PARTITION FUNCTION

IN UGM $E(x) = -\log \tilde{p}(x)$ IS UNNORMALIZED. $p(x;\theta) = \frac{1}{Z(\theta)} \tilde{p}(x,\theta)$. $Z(\theta)$ NORMALIZATION CONSTANT OR **PARTITION FUNCTION**. INTEGRAL OVER UNNORMALIZED PROBABILITY OF ALL STATES. OFTEN INTRACTABLE. MANY DEEP MODELS HAVE IT TRACTABLE OR DON'T REQUIRE COMPUTATION OF $p(x)$, OTHERS DIRECTLY ADDRESS AN INTRACTABLE $Z(\theta)$

LOGLIKELIHOOD GRADIENT FOR ENERGY MODELS

Z DEPENDS ON PARAMETERS → NLL GRADIENT HAS TERM OF Z GRADIENT $\frac{\partial \log p(x;\theta)}{\partial \theta} = -\frac{\partial E(x)}{\partial \theta} - \frac{\log Z(\theta)}{\partial \theta}$ Z INTRACTABLE → ∇Z INTRACTABLE.

- MONTE CARLO APPROXIMATION → $\frac{\partial}{\partial \theta} \log Z = E_{x \sim p(x)} \frac{\partial}{\partial \theta} \log \tilde{p}(x)$
  BECAUSE MATH
- $-\frac{\partial \log p(x,\theta)}{\partial \theta} = \frac{\partial E(x)}{\partial \theta} - E_{x \sim p(x)} \frac{\partial E(x)}{\partial \theta}$
  - **POSITIVE PHASE** → PUSH ENERGY DOWN ON POSITIVE CONTRIBUTIONS
  - **NEGATIVE PHASE** → PUSHES ENERGY UP EVERYWHERE, PROPORTIONAL TO CURRENT MASS
- ON A MINIMUM THE TWO TERMS CANCEL OUT

STOCHASTIC ML AND CONTRASTIVE DIVERGENCE

- EXPECTATION COMPUTED WITH MARKOV CHAINS BURN-IN EVERY TIME WE NEED GRADIENT, IF SGD → ONCE PER STEP, COMPUTATIONALLY INFEASIBLE
- **BALANCE** BETWEEN PUSHING UP (ON MODEL) WHERE DATA OCCURS AND PUSHING DOWN WHERE MODEL SAMPLES OCCUR. **HERE POSITIVE PHASE ASSUMED TRACTABLE**
  MAX $\log p$ **MIN $\log Z$**
- NEGATIVE PHASE APPROXIMATIONS MAKE IT CHEAPER TO COMPUTE BUT ALSO PUSH DOWN IN WRONG LOCATIONS, ARE POINTS MODEL CURRENTLY BELIEVES IN STRONGLY, ITS INCORRECT BELIEFS ABOUT THE WORLD. HALLUCINATIONS / DREAMS

CONTRASTIVE DIVERGENCE: MAIN COST IS BURNIN → LET'S DRAW SAMPLES FROM DATA DISTRIBUTION TO INITIALIZE MARKOV CHAIN, FREE BECAUSE THEY ARE ALREADY AVAILABLE. INITIALLY NEGATIVE PHASE NOT ACCURATE BECAUSE MODEL AND DATA DISTRIBUTION DIVERGE, THEN BETTER, MORE ACCURATE.
WEAKNESS! FAILS TO SUPPRESS SPURIOUS MODES → HIGH PROB REGIONS FAR FROM TRAINING EXAMPLES BECAUSE MCMC INITIALIZED ON TRAINING POINTS WON'T GO THERE

- CD IS BIASED FOR RBMs, SMALL BIAS. USE CD TO INITIALIZE MORE EXPENSIVE MCMC METHODS
- CD LIKE PENALIZING A MARKOV CHAIN CHANGING OUTPUT RAPIDLY WHEN DATA IS FROM THE INPUT
- USEFUL FOR PRETRAINING OF SHALLOW MODEL
- CD NOT GREAT FOR INITIALIZING DEEP MODELS RIGHT AWAY BECAUSE IT'S DIFFICULT TO SAMPLE HIDDEN UNITS GIVEN VISIBLE SAMPLES, HIDDEN ARE NOT IN THE DATA → WE'LL NEED BURNIN

STOCHASTIC ML / PERSISTENT CD

INITIALIZES THE MARKOV CHAINS AT EACH STEP WITH THEIR STATE FROM PREVIOUS STEPS. SHORT SGD STEPS → $M_t \approx M_{t-1}$, SO PREVIOUS SAMPLES ARE FAIR. SINCE MCMC ISN'T REINITIALIZED, IT WANDERS AND FINDS SPURIOUS / FAR MODES. **ALSO:** SAMPLES ARE STORED, SO WE CAN USE IT TO INIT / TRAIN DEEP MODELS TOO. SHORTEST MIXING TIME SML IS BEST. **WEAKNESS:** IF k TOO SMALL OR ϵ TOO LARGE → IF SGD MOVES TOO FAST WRT MARKOV MIXING RATE. NO FORMAL WAY TO CHECK FOR THIS BUT EMPIRICALLY LOOK FROM NEGATIVE PHASE SAMPLE VARIANCE. **WHEN DRAWING SAMPLE / GENERATIVE USE:** RESET MARKOV CHAIN FROM RANDOM, BECAUSE SAMPLES USED FOR TRAINING MIGHT DISTORT PERFORMANCE. **SML** HAS HIGHER VARIANCE THAN CD BECAUSE DIFFERENT TRAINING POINTS IN POS/NEG PHASES

- **MCMC METHODS** GENERALLY COOL BECAUSE THEY ALLOW DECOMPOSITION OF $\log \tilde{p}$ AND $\log Z$ TERMS → CAN COMBINE WITH OTHER METHODS PROVIDING A LOWER BOUND ON $\log \tilde{p}$ FOR POSITIVE PHASE

PSEUDOLIKELIHOOD

IDEA! LET'S AVOID COMPUTING THE PARTITION FUNCTION ALTOGETHER. RATIO OF UNNORMALIZED PROBABILITIES CANCELS PARTITIONS OUT.

$p(a|b) = \frac{\tilde{p}(a,b)}{\sum_{a,c} \tilde{p}(a,b,c)}$ a VARS WE WANT CONDITIONAL OF, b VARS TO CONDITION ON, c IRRELEVANTS • MOVE c INTO b TO REDUCE COST: $\sum^{N} \log p(x_i | x_{-i})$

- **REDUCTION** FROM k^N TO $k \times N$. OR FOR LARGE DATASETS **GENERALIZED PSEUDOLIKELIHOOD!** M SETS OF VARIABLES APPEARING TOGETHER LEFT OF CONDITIONING BAR
  $\sum_M \log p(x_{S} | x_{-S})$
  - $M = N$; $S^{i} = \{i\}$ → PSEUDOLIKELIHOOD VANILLA
- **POOR** WHERE WE NEED GOOD MODEL OF FULL JOINT → DENSITY ESTIMATION
- **GOOD** WHERE DATA HAS STRUCTURE ALLOWING S TO CAPTURE MOST CORRELATIONS; IE IMAGES
- **CANNOT** BE USED WITH VARIATIONAL INFERENCE OR OTHER LOWER BOUND TECHS BECAUSE HAS $\tilde{p}$ AT DENOMINATOR → LOWER BOUND ON DENOMINATOR IS UPPER BOUND ON EXPRESSION. MAXIMIZING UPPER BOUND MAKES NO SENSE.
- **STILL USEFUL** TO TRAIN SINGLE LAYER MODEL
- **PER STEP COST** IS LARGER THAN SML BECAUSE ALL CONDITIONALS COMPUTED; GENERALIZED PL CAN HAVE SIMILAR COST
- **IMPLICIT PRIOR:** ALL STATES HAVE MORE THAN ONE VARIABLE DIFFERENT FROM TRAINING EXAMPLES

SCORE / RATIO MATCHING

ALSO AVOIDS COMPUTING Z AND DERIVATIVES. MINIMIZES EXPECTED SQUARE DIFFERENCE BETWEEN DERIVATIVES OF MODEL LOG PDF AND DATA LOG PDF WRT INPUT.

- $\theta^* = \min_\theta J(\theta) = \frac{1}{2} E_x \| \nabla_x \log p_{model}(x,\theta) - \nabla_x \log p_{data}(x) \|_2^2$ REQUIRES KNOWLEDGE OF p_{data} $\longrightarrow$ $\hat{J}(\theta) = \frac{1}{M} \sum^M \sum^N \frac{\partial^2}{\partial x_j^2} \log p_{model}(x,\theta) + \frac{1}{2}\left(\frac{\partial}{\partial x_j} \log p_{model}(x,\theta)\right)^2$
  - **GSM** ALLOWS IT
  - $x_1 \dots x_M$ TRAINING DATA, DIM N
- DERIVATIVES WRT x → NO CAN USE ON DISCRETE DATA • NOT COMPATIBLE WITH VARIATIONAL INFERENCE
- OK FOR SHALLOW MODELS, NOT FOR DEEP OR PRETRAINING • LIKE CD WITH NON-GIBBS MARKOV CHAIN MOVING ON GRADIENT

RATIO MATCHING: FOR BINARY DATA $J^{RM}(\theta) = \frac{1}{M} \sum^M \sum^N \left(\frac{1}{1 + \frac{p_{model}(x^i,\theta)}{p_{model}(f(x,j),\theta)}} \right)^2$ j IS BIT AT POSITION j. $f(x_j)$ FLIPS THE j^{TH} BIT.

- RATIO → PARTITION FCN CANCELS OUT
- N TIMES HIGHER COST THAN SML
- **SAME PL IMPLICIT PRIOR** → HAMMING DISTANCE, ETC • USEFUL FOR HIGH-DIM SPARSE DATA IE WORD VECTORS

DENOISING SCORE MATCHING

REGULARIZING SCORE MATCHING WITH FITTING $p_{smooth}(x) = \int p_{data}(x+y)\, q(y|x)\, dy$ INSTEAD OF p_{data}. BECAUSE WE ONLY HAVE SAMPLES FROM IT AND GIVEN CAPACITY ANY ESTIMATOR WILL DEGENERATE TO SET OF DIRAC IMPULSES ON TRAINING POINT. SMOOTH WITH q NORMALLY DISTRIBUTED NOISE

- SOME DENOISING AUTOENCODERS CORRESPOND TO ENERGY MODELS WITH **DSM** BUT AE IS LESS EXPENSIVE TO CRUNCH • POSSIBLE TO DERIVE AE FOR ANY EBM ON REAL DATA

NOISE-CONTRASTIVE ESTIMATION

IDEA: MODEL REPRESENTED AS $\log p_{model}(x) = \log \tilde{p}_{model}(x,\theta) + c$. c IS APPROX OF $-\log(Z(\theta))$. TREATED AS ONE MORE PARAM, OPTIMIZED AT SAME TIME AND W/SAME ALGO AS θ. NOT A DISTRIBUTION BUT GETS BETTER AS c CONVERGES. **NO CAN DO WITH MLE**

- UNSUPERVISED ESTIMATION OF $p(x)$ $\longrightarrow$ SUPERVISED PROBLEM – INTRODUCE NOISE $p_{noise}(x)$ EASY TO EVAL/SAMPLE – INTRODUCE SWITCH VAR y → JOINT MODEL
- SWITCH DETERMINES WHETHER WE SAMPLE FROM $p_{data}(x) / p_{model}(x)$ OR $p_{noise}(x)$ $\longrightarrow$ MLE FOR FITTING $p_{joint-model}$ TO p_{train}.

→ **IS LOGISTIC REGRESSION** APPLIED TO LOG-PROBS DIFFERENCE OF MODEL AND NOISE $p_{joint-model}(y=1|x) = \sigma(\log p_{model}(x) - \log p_{noise}(x))$

- GOOD ON FEW RANDOM VARS → USED FOR MODELING CONDITIONAL WORD DISTRIBUTION GIVEN CONTEXT
- DOES NOT WORK WITH VARIATIONAL BOUNDS / METHODS

PARTITION FUNCTION ESTIMATION

FOR REALZ. WE NEED IT TO COMPUTE NORMALIZED LIKELIHOOD, MODEL EVALUATION, MONITORING PERFORMANCE, COMPARISON, ETC...

IDEA: TO COMPARE MODELS WE USE LIKELIHOOD RATIO → NOT STRICTLY NECESSARY TO HAVE $Z(\theta_M)$, BUT ONLY THEIR RATIO. $\frac{Z(\theta_A)}{Z(\theta_B)}$ $\longrightarrow$ WE NEED RATIO AND ONE OF THE Zs AND WE CAN GET THE REST

- $\sum_t \ln \frac{\tilde{p}_A(x^t,\theta_A)}{\tilde{p}_B(x^t,\theta_B)} - N_{TEST} \ln \frac{Z(\theta_A)}{Z(\theta_B)} > 0 \rightarrow A > B$
- $Z_1 = \int \tilde{p}_1(x)\,dx = Z_0 \int p_0(x) \frac{\tilde{p}_1(x)}{p_0(x)}\,dx$ $\longrightarrow$ $\frac{Z_1}{Z_0} \approx \sum_1^K \frac{\tilde{p}(x)}{p_0(x)},\ x \sim p_0$
  - **IMPORTANCE SAMPLING** — p_0 PROPOSAL
  - **MONTECARLO APPROXIMATION**
  - $KL(p_0 \| p_1)$ SMALL
- SAMPLE FROM p_0, WEIGHT ON RATIO AT SAME VALUE
- **WORKS** IF p_1 AND p_0 ARE CLOSE; BUT p_1 IS OFTEN MESSY (MULTIMODAL, HIGH DIM). IF p_1 AND p_0 NOT CLOSE → SAMPLES WILL MAKE NEGLIGIBLE CONTRIBUTIONS TO SUM

WAT DO? FIND INTERMEDIATE DISTS BETWEEN p_0 AND p_1

• **OBS:** $Z(p_0)$ IS KNOWN!!

ANNEALED IMPORTANCE SAMPLING

- LET'S INTRODUCE INTERMEDIATE DISTRIBUTIONS IN SEQUENCE $p_0 \rightarrow p_1$, THE RATIO THEN IS $\frac{Z_1}{Z_0} = \frac{Z_{\eta_1}}{Z_0} \cdots \frac{Z_1}{Z_{\eta_{N-1}}} = \prod_0^{N-1} \frac{Z_{\eta_{j+1}}}{Z_{\eta_j}}$
- LET'S ESTIMATE EACH FACTOR WITH IMPORTANCE SAMPLING → GET THE FINAL RATIO
- SEQUENCE IS DESIGNED TO SUIT THE PROBLEM • **POPULAR CHOICE:** WEIGHTED GEOMETRIC AVERAGE: $p_{\eta_j} \propto p_1^{\eta_j} \cdot p_0^{1-\eta_j}$
- **TO DO SAMPLING** DEFINE MARKOV CHAIN TRANSITION FCN $T_{\eta_j}(x',x)$ TRANSITION PROBS SO TO LEAVE $p_{\eta_j}(x)$ INVARIANT. $p_{\eta_j}(x) = \int p_{\eta_j}(x')\, T_{\eta_j}(x',x)\, dx'$
  → **GIBBS, MH, ...**

→ SAMPLE FROM p_0, USE TRANSITION CHAINS TO SAMPLE FROM INTERMEDIATES UNTIL WE GET TO p_1: $x_{\eta_1} \sim p_0(x)$, $x_{\eta_2} \sim T_{\eta_1}(x_{\eta_1}, x)$ ...

→ **IMPORTANCE WEIGHTS** FOR JUMPS ARE PARTIAL RATIOS OF TRANSITIONS $\frac{\tilde{p}_{\eta_1}(x_{\eta_1})}{\tilde{p}_0(x_0)} \cdots \frac{\tilde{p}_1(x_1)}{\tilde{p}_{\eta_{N-1}}(x_{\eta_{N-1}})} = w^k$

→ **FINAL RATIO:** $\frac{Z_1}{Z_0} \approx \frac{1}{K} \sum_1^K w^k$

- EQUIVALENT TO SIMPLE IMPORTANCE SAMPLING ON EXTENDED STATE SPACE • MOST COMMON WAY OF ESTIMATING UGM PARTITION FCNS, AT THE MOMENT

BRIDGE SAMPLING

RELIES ON INTERMEDIATE DISTRIBUTION p_* $\quad \frac{Z_1}{Z_0} \approx \sum_1^K \frac{\tilde{p}_*(x_0)}{\tilde{p}_0(x_0)} \Big/ \sum_1^K \frac{\tilde{p}_1(x_1)}{\tilde{p}_*(x_1)}$ OPTIMAL DISTRIBUTION IS $\propto \frac{\tilde{p}_0(x)\,\tilde{p}_1(x)}{R\,\tilde{p}_0(x) + \tilde{p}_1(x)}$. R IS $\frac{Z_1}{Z_0}$!!! → RECURSIVE ESTIMATE FROM COARSE START

- AIS > BRIDGE IF $KL(p_0 \| p_1)$ IS LARGE
- **LINKED IMPORTANCE SAMPLING:** USE BRIDGE TO 'INTERPOLATE' AIS SEQUENCE
- **PARTITION FCN TRACKING:** BRIDGE SAMPLING ESTIMATE OF RATIOS OF PARTITION FCNS OF NEIGHBORING PARALLEL TEMPERING CHAINS COMBINED WITH AIS ESTIMATES OVER TIME → LOW VARIANCE Z ESTIMATE AT EVERY ITERATION