VARIATIONAL INFERENCE

IDEA: PICK APPROXIMATION $Q(x)$, MAKE IT AS CLOSE AS POSSIBLE TO TRUE POSTERIOR $P^*(x) = P(x|D)$ | COMPUTE NODE FULL CONDITIONAL, AVERAGE OUT NEIGHBOURS.

→ MINIMIZE KL DIVERGENCE $KL(P^*||Q)$ OR ITS REVERSE $KL(Q||P^*)$, BECAUSE EXPECTATIONS WRT Q ARE TRACTABLE.

→ $J(Q) = \sum_x Q(x) \frac{\log Q(x)}{Z P^*(x)} = \sum Q(x) \log \frac{Q(x)}{P^*(x)} - \log Z = KL(Q||P) - \log Z$ BECAUSE Z IS CONSTANT

→ $J(Q) = KL(Q||P^*) - \log Z \geq -\log Z = -\log P(D)$, $J(Q)$ IS UPPER BOUND ON NLL. $Z = P(D)$, NORMALIZATION CONSTANT

→ ALTERNATIVE: $L(Q) = -J(Q) = \log P(D)$, MAXIMIZE LOWER BOUND ON DATA LL. ENERGY FUNCTIONAL

RELATED TO EM! LOWER BOUNDS!

→ $J(Q) = E_Q[\log Q(x)] + E_Q[-\log \tilde{P}(x)] = -H(Q) + E_Q[E(x)]$ • EXPECTED ENERGY – ENTROPY OF SYSTEM J = VARIATIONAL FREE ENERGY

$L\ E(x) = -\log \tilde{P}(x)$

→ $J(Q) = E_Q[-\log P(D|x)] + KL(Q(x)||P(x))$ • EXPECTED NLL + PENALTY TERM

REVERSE KL $KL(Q||P) = \sum_x Q(x) \ln \frac{Q(x)}{P(x)}$. INFORMATION, I-PROJECTION. ZERO-FORCING FOR Q UNDERESTIMATES P SUPPORT

FORWARDS KL $KL(P||Q) = \sum_x P(x) \ln \frac{P(x)}{Q(x)}$ MOMENT, M-PROJECTION. ZERO-AVOIDING FOR Q. OVERESTIMATES P SUPPORT

• IF TRUE DISTRIBUTION MULTIMODAL → FORWARDS KL IS GENERALLY BAD IDEA

ALPHA DIVERGENCE: FAMILY OF DIVERGENCE MEASURES $D_\alpha(P||Q) = \frac{4}{1-\alpha^2}\left(1 - \int P(x)^{(1+\alpha)/2} Q(x)^{(1+\alpha)/2} dx\right)$ 0 IFF P=Q

• NOT SYMMETRIC → NOT A TRUE METRIC

$\alpha \to 1 = KL(P||Q)$

$\alpha \to -1 = KL(Q||P)$

$\alpha \to 0 = D_H(P||Q) = \int \left(P(x)^{1/2} Q(x)^{1/2}\right)^2 dx$ ~ HELLINGER DISTANCE $\sqrt{D_H(P||Q)}$ IS TRUE DISTANCE METRIC

MEAN FIELD METHOD

ASSUMPTION: POSTERIOR IS FULLY-FACTORED APPROXIMATION $Q(x) = \prod_i Q_i(x_i)$ GOAL: $\min_{Q_1,...,Q_D} KL(Q||P)$ OPTIMIZING OVER PARAMS OF EACH MARGINAL Q_i

• COORDINATE DESCENT, UPDATE $\log Q_j(x_j) = E_{-Q_j}[\log \tilde{P}(x)] + const$ EACH STEP. $\tilde{P}(x) = P(x, D)$ UNNORMALIZED POSTERIOR, E_{-Q_j} EXPECTATION OVER $f(x)$ WRT ALL VARS EXCEPT x_j

• WHEN UPDATING Q_j WE REASON IN TERMS OF J'S MARKOV BLANKET, OTHERS ARE ABSORBED INTO CONSTANT.

• CAN BE USED FOR MANY MODELS

• UPDATE DERIVATION MINIMIZE $L(Q_j) \longrightarrow L(Q_j) = -KL(Q_j||f_j) \to Q_j = f_j \to \log Q_j(x_j) = E_{-Q_j}[\log \tilde{P}(x)] + const$ (MIN KL)

• SIMILAR TO GIBBS SAMPLING BUT SENDS MEAN MSGS.

MEAN FIELD FOR ISING MODEL

EXAMPLE → IMAGE DENOISING FACTORED APPROXIMATION IS $Q(x) = \prod_i Q(x_i, \mu_i)$. μ_i MEAN VALUE

• UPPER BOUND, REVERSE KL

• $\log \tilde{P}(x) = x_i \sum W_{ij} x_j + L_i(x_i) + const$

• $q_i(x_i) \propto \exp\left(x_i \sum_{j \in nbr} W_{ij}\mu_j + L_i(x_i)\right)$ → REPLACE NEIGHBOR W/ AVG VALUES

• $m_i = \sum W_{ij} \mu_j$

• $a_i = m_i + 0{,}5(L_i^+ - L_i^-)$ APPROX MARGINAL POSTERIOR

• $\mu_i = E_{Q_i}[x_i] = \tanh(a_i) \longrightarrow \mu_i = \tanh\left(\sum W_{ij}\mu_j + 0{,}5(L_i^+ - L_i^-)\right)$ CAN TURN INTO FIXED POINT

• DAMPED UPDATES $\mu_i^t = (1-\lambda)\mu_i^{t-1} + \lambda \tanh\left(\sum W_{ij}\mu_j^{t-1} + 0{,}5(L_i^+ - L_i^-)\right)$

STRUCTURED MEAN FIELD

EXPLOIT TRACTABLE SUBSTRUCTURE IN PROBLEM. GROUP VAR SETS TOGETHER AND UPDATE SIMULTANEOUSLY.

EXAMPLE: FACTORIAL HMM

M CHAINS, LENGTH T, K STATES $P(x,y) = \prod_M \prod_T (x_{TM} | x_{T-1,M}) P(y_t | x_{tM})$ EACH CHAIN IS APRIORI INDEPENDENT BUT COUPLED IN POSTERIOR DUE TO OBSERVATIONS. JUNCTION TREE IS $O(TMK^{M+1})$ MEAN FIELD IS $O(TM^2K^2I)$

- APPROXIMATION: PRODUCT OF CHAINS $Q(x_{TM} | x_{T-1,M}, \xi_{TM}) = \prod^{M} \left(\xi_{TMK} \prod_{j=1}^{K} (A_{MJK})^{x_{T-1,MJ}} \right)^{x_{TMK}}$ EACH CHAIN INDIVIDUALLY UPDATED WITH FWD-BWD

  $E_Q[x] = \langle \text{NO WAY} \rangle$, $E(x,y)$ EXACT POSTERIOR

- OBJECTIVE: $KL(Q \| P) = E(E] - E[E_Q] - \log Z_Q + \log Z$
- UPDATES $\xi_{TM} = \exp\left(W_M^T \Sigma^{-1} \tilde{y}_{tM} - \frac{1}{2} \delta_M\right)$

  $\delta_M = \mathrm{DIAG}(W_M^T \Sigma^{-1} W)$

  $\tilde{y}_{tM} = y_t - \sum_M W_\ell E[x_{t,\ell}]$
- ξ_{TM} IS LOCAL EVIDENCE, AVGES OVER NEIGHBOURING CHAINS.

  FWD-BWD UPDATE EACH CHAIN IN PARALLEL
- $O(TMK^2)$ FOR FULL UPDATE SWEEP.

VARIATIONAL BAYES

TO INFER MODEL PARAMETERS, NOT HIDDEN VARS. $P(\theta | D) \approx \prod_K Q(\theta_K)$ • LOWER BOUND

IF PARAMS + LATENT INFERENCE $\longrightarrow$ VARIATIONAL BAYES EM

VB FOR UNIVARIATE GAUSSIAN

CONJUGATE PRIOR: $P(\mu, \lambda) = N(\mu | \mu_0, (\kappa_0 \lambda)^{-1}) \, Ga(\lambda | a_0, b_0)$ APPROX FACTORED POSTERIOR: $Q(\mu, \lambda) = Q_\mu(\mu) Q_\lambda(\lambda)$

TARGET $\log \tilde{p}(\lambda, \mu) = \frac{N}{2} \log \lambda - \frac{\lambda}{2} \sum_{i}^{N} (x_i - \mu)^2 - \frac{\kappa_0 \lambda}{2} (\mu - \mu_0)^2 + \frac{1}{2} \log(\kappa_0 \lambda) + (a_0 - 1) \log \lambda - b_0 \lambda + \text{const}$

$q_\mu(\mu) = N(\mu | \mu_N, \kappa_N^{-1})$ OBTAINED FROM LOG = AVERAGE OVER λ $\quad q_\lambda(\lambda) = Ga(\lambda | a_N, b_N)$ • COMPUTE EXPECTATIONS, DERIVE EXPLICIT UPDATE FORMS $\mu_N, \kappa_N, a_N, b_N$. a_N, μ_N CONSTANTS $\rightarrow$ FIXED POINT UPDATES FOR OTHERS

- OBJECTIVE: MINIMIZE $L(q)$, LOWER BOUND ON LOG MARGINAL LL

$L(q) \leq \log p(D) = \log \iint p(D | \mu, \lambda) p(\mu, \lambda) \, d\mu \, d\lambda = \langle \text{TEDIOUS ALGEBRA} \rangle = \frac{1}{2} \log \frac{1}{\kappa_N} + \log \Gamma(a_N) - a_N \log b_N + \text{const}$ MONOTONICALLY INCREASES WITH UPDATES

- PUSH LOG INTO SUM
- INTEGRALS BECOME EXPECTATIONS
- ...

VB FOR LINEAR REGRESSION

PRIOR: $P(w, \alpha, \lambda) = N(w | 0, (\lambda \alpha)^{-1} I) \, Ga(\lambda | a_0^\lambda, b_0^\lambda) \, Ga(\alpha | a_0^\alpha, b_0^\alpha)$ FACTORED POSTERIOR: $Q(w, \alpha, \lambda) = Q(w, \lambda) Q(\alpha)$

OPTIMAL POSTERIOR: $Q(w, \alpha, \lambda) = N(w | w_N, \lambda^{-1} V_N) \, Ga(\lambda | a_N^\lambda, b_N^\lambda) \, Ga(\alpha | a_N^\alpha, b_N^\alpha)$ ⟨MURPHY P 747⟩

- CAN FORMULATE WITH ARD PRIORS
- ALTERNATE UPDATES $Q(w, \lambda)$ AND $Q(\alpha)$
- POSTERIOR PREDICTIVE ONCE PARAMS INFERRED $P(y | x, D) = T\left(y | w_N^T x, \frac{b_N^\lambda}{a_N^\lambda}(1 + x^T V_N x), 2a_N^\lambda\right)$
- EXACT MLL $P(D) = \iiint p(y | x, w, \lambda) p(w | \alpha) p(\lambda) \, dw \, d\alpha \, d\lambda$
- SIMILARITIES TO EMPIRICAL BAYES: MAX $\log p(D)$ WHILE VB MAXES LOWER BOUND ON IT

VARIATIONAL BAYES EM

WITH LATENTS AND PARAMS $z_i \to x_i \leftarrow \theta$. MIXTURE MODELS, PCA, HMM. VBEM IS MORE BAYESIAN, MODELS UNCERTAINTY IN θ TOO.
SAME COMPUTATIONAL COST AS REGULAR EM. **MEAN FIELD:** $p(\theta, z_{1:N}|D) \approx Q(\theta)Q(z) = Q(\theta)\prod_i Q(z_i)$

- **VARIATIONAL E STEP:** UPDATES $Q(z_i|D)$. AVERAGES OVER PARAMETERS INSTEAD THAN PLUGGING IN $\hat{\theta}$ MAP ESTIMATES. PLUG IN POSTERIOR MEAN THEN USE STUFF SUCH AS FWD-BWD
- **VARIATIONAL M STEP:** UPDATES $Q(\theta|D)$. UPDATES HYPERPARAMS USING EXPECTED SUFFICIENT STATISTICS. **VBEM = EM** $\to q(\theta|D) \approx \delta_{\hat{\theta}}(\theta)$
- **ADVANTAGE** MARGINALIZE PARAMS OUT $\longrightarrow$ CAN COMPUTE LOWER BOUND $\longrightarrow$ MODEL SELECTION

VBEM FOR MIXTURE OF GAUSSIANS

- **LIKELIHOOD** $p(z,x|\theta) = \prod_i \prod_k \pi_k^{z_{ik}} \mathcal{N}(x_i|\mu_k, \Lambda_k^{-1})^{z_{ik}}$
- **EXACT PRIOR:** $p(\theta) = \mathrm{Dir}(\pi|\alpha_0)\prod_k \mathcal{N}(\mu_k|m_0, (\beta_0\Lambda_k)^{-1})\,\mathrm{Wi}(\Lambda|L_0, \nu_0)$ ASSUME ALL PRIOR PARAMS ARE SAME FOR ALL CLUSTERS
- **FACTORED POSTERIOR:** $Q(z,\theta) = Q(z|\theta)Q(\theta) = \underbrace{\left[\prod_i \mathrm{Cat}(z_i|r_i)\right]}_{\text{E-STEP}}\underbrace{\left[\mathrm{Dir}(\pi,\alpha)\prod_k \mathcal{N}(\mu_k|m_k, (\beta_k\Lambda_k)^{-1})\,\mathrm{Wi}(\Lambda_k|L_k, \nu_k)\right]}_{\text{M-STEP}}$
- **ML LOWER BOUND** $L = \sum_z \int Q(z,\theta)\log\frac{p(x,z,\theta)}{Q(z,\theta)}\,d\theta \le \log p(D)$
- **POSTERIOR PREDICTIVE:** SUM OF WEIGHTED Ts
  $p(x|D) = \sum_z \int p(x|z,\theta)p(z|\theta)Q(\theta)\,d\theta = \sum_k \frac{\alpha_k}{\sum \alpha_{k'}} T(x|m_k, M_k, \nu_k, 1-D)$
- **MODEL SELECTION:** SELECT k. FIT SEVERAL MODELS AND COMPARE TO LOWER BOUND ON ML
  WATCH OUT FOR UNIDENTIFIABILITY, $k!$ EQUIVALENT MODES. $\log p(D|k) \approx \log(k!) + L(k)$
- **SPARSITY:** FIT A SINGLE MODEL WITH LARGE k AND $\alpha_0 \ll 1$, ENCOURAGES SPARSE MIXING VECTOR. IN VBEM MIXING WEIGHTS ARE SUBJECT TO A PENALTY MORE SEVERE FOR SMALL CLUSTERS (FEW WEIGHTED COMPONENTS) $\longrightarrow$ THEY WILL EMPTY OUT OVER ITERATIONS
  EFFICIENT WAY TO SEARCH OPTIMAL NO OF CLUSTERS, SPIKES NEAR EDGES OF SIMPLEX

VARIATIONAL MESSAGE PASSING

GEN PURPOSE METHOD FOR DGM WHOSE CPD ARE IN EXP FAMILY AND PARENTS ARE CONJ. SWEEP OVER GRAPH AND 1 AT TIME UPDATE. **REPARAM TRICK!**
GOOD FOR MODELS WHERE HIDDENS ARE CONTINUOUS. VIBES. **VIABES/MF OK FOR CONTINUOUS LATENTS**

LOCAL VARIATIONAL BOUNDS

WE NOW REPLACE A TERM IN THE JOINT WITH A SIMPLER ONE TO FACILITATE COMPUTING THE POSTERIOR. **LOCAL VARIATIONAL APPROXIMATION**
GENERALLY NEEDED WHEN GAUSSIAN PRIOR × MULTINOMIAL LIKELIHOOD: MULTI-TASK LEARNING, DISCRETE FACTOR ANALYSIS, CORRELATE TOPIC MODEL

GENERAL DIFFICULTY: $\mathrm{LSE}(\eta_i) = \log\left(1 + \sum^M e^{\eta_{i,M}}\right)$ LOG-SUM-EXP, **LIKELIHOOD:** $p(Y|X,w) = \prod^N \exp[y_i^T\eta_i - \mathrm{LSE}(\eta_i)]$ ← **LOGISTIC REGRESSION**

COMES FROM GAUSSIAN PRIOR × MULTINOMIAL LIKELIHOOD

MULTICLASS LOGISTIC REGRESSION

BOHNING QUADRATIC BOUND: $p(y_i|x_i,w) \ge f(x_i, \psi_i)\mathcal{N}(\tilde{y}_i|x_i, w, A_i^{-1}) \longrightarrow$ NOW EASY TO COMPUTE **APPROX POSTERIOR** $Q(w) = \mathcal{N}(m_N, V_N)$

- COMPUTE $L(q)$; INTRODUCE BOUNDS $\longrightarrow L(q) \ge$ [VERY LONG EXPRESSION] $\longrightarrow$ USE COORDINATE ASCENT $\longrightarrow$ UPDATE VARIATIONAL LIKELIHOOD ψ
  UPDATE VARIATIONAL POSTERIOR V_N, m_N

SIGMOID FUNCTION BOUND: JJ BOUND. HAS ADAPTIVE CURVATURE TERM, OR STILL BOHNING BOUND

$\log(1+e^\eta) \le \frac{1}{2}a(\xi)\eta^2 - b(\xi)\eta + c(\xi)$
- HAS ADAPTIVE CURVATURE

$\log(1+e^\eta) \le \frac{1}{2}a\eta^2 - b\eta + c$
- CONSTANT CURVATURE

- a, b, c, ARE STUFF.

OTHER BOUNDS, EVEN MORE MESSY:

*** SOMETIMES WE DO VARIATIONAL INFERENCE WITH UPPER BOUNDS INSTEAD ***

- PRODUCT OF SIGMOIDS (MURPHY P. 762-3)
- JENSEN'S INEQUALITY ALL ON LSE FUNCTION
- MULTIVARIATE DELTA

MOAR VARIATIONAL INFERENCE

Q NOW ISN'T FACTORIZED, NOT EVEN GLOBALLY VALID JOINT. JUST LOCALLY CONSISTENT → DIST OF TWO ADJACENT NODES AGREES WITH CORRESPONDING MARGINALS

LOOPY BELIEF PROPAGATION

IDEA: DISREGARD LOOPS, APPLY BELIEF PROPAGATION UNTIL 'CONVERGENCE'!

- ON PAIRWISE MODELS (BINARY ~~TREES~~ GRAPHS), WORKS WELL! BECAUSE THEY'RE TREE-LIKE LOCALLY. CYCLE IS $\log m$ LONG.
- ON FACTOR GRAPHS

  FACTOR GRAPH: WAY TO UNIFY DGM AND UGM. UNDIRECTED BIPARTITE GRAPH WITH ROUND NODES, VARIABLES, AND SQUARE NODES, FACTORS. EDGES FROM EACH VAR TO FACTOR THAT MENTION IT. IN UGM → FACTOR = POTENTIALS. IN DGM FACTORS → CPDs

$V_M \to F_{ACTOR}$ MSG: $m_{x\to f}(x) = \prod_{h\in nbr(x)\setminus\{f\}} m_{h\to x}(x)$ F→V MSG: $m_{f\to x}(x) = \sum_y f(x,y) \prod_{y\in nbr(f)\setminus\{x\}} m_{y\to f}(y)$ $BEL(x) \propto \prod_{f\in nbr(x)} m_{f\to x}(x)$

CONVERGENCE

NO GUARANTEES LBP WILL CONVERGE. INVESTIGATE WITH COMPUTATION TREES. T STEPS OF LBP → EXACT COMPUTATION IN TREE OF HEIGHT T+1

IF EDGE STRENGHTS ARE SUFF. WEAK → INFLUENCE DIMINISHES OVER DIFFERENT PASSES

DAMPING: SEND OUT DAMPED MESSAGES $M^n_{ts}(x_s) = \lambda M_{ts}(x) + (1-\lambda) M^{n-1}_{ts}(x_s)$, USUALLY $\lambda \sim 0.5$

MESSAGE SCHEDULING:

- SYNCHRONOUS: ABSORB/UPDATE ALL IN PARALLEL
- ASYNCHRONOUS: COMPUTE USING NEW MSG FROM EARLIER IN THE ORDERING, OLD FROM LATER IN THE ORDERING.

  HOW TO ORDER? FIXED, RANDOM, PICKING SPANNING TREES AND SWEEP ONE AT A TIME; ADAPTIVELY → BASED ON NORM OF DIFF FROM PREVIOUS VALUES. HIGHEST FIRST.

  TREE REPARAMETRIZATION RESIDUAL BELIEF PROPAGATION

ACCURACY: SINGLE LOOP → EXACT MAP ESTIMATES. ELSE NOT BUT ERROR CAN BE BOUNDED. GAUSSIAN MODELS → IF CONVERGES, MEANS EXACT, OVERCONFIDENT VARIANCES

SPEEDUP TRICKS:

- LARGE STATE SPACES: EACH LBP MSG IS $O(K^2)$. ~~IN~~ MANY STATES (IE, 256 FOR VISION) IS TOO MUCH. FFT MAKES IT $O(K\log K)$ BECAUSE MESSAGES ARE JUST CONVOLUTIONS. $\psi_{st}(x_s,x_t) = \psi(x_s - x_t)$. ALSO DISTANCE TRANSFORM IF APPLICABLE $O(K)$
- MULTI-SCALE: 2D LATTICES, VISION, INITIALIZE ACTUAL GRID BY COMPUTING VALS ON A STACK OF COARSER GRIDS.
- CASCADE: FILTER ON SPEED/ACCURACY TRADEOFF, PRUNE IMPROBABLE STATES.

LBP FROM VARIATIONAL POV

- WHY IS LBP VARIATIONAL??
- UGM IN EXP FAMILY REPRESENTATION: $P(x|\theta) = \frac{1}{Z(\theta)} \exp(-E(x))$, $E(x) = -\theta^T\phi(x)$. θ NODE/EDGE PARAMS ϕ EDGE INDICATORS (SUFF STATS)

  $\mu = E[\phi(x)]$ VECTOR MARGINALS. COMPLETELY CHARACTERIZES $P(x|\theta)$. MEAN PARAMS.

- MARGINAL POLYTOPE: $M(G)$, SPACE OF ALLOWABLE μ VECTORS. SET OF ALL MEAN PARAMS. GENERATED FROM VALID PROBABILITY DISTRIBUTION. OBTAINED BY TAKING CONVEX COMBINATION OF $\phi(x)$ → IS CONVEX HULL OF FEATURE SET. $M(G) = \mathrm{conv}\{\phi_1(x) \ldots \phi_D(x)\}$ IT DEFINES A VOLUME, OR INTERSECTION OF HALF-PLANES.
- INFERENCE AS VARIATIONAL PROBLEM: V.I. FINDS Q MAXIMIZING ENERGY FUNCTIONAL $L(Q) = \underbrace{E_Q[\log \tilde{P}(x)] + H(Q)}_{= -KL(q\|p)} \le \log Z$

  $\to \log \tilde{P} = \theta^T\phi(x)$, $Q = P \to \max_{\mu\in M(G)} \theta^T\mu + H(\mu)$ μ IS JOINT DISTRIBUTION OVER ALL STATE CONFIGURATION ENTROPY! YAY! $\mu = E_P[\phi(x)]$

  $\max_{\mu\in M(G)} \theta^T\mu + H(\mu) = \log Z(\theta)$

• MEAN FIELD AS VARIATIONAL: NATURAL PARAMS ASSOCIATED W/ SUFF STATS $\notin$ OUTSIDE CLASS $\to 0$

$M_F(G) = \{\mu \in \mathbb{R}^D : \mu = E_\theta[\phi(x)] \text{ for } \theta \in \Omega\}$ IS **INNER APPROXIMATION** OF MARGINAL POLYTOPE $M_F \subseteq M_G$

IS NON-CONVEX $\to$ MULTIPLE LOCAL OPTIMA

MEAN FIELD ENERGY FUNCTIONAL: $\max_{\mu \in M_F(G)} \theta^T\mu + H(\mu) \le \log Z(\theta)$

- MF MAXIMIZES A CONCAVE OBJECTIVE OVER A NON-CONVEX SET

• LBP AS VARIATIONAL: $M(G)$ HERE IS EXPONENTIALLY LARGE. WE RELAX CONSTRAINTS TO **LOCAL CONSISTENCY**: $\mu \to \tau$

PSEUDO-MARGINALS

- $\sum_{x_s} \tau_s(x_s) = 1$
- $\sum_{x_t} \tau_{st}(x_s, x_t) = \tau_s(x_s)$

SUM-TO-1, MARGINALIZA

- $L(G) = \{\tau \ge 0 : \text{ constraints hold}\}$ IS **CONVEX OUTER APPROXIMATION** ON $M(G)$, $M(G) \subseteq L(G)$
- IF GRAPH IS A TREE $M(G) = L(G)$
- **ENTROPY:** $H(\mu) = \sum_{s \in V} H_s(\mu_s) - \sum_{st \in E} I_{st}(\mu_{st})$. EXACT FOR TREES
- **BETHE APPROX:** $H_{BETHE}(\tau) = \sum H_s(\tau_s) - \sum I_{st}(\tau_{st}) \longrightarrow$ SCREW IT, USE IT EVEN IF GRAPH IS NOT TREE
- **BETHE FREE ENERGY:** $F_{BETHE}(\tau) = -\left[\theta^T\tau + H_{BETHE}(\tau)\right]$
- **LBP OBJECTIVE:** $\min_{\tau \in L(G)} F_{BETHE}(\tau) = \max_{\tau \in L(G)} \theta^T\tau + H_{BETHE}(\tau)$
  - OPTIMIZE NON-CONCAVE OBJECTIVE OVER CONVEX SET MULTIPLE LOCAL OPTIMA, APPROXES $\log Z(\theta)$, EXACT IF TREE
- ANY FIXED POINT OF LBP ALGO IS STATIONARY POINT FOR CONSTRAINED OBJECTIVE

MF VS LBP

LBP EXACT FOR TREES, MF NOT

LBP OPTIMIZES W/ NODE, EDGE MARGINALS; MF ONLY NODE $\to$ LBP MORE ACCURATE

IF EDGE MARGINALS FACTORIZE $\longrightarrow$ SAME FREE ENERGY APPROXIMATION

MF OBJECTIVE HAS MORE LOCAL OPTIMA $\to$ HARDER; CRAP IF STARTS FROM UNIFORM MSG BUT OK IF WE INIT MF WITH LBP MARGINALS

MF USEFUL BECAUSE GIVES A LOWER BOUND.

INIT MF WITH UNIFORM / RANDOM IS KINDA BAD $\to$ INIT WITH BP MARGINALS

GENERALIZED BP

CLUSTER VARIATIONAL METHOD: CLUSTER TOGETHER NODES FORMING A TIGHT LOOP $\to$ **HYPER EDGES** BETWEEN SETS OF VERTICES REPRESENT USING POSETS. IF HYPEREDGE SIZE = TREEWIDTH $\longrightarrow$ GRAPH IS TREE, METHOD IS EXACT.

ENTROPY: $H_{KIKUCHI}(\tau) = \sum_{g \in E} c(g) H_g(\tau_g)$. H ENTROPY OF JOINT OF VERTICES IN SET g. $c(g)$ IS OVERCOUNTING NUMBER

ENERGY FUNCTIONAL: $F_{KIKUCHI}(\tau) = -\left[\theta^T\tau + H_{KIKUCHI}(\tau)\right]$ **VARIATIONAL PROBLEM:** $\min_{\tau \in L(G)} F_{KIKUCHI}(\tau) = \max_{\tau \in L(G)} \theta^T\tau + H_{KIKUCHI}(\tau)$

- OBJECTIVE NOT CONCAVE. GENERALIZED BP ALGO $\to$ MORE ACCURATE THAN LBP, INCREASED COMPUTATIONAL COST

CONVEX BP

GETS A CONCAVE OBJECTIVE ON A CONVEX SET; TREES OR PLANAR GRAPHS

- WORK WITH SUBMODELS $F \subseteq G$. (TREES OR PLANAR GRAPHS).
- FEWER CONSTRAINTS → ENTROPY IS HIGHER THAN G. $H(\mu,\rho)$ IS CONCAVE WRT μ
- ENERGY $F_{CONVEX}(\mu,\rho) = -[\mu^T\theta + H(\mu,\rho)]$. WE HAVE ENTROPY UPPER BOUND
- POLYTOPE: SO THAT PROJECTION OF τ ON G IS ON PROJ OF M ON F.
  $L(G,F) = \{\tau \in \mathbb{R}^d : \tau(F) \in M(F) \ \forall F \in F\}$
- OBJECTIVE: $\min_{\tau \in L(G,F)} F_{CONVEX}(\tau,\rho) = \max_{\tau \in L(G,F)} \tau^T\theta + H(\tau,\rho)$

- TREE-REWEIGHTED BP:
  - CONSIDER SET OF ALL SPANNING TREES OF A GRAPH
  - FOR SINGLE NODES $\rho = 1$, FOR EDGES IS EDGE APPEARANCE PROBABILITY
  - IN THIS CASE $L(G,F) = L(G)$
  - GENERALLY DOES NOT CONVERGE, USE DAMPING OR DOUBLE-LOOP UPDATES
- OPTIMIZATION PROBLEM: $\max_{\tau \in L(G)} \left\{ \tau^T\theta + \sum_{s \in V} H_s(\tau_s) - \sum_{s,t \in E(G)} \rho_{st} I_{st}(\tau_{st}) \right\}$ SAME AS LBP BUT FOR ρ_s
- ALGO: TRBP MESSAGE $T \to S$ IS FCN OF ALL MSG FROM $V \to T + S \to T$
  MAIN DIFFERENCE IS STILL THE ρ WEIGHTS
  IF $\rho_{ST} = 1 \ \forall S,T$ IS STANDARD LBP BUT $\rho_{ST} = 1 \xleftrightarrow{IFF}$ ORIGINAL GRAPH IS ALREADY A TREE.

EXPECTATION PROPAGATION

BP WITH APPROXIMATED MSGS. GENERALIZES ASSUMED DENSITY FILTERING, APPROX POSTERIOR AT EACH TIME USING ASSUMED FUNCTIONAL FORM, EXTENDS ADF

CLUTTER PROBLEM: INFERRING UNKNOWN VECTOR x, WHEN OBSERVATION MODEL IS MIXTURE OF TWO GAUSSIANS, ONE AT x, ONE AT 0.

- $P(y|x) = (1-w)N(y|x,I) + wN(y|0,aI)$:
- WITH FIXED PRIORS: NOT EXPONENTIAL FORM
- INFERENCE SPACE $M(\phi,\Phi)$ IS SET OF MEAN PARAMS REALIZABLE BY ANY PROB. DISTRIBUTION 'SEEN' THROUGH SUFFICIENT STATISTICS
  → IS INTRACTABLE, 2^N MODES.
- IDEA: USE Φ AUGMENTED DISTRIBUTIONS, INCORPORATE INTRACTABLE TERMS AND WORK ITERATIVELY $P(x|\theta,\tilde{\theta}_i) \propto f_0(x)\exp(\theta^T\phi(x))\exp(\tilde{\theta}_i^T\Phi_i(x))$
  → $P(x|\theta,\tilde{\theta}_i) = \exp(-\frac{1}{2}x^T\Sigma^{-1}x)[wN(y_i|0,aI) + (1-w)N(y_i|x_i,I)]$ NOW IS TRACTABLE!!
  → APPROX $M(\phi,\Phi)$ WITH $L(\phi,\Phi)$, WHICH STILL CONVEX
  → VARIATIONAL PROBLEM: $\max_{(\tau,\tilde{\tau}) \in L(\phi,\Phi)} \tau^T\theta + \tilde{\tau}^T\tilde{\theta} + H_{EP}(\tau,\tilde{\tau})$

ENTROPY: $H_{EP}(\tau,\tilde{\tau}) = H(\tau) + \sum_{1}^{d_I}[H(\tau,\tilde{\tau}_i) - H(\tau)]$

OPTIMIZE VIA MOMENT-MATCHING:
PICK A TERM i,
COMPUTE Q_i (REMOVE OLD APPROX Φ_i) → FACTOR
COMPUTE NEW Q WITH MOMENT MATCHING. ~~AN~~ SOLVING VAR. OBJECTIVE
COMPUTE NEW FACTOR MESSAGE
AT (IF) CONVERGENCE: APPROX MARGINAL LIKELIHOOD $P(D) \approx \int \prod \tilde{f}_i(x)dx$

- BETTER ACCURACY THAN VARBAYES OR MCMC PER UNIT OF CPU TIME
- LBP IS SPECIAL CASE OF EP WHERE BASE DISTRIBUTION HAS NODE MARGINALS AND INTRACTABLE TERM ARE EDGE POTENTIALS
- APPLICATION: XBOX TRUESKILL

MAP STATE ESTIMATION

FINDING MOST PROBABLE CONFIGURATION OF VARIABLES FOR DISCRETE-STATE GM. IF TREEWIDTH LOW → CAN USE JUNCTION TREE.

$x^* = \arg\max_{x \in X} \theta^T\phi(x)$, θ = NODE + FACTOR POTENTIALS.

LINEAR PROGRAMMING RELAXATION: VARIATIONAL FORM $\max_{\mu \in M(G)} \theta^T\mu \le \max_{\tau \in L(G)} \theta^T\tau$ L IS CONVEX OUTER BOUND OF POLYTOPE

- OPTIMIZE WITH DISTRIBUTED MESSAGE-PASSING

OBS! OBJECTIVE IS SAME AS STD VARIATIONAL OBJECTIVE WITHOUT ENTROPY TERM → ZERO TEMPERATURE LIMIT: DISTRIBUTION HAS ALL ITS MASS CENTERED ON MODE. BETHE OBJECTIVE DOES NOT WORK, BECAUSE NOT CONCAVE.
USE TRBP, WITH SPECIAL PARALLEL SCHEDULING ALWAYS CONVERGES

GRAPHCUTS

FIND MAP STATE ESTIMATES / MIN ENERGY CONFIGURATIONS USING MAX FLOW/MIN CUT ALGORITHMS ON GRAPHS

- ADD SOURCE S AND SINK T TO GRAPH.
- COMPUTE MINIMUM S–T CUT: PARTITION OF S, T CONNECTED NODES MINIMIZING SUM OF EDGE COSTS BETWEEN NODES ON DIFFERENT SIDES OF PARTITION
  - → EQUIVALENT TO MINIMIZING ENERGY
  - → SUBMODULAR ENERGIES: SUM OF DIAGONAL ENERGIES $<$ SUM OF OFF-DIAGONAL ENERGIES. E.G. STD ISING MODEL WITH $\lambda > 0$. EXACT MAP.
  - → NONBINARY METRIC MRF: REQUIRES PAIRWISE ENERGIES FORM A METRIC, METRIC MRF. PICKS ONE STATE, EACH VAR STAYS SAME OR MOVES TO α. α-EXPANSION: STRONG LOCAL OPTIMUM AT EVERY STEP, BETTER THAN GREEDY SEARCH
    α-β SWAP: VARS FLIP STATE IF $<$ ENERGY

GRAPHCUTS VS BP

TRBP AND CUTS RULE. VANILLA BP NOT SO MUCH. CUTS ARE FASTEST BY LITTLE MARGIN.

DUAL DECOMPOSITION:

$P^* = \max \sum \theta_i(x_i) + \sum \theta_f(x_f)$ WE CAN LOCALLY OPTIMIZE EACH FACTOR BUT COMBINATION MAKES THINGS UNTRACTABLE.

IDEA: OPTIMIZE EACH TERM INDEPENDENTLY, THEN INTRODUCE CONSTRAINTS FORCING STUFF TO AGREE.

- LAGRANGIAN MULTIPLIERS EVERYWHERE $\xrightarrow{\delta}$ $L(\delta) = \max L(\delta, x, x')$. IS DUAL OF LP RELAXATION. ALLOWS TO MIX AND MATCH DIFFERENT OPTIMIZATION ALGOS
- GRADIENT DESCENT: MUST USE SUBGRADIENT BECAUSE $L(\delta)$ IS CONVEX BUT NONDIFFABLE AT POINTS. PARALLEL UPDATES.
- COORD DESCENT: UPDATE ALL δ VECTOR AT ONCE. WITH MAX PRODUCT LINEAR PROGRAMMING, USUALLY FASTER THAN GRAD. NO CONVERGENCE GUARANTEES THO.
- X^* RECOVERY: GENERALLY NP-HARD, OR IF LOCALLY DECODABLE → EACH $\theta_i^{-\delta}$ HAS UNIQUE MAXIMUM x_i^*, LP RELAXATION IS UNIQUE

VARIATIONAL INFERENCE - FROM BISHOP

LOG-MARGINAL

BASIC: FUNCTIONALS. IE, ENTROPY OF PROB DISTR IS A FUNCTIONAL. $\log p(x) = L(q) + KL(q\|p)$. $L(q) = \int Q(z) \cdot \log \frac{P(x,z)}{Q(z)}$ $KL = -\int Q(z) \cdot \log \frac{P(z|x)}{Q(z)}$

LOWER BOUND

MEAN FIELD: $\ln q_j(z_j) = E_{i\neq j}[\ln p(x,z)] + \text{const}$ ←

LOG OF OPTIMAL FOR FACTOR Q IS EXP OF JOINT OVER ALL OTHERS

EXPECTATION PROPAGATION: USES REVERSE KL. $KL(p\|q) = -\int p(z)\left[\sum_1^M \log q_i(z_i)\right] dz + \text{const}$

LOWER BOUND: USEFUL TO MONITOR THE PROGRESS

VARIATIONAL MESSAGE PASSING: IN DAG. $p(x) = \prod_i p(x_i|pa_i)$, $Q(x) = \prod_i q_i(x_i)$, $\ln Q_j^*(x_j) = E\left[\sum_i \ln p(x_i|pa_i)\right] + \text{const}$

MARKOV BLANKET

LOCAL VARIATIONAL METHODS: PB ON LOGISTIC SIGMOID

EXPECTATION PROPAGATION: // USES MOMENT MATCHING

VI - OTHER FORMULATIONS

EVIDENCE LOWER BOUND

- JENSEN'S INEQUALITY: f CONCAVE → $f(E[x]) \geq E[f(x)]$
- JENSEN'S ON OBSERVATION LOG PROBS: $\log p(x) = \log \int_z p(x,z) = \log \int_z p(x,z)\frac{Q(z)}{Q(z)} = \log E_Q\left[\frac{p(x,z)}{Q(z)}\right] \geq \underbrace{E_Q[\log p(x,z)]}_{\text{EXPECTED LOG JOINT}} - \underbrace{E_Q[\log Q(z)]}_{H(Q)\text{ ENTROPY}}$

EVIDENCE LOWER BOUND → MAXIMIZE IT, ELBO IS ALSO FREE ENERGY

- RELATION TO KL

$$KL(Q(z)\|P(z|x)) = E_Q\left[\log \frac{Q(z)}{P(z|x)}\right] = E_Q[\log Q(z)] - E_Q[\log P(z|x)] = E_Q[\log Q(z)] - E_Q[\log P(z,x)] + \log P(x) = -\text{ELBO} + \log P(x)$$

$P(z|x) = \frac{P(z,x)}{P(x)}$

→ $\log p(x)$ DOES NOT DEPEND ON Q → MAXIMIZE ELBO = MINIMIZE KL

- $\ln(P(Y)) = KL(Q(z)\|P(z|x)) + F(z,x)$

$F(z,x)$: ELBO/FREE ENERGY $= \underbrace{E_Q[\log P(x,z)]}_{\text{X LOG JOINT}} + \underbrace{H(Q)}_{\text{ENTROPY}}$, ENTROPY $= -E_Q[\log Q(z)]$

IN MEAN FIELD

$Q(z) = \prod Q(z_i)$ → $F(z,x) = -KL(Q_i\|\text{Exp}(\langle\log P(z,x)\rangle_{Q_{-i}})) + C$ → GIBBS LIKE SCHEME 1 PARAM AT A TIME → VARIATIONAL EM

- EXPLOIT $\log P(x|z) - F(Q,x) = KL(Q\|P)$

- E MAX F WRT Q, θ FIXED
- M MAX F WRT θ, Q FIXED

VARIATIONAL BAYES → EVIDENCE → MODEL SELECTION

- PUT A LOWER BOUND ON MARGINAL LIKELIHOOD . $\log P(y|M) = \log \int P(x,y,\theta|M)\,dx\,d\theta \overset{\text{JENSENS}}{\geq} \int Q(x,\theta) \log \frac{P(x,y,\theta|M)}{Q(x,\theta)}\,dx\,d\theta = F_M$
- FACTORIZE Q , $\log P(y|M) - F_M(Q_x(x), Q_\theta(\theta), y) = KL(Q\|P)$
- MODEL SELECTION METRICS: BIC, AIC, ANNEALED IMPORTANCE SAMPLING, CONTINUATION METHOD FOR BRIDGE OF $Q(x)$,

FOR CONJUGATE - EXPONENTIAL MODELS

- JOINT IS IN EXPONENTIAL FAMILY $P(x,y|\theta) = f(x,y)g(\theta)\,\text{Exp}\left[\underbrace{\phi(\theta)}_{\text{NATURAL PARAMS}}{}^T \overbrace{u(x,y)}^{\text{SUFF STATS}}\right]$
- CONJUGATE PRIOR $P(\theta|\eta,\nu) = h(\eta,\nu)g(\theta)^\eta\,\text{Exp}\left[\phi(\theta)^T\nu\right]$

η = NO OF OBSERVATIONS , ν = VALUE OF OBSERVATIONS

- IS E-M VANILLA IF $Q_\theta(\theta) = \delta(\theta - \theta^*)$

GENERAL VBEM

E $Q_x^{(t+1)}(x) = P(x|y,\bar\phi^t)$

M $Q_\theta^{t+1}(\theta) = \text{Exp}\left[\int Q_x^{(t+1)}(x) \cdot \log P(x,y,\theta)\,dx\right]$

VI – RECONSTRUCTION FORMULATION

ELBO / $\mathcal{F}(Y,Q)$

- $\log P(Y) = \log \int P(Y|Z) P(Z)\, dZ \xrightarrow{\text{JENSEN}} \log P(Y) \geq \int Q(Z) \log\left(P(Y|Z)\frac{P(Z)}{Q(Z)}\right) dZ = \int Q(Z) \log P(Y|Z) - \int Q(Z) \log \frac{Q(Z)}{P(Z)} = \underbrace{E_{Q(Z)}[\log P(Y|Z)]}_{\text{RECONSTRUCTION COST}} - \underbrace{KL(Q(Z)||P(Z))}_{\text{PENALTY}}$

- ENCODER VIEW
  - ENCODER $Q(Z|Y)$ → VARIATIONAL DISTRIBUTION
  - DECODER $P(Y|Z)$ → LIKELIHOOD
  - $E[Q(Z)][\log P(Y|Z)]$ DATA CODE LENGTH UNDER Q
  - $KL(Q(Z)||P(Z))$ = HYPOTHESIS CODE, PENALTY = $\Omega(Z,Y)$

- VARIATIONAL INFERENCE LATENTS ARE DISTRIBUTIONS, ML OTHER PARAMS $Q(Z)$
- VARIATIONAL BAYES: ALL IS A DISTRIBUTION $Q(Z,\theta|Y)$

- FREE-FORM: SOLVES FOR THE EXACT DISTRIBUTION $\longrightarrow \frac{\delta F(Y,Q)}{\delta Q(Z)} = 0$ ST $\int Q(Z) dZ = 1 \longrightarrow Q(Z) \propto P(Z) \exp(\log P(Y|Z,\theta))$
- FIXED FORM EXPLICIT FORM OF DISTRIBUTION $Q_\phi(Z) = f(Z,\phi)$

VARIATIONAL EM

- FOR $I = 1 .. N$ REPEAT UNTIL CONVERGED
- E $\underset{\text{VARIATIONAL PARAMS}}{\phi} \propto \nabla_\phi F(Y,Q) = \nabla_\phi E_{Q_\phi(Z)}[\log P_\theta(Y_i|Z)] - \nabla_\phi KL(Q(Z_i)||P(Z_i))$
- M $\underset{\text{MODEL PARAMS}}{\theta} \propto \nabla_\theta F(Y,Q) = \frac{1}{N} E_{Q_\phi(Z)}[\nabla_\theta \log P_\theta(Y_N|Z_N)]$

STOCHASTIC VARIATIONAL INFERENCE

LIKE EM, BUT ON DATA MINIBATCH

DOUBLY STOCHASTIC VI

LIKE EM BUT KL AND EXPECTED LL HARD TO COMPUTE/GRADIENT → USE MONTECARLO

- M $Z_N \sim Q(Z_N|Y_N)$, $\phi_N \propto \nabla_\phi \frac{1}{S} \sum_S \left[\log P_\theta(Y_N|Z_N(\phi)^{(S)}) - \log \frac{Q(Z_N(\phi)^{(S)})}{P(Z)}\right]$

AMORTIZED VI

INFERENCE NETWORK → Q IS ENCODER / INVERSE MODEL, PARAMS OF Q ARE GLOBAL ON TEST AND TRAIN POINTS. COST OF INFERENCE AMORTISED ON ALL DATA

JOINT OPTIMIZATION OF MODEL AND VARIATIONAL PARAMS, NO ALTERNATION.

$$Z \longrightarrow \underset{\text{MODEL}}{P(Y|Z)} \longrightarrow Y \sim P(Y|Z) \longrightarrow \underset{\text{INFERENCE NETWORK}}{Q(Z|Y)} \longrightarrow Z \sim Q(Z|Y)$$

HOW TO GRADIENT

a • LOCAL VARIATIONAL METHODS/APPROXIMATIONS

b • STOCHASTIC BACKPROPAGATION [REPARAMETRIZATION TRICK]

$$\nabla_\xi E_{Q(Z)}[f(Z)] \longrightarrow Z \sim N(\mu,\sigma^2),\ Z = \mu + \sigma\epsilon,\ \epsilon \sim N(0,1) \longrightarrow \nabla_\xi E_{N(0,1)}[f(\mu+\sigma\epsilon)] \to E_{N(0,1)}[\nabla_{\xi=\{\mu,\sigma\}} f(\mu+\sigma\epsilon)]$$

- NO NEED FOR LOWER BOUNDS, LOW VARIANCE, CAN USE MANY DISTRIBUTIONS

c • MONTE CARLO VARIATE ESTIMATORS

- SCORE FUN: $\nabla_\xi \log q_\xi(Z|X) = \frac{\nabla_\xi q_\xi(Z|X)}{q_\xi(Z|X)}$
- PROBLEM $\nabla_\phi E_{Q_\phi(Z)}[\log P_\theta(Y|Z)]$
- MCCV $E_{Q_\phi(Z)}[(\log P_\theta(Y|Z) - C)\nabla_\phi \log Q(Z|Y)]$
  ↳ CONTROL VARIATE FOR ESTIMATOR VARIANCE CONTROL

STOCHASTIC VARIATIONAL INFERENCE <PAPER>

- NATURAL GRADIENT VANILLA GRADIENT $\theta^{t+1} = \theta^{t} + \rho \nabla_\lambda f(\theta)$ • IN DISTRIBUTION SPACE EUCLIDEAN DISTANCE OF PARAMS IS NOT REPRESENTATIVE OF DISTRIBUTION DISSIMILARITY
- SYMMETRIZED KL $D_{KL}(\theta, \theta') = E_\theta\left[\log \frac{q(\phi|\theta)}{q(\phi|\theta')}\right] + E_{\theta'}\left[\log \frac{q(y|\theta')}{q(y|\theta)}\right]$
- RIEMANNIAN METRIC TO MAP THE TWO SPACES → LINEAR TRANSFORMATION OF θ MAPPING DISTANCE θ = DISTANCE D_{KL}. $d\theta^T G(\theta) d\theta = D_{KL}^{sym}(\theta, \theta + d\theta)$

  → $\overset{NAT}{\nabla_\theta} f(\theta) = G(\theta)^{-1} \nabla f(\theta)$, $G(\theta)$ = FISHER INFO OF $q(\theta)$ $= E_\theta[(\nabla_\theta \log q(y|\theta))(\nabla_\theta \log q(y|\theta))^T]$
- $D_{KL}(\theta, \theta + d\theta) = d\theta^T G(\lambda) d\theta$ q IN EXP FAMILY: $G(\theta) = \nabla_\theta^2 a_g(\theta)$ HESSIAN OF LOG-NORMALIZER FCN $[a]$ IS COV MATRIX OF SUFFICIENT STATISTICS $[\theta]$
- $\overset{NAT}{\nabla_\theta} L = E_\phi[\eta_g(x, z, a)] - \theta$ • CAN DO PARALLEL! • IS PROJECTED GRADIENT

  $\overset{NAT}{\nabla_\phi} L = E_{\theta\phi}[\eta_l(x_n, z_{n,-j}, \theta)] - \phi_{ij}$ • η NATURAL PARAMETERS ; COMPUTE CONDITIONAL

ALGORITHM:

- SAMPLE DATAPOINT X_i
- $\phi = E_{\theta^{t-1}}[\eta_g(x_i^*, z_i^*)]$ GLOBAL VAR PARAMS
- $\tilde{\theta} = E_\phi[\eta_g(x_i, z_i)]$ INTERMEDIATE LOCAL PARAMS
- $\theta^t = (1-\rho_t)\lambda^{t-1} + \rho_t \tilde{\theta}$ FINAL LOCAL PARAMS

VARIANTS

- MINIBATCHES $\theta^t = (1-\rho_t)\theta^{t-1} + \frac{\rho_t}{N} \sum \hat{\theta}_t$ – FOR EMPIRICAL BAYES