Bridges: Inference and the Monte Carlo Method

Jingcheng Liu

SJTU-ACM Class 2010

December 26, 2013

Outline

1. Statistical Inference
   - Bayesian Network
   - Inference
2. Monte Carlo method
3. Free-energy barriers
   - Geometric lowerbound on mixing time
   - Applications and Inapproximability

Factorizing Probability Distributions

There are two ways to encode the factorization structure of a probability distributions:

- Factor graph, which we have seen before.

Factorizing Probability Distributions

There are two ways to encode the factorization structure of a probability distributions:

- Factor graph, which we have seen before.
- Bayesian network.

Factorizing Probability Distributions

There are two ways to encode the factorization structure of a probability distributions:

- Factor graph, which we have seen before.
- Bayesian network.

Definition

A Bayesian network (BN) is a directed graph with a set of conditional probability distributions (CPD). We say a joint distribution factorizes over a BN G if

$$p(\underline{x}) = \prod_{v \in \pi(G)} p(x_v) \prod_{v \in G \setminus \pi(G)} p(x_v \big| \underline{x}_{\pi(v)}),$$

where $\pi(G)$ denotes the set of vertices with no parent, and $\pi(v)$ is the set of parents of the vertex v.

Bayesian Network

Example

See the book. . .

Bayesian Network

Example

See the book. . .

BN is indeed a good structure as representation, and it's usually served as a compact (sparse) representation for the set of conditional independence structure of form $\{(X \perp Y \big| Z)\}$.

Theorem

For almost all distributions P that factorizes over G, i.e. except for a set with measure zero in the space of CPD parameterizations, we have $I(P) = I(G)$, where $I(P)$ is the set of conditional independence structure in P and $I(G)$ is determined via the independence test known as d-separated test.

Inference in coding, physics, and optimization

In the language of inference:

- Symbol MAP decoding, LDPC: factor graphs to BN.
- Statistical mechanics: expectations and covariances.
- **Combinatorial optimization:** A little bit more detailed as follows.

Cost (energy) function $E(\underline{x}) = \sum_a E_a(\underline{x}_a)$.
Let $\mu_*(\underline{x})$ be the uniform distribution over optimal solutions. Thus minimum energy is just $E_* = \sum_a \left(\sum_{\underline{x}} \mu_*(\underline{x}) E_a(\underline{x}_a) \right)$.

Inference in coding, physics, and optimization

In the language of inference:

- Symbol MAP decoding, LDPC: factor graphs to BN.
- Statistical mechanics: expectations and covariances.
- **Combinatorial optimization:** A little bit more detailed as follows.

Cost (energy) function $E(\underline{x}) = \sum_a E_a(\underline{x}_a)$.
Let $\mu_*(\underline{x})$ be the uniform distribution over optimal solutions. Thus minimum energy is just $E_* = \sum_a \left(\sum_{\underline{x}} \mu_*(\underline{x}) E_a(\underline{x}_a)\right)$.
Hard! As we usually don't know $\mu_*(\underline{x})$ explicitly!

Inference in coding, physics, and optimization

In the language of inference:

- Symbol MAP decoding, LDPC: factor graphs to BN.
- Statistical mechanics: expectations and covariances.
- **Combinatorial optimization:** A little bit more detailed as follows.

Cost (energy) function $E(\underline{x}) = \sum_a E_a(\underline{x}_a)$.
Let $\mu_*(\underline{x})$ be the uniform distribution over optimal solutions. Thus minimum energy is just $E_* = \sum_a \left(\sum_{\underline{x}} \mu_*(\underline{x}) E_a(\underline{x}_a) \right)$.
Hard! As we usually don't know $\mu_*(\underline{x})$ explicitly!
Softening as $\mu_\beta(\underline{x}) = e^{-\beta E(\underline{x})}/Z$.
WLOG rescale the energy so that minimum is zero.

Inference in optimization

Let $U(\beta) = \sum_{\underline{x}} \mu_\beta(\underline{x}) E(\underline{x})$. Clearly if we let $\beta \to \infty$ we expect that $U(\beta) \to E_*$. But how large should β be to get good estimate of E_*? Suppose the second smallest energy is bounded away from the smalest energy by a constant C. (e.g. the integer-valued condition on the book).

$$0 \leq \frac{\partial U}{\partial T} \leq \frac{1}{T^2} \Delta_{max}^2 \left|\chi\right|^N e^{-C/T}.$$

Therefore, if we let $T = C'/N$, with $C' \to \infty$ we have $\frac{\partial U}{\partial T} \to 0$, hence $U(\beta) \to E_*$.

Inference via sampling

Computing the marginals gets reduced to almost i.i.d. samples. Always soften the constraint with a temperature.

- LDPC:

$$\mu_{y,\beta}(\underline{x}) = \frac{1}{Z(\underline{y},\beta)} \prod_{a=1}^{M} e^{-\beta E_a(x_{i_1^a} \cdots x_{i_k^a})} \prod_{i=1}^{N} Q\left(y_i \middle| x_i\right).$$

- Ising model: slow mixing in low-temperature phase (for random graph, square lattice, tree).
- MAX-SAT: some numerical experiments on the book...

Geometric lowerbound on mixing time

How to formalize the intuition of Arrhenius law? Let's consider the discrete time case and derive the well-known geometric lowerbound.

Theorem

For $\mathcal{A}$ with $\mu(\mathcal{A}) \leq 1/2$, we have

$$\tau(1/4) \geq \frac{\mu(\mathcal{A})}{4W(\mathcal{A} \to \mathcal{X} \backslash \mathcal{A})}.$$

Proof.

Consider the initial distribution

$$\mu_0(x) = \begin{cases} \frac{\mu(x)}{\mu(\mathcal{A})}, \text{if } x \in \mathcal{A} \\ 0, \text{otherwise} \end{cases}$$

□

Geometric lowerbound on mixing time contd.

Contd.

$$\|\mu_1 - \mu_0\|_{TV} = \frac{W(\mathcal{A} \to \mathcal{X} \backslash \mathcal{A})}{\mu(\mathcal{A})}.$$

$$\frac{1}{4} \geq \|\mu_t - \mu\|_{TV} \geq \|\mu_0 - \mu\| - \|\mu_t - \mu_0\| \geq \frac{1}{2} - t\,\|\mu_1 - \mu_0\|_{TV}$$

This concludes the proof. □

Geometric lowerbound on mixing time contd.

Contd.

$$\|\mu_1 - \mu_0\|_{TV} = \frac{W(\mathcal{A} \to \mathcal{X} \backslash \mathcal{A})}{\mu(\mathcal{A})}.$$

$$\frac{1}{4} \geq \|\mu_t - \mu\|_{TV} \geq \|\mu_0 - \mu\| - \|\mu_t - \mu_0\| \geq \frac{1}{2} - t\,\|\mu_1 - \mu_0\|_{TV}$$

This concludes the proof. □

Example

Applications in inapproximability of counting independent sets in an a.a.s 6-regular graph.

Thank you!