WORLD

$$\inf_{\theta} \sup_{w} f(\theta, w)$$ ↳ HIGH-DIM. + CONSTRAINT

BEST CASE: f CONVEX IN θ, CONCAVE IN w

PART 1: CONVEX-CONCAVE OBJECTIVES

- MIN-MAX TH.: IF $X \subset \mathbb{R}^n$, $Y \subset \mathbb{R}^n$ ARE COMPACT & CONVEX, AND $f: X \times Y \to \mathbb{R}$ IS CONVEX-CONCAVE, THEN

$$\min_{x \in X} \max_{y \in Y} f(x,y) = \max_{y \in Y} \min_{x \in X} f(x,y)$$

  - ↳ OPTIMAL POINT IS UNIQUE [IF STRICT] ← VALUE IS ALWAYS UNIQUE

POINTS = EQUILIBRIA OF ZERO-SUM GAME

- BROWN 49' → FICTITIOUS PLAY!
- NO-REGRET-LEARNING

$$\frac{1}{T} \sum_t \ell_t(x_t) - \frac{1}{T} \min_x \sum_t \ell_t(x) \to 0$$

  - ↳ FOLLOW-THE-REGULARIZED LEADER

$$x_t \in \arg\min \left[\sum_{\tau < t} \ell_\tau(x) + \frac{1}{\eta} R(x) \right]$$

  - ↳ GAME VERSION: X, Y PLAYERS, $f(\cdot, y_t)$, $-f(x_t, \cdot)$ — L_2 → SGD!

⇒ NO-REGRET LEARNING CONVERGENCE → AVG. STRATEGY APPROX. NASH

GANS [GOODFELLOW ET AL. 14' / ARJOVSKY ET AL. 17']

GENERATOR ⟷ DISCRIMINATOR

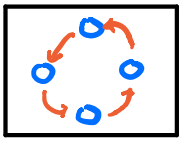
$$\min_{\theta} \max_{w} \mathbb{E}_{x \sim \text{Preal}} [D_w(x)] - \mathbb{E}_{z \sim \omega} [D_w(G_\theta(z))]$$

Preal: REAL; $G_\theta(z)$: HALLUCINATED

→ TRAINING OSCILLATIONS [1st ORDER]
→ HIGH-DIM. STATISTICS PROBLEM
→ COMMON: MORE FREQ. UPDATES FOR GEN.

TRAINING OSCILLATIONS

- METZ ET AL. 17' ⇒ MODE COLLAPSE
  - ↳ SGD CYCLES BETWEEN



- DASKALAKIS ET AL. 18'
  - ↳ EVEN SIMPLER 2D GAUSS!
  - ↳ PURE OPTIMIZ. PROBLEM

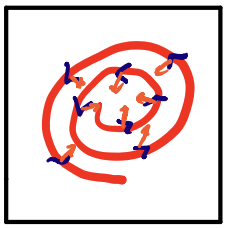
⇒ EVEN FOR CONVEX-CONCAVE OBJECTIVES!
⇒ EVEN WHEN FCT. IS PERFECTLY KNOWN
⇒ EVEN WHEN THERE ARE NO STATES

REMOVING OSCILLATIONS VIA NEGATIVE MOMENTUM

- OPTIMISTIC GD [POPOV 80']: $x_{t+1} = x_t - \eta \nabla f(x_t) + \eta/2 \nabla f(x_{t-1})$
- FTRL: UNIFORM POLICY HISTORY → UPWEIGHT MOST IMMEDIATE HISTORY
- EXTRA-GRAD. [KORPELEVICH 76']:

$$x_{t+1/2} = x_t - \eta \nabla f(x_t)$$
$$x_{t+1} = x_t - \eta \nabla f(x_{t+1/2})$$

$f(x,y) = x \cdot y$



NEG. MOM. "GRAVITY PULL"

⇒ SPIRALLING OUT

↳ OGDA: LAST-ITERATE CONV. + LINEAR RATES FOR UNCONSTR. BILINEAR GAMES W. WELL. COND. PAYOUT MATRIX

↳ EG: SAME FOR UNCONSTR. FOR STRONG CASE!

- CONSTRAINED CASE ⇒ PROJECTED VERSIONS!

PART II: NONCONVEX - NONCONCAVE OBJECTIVES

- DASKALAKIS ET AL. 18': OPTIMISTIC ADAM → CIFAR 10 GAN
  - ↳ DOES NOT NEED DIFFERENT # STEPS GENERATOR VS. DISCRIMINATOR!
- $f(x,y)$ NOT CONVEX-CONCAVE ⇒ MIN-MAX THEOREM BREAKS! ⇒ GOAL?
  - ↳ MORE GENERAL CASE: [GENERAL CONSTRAINED MIN-MAX PROBLEM]

GIVEN CONTINUOUS f & CONVEX, COMPACT SETS S:

$$\min_x \max_y f(x,y) \quad \text{s.t. } (x,y) \in S$$

DEF.: $S(x,\cdot) = \{y \mid (x,y) \in S\}$

$S_x = \{x \mid \exists y \text{ s.t. } (x,y) \in S\}$

1 LOCAL MIN-GLOBAL MAX SOLUTION

$(x^*, y^*) \in S$ s.t., for some $\delta > 0$,

$f(x^*, y) \le f(x^*, y^*) \le \max_{y' \in S(x^*,\cdot)} f(x, y')$,

$\forall y \in S(x^*,\cdot)$ and $x \in N_\delta(x^*) \cap S_x$

↳ ASYMMETRY INNER/OUTER PLAYER

2 LOCAL MINIMAX

↳ y PLAYER IS ALSO CONSTRAINED TO MOVE LOCALLY

3 LOCAL MIN-MAX EQ.

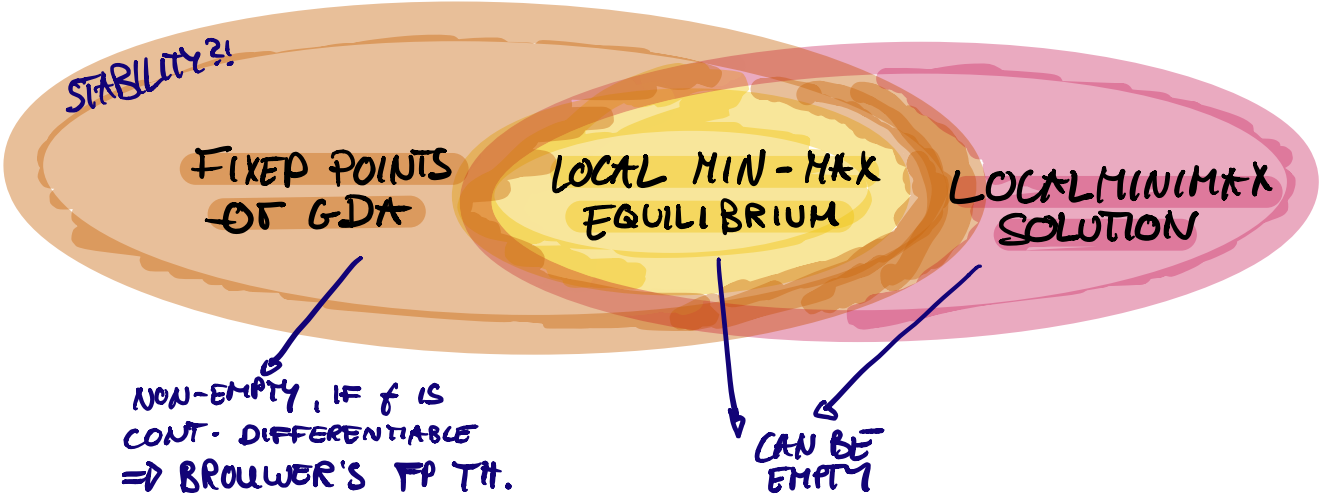
$(x^*, y^*) \in S$ s.t., for some $\delta > 0$,

$f(x^*, y) \le f(x^*, y^*) \le f(x, y^*)$,

$\forall y \in N_\delta(y) \cap S(x^*,\cdot)$ & $x \in N_\delta(x^*) \cap S(\cdot, y^*)$

1 + **2**: SEQ. MOVES VS. **3**: SIMULT.

1: GUARANTEED TO EXIST → BUT NP-HARD TO FIND VS. **2** + **3**: NOT ALWAYS EXISTING

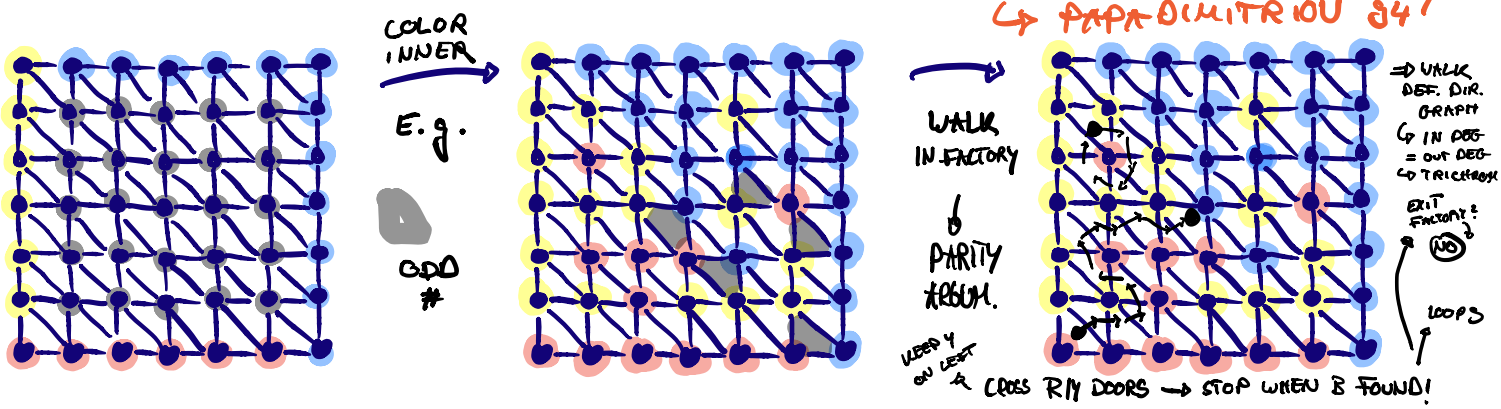


☐ GOAL: FIND APPROX. FP: $\|(x^*, y^*) - F_{GDA}(x^*, y^*)\|_2 \leq \varepsilon$

↳ APPROX. BROUWER FP ⇒ PPAD COMPLETE → BETWEEN P & PPAD

PPAD: THE CLASS OF ALL PROBLEMS IN NP W. GUARANTEED SOLUTION BY GRAPH-THEORETIC LEMMA: IF ∃ UNBALANCED NODE, THEN ∃ ANOTHER

↳ PAPADIMITRIOU 94'



SPERNER'S LEMMA: NO MATTER HOW THE INTERNAL NODES ARE COLORED THERE EXISTS A TRI-CHROMATIC TRIANGLE. → AN ODD # OF THEM.

↳ BOILS DOWN TO DIRECTED PARITY ARGUMENT → LIKE PPAD!

↳ WHAT MAKES PROBLEM HARD? ⇒ COLORING CIRCUIT MAY BE SMALL BUT IMPLIED GRAPH EXP. LARGE!

↳ CAN BE USED TO PROOF BROUWER!

DASKALAKIS ET AL. 20': COMPUTING LOCAL MIN-MAX EQ. IN ZERO-SUM GAMES W. NONCONVEX-NONCAVE OBJ. ⇒ AS HARD AS NASH-EQ. IN GENERAL SUM GAMES

☐ **MULTI-AGENT RL** ⇒ STRUCTURE ALLOWS NASH IN NONCONVEX-NONCONCAVE CASE

$\max_{\pi_i} V^i_{\pi_i, \pi_{-i}} = \mathbb{E}\left[\sum_{t \geq 0} R^i(s_t, a_t, s_{t+1})\right]$

⇒ NASH EQ.: $\pi^{1,\dots,N}$ s.t. $\forall i: V^i_{\pi_i, \pi_{-i}} \geq V^i_{\pi', \pi_{-i}}$

↳ LOTS OF OPEN QUESTIONS!

⇒ DECENTRALIZED CONTROL

⇒ TOWARDS GENERAL SUM!

SHAPLEY 53'