Francis Bach (INRIA): Optimization

LARGE-SCALE SUPERVISED ML

↙ LARGE n ↘ LARGE d

⇒ IDEAL: RUNNING-TIME COMPLEX. $O(dn)$

OPTIM.: $\frac{1}{n}\sum_{i=1}^{n} f_i(\theta)$ ↳ TRAINING ERROR

STATISTICS: $\mathbb{E}_{p(x,y)}\, \ell(y, h(x,\theta))$ ↳ TESTING COST

⇒ NEED TO RESPECT BOTH!

I. INTRO: SUPERVISED ML

- DATA: $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \dots, n$
  PRED. FCT.: $h(x, \theta) \in \mathbb{R}$ w. $\theta \in \mathbb{R}^d$
- (REG.) EMPIRICAL RISK MINIMIZATION

$$\min_{\theta \in \mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, h(x_i,\theta)) + \lambda \Omega(\theta)$$

$g(\theta)$; $h(x,\theta) = \theta^\top \Phi(x)$

- HINGE: $\ell(\cdot,\cdot) = \max\{1 - y h(x,\theta), 0\}$
- LOGISTIC: $\ell(\cdot,\cdot) = \log(1 + \exp(-y h(x,\theta)))$
- LS REG: $\ell(\cdot,\cdot) = \frac{1}{2}(y - h(x,\theta))^2$

↳ BIG ASS.: LINEARITY $h(\cdot)$ NOT CONVEX. LOSS

SMOOTHNESS + CONVEXITY + GRADIENT DESCENT

- $g: \mathbb{R}^d \to \mathbb{R}$ IS L-SMOOTH IFF ITS TWICE DIFF. + HESSIAN IS BOUNDED:

$$\forall \theta \in \mathbb{R}^d, \ \text{largest eigenv.}[g''(\theta)] \le L$$

- TWICE DIFF. $g: \mathbb{R}^d \to \mathbb{R}$ μ-STRONGLY CONVEX IFF

$$\forall \theta \in \mathbb{R}^d, \ \text{smallest eigenv.}[g''(\theta)] \ge \mu$$

↳ CONDITION NUMBER DETERMINES DIFFICULTY OF PROBLEM:

$$\kappa = L/\mu \ge 1$$

↳ CONVEXITY MAINLY ANALYSIS TOOL → ALGOS SHOULD ALSO RUN IN NON-CONVEX PROBLEM

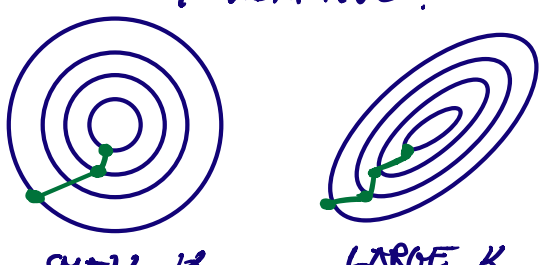
↳ LINEAR CASE $[\Phi(x)\Phi(x)^\top]^{-1} \Rightarrow n \gg d$

- $g(\theta)$: OFTEN WEAKLY CONVEX
  ↳ STRONG-CONVEXITY RESULTS FROM ADDING REGULARIZER!
- GRADIENT DESCENT:
  ↳ g CONVEX + L-SMOOTH ON $\mathbb{R}^d$

$$\theta_t = \theta_{t-1} - \gamma_t g'(\theta_{t-1})$$

γ_t: LINE SEARCH; $g'(\theta_{t-1}) = \frac{1}{n}\sum_{i=1}^{n} f_i'(\theta_{t-1})$

↳ FULLY ADAPTIVE!



SMALL κ LARGE κ

⇒ LINEAR CONVERGENCE $O(1/t)$
⇒ $O(e^{-t/\kappa})$ IF STRONGLY CONVEX!

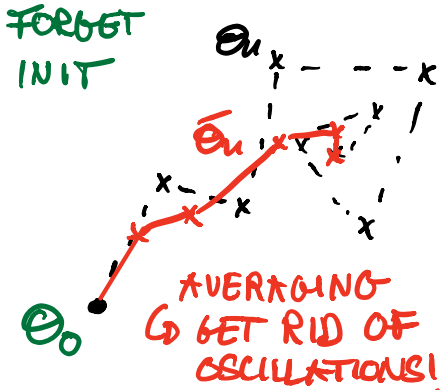
- CHOOSE t (ITERS) PROP. TO CONDITION NUMBER κ!
- ANALYZE $\|\theta_t - \theta_*\|^2$ VS. $g(\theta_t) - g(\theta_*)$
- NEWTON 2ND ORDER METHOD: $\theta_t = \theta_{t-1} - g''(\theta_{t-1})^{-1} g'(\theta_{t-1})$
  - "RECONDITION PROBLEM TO LOOK MORE LIKE BALL" → NATURAL GRADIENTS

ALGO	L-SMOOTH + μ-CONVEX	NON-STRONG CONV.
SGD	$d \times L/\mu \times 1/\varepsilon$	$d \times 1/\varepsilon^2$
BGD	$d \times n \times L/\mu \times \log 1/\varepsilon$	$d \times n/\varepsilon$
ACC. GD	$d \times n\sqrt{L/\mu} \times \log 1/\varepsilon$	$d \times n/\sqrt{\varepsilon}$
SAG	$d \times (n + L/\mu) \times \log 1/\varepsilon$	$d \times \sqrt{n}/\varepsilon$
ACC.	$d \times (n + \sqrt{n L/\mu}) \times \log 1/\varepsilon$	

BOTTOU & BOUSQUET 08':
* NO NEED TO OPTIM BELOW STAT. ERR
* COST FUNCTIONS ARE AVERAGES
* TESTING ERROR

II. FAST STOCHASTIC GRADIENT METHODS FOR CONVEX PROBLEMS

- SGD → POLYAK-RUPPERT AVG: $\bar{\theta}_t = \frac{1}{t+1} \sum_{u=0}^{t} \theta_u$
  - $E\, g(\bar{\theta}_t) - g(\theta_*) \leq \begin{cases} O(1/\sqrt{t}) & \text{if } \gamma_t = 1/(L\sqrt{t}) \\ O(\kappa/t) & \text{if } \gamma_t = 1/(\mu t) \end{cases}$ → MORE ITERS BUT SMALLER BATCH!



  - RUNNING TIME: $O(d \cdot \kappa/\varepsilon)$
  - → PROBLEM: NOT ADAPTIVE IN GENERAL!
- GOAL: BEST OF STOCHASTIC + DETERM. WORLD
  - LINEAR RATE w. $O(d)$ ITER COST
  - ⊕ SIMPLE CHOICE OF STEP CHOICE

- ACCELERATION OF GRADIENT METHODS
  - → NESTEROV 83', 04': $\theta_t = \eta_{t-1} - \gamma_t g'(\eta_{t-1})$ w. $\eta_t = \theta_t + \delta_t(\theta_t - \theta_{t-1})$
  - $g(\theta_t) - g(\theta_*) \leq O(1/t^2)$
  - $g(\theta_t) - g(\theta_*) \leq O(e^{-t/\sqrt{\kappa}})$

} STILL PROBLEM OF COMP. COMPLEXITY!

→ STOCHASTIC AVG. GRADIENT [LE ROUX ET AL., 12']

$$\theta_t = \theta_{t-1} - \frac{\gamma_t}{n} \sum_{i=1}^{n} y_i^t \quad \text{w. } y_i^t = \begin{cases} f_i'(\theta_{t-1}) & \text{if } i = i(t) \\ y_i^{t-1} & \text{, else.} \end{cases}$$

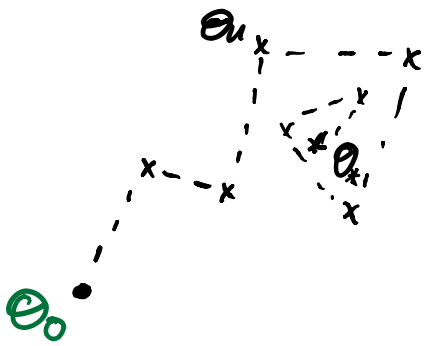
RANDOM SAMPLE W. REPLACEM. → ADD. MEMORY REQ.

↳ DIFFERENT INTERP.: LAZY GRAD EVAL. ⊕ VARIANCE REDUCTION

□ SGD MINIMIZES TEST COST! [INDEP. SAMPLES + SINGLE PASS ⇒ BOUND]

→ OPTIMAL CONV. RATE: $O(1/\sqrt{n})$ & $O(1/n\mu)$

→ CONSTANT STEP-SIZE SGD: LINEAR CONVERGENCE FOR STRONGLY CONV.



⇒ CONSTANT STEP-SIZE FOR LS:

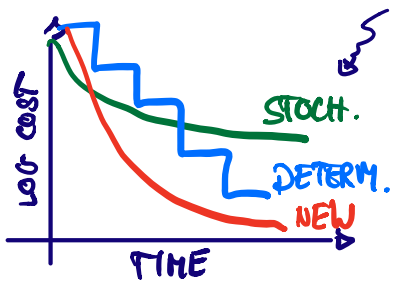
$(\theta_n)_n$ ↔ HOMOGENOUS MARKOV CHAIN

↳ CONVERGENCE TO STATIONARY DISTR.

↳ ERGODIC THEOREM: AVG. ITERATES CONVERGE AT RATE $O(1/n)$

III. BEYOND CONVEX PROBLEMS

□ SO FAR: CONVEX LOSS + LINEAR PREDICTOR $h(x,\theta) = \theta^T \Phi(x)$



□ HOW TO TRANSFER SUCCESS TO DEEP LEARNING?

↳ NON-CONVEX OPTIMIZATION PROBLEM

↳ GENERALIZATION GUARANTEES IN OVERPARAMETRIZED REGIME?!

LOCAL MINIMA

STATIONARY POINTS

PLATEUS

BAD INIT

. . . .

⇒ NO GLOBAL GUARANTEES POSSIBLE!

□ LOCAL THEORETIC GUARANTEES

↳ REPLACE ITERATES BY FUNCTION VALUES!

↳ CHOROMSKA ET AL. '15: MOST LOCAL MINIMA ARE EQUIVALENT! → MANY ASS.

↳ SOLTANOLKOTABI ET AL. '18: NO SPURIOUS LOCAL MINIMA → NOT USED ACTIVATIONS

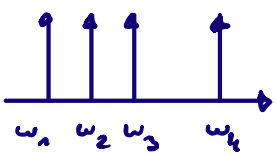
SINGLE HIDDEN LAYER FC NETWORK:

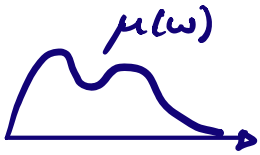
$$h(x) = \frac{1}{m} \theta_2^T \sigma(\theta_1^T x) = \frac{1}{m} \sum_{j=1}^{m} \theta_2(j) \cdot \sigma[\theta_1(\cdot,j)^T x]$$

↳ NO SHARING OF PARAMS

MEAN-FIELD LIMIT PERSPECTIVE! [WEIGHTS = PARTICLES]

$$\sum_{j=1}^{m} \Phi(w_j) = \int_\Omega \Phi(w)\, d\mu(w) \text{ w. } d\mu(w) = \frac{1}{m} \sum_{j=1}^{m} \delta_{w_j}$$





$$\min F(\mu) = R\left(\int_\Omega \Phi(w)\, d\mu(w)\right)$$

↳ GRADIENT FLOW: $\dot{w} = -m \nabla F_m(w)$

↳ LIMIT $m \to \infty$: WASSERSTEIN GRADIENT FLOW

- CHIZAT & BACH '18a : GRADIENT FLOW CONVERGES TO GLOBAL OPTIM.!
- KEY ASSUMPTIONS: HOMOGENEITY [RELU/SIGMOID] + INIT MEASURE SPREAD
  - ↳ NOTE: ONLY QUALITATIVE RESULT! ⇒ CAN WE GET MORE?
- CHIZAT & BACH '18b : LAZY TRAINING [NOT TOO MUCH MOVEMENT F. INIT]
  - ↳ SCALE FACTOR $\alpha > 0$: $G_\alpha(w) = R(\alpha h(w))/\alpha^2$
  - ↳ INIT $w(0)$ S.T. $\alpha h(w(0))$ IS BOUNDED!
  - ↳ AROUND $w(0)$, $G_\alpha(w)$ has
    - * ∇ "prop" to $\nabla R(\alpha h(w_0))/\alpha$
    - * HESSIAN "prop" to $\nabla^2 R(\alpha h(w_0))$

⇒ MAKES MODEL LINEAR! → 1ST ORDER TAYLOR ARGUMENT
↳ NOT TO FAR FROM INIT!!!

- JACOT ET AL. '18: NEURAL TANGENT KERNEL

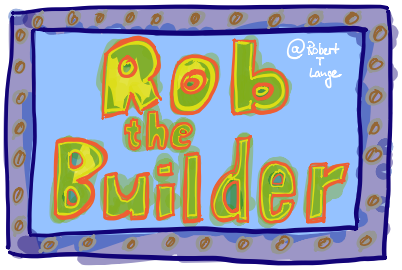
$$h(x,w) \approx h(x,w(0)) + [w - w(0)]^T \nabla h(x, w(0))$$

KERNEL: $k(x,x') = \nabla h(x, w(0))^T \nabla h(x', w(0))$

↳ EFFECT OF SCALE + APPLIES TO ALL DIFF. MODELS

OPEN Qs:
- STEPSIZE SCHEDULES
- NON-CONVEXITY
- DISTRIB. LEARNING

- PHYSICAL REVIEW: "NEW" / "FIRST" SHOULD NOT BE USED IN PAPERS
  - ↳ DON'T OVERHYPE YOUR OWN WORK!

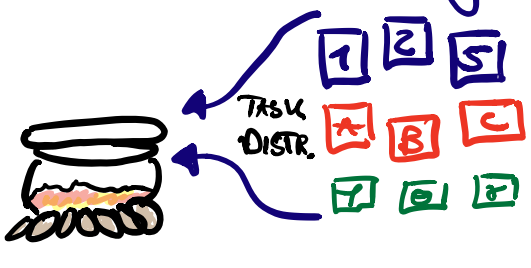


Yee Whye Teh (Ox+DeepMind): Meta-Learning

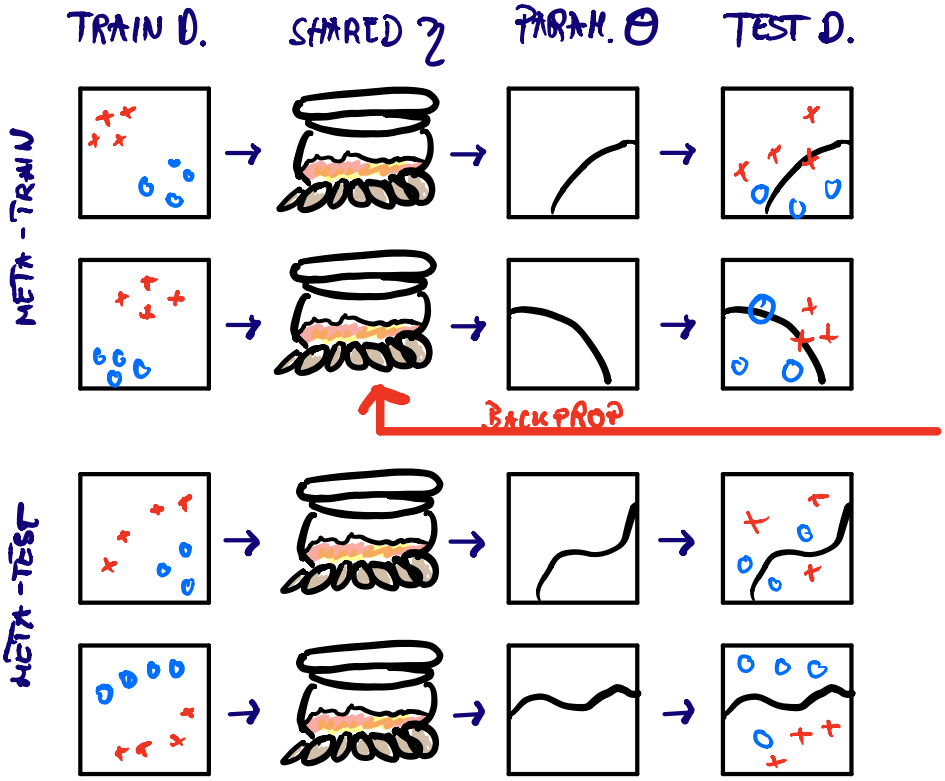
MOTIVATION: SMALL DATA PROBLEMS

- FEW-SHOT LEARNING
- RECOMMENDER SYSTEMS
- ROBOTICS

} LEARN INDUCTIVE BIASES!



OPTIMISATION PERSPECTIVE ON META-LEARNING



INNER LEARN θ:

$$\theta_j = \operatorname{argmin}_{\theta_j} \sum_{i=1}^{n} \text{Loss}\left(f_{\eta,\theta_j}(x_{ji}), y_{ji}\right) =: \text{Learner}(\eta, \text{TrainData}_j)$$

TEST PERFORMANCE:

$$\sum_{i=n+1}^{n+m} \text{Loss}\left(f_{\eta,\theta_j}(x_{ji}), y_{ji}\right) =: \alpha(\eta, \theta_j, \text{TestData}_j)$$

OPTIM. SHARED PARAMS:

$$\operatorname{argmin}_{\eta} \sum_{j=1}^{T} \alpha(\eta, \theta_j, \text{TestData}_j) =: \operatorname{argmin}_{\eta} \sum_{j=1}^{T} \sum_{i=n+1}^{n+m} \text{Loss}\left(f_{\eta,\text{LEARNER}(\eta,\text{Train})}(x_{ji}), y_{ji}\right)$$

DIFFERENT KINDS OF META-LEARNING

- OPTIM-BASED
- BLACK-BOX
- METRIC-BASED
- MEMORY-BASED

DIFFERENT INDUCTIVE BIASES

- PSYCHOLOGICAL THEORIES
- SYMMETRIES/ INVARIANCES/ EQUIVARIANCES
- COMP./ LEARNABILITY CONSTRAINTS

OPTIMISATION-BASED META-LEARNING

- BASE-LEARNER: θ_0

$$\theta_1 \leftarrow \theta_0 + \text{Update}_{\eta}\left[\nabla_{\theta_0} L(.)\right]$$

$$\theta_2 \leftarrow \theta_1 + \text{Update}_{\eta}\left[\nabla_{\theta_1} L(.)\right]$$

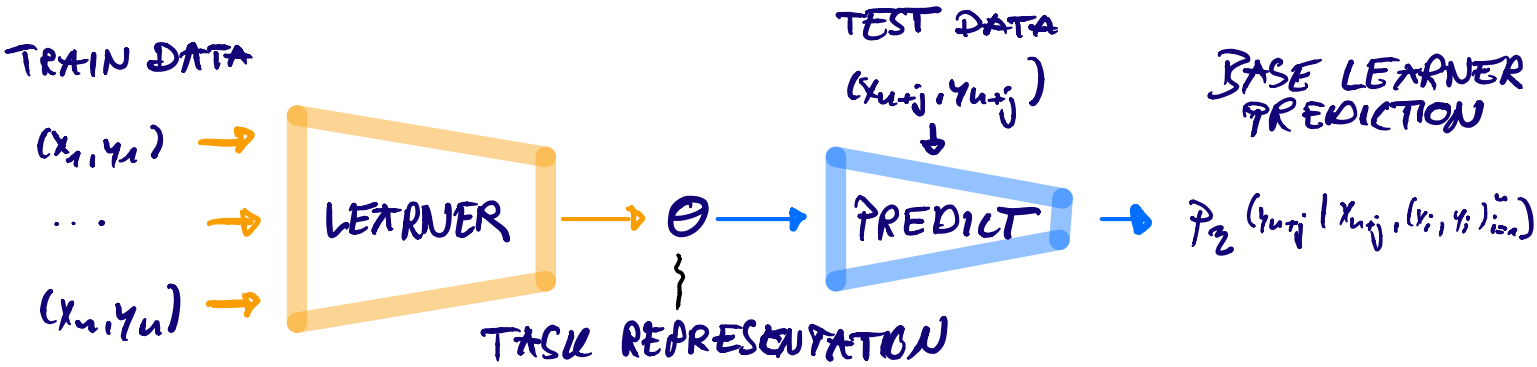
...

- META-LEARN θ_0, η
- L-to-L GD-by-GD, MAML (Andrychowicz '16, Finn '17)

□ 2-LEVEL OPTIMISATION PROBLEM [INNER + OUTER]

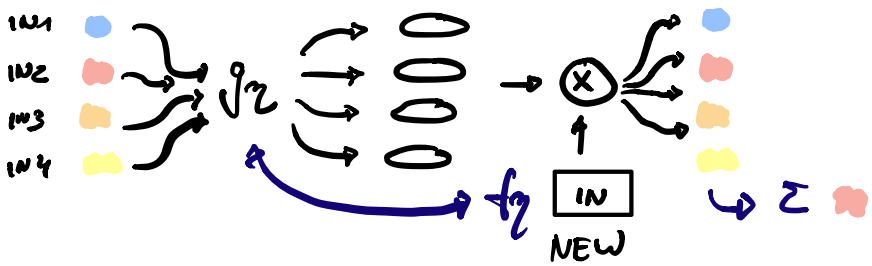
↳ FLEXIBLE + WIDELY APPLICABLE → CAN BE SENSITIVE TO ARCHITECTURE
↳ POSITIVE INDUCTIVE BIAS → CAN BE EXPENSIVE!

BLACK-BOX META-LEARNING → BASE-LEARNER = DIFFERENTIABLE PROGRAM



↳ REDUCES META-LEARNING TO SUPERVISED LEARNING ⇒ BROAD
↳ HARDER 2 LEARN → NO INDUCTIVE BIAS ⇒ LESS GENERALIZATION

□ MATCHING NETWORKS [VINYALS ET AL. 16']

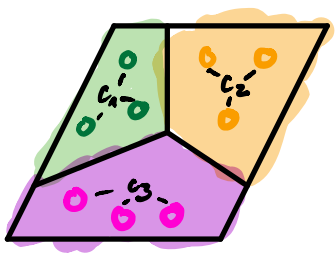


[1] EMBED TRAIN / TEST IN P.
$x_i \mapsto g_\theta(x_i)$
$x_{n+j} \mapsto f_\eta(x_{n+j})$

[2] SOFT-1NN CLASSIFIER

$$\sum_{i=1}^{n} \frac{e^{g_\theta(x_i)^T f_\theta(x_{n+j})}}{\sum_{i=1}^{n} e^{g_\theta(x_i)^T f_\theta(x_{n+j})}} y_i$$

□ PROTOTYPICAL NETWORKS [SNELL ET AL. 17']



[1] EMBED $x_i \mapsto f_\eta(x_i)$

[2] FORM PROTOTYPES $c_k = \sum_{i:y_i=k} f_\eta(x_i) / \sum_{i:y_i=k} 1$

[3] PREDICT: $\frac{e^{-\|f_\eta(x_{n+j}) - c_k\|^2/\sigma^2}}{\sum_{k'=1}^{m} e^{-\|f_\eta(x_{n+j}) - c_{k'}\|^2/\sigma^2}}$

□ MEMORY-AUGMENTED NNS [SANTORO ET AL. 16']

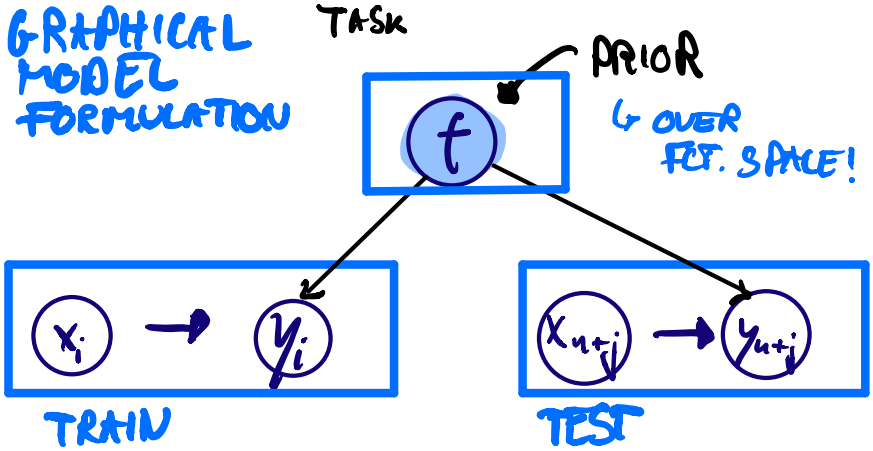
→ TASK = SEQ. $(x_1, y_1), \ldots, (x_T, y_T)$ ⇒ BASE: $p_\theta(y_t \mid (x_i, y_i)_{i=1}^{t-1}, x_t)$
↳ NEURAL TURING MACHINE WITH EXTERNAL MEMORY

□ CONDITIONAL NPs [GARNELO ET AL., 18'] → PERMUTATION INVARIANCE

→ EMBED + AGGREGATE: $\theta = 1/n \sum_{i=1}^{n} f_\eta(x_i, y_i)$ ↔ TASK REPR.
(IN-OUT PAIRS) (INVARIANT!)

PROBABILISTIC PERSPECTIVE ON META-LEARNING

GRAPHICAL MODEL FORMULATION



- STOCH. PROCESS = JOINT DISTR. OVER INF. COLLECTION OF RANDOM VARS. $[f(x), x \in \mathcal{X}]$
- KOLMOGOROV EXTENSION TH.: CONSTRUCT STOCH. PROCESS BY SPECIFYING ITS FINITE DIM. MARGINAL DISTR.

- FAMILY OF FINITE DIM. JOINT DISTR. $\rho_{x_{1:n}}$, ONE FOR EACH $n \in \mathbb{N}$ AND FINITE SEQ. $x_{1:n} \in \mathcal{X}$ → WANT THESE TO FORM MARGINALS:

$$\rho_{x_{1:n}}(y_{1:n}) = p(f(x_1) = y_1, \dots, f(x_n) = y_n)$$

  - EXCHANGEABILITY: FOR EACH n, $x_{1:n}$ AND PERMUTATION π OF $\{1,\dots,n\}$

$$\rho_{x_{1:n}}(y_{1:n}) = \rho_{x_{\pi(1)},\dots,x_{\pi(n)}}(y_{\pi(1)},\dots,y_{\pi(n)})$$

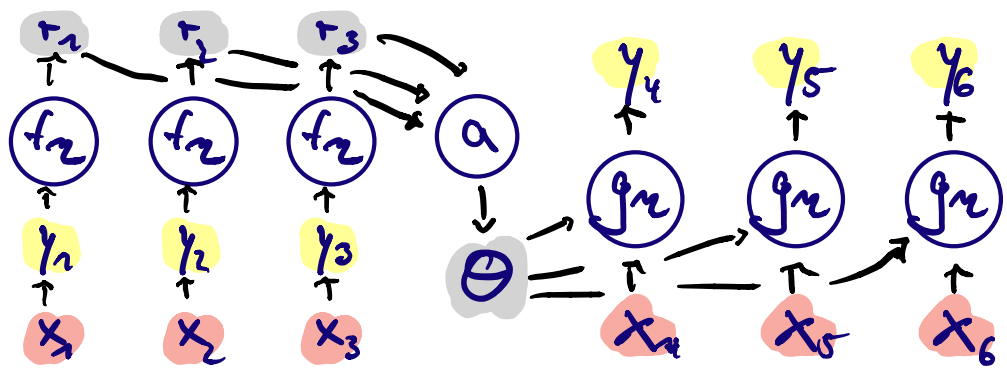
  - CONSISTENCY: FOR EACH n, m, $y_{1:n+m}$

$$\rho_{x_{1:n}}(y_{1:n}) = \int \rho_{x_{1:n+m}}(y_{1:n+m})\, dy_{n+1:n+m}$$

- GENERALLY: STOCH. PROCESS CAN BE DESCRIBED USING CONSISTENT FAMILY OF CONDITIONAL DISTRIBUTIONS:

$$p(f(x_{n+j}) = y_{n+j} \mid f(x_1) = y_1, \dots, f(x_n) = y_n)$$

- CONDITIONAL NEURAL PROCESSES (II)



AGGREGATE

→ PREDICT

$$p_\phi(y_{n+j} \mid x_{n+j}, (x_i, y_i)_{i=1}^n) = \mathcal{N}(y_{n+j}; g_\phi(\theta, x_{n+j}))$$

→ PROBLEM: ARCH. CANNOT MODEL DEP. AMONG TEST OUTPUT

↓

$$p_\phi(z, y_{n+1:n+m} \mid x_{1:n+m}) = p_\phi(z) \prod_{i=1}^{n+m} p_\phi(y_i \mid z, x_i)$$

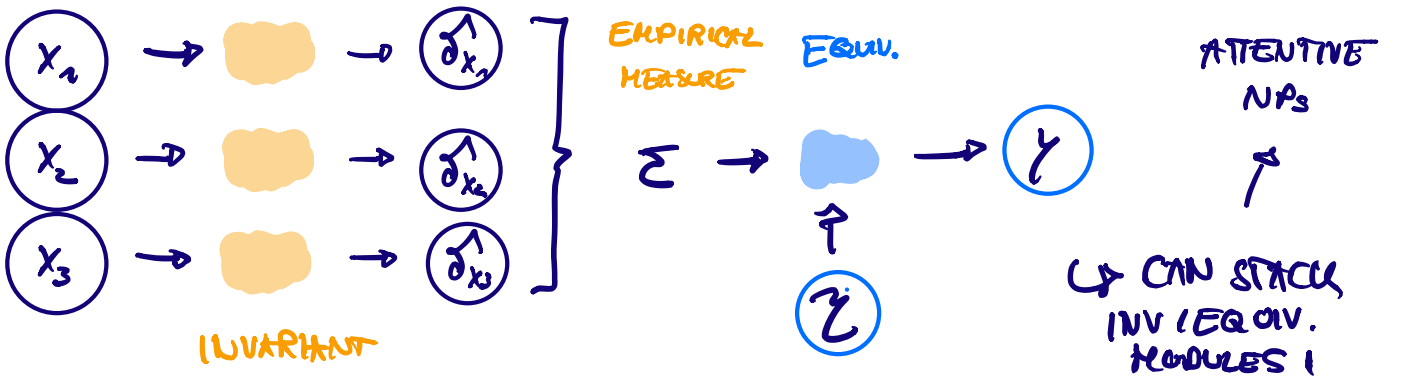
← GEN. LATENT VAR. MODEL ← AUTOREGRESSIVE

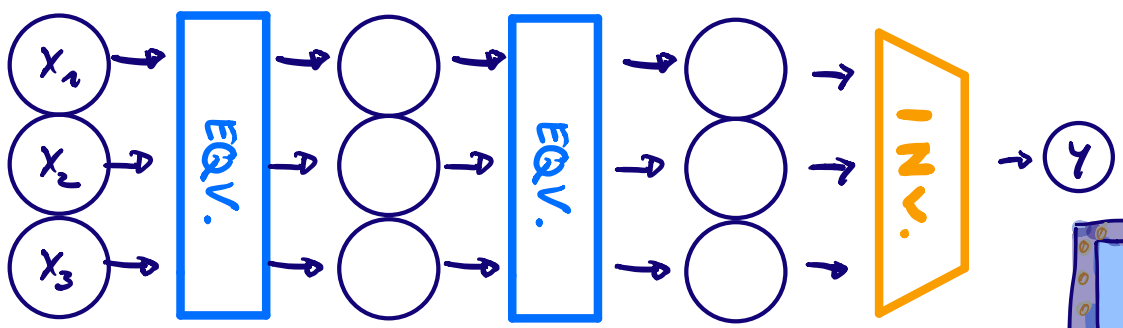
↳ ELBO ON $\log p(y_{n+1:n+m} \mid x_{1:n+m}, y_{1:n})$

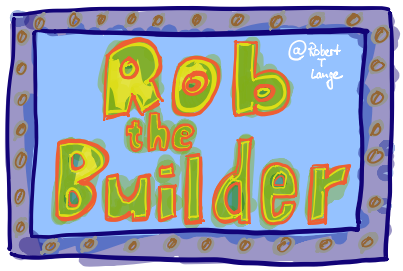
- META-LEARNING AS LEARNING STOCH. PROCESS → PRIOR OVER FCTS.
  - ↳ THINK OF BASE LEARNER AS FORM OF AMORTIZED LEARNING
  - ↳ IMPORTANCE OF UNCERTAINTY: ACTIVE LEARNING, BO, RL
- EFFICIENT MODEL-BASED RL → META-LEARN $\hat{p}(s'|s,a)$, $\hat{r}(s,a)$
  - ↳ ADVERSARIAL TESTING OF RL AGENTS → BAYES OPTIMIZATION PROBLEM ⇒ GALASHOV ET AL. '19 → MAZE CONSTRUCTION

PROBABILISTIC SYMMETRIES AND NEURAL ARCHITECTURES

- CHARACTERISING PERMUTATION-INVARIANT FCTS. $h: \mathcal{X}^n \to \mathcal{Y}$
  - → IS PERM.-INV.: $h(\pi \cdot (x_1, \dots, x_n)) = h(x_1, \dots, x_n)$
  - → IS PERM.-EQUIV.: $h(x_1, \dots, x_n) = (y_1, \dots, y_n) \to h(\pi \cdot (x_1, \dots, x_n)) = \pi \cdot h(x_1, \dots, x_n)$
- + COND. DISTR. $P(y|x)$ IS STOCH. RELAX. FOR FCT. $y = h(x)$
  - → $P(y|x)$ IS G-INVARIANT IF: $P(y|x) = P(y|g \cdot x)$
  - → $P(y|x)$ IS G-EQUIVARIANT IF: $P(y|x) = P(g \cdot y | g \cdot x)$
- EMP. MEASURE OF x_n: $M_{x_n}(\cdot) = \sum_{i=1}^{n} \delta_{x_i}(\cdot)$
  - ↳ SUFFICIENT STATS → ADEQUATE STATS → NOISE OUTSOURCING







Shakir Mohamed (DeepMind): Bayesian Learning

I. BAYESIAN BASICS

$p(x) \geq 0, \int p(x)dx = 1$ — $\mathbb{E}_{p_\theta(x|z)}[f(x;\theta)]$ — $\nabla_\phi f(x;\phi)$

PROB. — EXPECTATION — GRADIENT

PROBABILISTIC QUANTITIES

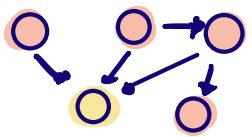
BAYES RULE: $p(z|x) = p(x|z)\,p(z)/p(x)$ (INVERSE PROB.)

REPARAM.: $p_\theta(x|z) = p(x|z;\theta)$

PROBABILITY
- STATISTICAL
- SUBJECTIVE → DEGREE OF BELIEF! ⇒ OBSERVER
- LOGICAL
- PROPENSITY

PROBABILISTIC MODEL

→ MODEL = DESCRIPTION OF WORLD, DATA, POTENTIAL SCENARIOS, PROCESSES
- ↳ PROBABILISTIC: USES LANGUAGE OF PROB. TO WRITE OUT!
- ⇒ AIM = LEARN PROB. DISTR. OF DATA → REASONING



PROBABILITY OF A SEQUENCE

COIN TOSS SEQUENCE: 1, 0, 1, 1, ...

→ EXCHANGEABILITY: $p(x_1,\dots,x_N) = p(x_{\pi_1},\dots,x_{\pi_N})$ ⇒ INFINITE – JOINT INV. TO PERMUTATION

→ DEFINETTI: $p(x_1,\dots,x_N) = \int \prod_{n=1}^{N} p(x_n|\theta)\,P(d\theta)$ ⇒ PARAMETERS, PRIORS, AVERAGE

ASSUMPTION OF DENSITY!

$$p(x_1,\dots,x_N) = \int \prod_{n=1}^{N} p(x_n|\theta)\,p(\theta)\,d\theta$$

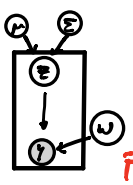
- $p(x_1,\dots,x_N)$: DATA $x_1,\dots,x_N$ IS CONDITIONALLY INDEP.
- $p(x_n|\theta)$: LIKELIHOOD
- $p(\theta)$: PRIOR
- θ: PARAMETER

MODEL-BASED! JOINT PROB. VIA PARAM. LIKELIHOOD

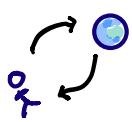
→ EVIDENCE: $p(y|x) = \int p(y|h(x);\theta)\,p(\theta)\,d\theta$ → INTEGRATE OUT UNOBS.
- ↳ OFTEN INTRACTABLE

→ POSTERIOR: $p(\theta|y,x) \propto p(y|h(x);\theta)\,p(\theta)$

REGRESSION & CLASSIFICATION — PROB. MODELS OVER FCT.



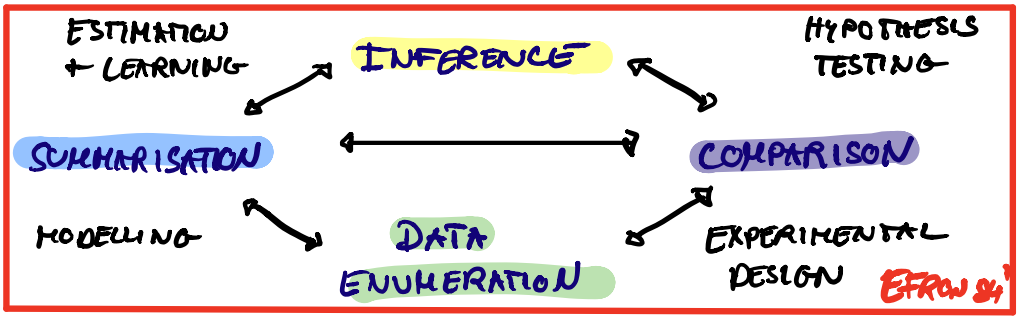
DENSITY ESTIMATION — PROB. DISTR. OVER DATA



DECISION MAKING — PROB. MODELS OVER ENVS / ROBOTS

MARR'S LEVELS OF ANALYSIS:
- COMPUTATIONAL
- ALGORITHMIC
- IMPLEMENTATION

SHAKIR'S 4 LEVELS:
- APPLICATIONS
- REASONING
- INFORMATION
- PRINCIPLES



MODELS — LEARNING PRINCIPLES — ALGORITHMS

1 MODELS → 2 LEARNING PRINCIPLES → 3 ALGORITHMS

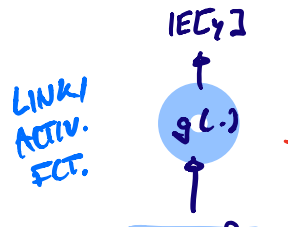
- DIRECTED / UNDIRECTED
- PARAM. / NON-PARAM. / SEMI-PARAM.
- FULLY-OBS. / LATENT VARS.

2 LEARNING PRINCIPLES
- DIRECT
  - LAPLACE APPROX.
    - MAP
    - VI
    - ...
  - MAX LH.
- INDIRECT
  - 2-SAMPLE COMP.
    - ABC
  - METHOD OF MOMENTS
  - TRANSPORT METHODS

3 ALGORITHMS (MANY DIFF. IMPLEMENT.)
- CNNs + REG.
  - RBMs + MAX LH
- IMPLICIT GEN. MODEL + 2 SAMPLE COMP.
- ...

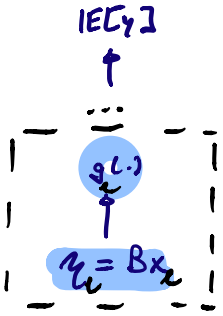
II. BAYESIAN COMPUTATION

LINEAR REGRESSION



OPTIM. NEG. LOG-LH.
$-\log p(y \mid g(z); \theta)$

→ NON-LINEAR COMPOSITION →



DEEP LEARNING

HIERARCHICAL MODELS

↳ DECOMP. OF PROB. DISTR. INTO SEQ. OF CONDITIONAL DISTRIBUTIONS
- DIR. GEN. MODEL
- HIERARCH. MoE
- INFO BOTTLEN.

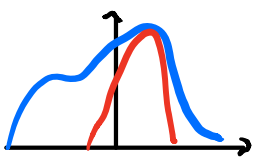
→ LIKELIHOOD FUNCTION: $\mathcal{L}(\theta) = \sum_u \log p(y_u \mid x_u; \theta)$ → LIKELIHOOD OF PARAMS.

- EFFICIENT ESTIMATORS
- TESTS W. GOOD POWER
- POOL INFORMATION
- WIDELY-APPLICABLE
- MISSPECIFICATION POSSIBLE

↳ MAX LH: $\max_\theta \mathcal{L}(\theta)$ ⇒ MAP: $\max_\theta \mathcal{L}(\theta) + \frac{1}{2} R(\theta)$ ⇒ SHRINKAGE

→ INVARIANT MAP: $p(y \mid h(x); \theta)\, p(\theta)\, |I(\theta)|^{1/2}$ → FISHER INFO → REMOVE PARAM. SENSITIVITY

→ BAYESIAN INFERENCE: EVIDENCE ⇒ $\boxed{p(x) = \int p(x \mid \theta)\, p(\theta)\, d\theta}$ & POSTERIOR: $\boxed{p(\theta \mid x)}$



1 CONJUGACY = POSTERIOR IS IN SAME FORM OF OBS. DISTRIB.

↳ TYPICALLY: CONJUGATE PRIORS FALL INTO EXPONENTIAL DISTRIB.

2 INTEGRAL APPROX. → E.G. 1ST ORDER TAYLOR ⇒ LAPLACE!

→ LEARNING AND INFERENCE IN MACHINE LEARNING

→ INFERENCE = REASON ABOUT / COMPUTE UNKNOWN PROB. DISTR.

→ LEARNING = FINDING GOOD POINT ESTIMATES OF MODEL QUANTITIES

→ INFERENTIAL Qs!

EVIDENCE ESTIMATION
$p(x) = \int p(x, z)\, dz$

MOMENT COMPUTATION
$\mathbb{E}[f(z) \mid x] = \int f(z)\, p(z \mid x)\, dz$

PARAMETER ESTIMATION
$p(\theta \mid x_{0:N})$

PREDICTION
$p(x_{t+1} \mid x_{0:t})$

PLANNING
$\mathbb{E}_p\left[\int_0^\infty c(x_t)\, dt \mid x_0, u\right]$

HYPOTHESIS T.
$\log p(x \mid H_1) - \log p(x \mid H_2)$

EXPERIMENTAL DESIGN
$D[p(x_{t:T} \mid u) \,\|\, p(x_{0:t})]$

SOLUTIONISM — FORMALISM

!!! NEUTRALITY TRAPS !!!

PORTABILITY — RIPPLE EFFECT

III. BAYESIAN APPROXIMATION

DIRECT STAT. INFERENCE

- EM
- SEQ. MC
- INF. LAPLACE
- MCMC
- NOISE CONTRASTIVE
- DIRECT COMP. $p(x)$
- LAPLACE APPROX.
- MEAN LH
- VI
- MAP
- CAVITY METHODS

ANALYTIC — SAMPLING PROC.

DISTRIBUTION REPRESENTATION

APPROX. — SAMPLES

→ MONTE CARLO METHODS:

↳ INTEGRAL: $\mathcal{F}(\theta) = \int f(x)\, p(x|\theta)\, dx$

↳ SAMPLE ESTIM.: $\hat{\mathcal{F}}(\theta) = \frac{1}{N} \sum_{i=1}^{N} f(\hat{x}^{(i)}), \; \hat{x}^{(i)} \sim p(x|\theta)$

DRAW EVAL AVG.

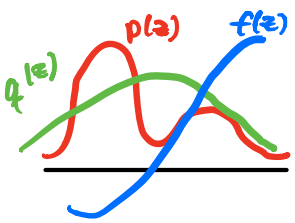
CONSISTENCY — ESTIMATE → TRUE

UNBIASEDNESS — TRUE CENTRING

LOW VARIANCE — ESTIM. ↔ RV

COMPUTATION — EFFICIENT

IMPORTANCE SAMPLING



$$\mu = \mathbb{E}_{q(z)}\left[f(z)\,\frac{p(z)}{q(z)}\right]$$

↳ IDENTITY TRICK ⇒ EASY SAMPLING!

MARKOV CHAIN MC

⇒ GEN. SAMPLES USING SEARCH PROCESS

1ST-ORDER MC: $p(x_t | x_{1:t-1}) = p(x_t | x_{t-1})$

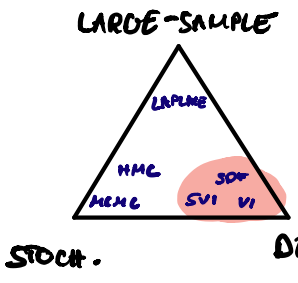
REVERSIBILITY: $Q(x|y)\,\pi(y) = Q(y|x)\,\pi(x)$

- IRREDUCIBILITY ⇒ DON'T GET STUCK!
- APERIODICITY ⇒ NO LOOP BETW. STATES
- ⇒ ERGODICITY ⇒ STAT. DISTRIB.

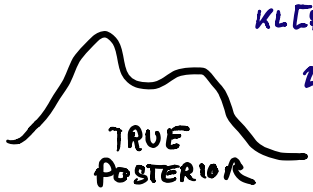
LIMITATIONS: × COMP. INTENSIVE ← & MIXING GUARANTEES

× SOLVE BAYES PROBLEM USING FREQ. METHOD // PAST DEP. UNKNOWN

↳ MANY DIFFERENT VERSIONS ⇒ QUASI-MC, BAYESIAN MC



VARIATIONAL METHODS



→ DETERM. APPROX. PROCEDURE

↳ FIT PARAMS!

↳ $p(x) = \int p(x|z)\, p(z)/q(z)\, dz$ [IS] ← FROM IMPORTANCE SAMPL. TO VARIATIONAL INF.

$$\log p(x) \geq \int q(z) \log\left(p(x|z)\, p(z)/q(z)\right) dz \quad \text{[JENSEN'S INEQ.]}$$

$$= \int q(z) \log p(x|z) - \int q(z) \log \frac{q(z)}{p(z)}$$

↳ TURN INTEGRATION INTO OPTIM. PROBLEM

$F(x, q)$ ⇒ $\mathbb{E}_{q(z)}[\log p(x|z)] - KL[q(z) \| p(z)]$

APPROX. POSTERIORS — RECONSTRUCTION — PENALTY → SMOOTH

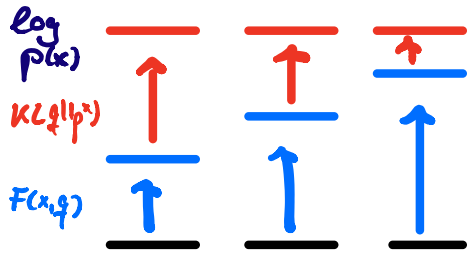
[VARIATIONAL LOWER BOUND]

CLASSICAL INFERENCE APPROACH

SVI ← MODERN METHODS

$p(x,z)$, $q_\phi(z|x)$ → $\int(\dots) q_\phi(z|x)\,dx$ → ∇_ϕ

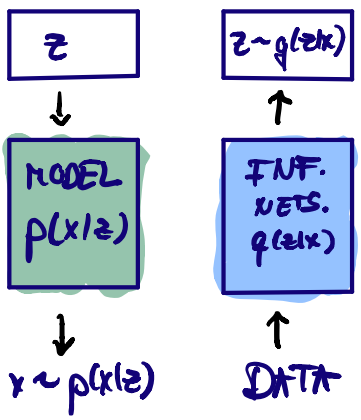
E-STEP M-STEP



LOG-DERIVATIVE TRICK: $\nabla_\phi \log q_\phi(z) = \frac{\nabla_\phi q_\phi(z)}{q_\phi(z)}$ → REWRITE SO THAT 1. SAMPLE + 2. EVALUATE

↳ SCORE FCT. GRADIENT: $\nabla_\phi \mathbb{E}_{q_\phi(z)}[f_\theta(z)] = \mathbb{E}_{q_\phi(z)}[f(z)\nabla_\phi \log q_\phi(z)]$

VAES



⇒ SPEC. COMB. OF VI IN LATENT VAR. MODEL USING INF. NETS

IV. BAYESIAN FUTURES

INFERENCE ⟷ DECISION MAKING

↓ KNOW ABOUT DATA ↓ DO WITH DATA

BAYESIAN RL

↳ ENV AS GEN. PROCESS

$u(s,a) \sim Env(a)$

↳ $p(R(s)|a) \propto \exp(u(s,a))$

$\mathcal{F}(\theta) = \mathbb{E}_{\pi(a|s)}[R(s,a)] - KL[\pi_\theta(a|s)\|p(a)]$

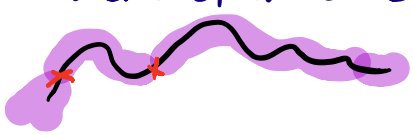
↳ $\nabla_\theta \mathcal{F}(\theta) = \mathbb{E}_{\pi(a|s)}[(R(s,a)-c)\nabla_\theta \log \pi_\theta(a|s)] + \nabla_\theta \mathcal{H}[\pi_\theta]$

⇒ HIERARCHICAL PLANNING VIA ACTION PLANNING

DECOLONIALIZE → CONSIDER CONTEXT/HISTORICAL BASIS OF OUR WORK ⇒ BE CRITICAL!

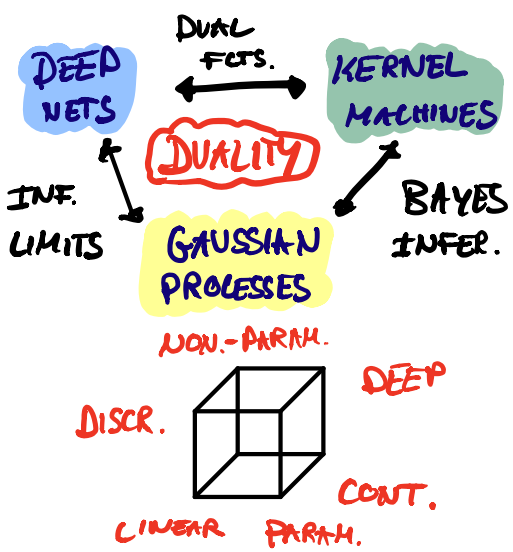
CONTEXTUAL VALUES → RESPECT, BENEFICIENCE, JUSTICE, DO NO HARM!, EXPLICABILITY

BAYESIAN OPTIMIZATION ⇒ GLOBAL OPTIM. OF BLACK-BOX $f(x)$



$p(f|D) \leftrightarrow$ GP

↳ ACQ. FCT.



PROB. NUMERICS

⇒ THINK OF NUMERICAL METHODS AS BAYESIAN LEARNING ALGORITHMS

↳ ODEs / PDEs / LINE SEARCH