Importance Sampling over Sets: A New Probabilistic Inference Scheme

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

Importance Sampling
- Proposal single points \( x_i \in X \)
  \[ x_1 \sim q(x) \quad x_2 \sim q(x) \quad \ldots \quad x_M \sim q(x) \]
- Rescale by the importance weight \( q(x) \)
  \[ \hat{Z} = \frac{1}{M} \sum_{i=1}^{M} \frac{w(x_i)}{q(x_i)} \]
- \( \gamma \) = total probability \( x \) exists in some set \( S \) sampled from \( q \)

Importance Sampling over Sets (ISS)
- Sample sets of points \( S \subseteq X \) from Set-Proposal Distribution \( q(S) \)
  \[ S_1 \sim q(S) \quad \ldots \quad S_M \sim q(S) \]
- Set size can be exponential, sampled implicitly
  - E.g. sample \( S \sim q \) by randomly constraining variables \( x \in x \)
  - Approximate \( Z \) with sum of all sampled configurations \( x \in S \):
    \[ \hat{Z} = \sum_{x \in S} \frac{w(x)}{\gamma(x,q)} \]
    \[ \hat{Z} = \max_{x \in S} \frac{w(x)}{\gamma(x,q)} \]

Multiple Set-Proposal Distributions
- When is the mode a good approximation to the total area?
  - If many \( x \in S \) have \( w(x) \) similar to the mode, subsample to obtain a smaller set \( S' \) since \( \min_{x \in S} w(x) \approx \min_{x \in S'} w(x) \)
  - By exponentially varying the typical size of a sampled set, estimate \( \log Z \) for distributions peaked to various degrees
  - Even faster by searching for only the most “peaked” \( q(S) \):
    \[ \hat{Z} = \max_{i=1}^{k} \max_{x \in S} \frac{w(x)}{\gamma(x,q)} \]
    \[ \hat{Z} = \max_{x \in S} \frac{w(x)}{\gamma(x,q)} \]
    \[ \gamma \] = total probability \( x \) exists in some set \( S \) sampled from \( q \)

Adaptive Set Importance Sampling
- While searching for optimal \( q(S) \), define each new set-proposal distribution \( q_{i+1} \) with empirical marginals of previous iteration \( q_i \)
- Proposal Distribution

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

Learning Algorithm based on Cutting_planes
- Formulate learning as a Linear Program:
  \[ \max_{\theta, \alpha} \frac{1}{M} \sum_{i=1}^{M} \phi(x_i) - \alpha \]
  subject to \( \alpha \geq \log Z(\theta) \)
- Approximate \( \log Z \) using importance sampling over sets (ISS)
- Each iteration, add LP constraint for heaviest \( x \) from ISS (most violated constraint) such that \( \alpha \rightarrow \log Z \)

Learning Algorithm based on Cutting Planes

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