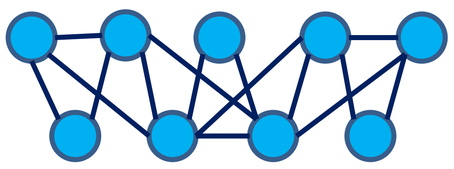
Importance Sampling over Sets: A New Probabilistic Inference Scheme

uai2015

Stefan Hadjis, Stefano Ermon
Stanford University

All code and data available at: cs.stanford.edu/~shadjis

Partition Function

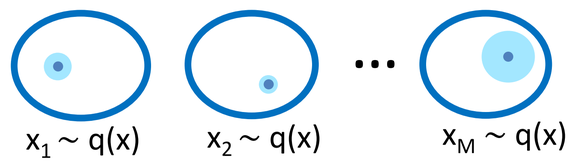


- Undirected graphical model
- n binary variables, $X = \{0,1\}^n$
- Product of factors

$$w(x) = \prod_{\alpha \in \mathcal{I}} \psi_\alpha(\{x\}_\alpha)$$
$$p(x) = w(x)/Z$$
$$Z = \sum_{x \in \mathcal{X}} w(x)$$

Importance Sampling

- Proposal Distribution q(x)
- Sample single points $\mathbf{x}_i \in X$

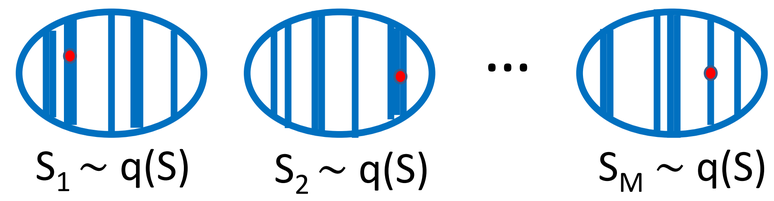


- Rescale by the **importance weight** $q(x)$

$$\widehat{Z} = \frac{1}{M} \sum_{i=1}^{M} \frac{w(x_i)}{q(x_i)}$$

Importance Sampling over Sets (ISS)

- Sample sets of points $S \subseteq X$ from Set-Proposal Distribution $\mathbf{q}(S)$



Set size can be **exponential**, sampled implicitly

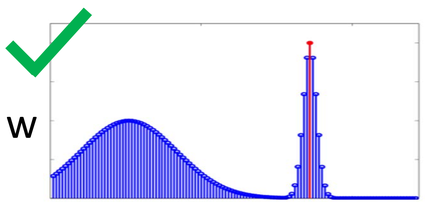
- E.g. sample $S \sim q$ by randomly constraining variables $x \in \mathbf{x}$
- Approximate Z with sum of all sampled configurations $\mathbf{x} \in S$
- Sum dominated by a few terms: use the **mode** instead

$$\widehat{Z} = \sum_{x \in S} \frac{w(\mathbf{x})}{\gamma(\mathbf{x}, q)} \rightarrow \widehat{Z} = \max_{x \in S} \frac{w(\mathbf{x})}{\gamma(\mathbf{x}, q)}$$

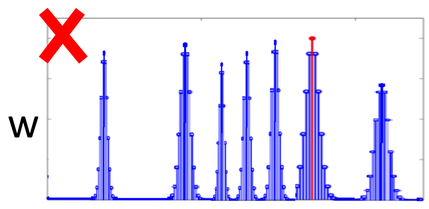
- γ = total probability $\mathbf{x}$ exists in some set S sampled from q

Multiple Set-Proposal Distributions

- When is the mode a good approximation to the total area?

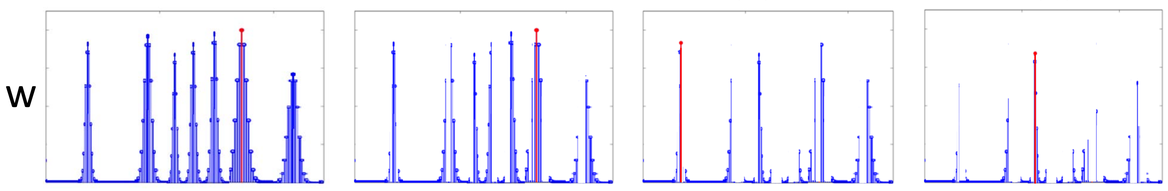


Configurations of $\mathbf{x} \in S$



Configurations of $\mathbf{x} \in S$

- If many $\mathbf{x} \in S$ have $w(\mathbf{x})$ similar to the mode, subsample to obtain a smaller set S' since $\max_{\mathbf{x} \in S} w(\mathbf{x}) \approx \max_{\mathbf{x} \in S'} w(\mathbf{x})$



$S \sim q_1(S)$ $S \sim q_2(S)$ $S \sim q_3(S)$ $S \sim q_4(S)$

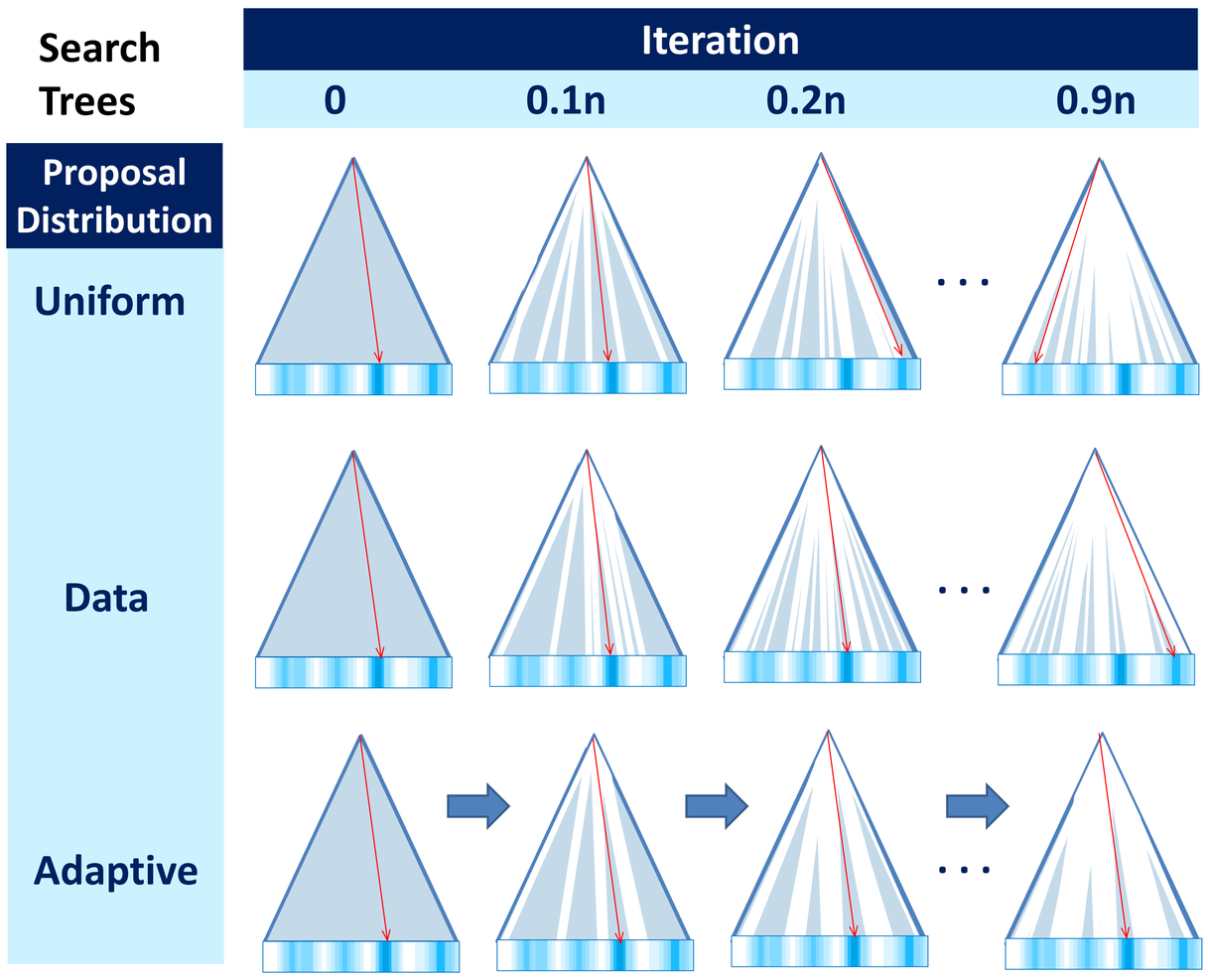
- By exponentially varying the typical size of a sampled set, estimate logZ for distributions peaked to various degrees
- Even faster by searching for only the most "peaked" $q_i(S)$:

$$\widehat{Z} = \max_{i=1}^{k} \max_{\mathbf{x} \in S_i} \frac{w(\mathbf{x})}{\gamma(\mathbf{x}, q_i)}$$

- Guaranteed to estimate logZ at least as well as MAP/MPE configuration and importance sampling

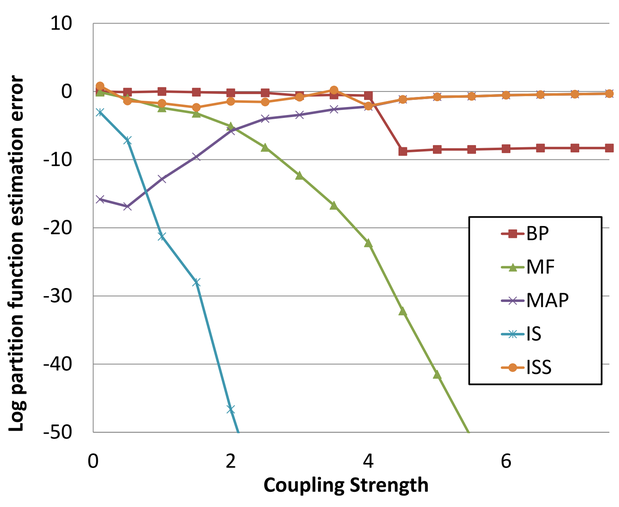
Adaptive Set Importance Sampling

- While searching for optimal $q_i(S)$, define each new set-proposal distribution q_{i+1} with empirical marginals of previous iteration q_i



Partition Function Experiments

- When MAP configuration can be computed efficiently (e.g. using graphcuts), Importance Sampling over Sets can scale to models with a million variables



10x10 Attractive Ising Models
Field factor 2.0, various coupling strengths

Algorithm	$\log Z$
AIS (no data)	446.2
AIS (data)	451.1
BP	–
MF	437.5
MAP	71.6
IS	447.2
ISS (no data)	450.2

Restricted Boltzmann machine trained on MNIST
784 visible, 500 hidden units

Learning Algorithm based on Cutting Planes

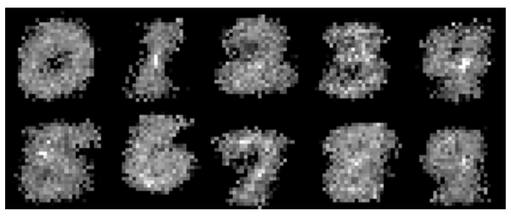
Formulate learning as a **Linear Program:**

$$\underset{\theta, \alpha}{\text{maximize}} \quad \theta \frac{1}{M} \sum_{i=1}^{M} \phi(\mathbf{x}_i) - \alpha$$
$$\text{subject to} \quad \alpha \geq \log Z(\theta)$$

- Approximate logZ using importance sampling over sets (ISS)
- $\alpha \rightarrow$ true logZ with an exponential number of constraints
- Each iteration, add LP constraint for heaviest $\mathbf{x}$ from ISS (most violated constraint) such that $\alpha \rightarrow$ logZ

Iteration 5

Iteration 1000



Visualization of weights of a Naïve Bayes model trained generatively on MNIST using the cutting planes algorithm