DEEP BOLTZMANN MACHINES

COMPLETELY UNDIRECTED. MULTIPLE HIDDEN LAYERS. IN EACH LAYER VARS ARE MUTUALLY INDEPENDENT, CONDITIONED IN NEIGHBOURING LAYERS. BINARY BUT VISIBLES MAY BE REALS

- $P(v, h^1, h^2, h^3) = \frac{1}{Z(\theta)} \exp\left(-E(v, h^1, h^2, h^3; \theta)\right)$ • $E(v, h^1, h^2, h^3, \theta) = -v^T W^1 h^1 - h^{1T} W^2 h^2 - h^{2T} W^3 h^3$ • CONNECTIONS BETWEEN HIDDEN UNITS

- DBM LAYERS CAN BE ORGANIZED IN BIPARTITE GRAPHS, ODD/EVEN LAYERS. CONDITIONING ON ONE PARTITION MAKES VARS OF THE OTHER BECOME CONDITIONALLY INDEPENDENT → $P(h^1 | v, h^2) = \prod_j P(h^1_j | v, h^2)$

→ **CONDITIONALS** ON NEIGHBOURING LAYERS ARE FULLY FACTORIAL AND SIGMOID → $P(h^1_j = 1 | v, h^2) = \mathrm{sigm}(v^T W^1_{:,j} + W^2_{j,:} h^2)$

PROPERTIES:
- SIMPLER POSTERIOR THAN DBN → CI OF LAYERS WRT OTHERS ALLOWS FIXED POINT OPTIMIZATION OF VARIATIONAL LOWER BOUND, PROPER MEAN FIELD
- INEFFICIENT SAMPLING BECAUSE MCMC AT EACH LAYER **WHAT DO?** GENERATIVE OR CLASSIFIER (+ MLP)

MEAN FIELD INFERENCE

POSTERIOR OVER ALL h IS COMPLICATED $P(h^1, h^2 | v)$ BECAUSE INTERACTION WEIGHTS W^2 → MAKE h_1 h_2 MUTUALLY DEPENDENT GIVEN V

→ MEAN FIELD OK BECAUSE CONDITIONALS ARE FULLY FACTORIZED $Q(h^1, h^2 | v) = \prod^{N} Q(h^1_j | v) \prod^{N} Q(h^2_k | v)$ FULLY FACTORED APPROXIMATION

→ **MINIMIZE** $KL(Q \| P) = \sum_h Q(h^1, h^2 | v) \log \frac{Q(h^1, h^2 | v)}{P(h^1, h^2 | v)}$ • NOT NECESSARY TO SPECIFY FORM OF APPROXIMATION BUT HERE WE CAN AS A PRODUCT OF BERNOULLIS

→ MINIMIZE THE VARIATIONAL / EVIDENCE LOWER BOUND (OR EQUIVALENTLY $KL(Q \| P)$ EXPLICITLY) $L(Q) = -\sum_{h^1 h^2} Q(h^1 h^2 | v) E(v, h^1, h^2, \theta) - \log Z(\theta) + H(Q)$

→ SUBSTITUTE FORM OF Q IN L → DERIVE WRT EACH h^n AND SET TO 0:
$$\begin{cases} \hat{h}^1_j = \mathrm{sigm}\left(\sum_i v_i W^1_{ij} + \sum_k W^2_{jk} \hat{h}^2_k\right) \ \forall j \\ \hat{h}^2_k = \mathrm{sigm}\left(\sum_j W^2_{jk} \hat{h}^1_j\right) \ \forall k \end{cases}$$

→ ITERATE UPDATES UNTIL CONVERGENCE ≈ 10 ITERATIONS

PARAMETER LEARNING

- EVALUATING PMF REQUIRES APPROXIMATE METHODS (AIS) • TRAINING REQUIRES LOG PARTITION GRADIENT APPROXIMATIONS. • POSTERIOR ALSO INTRACTABLE → APPROX DURING LEARNING TOO
- MF / CD WON'T WORK / OFFER SPEEDUPS → **USUALLY STOCHASTIC ML IS USED** NEGATIVE PHASE SAMPLES WITH GIBBS ALTERNATING EVEN AND ODD LAYERS

→ **VARIATIONAL E-M** CAN BE SEEN AS
- **E:** OPTIMIZE $L(Q, \theta)$ WRT VARIATIONAL PARAMS - EXPLAINED ABOVE $h^1 h^2$ • NO PROVEN GUARANTEE LIKE FOR STANDARD EM, BUT FOUND TO WORK TO MAKE TRUE LIKELIHOOD BETTER
- **M:** OPTIMIZE $L(Q, \theta)$ WRT θ → WE HAVE INTRACTABLE Z. WE ONLY TAKE SMALL STEP ALONG DIRECTION OF GRADIENT, NO $=0$ OFF THE BAT.

PRACTICALITIES: TRAINING DBM STARTING FROM RANDOM INIT IS USUALLY FAILURE. - FAIL TO REPRESENT DISTRIBUTION - NO BETTER LIKELIHOOD THAN SHALLOW RBM **NO IDEA WHY**

- INITIALIZATION FROM PRETRAINED CONFIGURATION DOES WORK INSTEAD. MAYBE BECAUSE LEARNING RATE FOR GRADIENT DESCENT DOESN'T PLAY WELL WITH NO OF GIBBS STEPS IN NEGATIVE PHASE. NOT GOOD MIXING. OR MAYBE BECAUSE ILL-CONDITIONED HESSIAN.

LAYERWISE PRETRAINING EACH LAYER TRAINED AS RBM IN ISOLATION. ? RBM TRAINED ON SAMPLES DRAWN FROM $L-1$ POSTERIOR | COMBINED INTO A DBM. PERSISTENT CD TRAINING | SML TOO

BUT WATCH OUT WITH WEIGHTS $\times 2 / \frac{1}{2}$ WHEN RBM → DBM. ALSO REMOVE y IN MLP

JOINT TRAINING TO MAKE PERFORMANCE TRACKING EASIER. (HYPERPARAMS LATE)

- **CENTERED DBM:** REPARAMETRIZATION OF THE MODEL TO MAKE HESSIAN BETTER CONDITIONED → NO MORE PRETRAINING. GREAT FOR GENERATIVE USE CASE. NOT SO MUCH IN CLASSIFICATION
  - $E(s; W, b) = -\frac{1}{2} s^T W s - s^T b$ ⟶ • $E(s'; W, b) = -\frac{1}{2}(s-\mu)^T W (s-\mu) - (s-\mu)^T b$ W WEIGHTS; b BIASES; s UNITS. $s - \mu \approx$ AT BEGINNING AFFECTS SML DYNAMICS

- **MULTI-PREDICTION DBM (MP-DBM)** VIEWS MEAN FIELD EQUATIONS AS DEFINING FAMILY OF RECURRENT NETS FOR APPROX SOLVING EVERY INFERENCE PROBLEM
- TRAIN TO MAKE RNN OBTAIN ACCURATE ANSWERS TO INFERENCE PROBLEM. SAMPLE TRAIN EXAMPLE, SAMPLE INPUT SUBSET FOR INFERENCE NETWORK, TRAIN INFERENCE NETWORK TO PREDICT VALUES OF REMAINING UNITS. BACKPROP THROUGH INFERENCE GRAPH. LOSS IS NOT LOWER BOUND BUT APPROX CONDITIONAL THAT INFERENCE NET IMPOSES ON MISSING VALUES. TRAINS MODELS AS THEY'RE USED. BETTER CLASSIFICATION PERFORMANCE. BETTER LIKELIHOOD MEAN FIELD. LOSS EXACT GRADIENTS, WRT SML IS APPROX GRADIENT. DROPOUT BY MASKING INPUTS. INPUTS NOT GIVEN / ACTIVE BECOME TARGETS

BOLTZMANN MACHINES FOR REAL DATA <MISSING: mcRBM, mPoT>

SPIKE-AND-SLAB RBM

• TWO SETS OF HIDDEN UNITS: SPIKE (BINARY), SLAB (REAL). MEAN OF $P(V|H) = (h \odot s) W^T$. SPIKE = DETERMINE COMPONENT PRESENCE. SLAB = COMPONENT INTENSITY
→ MODEL INPUT COVARIANCE. NO NEED FOR MATRIX INVERSION; CD AND PERSISTENT CD WORK.
ISSUE: SOME PARAM SETTINGS RESULT IN NON POSITIVE DEFINITE COV MATRIX, INTEGRAL DIVERGES. • EXTENDED TO CONVOLUTION, HIGHER ORDER INTERACTION AND AVG-POOLING OF SLAB FEATURES
*1

CONVOLUTIONAL BOLTZMANN MACHINES

IN STANDARD CONVNETS POOLING REDUCES SPATIAL INPUT SIZE AT EACH LAYER, WHAT DO IN ENERGY MODELS? PROBABILISTIC MAX POOLING: CONSTRAIN DETECTOR UNITS SO ONLY 1 AS AT MOST ACTIVE AT A TIME. POOLING ON IFF DETECTOR IS ON, NO ACTIVES → ENERGY 0. DRAWBACK! DETECTORS MUTUALLY EXCLUSIVE, NON OVERLAPPING POOLING
→ cRBM CAN DO IMAGE INPUTATION BUT NO BETTER THAN SHALLOW MODELS FOR PRETRAINING

<MISSING: CONDITIONAL BM, RNN-RBM, DISCRIMINATIVE BM, ...>

DIRECTED GENERATIVE NETS

UNTIL RECENTLY IN DL ONLY UNDIRECTED GENERATIVE MODELS WERE POPULAR (RBM AND FRIENDS)
MISSING: DIFFERENTIABLE GENERATOR NETS, VARIATIONAL AUTOENCODER, CONVOLUTIONAL GENERATIVE NETWORK>

GENERATIVE ADVERSARIAL NETWORKS

DIFFERENTIABLE MAPPINGS FROM INPUT NOISE TO SAMPLES RESEMBLING DATA. SIMILAR TO VARIATIONAL AUTOENCODERS BUT DIFFERENT TRAINING PROCEDURE, NO INFERENCE NETW
→ BASED ON GAME THEORY. GENERATOR NETWORK: TRAINED TO MAP INPUT NOISE z TO SAMPLES X. $g(z)$ GENERATIVE MODEL. $p(z)$ NOT LEARNED. g DEFINES CONDITIONAL $p(x|z) = N(x|g(z), 1/\beta I)$. DISCRIMINATOR NETWORK: ESTIMATE PROBABILITY OF x TO HAVE BEEN SAMPLED FROM THE DATA VS THE MODEL
→ TRAINING: D MAXIMIZES, G MINIMIZES ~~PROBABILITY~~ VALUE FCN OF D BEING CORRECT. $g^* = \arg\min_G \max_D V(g,d)$. $V(g,d) = \mathbb{E}_{x\sim \text{DATA}} \log d(x) + \mathbb{E}_{x\sim \text{MODEL}} \log(1-d(x))$
SIMPLY TRAIN W/ BACKPROP, NO APPROXIMATE INFERENCE REQUIRED

* GAUSSIAN-BERNOULLI RBM

BINARY HIDDEN UNITS, REAL VISIBLE UNITS. CONDITIONAL OVER VISIBLES IS GAUSSIAN WITH MEAN IS FUNCTION OF HIDDENS. DIFFERENT PARAMETRIZATIONS: COVARIANCE/PRECISION FORM.
→ $P(v|h) \propto N(v|Wh, \beta^{-1})$, $\log N(v|Wh, \beta^{-1}) = -\frac{1}{2}(v-Wh)^T \beta (v-Wh) + f(\beta)$ — $P(h|v)$ WE HAVE SOME CHOICE, NOT FULL COVARIANCE MATRIX BECAUSE ELSE WE GET LINEAR FACTOR MODEL, B/C CONNECTION BETWEEN EDGES
→ USUALLY DIAGONAL β ~~with~~ VALUED. BIASES OR NOT

OTHER BOLTZMANN MACHINES

MANY OTHER VARIANTS. • CONDITIONAL RBM WITH COND. TERMS IN ENERGY FCN. • RNN-RBM → RNN EMITS RBM PARAMS AT EACH TIMESTEP
• OTHER TRAINING CRITERIONS IE TRAIN ON $\log p(y|v)$ • TRAIN WITH HIGHER (>2) ORDER INTERACTIONS
*

SIGMOID BELIEF NETS

DGM WITH SPECIFIC TYPE OF CONDITIONAL. VECTOR OF BINARY S. $P(S) = \text{sigm}\left(\sum_{j<i} W_{ji} s_j + b_i\right)$ S INFLUENCED BY ANCESTORS. MANY LAYERS. ANCESTRAL SAMPLING SIMILAR TO DBN BUT SAMPLES ARE INITIALLY INDEPENDENT. SAMPLING THE VISIBLES IS EASY. OTHER OPERATIONS NOT SO MUCH. INFERENCE INTRACTABLE
RECENT BREAKTHROUGHS: IMPORTANCE SAMPLING, WAKE-SLEEP, HELMHOLTZ MACHINES MADE IT FEASIBLE, GOOD PERFORMANCE. INFERENCE WITH SPECIAL LOWER BOUND OR INFERENCE NETWORK. SBN WITHOUT LATENTS IS A TYPE OF AUTOREGRESSIVE NET

DIFFERENTIABLE GENERATOR NETS

IS FAMILY OF VARIATIONAL AUTOENCODER. WE HAVE A GENERATOR NETWORK. TRANSFORM SAMPLES OF LATENTS z TO SAMPLES OF X OR DISTRIBUTION OF SAMPLES OVER X VIA DIFFERENTIABLE FUNCTION $g(z, \theta)$, THE GENERATOR NETWORK. • NET ARCHITECTURES GIVES FAMILY OF DISTRIBUTIONS TO SAMPLE FROM, PARAMS SELECT THE SPECIFIC DISTRIBUTIONS
THINK BOX-MULLER SCHEME! IS OF VERY SIMPLE GEN NETWORK. SAMPLE GEN DISTRIBUTION → DISCRETE AND CONTINUOUS DATA

DIRECT SAMPLES → ONLY ~~DISCR~~ CONTINUOUS PRO, WE DON'T HAVE CONSTRAINTS TO CARE ABOUT

• GENERATIVE MODELING IS (OBV) MORE DIFFICULT; BUT NOWADAYS COMPUTATIONAL POWER IS SUFFICIENT.
→ DIFFICULTY IS IN TRAINING WHEN VALUES OF z FOR EACH X ARE NOT FIXED AND KNOWN BEFOREHAND

*2 [NEXT PAGE]

AUTOREGRESSIVE NETWORKS

SIMILAR TO RECURRENT NETS BUT NO MORE PARAMETER SHARING ACROSS TIME, ELEMENTS NOT A TRANSLATION EQUIVARIANT SEQUENCE, BUT ARBITRARY TUPLE.

LOGISTIC A/R NETS: NO HIDDEN UNITS, NO SHARING. $P(x_i | x_{i-1} .. x_1)$ PARAMETERIZED AS LOGISTIC REGRESSION, FIXED CAPACITY MODEL. CONTINUOUS VARIABLES → LINEAR A/R MODEL.

NEURAL A/R NETS: MADE TO AVOID COD ISSUES IN NONPARAMETRIC GRAPHICAL MODELS: ESTIMATION OF CONDITIONAL PROBABILITIES W/O EXPONENTIAL NO. OF PARAMS. LEFT-TO-RIGHT CONNECTIVITY. HIDDEN LAYER FEATURES FOR x_T ARE REUSED FOR x_{T+K}. ALL UNITS FOR T DEPEND ONLY FROM $x_1 ... x_T$. MULTITASK / TRANSFER LEARNING LIKE

- NN OUTPUTS PREDICT PARAMETERS OF CONDITIONAL OF x_T

NADE: LIKE NEURAL A/R NET BUT HAS WEIGHT SHARING. ALL $x_t \to h_i$, $i = 1..T$ HAVE SAME WEIGHT MATRIX. RATIONALE IS SIMILARITY TO MEAN-FIELD INFERENCE IN RBM → **RNADE:** FOR CONTINUOUS PROB DISTRIBUTIONS: MODELED AS GAUSSIAN MIXTURE. • **GETTING RID OF CHOOSING ARBITRARY ORDER:** RANDOMLY SAMPLE ANY ORDER + TELL TO HIDDENS WHAT IS BEING OBSERVED, AND WHAT IS TO BE PREDICTED / MISSING → ANY INFERENCE EFFICIENTLY.

ENSEMBLE ON ORDER: $P_{ENS}(x) = \frac{1}{M} \sum_{1}^{M} P(x|O)$. **DOESN'T SCALE WELL FOR DEEP MODELS**

AUTOENCODERS AS GENERATIVE MODELS

DENOISING AE CAN BE SAMPLED FROM WITH MCMC FOR GAUSSIAN RBM. CONTRACTIVE AE ESTIMATE TANGENT MANIFOLD → TO SAMPLE DO ENCODE / DECODE AND INJECT NOISE.

DAE MARKOV CHAIN: FROM PREVIOUS $x \to C(\tilde{x}|x)$ • $h = f(\tilde{x})$ • DECODE $w = g(h)$ OF $P(x|w = g(h)) = P(x|\tilde{x})$ • SAMPLE NEXT STATE x FROM $P(x|w = g(h)) = P(x|\tilde{x})$

- IF AE IS CONSISTENT ESTIMATOR WRT TRUE CONDITIONAL → STATIONARY DISTRIBUTION OF SUCH MARKOV CHAIN IS A CONSISTENT ESTIMATOR FOR x

CONDITIONAL SAMPLING: CLAMP OBSERVED UNITS x_f, ONLY RESAMPLE FREE UNITS $x_0 | x_f$ AND SAMPLE LATENT VARS

WALK BACK TRAINING: SPEEDS UP CONVERGENCE FOR GENERATIVE DAE. **IDEA:** MULTIPLE ENCODE-DECODE STEPS FROM CHAIN INITIALIZED AT TRAINING EXAMPLE, WITH PENALTIES ON THE LAST RECONSTRUCTIONS. LIKE IN CONTRASTIVE DIVERGENCE, BETTER AT REMOVING SPURIOUS MODES

GENERATIVE STOCHASTIC NETWORKS

ARE GENERALIZATIONS OF DAE, INCLUDE HIDDEN VARIABLES IN GENERATIVE MC. TWO CONDITIONALS: • $P(X_k|H_k)$ RECONSTRUCTION DISTRIBUTION

- $P(H_k | H_{k-1}, X_{k-1})$ LATENT STATE UPDATE. **GSN AND DAE** MODEL THE GENERATIVE PROCESS ITSELF, NOT THE JOINT DISTRIBUTION OF H AND V. IF IT EXISTS IT'S IMPLICIT AND ITS THE STATIONARY MC DISTRIBUTION
- TRAIN WITH RECONSTRUCTION LOG PROBS ON VISIBLES, WALKBACK, ETC...

DISCRIMINANT GSN

LET'S USE GSN TO OPTIMIZE $P(y|x)$. BACKPROP LOG-PROBS OVER OUTPUT VARS, KEEPING INPUTS FIXED: STRUCTURED OUTPUT → MODELS CHAIN OVER OUTPUT, INPUTS ARE CONDITIONING.

*2 [PREVIOUS PAGE]

GENERATIVE ADVERSARIAL NETWORKS

IS DIFFERENTIABLE GENERATOR NETWORK. GAME-THEORETIC SCENARIO, GEN NETWORK COMPETES AGAINST ADVERSARY → **DISCRIMINATOR NETWORK** HAS TO DISTINGUISH TRAINING SAMPLES FROM GENERATED ONE

USUALLY ZERO-SUM GAME WITH PAYOFF $V(\theta^g, \theta^d) = E_{x \sim p_{data}} \log d(x) + E_{x \sim p_{model}} \log(1 - d(x))$. AT CONVERGENCE SAMPLES ARE INDISTINGUISHABLE

- DISCRIMINATOR GETS $V(\theta^g, \theta^d)$, GENERATOR GETS $-V(\theta^g, \theta^d)$ • NO GUARANTEES FOR CONVERGENCE, SADDLE POINTS, PLAYERS EQUILIBRIA MAY NOT BE MINIMA FOR V
- CAN TRAIN CONDITIONAL GAN TO SAMPLE FROM $P(x|y)$ → LAPGAN • **UNUSUAL PROPERTY:** CAN FIT PROBS ASSIGNING ZERO MASS TO CERTAIN POINTS, BECAUSE IT DOESN'T MAXIMIZE LOG PROBS OF POINTS BUT FITS A MANIFOLD

GENERATIVE MOMENT-MATCHING NETWORKS

ONLY GENERATOR NETWORK **NO INFERENCE / DISCRIMINATOR**. MOMENT-MATCHING OF TRAINING SAMPLES STATISTICS AND GENERATED SAMPLES STATISTICS

- COST FCN: MAXIMUM MEAN DISCREPANCY. KERNEL TRICK, SO KERNEL FCN, TO IMPLICITLY MATCH 1ST MOMENT IN ∞-DIM SPACE. **WORKS ON BATCH OBVIOUSLY**

CONVOLUTIONAL GENERATIVE NETWORKS

CONVOLUTIONAL STRUCTURE **BUT ADD INFORMATION THROUGH LAYERS** INSTEAD OF REMOVING IT. POOLING IS NOT INVERTIBLE → SO INCREASE FEATURE MAPS SPATIAL SIZE

HOPFIELD NETWORKS

- FULLY CONNECTED ISING WITH SYMMETRIC WEIGHTS • PERFORMS AS **ASSOCIATIVE MEMORY**, CONTENT ADDRESSABLE • DENOISING / RECONSTRUCTION TASKS
- S_i: $\mathrm{sign}\left(\sum_j W_{ij} S_j^{t-1}\right)$ THRESHOLD AT 0, +1/-1 ENCODING. $W_{ij} = \frac{1}{N} \sum_{1}^{N} S_i(n) S_j(n)$ ← **TRAINING / UPDATE**, SYNCH / ASYNCH.
  UNTIL CONVERGENCE
- **INFERENCE**: PRESENT INPUT UNTIL CONVERGENCE, READ S_i AS OUTPUT
- **ENERGY FUNCTION**: $H = -\frac{1}{2} \sum_i \sum_j W_{ij} S_i S_j = -\frac{1}{2} S W S^T$. DESCRIBES 'AMOUNT OF MISMATCH' • $H(S) - H(S') = -\frac{1}{2}(S - S') \sum^{d} W_{ij} S_j$
- **CAPACITY**: $N \approx 0{,}138\, d$ PATTERNS
- **CONTINUOUS HOPFIELD NET**: FOR IE GRAYSCALE IMAGES: +1/-1 OUTPUTS VIA SIGMOID. $P(x|W) = \frac{1}{Z(W)} \exp\left[\frac{1}{2} x^T W x\right]$ IS BOLTZMANN MACHINE

VARIATIONAL AUTOENCODER

MADE TO WORK ON: • INTRACTABLE MARGINAL LIKELIHOOD, POSTERIOR, INTRACTABLE MF VARIATIONAL BAYES • LARGE DATASET

ASSUMPTION: $p_\theta(z)$, $p_\theta(z|x)$ ARE PARAMETRIC AND DIFFERENTIABLE ALMOST EVERYWHERE WRT θ, z.

WHAT IT SOLVES: 1 - ML, MAP FOR θ 2 - POSTERIOR INFERENCE OF $z|x,\theta$ 3 - EFFICIENT MARGINAL INFERENCE OF X

$Q_\phi(z|X)$: VARIATIONAL APPROXIMATION, ENCODER z: CODE

$P_\theta(X|Z)$: DECODER

LOWER BOUND: $\log p_\theta(x) \geq \underbrace{E_{Q\phi(z|x)}\left[-\log q_\phi(z|x) + \log p_\theta(x,z)\right]}_{A} = \underbrace{-KL\left(Q_\phi(z|x) \| P_\theta(z)\right) + E_{Q\phi(z|x)}\left[\log p_\theta(x|z)\right]}_{B}$ $= L(\theta,\phi,x) =$

• DIFFERENTIATE AND OPTIMIZE WRT θ, ϕ → MC ESTIMATOR IS BAD, HIGH VARIANCE

REPARAMETRIZATION!

$\hat{z} \sim Q_\phi(z|X) \longrightarrow G_\phi(\epsilon, x)$ TRANSFORMATION + NOISE $\epsilon \sim p(\epsilon)$ MONTE-CARLO ESTIMATE OF $f(z)$ WRT $Q_\phi(z,x) \rightarrow E_{Q\phi}[f(z)] = E_{p(\epsilon)}\left[f(g_\phi(\epsilon, x'))\right] \simeq \frac{1}{L}\sum^{L} f(g_\phi(\epsilon, x))$

SGVB ESTIMATOR! $\hat{L}^A(\theta,\phi,x) = \frac{1}{L}\sum^{L} \log p_\theta(x,z) - \log_{Q\phi}(z|x)$, $z = g_\phi(\epsilon, x)$ a

$\hat{L}^B(\theta,\phi,x) = -KL(Q_\phi(z|x) \| p_\theta(z)) + \frac{1}{L}\sum_{L}\left(\log p_\theta(x|z)\right)$ b WHEN KL IS COMPUTABLE ITSELF.

KL: REGULARIZER ON ϕ; MAKES IT CLOSE TO PRIOR P

E_Q ... RECONSTRUCTION ERROR

NOISE 'INJECTED' IN SAMPLES FROM APPROXIMATE POSTERIOR

AEVB ALGORITHM:

- θ, ϕ INIT
- PICK MINIBATCH
- SAMPLE NOISE ϵ

E • $g = \nabla_{\theta,\phi} \hat{L}(\theta,\phi; X, \epsilon)$ GRADIENTS WRT MINIBATCH

M • UPDATE ϕ, θ WITH GRAD DESCENT ALGO → SGD, ADAGRAD, ...

REPARAMETRIZATION TRICK!

• IS ALTERNATE SAMPLE GENERATION METHOD FOR $Q(z|x)$ • MAKES MONTECARLO EXPECTATION DIFFERENTIABLE WRT ϕ

TYPICALLY $z \sim N(\mu, \sigma^2) \longrightarrow z = \mu + \sigma\epsilon$, $\epsilon \sim N(0,1)$ OTHERS: INVERSE CDFs, CONJUGATE PRIORS, COMPOSITES

VARIATIONAL AUTOENCODER

• THE ENCODER IS A NEURAL NETWORK

- $P_\theta(z) = N(z; 0, I)$ ISOTROPIC MVN - $P_\theta(x|z)$: MVN OR BERNOULLI, THEIR PARAMS COMPUTED FROM z WITH A MLP NETWORK
- APPROX $\log q_\phi(z|x) = \log N(z; \mu, \sigma^2 I)$ MVN DIAGONAL COVARIANCE; μ, σ ARE OUTPUTS OF MLP, AND VARIATIONAL ϕ
- $L(\theta,\phi,x) = \frac{1}{2}\sum^{J}\left(1 + \log((\sigma_j)^2) - \mu_j^2 - \sigma_j^2\right) + \frac{1}{L}\sum^{L} \log p_\theta(x|z)$

RELATED TO: WAKE-SLEEP, STOCHASTIC VARIATIONAL INFERENCE, PROBABILISTIC MODELS ARE FUNDAMENTALLY AUTOENCODERS

STOCHASTIC BACKPROPAGATION

DEEP LATENT GAUSSIAN MODELS: GENERAL CLASS OF DEEP MODELS, GAUSSIAN LATENTS AT EACH LAYER

GENERATIVE USE: TOP → BOTTOM, PERTURB LAYERS ABOVE WITH GAUSSIAN NOISE ⇒ GENERATE INPUT DATA. PARAMS θ^g, $P(\theta^g) = N(\theta|0, kI)$ GAUSSIAN PRIOR

$$\begin{cases} P(v|h) = P(v|h_1, \theta^g) P(h_L|\theta^g) P(\theta^g) \prod^{L-1} P_l(h_l|h_{l+1}, \theta^g) \\ P(v|\xi(0,1)) = P(v|h_1(\xi_{1\ldots L}), \theta^g) P(\theta^g) \prod^{L} N(\xi|0, I) \end{cases}$$

- $\xi(0,1)$ IS IID MVN

GAUSSIAN BACKPROP:

$$\begin{cases} \nabla_\mu E_{N(\mu,C)}[f(\xi)] = E_{N(\mu,C)}[\nabla_\xi f(\xi)] \\ \nabla_C E_{N(\mu,C)}[f(\xi)] = \frac{1}{2} E_{N(\mu,C)}[\nabla^2_{\xi\xi} f(\xi)] \end{cases}$$

- BECAUSE EXP FAMILY

$\theta_g = \mu, R$

FULL: $\nabla_\theta E_{N(\mu,C)}[f(\xi)] = E_{N(\mu,C)}\left[g^T \frac{\partial \mu}{\partial \theta} + \frac{1}{2} \mathrm{TR}\left(H \frac{\partial C}{\partial \theta}\right)\right]$ IS 2ND ORDER

WITH REPARAM TRICK: $\nabla_R E_{N(\mu,C)}[f(\xi)] = \nabla_R E_{N\sim(0,1)}[f(\mu + R\epsilon)] = E_{N(0,1)}[\epsilon g^T]$, $g = \nabla f$ AT $\mu + R\epsilon$, $y = \mu + R\epsilon$, $\cancel{R} = RR^T$

INFERENCE

LOWER BOUND: $L(V) = -\log P(V) \leq KL(Q(\xi) \| P(\xi)) - E_Q[\log P(V|\xi, \theta^g) P(\theta^g)]$, $Q(\xi|V)$ GAUSSIAN FACTORIZED ACROSS LAYERS (NOT WITHIN LAYER)

V INPUT DATA

- RECOGNITION INDEPENDENT FROM GENERATIVE MODEL

$$\underbrace{Q(\xi|V, \theta^R) = \prod^{N} \prod^{L} N(\xi_{N,L} | \mu_L(V_N), C_L(V_N))}_{\text{RECOGNITION MODEL}}$$

μ, C FROM NEURAL NETS $= \theta^R$

GRADIENTS

$$\begin{cases} \nabla_{\theta^g} F(V) = -E_Q[\nabla_{\theta^g} \log P(V|h)] + \frac{1}{K}\theta^g, \\ \nabla_{\theta^R} F(V) = \nabla_\mu F(V)^T \frac{\partial \mu}{\partial \theta^R} + \mathrm{TR}\left[\nabla_R F(V) \frac{\partial R}{\partial \theta^R}\right] \end{cases}$$

- GRAD DESCENT WITH RMSPROP