Northeastern

CS 7140:
ADVANCED MACHINE LEARNING

Recap: Importance Sampling

Problem: Estimate expectations of functions under the distribution

$$\Phi = \mathbb{E}_{\mathbf{x}\sim P(\mathbf{x})}[\phi(\mathbf{x})] \equiv \int P(\mathbf{x})\phi(\mathbf{x})d\mathbf{x}$$

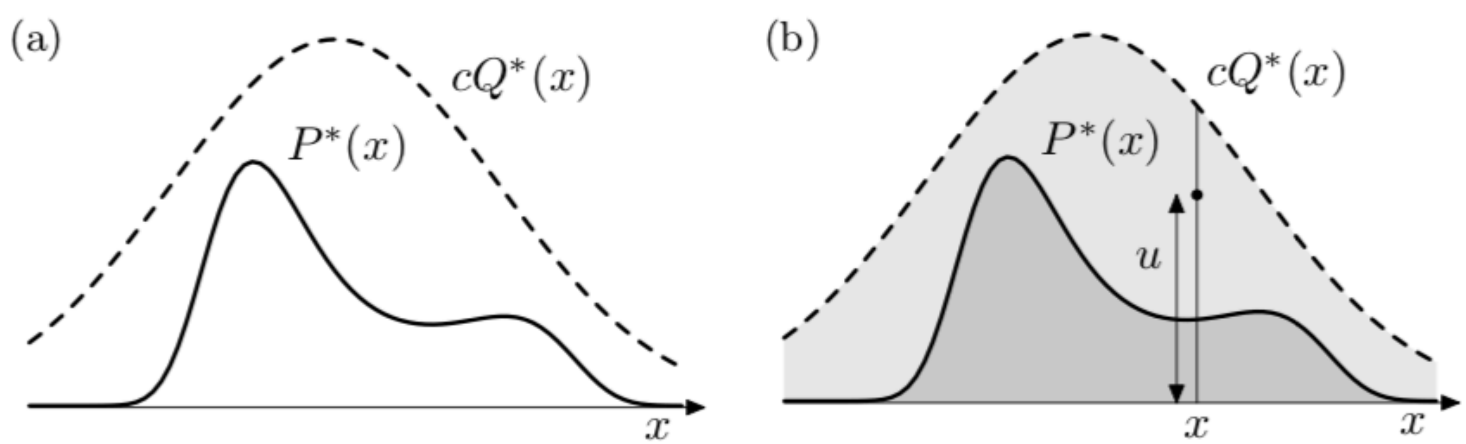
- Sample from $P(x)$ is **hard**, sample from $Q(x)$ is **simple**

Importance: $w_r \equiv \dfrac{P^\star(x^{(r)})}{Q^\star(x^{(r)})}$ target / proposal

$$\boxed{\hat{\Phi} \equiv \frac{\sum_r w_r \phi(x^{(r)})}{\sum_r w_r}}$$

Recap: Rejection Sampling

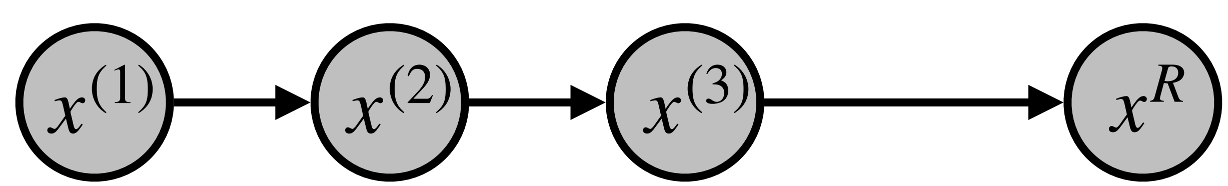
- Sampling from $P(x)$ is **hard**, sampling from $Q(x)$ is **simple**
- Assume we know c, such that $cQ^\star(x) > P^\star(x)$



1. Draw a sample x from $Q(x)$
2. Draw a point u from uniform $[0,\ cQ^\star(x)]$
3. Reject x is $u > P^\star(x)$, accept otherwise

Recap: Metropolis-Hastings Method

- Draw a sample x from $Q(x; x^{(t)})$
- Evaluate $a = \dfrac{P^\star(x)Q(x^{(t)}; x)}{P^\star(x^{(t)})Q(x; x^{(t)})}$
  - If $a \geq 1$,accept, set $x^{(t+1)} = x$
  - Otherwise, reject, set $x^{(t+1)} = x^{(t)}$



APPROXIMATE INFERENCE:VARIATIONAL METHOD

Approximate Inference

- A method for approximating a complex distribution

- Gibbs inequality

$$D_{KL}(Q||P) = \sum_{x} Q(x) \log \frac{Q(x)}{P(x)} \geq 0$$

  - relative entropy is always nonnegative

  - non symmetric: $D_{KL}(P||Q) \neq D_{KL}(Q||P)$

- Applications in statistical physics, machine learning and data science

Variational Free Energy

- Approximating the **complex** $P(x)$ with a **simple** $Q(x;\theta)$
- Probability distribution $P(x) = \dfrac{1}{Z} P^{\star}(x) = \dfrac{1}{Z} \prod_{m=1}^{M} \phi(x_m)$
- By Gibbs inequality

$$\begin{aligned} D_{KL}(Q\,||\,P) &= \log Z - \sum_{m} \mathbb{E}_Q[\log \phi] - H_Q \\ &= \log Z + F[P^{\star}, Q] \geq 0 \end{aligned}$$

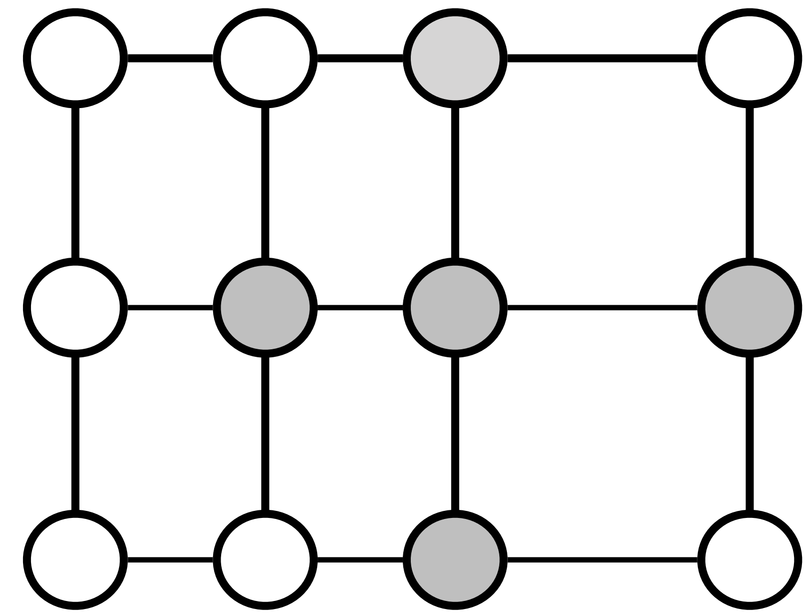
variational free energy

- Minimizing the relative entropy is equivalent to minimizing the variational free energy

Variational Inference

- Finding a good approximation $Q(x;\theta)$ to minimize the relative entropy $D_{KL}(Q||P)$
- Equivalent to minimizing the variational free energy
- Energy functional is a *lower bound* of the partition function $\log Z \geq -F[P^\star, Q]$
- Approximation quality depends on the choice of Q and variational parameters θ

Ising Model



- Binary probability distribution

$$P(x \mid \beta, J) = \frac{1}{Z} \exp[-\beta E(x; J)]$$

- Rewrite the expression

$$F(P^{\star}, Q) = \beta \sum_{x} Q(x; \theta) E(x; J) - H_Q$$

$$\equiv \beta \mathbb{E}_Q[E(x; J)] - H_Q$$

mean energy

entropy

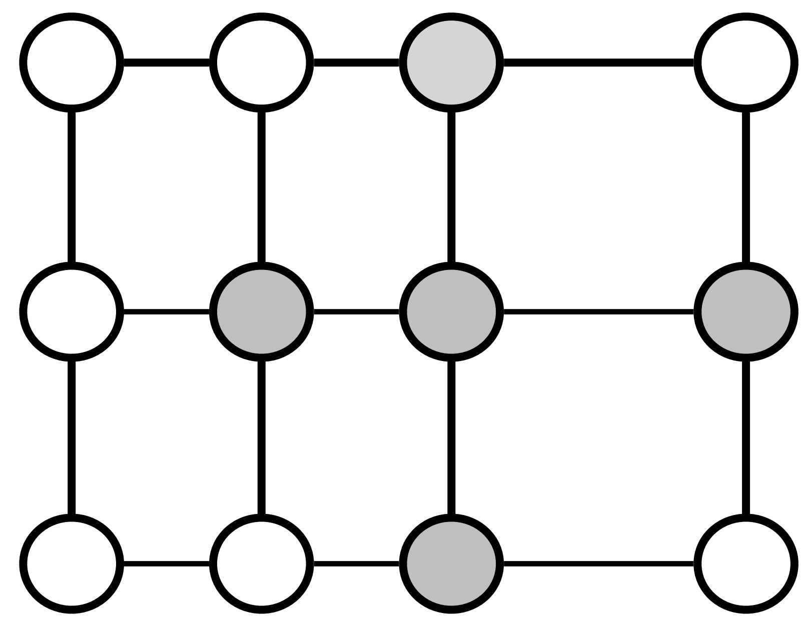
- Energy function

$$E(x; J) = -\frac{1}{2} \sum_{m,n} J_{mn} x_m x_n - \sum_{n} h_n x_n$$

Mean Field Theory

- Choose a **separable** approximating distribution

$$Q(x;a) = \prod_n Q_n(x_n;a) = \frac{1}{Z}\exp(\sum_n a_n x_n)$$



- The entropy

$$H_Q = \sum_x Q(x;a)\log\frac{1}{Q(x;a)}$$

- For a single node x_n

$$q_n = \frac{e^{a_n}}{e^{a_n}+e^{-a_n}} = \frac{1}{1+\exp(-2a_n)} \quad H(q) = q\log\frac{1}{q} + (1-q)\log\frac{1}{1-q}$$

Mean Field Theory

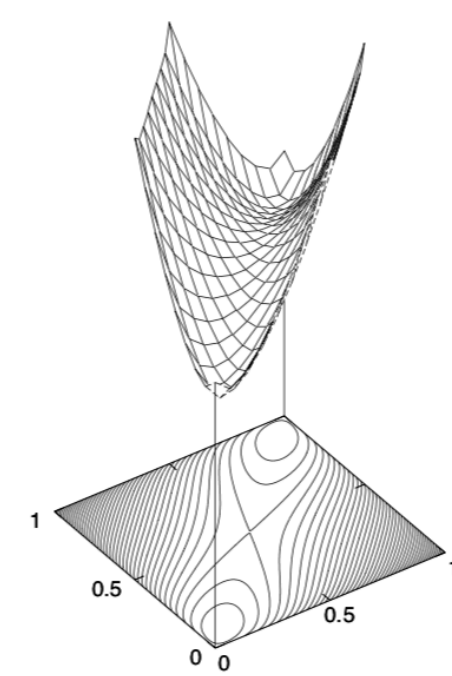
- The mean energy

$$\mathbb{E}_Q[E(x;J))] = \mathbb{E}_Q[-\frac{1}{2}\sum_{mn} J_{mn}x_m x_n - \sum_n h_n x_n]$$

$$= -\frac{1}{2}\sum_{mn} J_{mn}\mathbb{E}_Q[x_m]\mathbb{E}_Q[x_n] - \sum_n h_n \mathbb{E}_Q[x_n]$$

- The mean value for a single node x_n

$$\mathbb{E}_Q[x_n] = \frac{e^{a_n} - e^{-a_n}}{e^{a_n} + e^{-a_n}} = 2q_n - 1$$



Ising Model

- Minimize the variational free energy

$$F(P^\star, Q) = \beta\Big(-\frac{1}{2}\sum_{mn} J_{mn}\mathbb{E}_Q[x_m]\mathbb{E}_Q[x_n] - \sum_n h_n\mathbb{E}_Q[x_n]\Big) - \sum_n H(q_n)$$

- Find the parameter by taking the derivative

$$\frac{\partial}{\partial a_m}F = \beta\Big[-\sum_n J_{mn}\mathbb{E}[x_n] - h_m\Big]\Big(2\frac{\partial q_m}{\partial a_m}\Big) - \log\Big(\frac{1-q_m}{q_m}\Big)\Big(\frac{\partial q_m}{\partial a_m}\Big)$$

- Setting the derivative to zero

$$a_m = \beta\Big(\sum_n J_{mn}\mathbb{E}_Q[x_n] + h_m\Big)$$