DEEP IMPLICIT LAYERS: NEURAL ODEs, EQUILIBRIUM MODELS, ...

D. DUVENAUD (UoT), J. Z. KOLTER (CMU), M. JOHNSON (BRAIN)

LAYER = DIFFER. PARAMETRIC FCT.

- **EXPLICIT** — PROVIDE COMP. GRAPH → FORWARD: $x \to$ $\boxed{\text{COMPUTE } y = f(x)}$ $\to y$
- **IMPLICIT** — SATISFY IN-OUT CONDITION: $x \to$ $\boxed{\text{FIND } y \text{ s.t. } g(x, y) = 0}$ $\to y$

WHY USE IMPLICIT LAYERS?

1. POWERFUL REPRESENTATION
2. MEMORY EFFICIENCY
3. DESIGN SIMPLICITY
4. ABSTRACTION: "WHAT" - "HOW"

① DEEP EQ. M.	② NEURAL ODEs	③ DIFF. OPTIM.
IMAGE CLASS. SEMANTIC SEG. LANGUAGE M.	CONT.-TIME SYSTEMS GENER. MODELS SMOOTH DENSITY EST.	CONSTRAINED OPTIMIZAT.

HISTORICAL BACKGROUND

- RECURRENT BACKPROP [PINEDA 87', ALMEIDA 87']
- APPLIED ENGINEERING [RICO-MARTINEZ ET AL. 92', 95']
- DIFFERENTIABLE OPTIMIZATION
  - ↳ STRUCTURED VAEs [JOHNSON ET AL. 16']
  - ↳ D. DECLARATIVE NETS [GLOUD ET AL. 16', 19']
  - ↳ OPTNET [AMOS + KOLTER 17']
  - ↳ CVXPY LAYERS [AGARWAL ET AL. 19']
  - ↳ SatNet [WANG ET AL., 19']
  - ↳ SUBMODULAR OPTIM. [DJOLONGA + KRAUSE, 19']

SOLVING ODEs ↔ IMPLICIT LAYER

$$\frac{dz}{dt} = f(z(t), t, \theta)$$

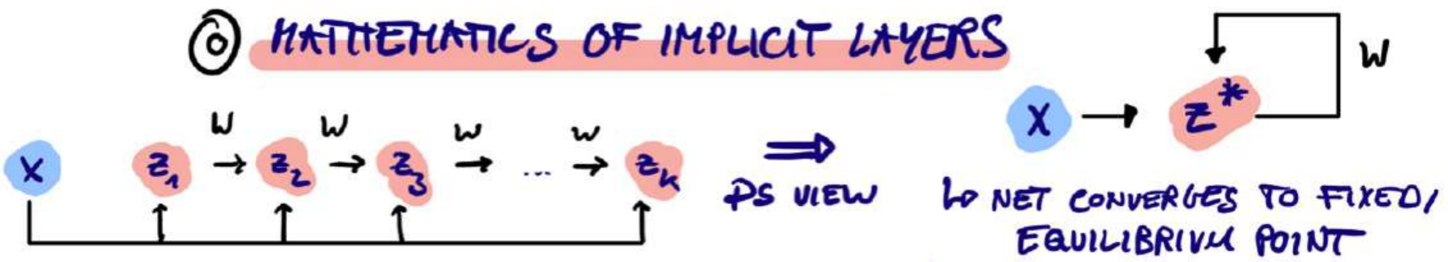
$$\Rightarrow z(t_1) = z(t_0) + \int_{t_0}^{t_1} f(z(t), t, \theta)\, dt$$

$$\Rightarrow y = \text{odeint}(f, x, t_0, t_1, dt)$$

⇒ DROP-IN REPLACEMENT RESNET

- CONTINUOUS-TIME PHYSICS MODELS
  - ↳ INCORPORATE KNOWN STRUCTURES!
  - ↳ HAMILTONIANS / LAGRANGIANS
- CONTINUOUS NORMALIZING FLOWS
  - ↳ EASIER CHANGE OF VARS. COMPUTATION
  - ↳ AMORTIZATION FOR DENSITY COMPUTATION
  - ↳ HOMEOMORPHISMS ON POINT CLOUDS

⓪ MATHEMATICS OF IMPLICIT LAYERS

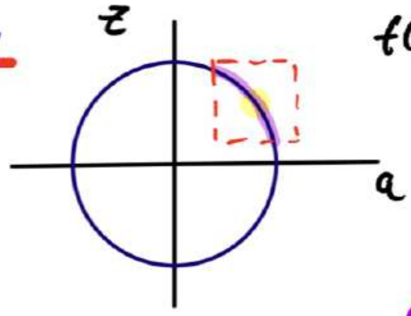


WEIGHT-TYING + RE-INJECT! $\quad z_{i+1} = \sigma(W z_i + x)$

↳ NET CONVERGES TO FIXED/EQUILIBRIUM POINT

↳ HOW TO BACKPROP THROUGH? ⇒ MEMORY PROBLEM!

SOLUTION: IMPLICIT FUNCTION THEOREM



$f(a,z) = a^2 + z^2 - 1 = 0$

LOCAL SOLUTION MAPPING FCT.

$f: \mathbb{R}^p \to \mathbb{R}^n$, $a_0 \in \mathbb{R}^p$, $z_0 \in \mathbb{R}^n$ s.t.

1. $f(a_0, z_0) = 0$
2. f CONT. DIFF. W. NON-SING. JACOBIAN $\partial_1 f(a_0, z_0)$ wrt. z

$\Rightarrow \exists$ OPEN SET $S_{a_0} \subset \mathbb{R}^p$, $S_{z_0} \subset \mathbb{R}^n$ CONTAINING a_0, z_0 & UNIQUE CONT. FCT. $z^*: S_{a_0} \to S_{z_0}$ S.T.

1. $z_0 = z^*(a_0)$,
2. $f(a, z^*(a)) = 0 \ \forall a \in S_{a_0}$
3. z^* DIFF. ON S_{a_0}

↳ FROM EXISTENCE TO DERIVATIVE EXPRESSION:

$$f(a, z^*(a)) = 0, \ \forall a \in S_{a_0}$$

CAN BE z, ETC.

$$\partial_0 f(a, z^*(a)) + \partial_1 f(a, z^*(a)) \partial z^*(a) = 0, \ \forall a \in S_{a_0}$$

$$\Rightarrow \partial z^*(a_0) = -[\partial_1 f(a_0, z_0)]^{-1} \partial_0 f(a_0, z_0)$$

CAN DIFFERENTIATE SOLUTION MAPPING FCT. WRT. PARAMS BY EVALUATING DERIV. OF f CLOSE TO SOLUTION POINT!

$\Rightarrow$ FIXED POINT VERSION: $\partial z^*(a_0) = [I - \partial_0 f(z_0, a_0)]^{-1} \partial_1 f(z_0, a_0)$

AUTODIFF. CONNECTION

1. JCP/FORWARD: $v \mapsto \partial f(x) v$
2. VJP/REVERSE: $w \mapsto w^T \partial f(x)$

↳ $w^T \partial z^*(a_0) = u^T f(z_0, a_0)$

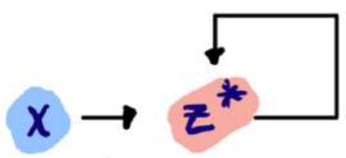
with $u^T = w^T + u^T \partial_0 f(z_0, a_0)$

$\Rightarrow$ BACKWARD PASS $\Leftrightarrow$ LINEAR FIXED POINT IN TERMS OF VJPs!

1 DEEP EQUILIBRIUM MODELS [BAI ET AL., 19'] — DEQ

$\Rightarrow$ REPLACE ENTIRE NET BY EQ. LAYER: $z^* = f(z^*; x, \theta)$

↳ USE STABLE FP SOLVER ALGORITHM



$z^* = f(z^*; x, \theta)$

BACKWARD VIA IMPLICIT GRADS:

$$\partial \ell(\theta) = \partial_0 \ell(z^*; y) \cdot [I - \partial_0 f(z^*; x, \theta)]^{-1} \cdot \partial_2 f(z^*; x, \theta)$$

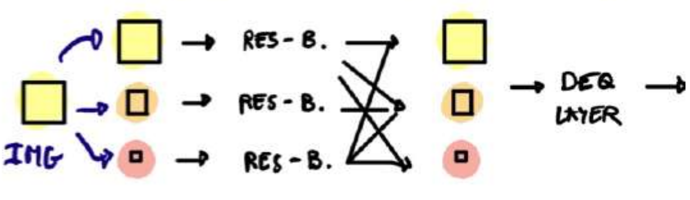
→ PRACTICAL DETAILS: ANDERSON ACCELARATION FOR FP ITERATIONS → FORWARD + BACKWARD!

→ THEORY RESULT I: SINGLE-L. DEQ CAN REPRESENT ANY FFW DEEP NET.

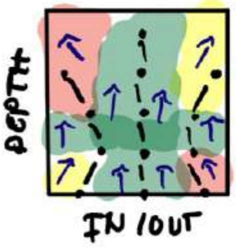
→ THEORY RESULT II: SINGLE-L. DEQ CAN REPRESENT ANY MULTI-LAYER DEQ.

→ BUT: EXISTENCE EQ. POINT, UNIQUENESS, STABILITY?!

→ MULTI-SCALE REPRESENTATIONS ↳ MIXING!



② NEURAL ORDINARY DIFFERENTIAL EQ. [CHEN ET AL., 18']



$y = odeint(f, x, t_0, t_1, dt)$ → USE ANY SOLVER: EULER, RK, ADAPTIVE
↳ FIT LOCAL POLYNOMIAL!

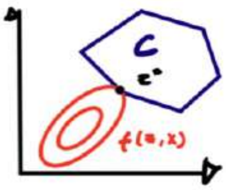
↳ f CONT. DIFF. + LIPSCHITZ

↳ "PROBLEM": "INPUT PATHS" CAN'T OVERLAP! → ONLY BIJECTIVE TRAFOS
↳ OFTEN DYNAMICS BECOME MORE COMPLEX W. TRAINING → MORE FCT. EVALS

$$\frac{\partial}{\partial t}\frac{\partial \mathcal{L}}{\partial z(t)} = \frac{\partial \mathcal{L}}{\partial z(t)}\frac{\partial f(z(t),\theta)}{\partial z} \Rightarrow \frac{\partial \mathcal{L}}{\partial \theta} = \int_{t_0}^{t_1} \frac{\partial \mathcal{L}}{\partial z(t)}\frac{\partial f(z(t),\theta)}{\partial \theta}\, dt$$

↳ CONTINUOUS-TIME BP VIA ADJOINT METHOD: AUGMENT TRACE W. VJP → $O(1)$
↳ USE WHEN YOU CARE ABOUT TRAJECTORY → CHANGE OF VAR. → NORM. FLOWS
↳ TRICK TO GET AROUND JACOBIAN TRACE: HUTCHINSON'S TRACE ESTIMATOR $tr(A) = E_{v \sim N(0,1)}[v^T A v]$
↳ NUMERICAL ERROR ⇒ PROPORTIONAL TO SOLVER TOLERANCE!

③ DIFFERENTIABLE OPTIMIZATION



$$\boxed{\text{LAYER: } z^* = \arg\min_{z \in C} f(z, x)}$$

↳ SOLUTION TO CONSTRAINED OPT. ⟺ SOLUTION TO KKT CONDITIONS
↳ VIEW OPTIM. PROCEDURE AS FP ITERATION ⇒ E.G. PROJECTED GD
$z_{k+1} = Proj_C[z_k - \alpha \partial f(z_k, x)]$ → AGAIN: IMPLICIT FCT. THEOREM
↳ APPLICATIONS: LEARN CONVEX POLYTOPES, MNIST SODUKO, HVAC MPC
↳ CVXPY LAYERS → DIFFERENTIABLE CONVEX MODELING ⇒ EMBED OPTIM. PROBLEM!

FUTURE DIRECTIONS AND OPEN PROBLEMS

DEQ	N.-ODE
o DROP-IN IMPLICIT REPLACEMENT	o CONT. TS → IRREGULAR → PHYSICS
o SUPERVISED LEARNING	o FLEXIBLE DENSITY
o STANDARD UNSUPERVISED	o HOMEOM.

??? REGULARIZING DEQs / NEURAL ODES FOR FASTER SOLVING → PENALIZE DYNAMICS

??? ADAPT ARCHITECTURES TO EXPLOIT MEMORY ADVANTAGES → DEQ "CELL" NAS

??? SCALING / APPLYING LATENT SDEs

??? PDE SOLUTIONS AS A LAYER

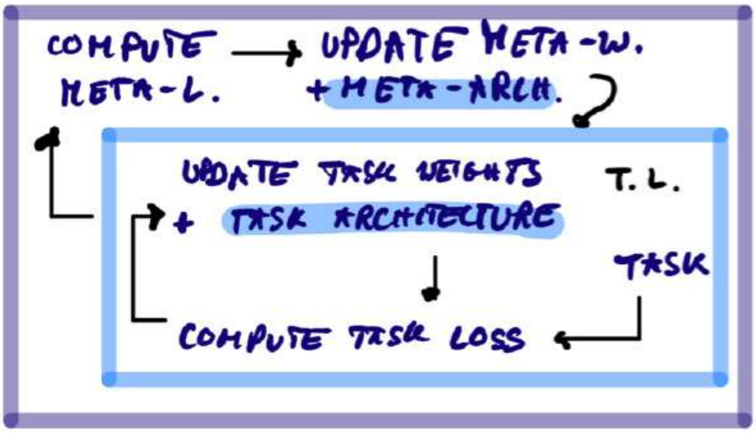
1–3 of 3

F. HUTTER (UoFREIBURG): 'META-LEARNING NEURAL ARCHITECTURES, INITIAL WEIGHTS, HYPERPARAMETERS AND ALGORITHM COMPONENTS'

1 SAMPLE-EFFICIENT JOINT META-LEARNING OF MULTI-COMPONENTS

ELSKEN ET AL. 20' ⇒ METANAS

↳ SIMULTANEOUS WEIGHTS + ARCH.



ZIMMER ET AL. 20' ⇒ AUTO-PYTORCH

↳ MULTI-FIDELITY + ACROSS DATA

↳ OFF-P. RL: SEARL, FRANKE ET AL. 20'

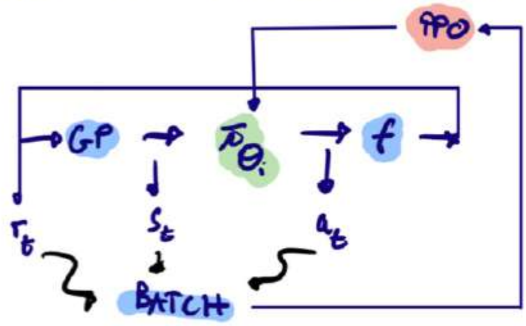
2 META-LEARNING TO IMPROVE EXISTING ALGOS

CHALLENGES WHEN LEARNING FROM SCRATCH:

GENERALIZATION
TH. GUARANTEES
SOTA PERF.
⇒ INSTEAD: IMPROVE EXISTING

↳ ALSO: APPLICATION TO COMB. SOLVER ALGO SELECTION

VOLPP ET AL. 20': META-L. ACQ. FCT. FOR BAYES OPT.



↳ ALSO: DYNAMIC ALGO CONFIG. → ADAPT PARAMS TO CONTEXT

3 BENCHMARKS AS FOUNDATIONS OF MEASURABLE PROGRESS

MDP PLAYGROUND

RAJAN ET AL. 20'

↳ FAST TO RUN

↳ MANY DIMS TO CONTROL

NAS-BENCH-XYZ

↳ FAST EVALUATION BASED ON TABULAR LOOK-UPS ⇒ SINGLE RUN

↳ SIEMS ET AL. 2020 ⇒ SURROGATE MODELS

L. ZINTGRAF (UoOXFORD): 'EXPLORATION IN META-RL'

MOTIVATION: LEARNING A NEW TASK OFTEN REQUIRES EXPLORATION

① EXPLORE @ META-TEST TIME VS. EXPLORE @ META-TRAIN TIME

EP. BEFORE EVAL

A MANY-EPIS.-LEARNING

B FEW-EPIS.-LEARNING

C ONLINE ADAPTATION

EXPLORATION INDUCTIVE BIAS →

*. MANY-EPISODE-LEARNING

- ↳ EXPLORATION "FREE" DURING LEARNING
- ↳ GOAL: EFFICIENT DATA COLLECTION
- ↳ LPG [OH ET AL. 20'], METAGENRL [KIRSCH ET AL. 20']
- ↳ MYOPIC EXPLORATION

B. FEW-EPISODE-LEARNING

- ↳ MAX. RETURN AFTER N EPISODES
- ↳ GRADIENT-BASED ADAPTATION
- ↳ AGGREGATION-BASED ADAPT.

C. ONLINE ADAPTATION

- ↳ MAX. EXP. RETURN ONLINE
- ↳ SOLUTION VIA BAYES-ADAPTIVE AGENTS ⇒ DUFF & BARTO 02'

TASK BELIEF ⟶ OPTIMAL ACTION UNDER UNCERT.

→ APPROX. INFERENCE - variBAD [ZINTGRAF ET AL 20'a]

BELIEF INFERENCE VIA VAE ↙ ↘ BAYES-ADAPTIVE POLICY

$b_t \leftarrow q(m \mid \tau_{:t})$ $\quad$ $\pi(a_t \mid s_t, b_t)$

⇒ PROBLEM W. HARD EXPLORATION!

CHALLENGE: NEED TO EXPLORE META-TEST STRATEGIES AT META-TRAIN TIME

↙ CAN BE HARD TO FIND!

↘ STATE VALUE MAY VARY ACROSS TASKS

C. ONLINE ADAPTATION

↳ EXPLORATION BONUS IN HYPER-STATE SPACE ⇒ HyperX [ZINTGRAF ET AL. 20'b]

$$\boxed{r_{env} + r_{hyper} + r_{error}}$$

EXTRINSIC REWARD | HYPER-STATE NOVELTY BONUS | BONUS FOR WRONG B. INF.

B. FEW-EPISODE-LEARNING

→ METACURE [ZHANG ET AL. 20']

↳ MAX. INFO GAIN OF EXPLORATION POLICY

→ EXPLORE-THEN-EXECUTE [LIU ET AL. 20']

↳ EXPLICIT LEARNING OF EXPLORATION π TO RECOVER TASK ID EMBEDDINGS

*. MANY-EPISODE-LEARNING

→ OPEN Q! FUTURE RESEARCH

- ↳ HOW TO DEFINE + SEARCH?!
- ↳ WOULD LIKE TO LEARN HOW TO SET EXPLORATION BONUSES!

L. KIRSCH (IDSIA): 'GENERAL META-LEARNING'

MOTIVATION: MINIMIZE HAND-CRAFTED INDUCTIVE BIASES
↳ NEED FOR BROAD GENERALIZATION TO BE APPLICABLE

① METAGENRL - KIRSCH ET AL. 2020

DIVERSE POPULATION

$r, f(\pi_\phi), V$ INPUTS → L_α MetaGenRL LSTM → LOSS → POLICY UPDATE $\phi' = \phi - \nabla_\phi L_\alpha$

ROLLOUT AGENT ↑ INPUTS

UPDATE OBJECTIVE ↑ L_α: $\alpha \leftarrow \alpha + \nabla_\alpha \sum_i Q_\theta^{(i)}(s, \pi_{\phi_i}^{(i)}(s))$ ← EVALUATE UPDATE $Q_\theta^{(i)}(s, \pi_{\phi_i}^{(i)}(s))$

OFF-POLICY

⇒ GENERALIZATION TO UNSEEN ENVS + OUTPERFORMING PPO VIA
⇒ EXTENSION BY LEARNED PG → OH ET AL 20' 2ND ORDER-GRADS!

② VARIABLE SHARED META-LEARNING (VS-ML)

LEARNED LRULES / FAST WEIGHTS ←→ LEARNING IN RNN ACTIVATIONS

$\Delta w_{ab}(t+1) = f_\theta(x_a(t), x_b(t+1), w_{ab}(t), r_{ab}(t))$

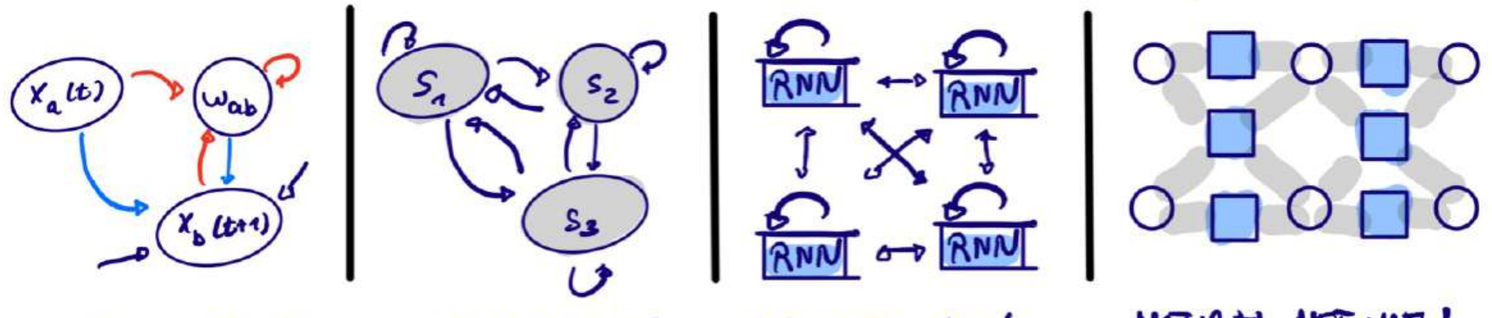
V_M (META) → SLOW
V_L (LEARNED) → FAST

RNN, inputs O_t, r_t → FEEDBACK

V_M: W_{rec}
V_L: h_t-act.

RATIO PROBLEM V_L / V_M
↳ WANT $V_L \gg V_M$!

↳ VS-ML: $s_{muj} \leftarrow \sigma(b_j + \sum_i s_{mui} w_{ij} + \sum_{ci} s_{cmi} V_{ij})$ INTERACTION



FAST WEIGHTS · SINGLE RNN · VS-META-RNN · NEURAL NET VIEW

↳ META-LEARN GENERAL LEARNING ALGO ⇒ MNIST vs. FASHION MNIST

③ BOOTSTRAPPING AI CONJECTURE

GENERAL META-LEARNER ⇄ UNSUPERVISED OBJECTIVE / ENV

INIT + REVERS. LEARN. ALGO · INIT + REVERS. HUMAN ENG.