CS 7140: ADVANCED MACHINE LEARNING

Recap: 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

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

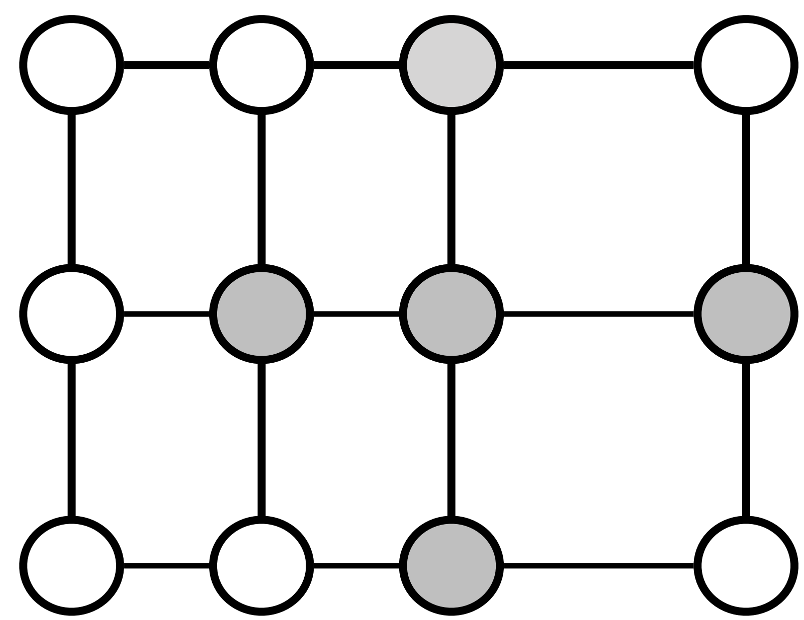
variational free energy

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

Recap: 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

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

PARAMETER LEARNING

Maximum Likelihood Estimation

- Finding a hypothesis that fits the data well
  $\theta^\star = \text{argmax}_\theta \log P(D \mid \theta, H)$
- Work with the *logarithm* of the likelihood
  - products of probabilities tends too be small
  - likelihood multiples, log likelihood adds
- MLE is equivalent to minimize the relative entropy
  $KL(P(x \mid \theta^\star) \,||\, P(x \mid \theta)) = \mathbb{E}[\log P(x \mid \theta^\star)] - H[P(x \mid \theta^\star)]$

Example: One Gaussian

- Data $\{x_n\}_{n=1}^{N}$

- Log likehood
  $\log P(\{x_n\}_{n=1}^{N} \,|\, \mu, \sigma) = -N \ln(\sqrt{2\pi}\sigma) - \sum_{n} (x_n - \mu)^2/(2\sigma^2)$

- Sample mean $\bar{x} \equiv \sum_{n=1}^{N} x_n/N$ $\boxed{\text{sufficient statistics}}$
  Sum of deviation $S \equiv \sum_{n} (x_n - \bar{x})^2$

- MLE: $\mu = \bar{x} \quad \sigma^2 = S/N$ (hint: use $du^n/d(\ln u) = nu^n$)

Example: One Gaussian

- Log likehood
$$\log P(\{x_n\}_{n=1}^{N} \,|\, \mu, \sigma) = -N \ln(\sqrt{2\pi}\sigma) - \sum_n (x_n - \mu)^2/(2\sigma^2)$$

- Maximum likelihood mean μ is the sample mean, for any value of σ
$$\frac{\partial}{\partial \mu} \log P = -\frac{N(\mu - \bar{x})}{\sigma^2} = 0$$

- Maximum likelihood standard deviation σ
$$\frac{\partial \ln P}{\partial \ln \sigma} = -N + \frac{S}{\sigma^2} = 0$$

Example: Mixture of Gaussian

- Data $\{x_n\}_{n=1}^N$
- Probability

$$P(x \mid \mu_1, \mu_2, \sigma) = \Big[\sum_{k=1}^{2} p_k \frac{1}{\sqrt{2\pi\sigma}} \exp\Big(-\frac{(x-\mu_k)^2}{2\sigma^2}\Big)\Big]$$

Parameters $\theta = \{\{\mu_k\}_{k=1}^2, \sigma\}$

- Take the log likelihood L

$$\frac{\partial}{\partial \mu_k} L = \sum_n p_{k|n} \frac{(x_n - \mu_k)}{\sigma^2}$$ where $p_{k|n} = P(k_n = k \mid x_n, \theta)$

Soft K-means

- Fitting a mixture of spherical Gaussian
- Variance is the same in all directions
- Can take a long time to converge

Assignment step. The responsibilities are

$$r_k^{(n)} = \frac{\pi_k \frac{1}{(\sqrt{2\pi}\sigma_k)^I} \exp\left(-\frac{1}{\sigma_k^2} d(\mathbf{m}^{(k)}, \mathbf{x}^{(n)})\right)}{\sum_{k'} \pi_k \frac{1}{(\sqrt{2\pi}\sigma_{k'})^I} \exp\left(-\frac{1}{\sigma_{k'}^2} d(\mathbf{m}^{(k')}, \mathbf{x}^{(n)})\right)} \tag{22.22}$$

where I is the dimensionality of $\mathbf{x}$.

Update step. Each cluster's parameters, $\mathbf{m}^{(k)}$, π_k, and σ_k^2, are adjusted to match the data points that it is responsible for.

$$\mathbf{m}^{(k)} = \frac{\sum_n r_k^{(n)} \mathbf{x}^{(n)}}{R^{(k)}} \tag{22.23}$$

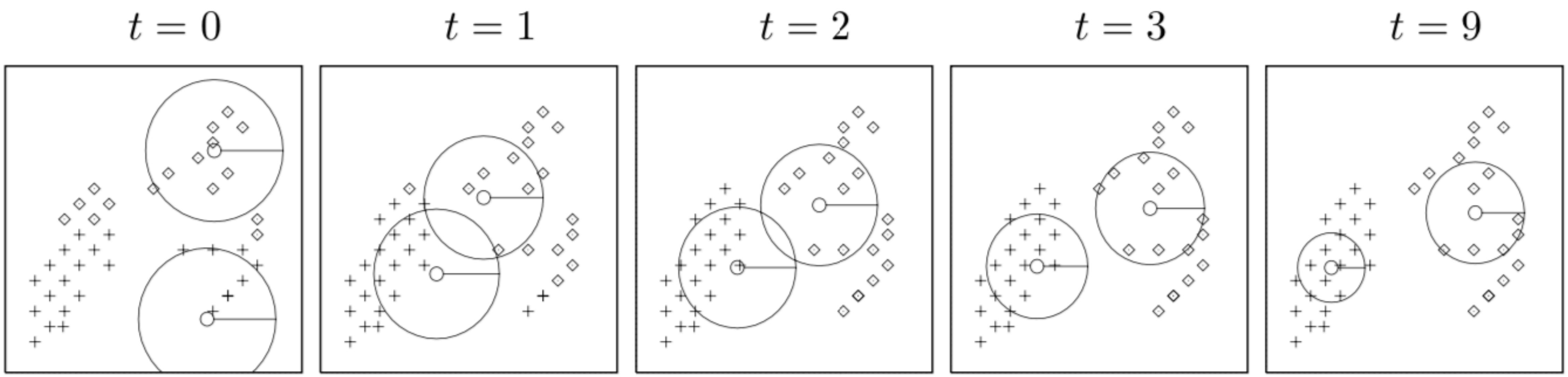
$$\sigma_k^2 = \frac{\sum_n r_k^{(n)} (\mathbf{x}^{(n)} - \mathbf{m}^{(k)})^2}{I R^{(k)}} \tag{22.24}$$

$$\pi_k = \frac{R^{(k)}}{\sum_k R^{(k)}} \tag{22.25}$$

where $R^{(k)}$ is the total responsibility of mean k,

$$R^{(k)} = \sum_n r_k^{(n)}. \tag{22.26}$$

Soft K-means

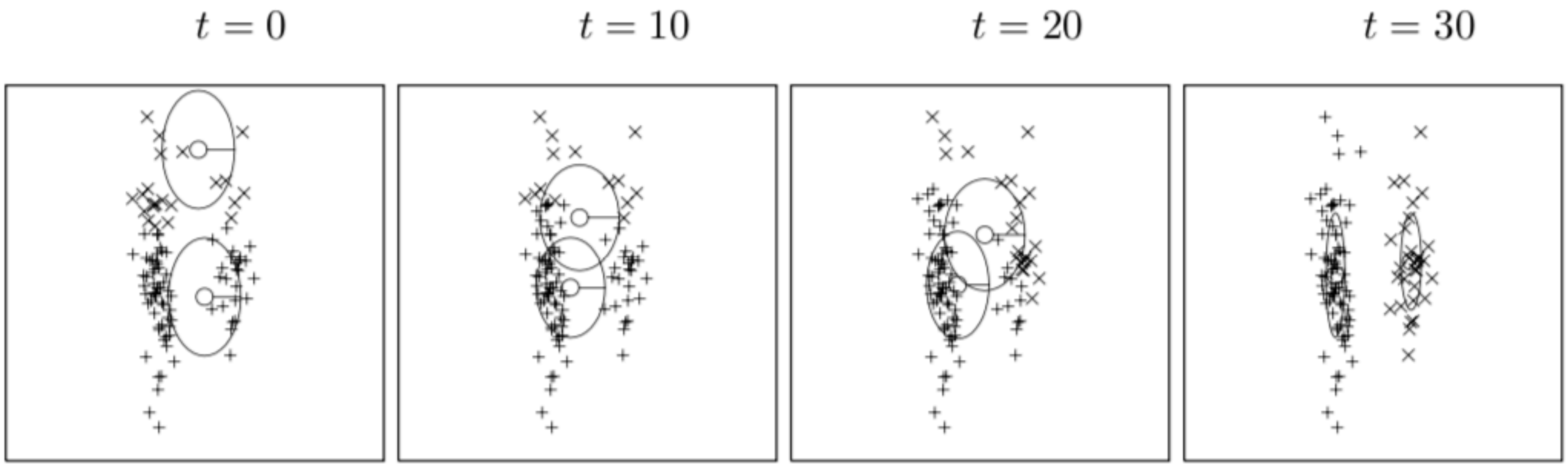


- model clusters with axis-aligned Gaussian
- with possibly -unequal variances

$$r_k^{(n)} = \frac{\pi_k \frac{1}{\prod_{i=1}^{I} \sqrt{2\pi}\sigma_i^{(k)}} \exp\left(-\sum_{i=1}^{I} (m_i^{(k)} - x_i^{(n)})^2 \Big/ 2(\sigma_i^{(k)})^2\right)}{\sum_{k'} \text{(numerator, with } k' \text{ in place of } k)} \qquad (22.27)$$

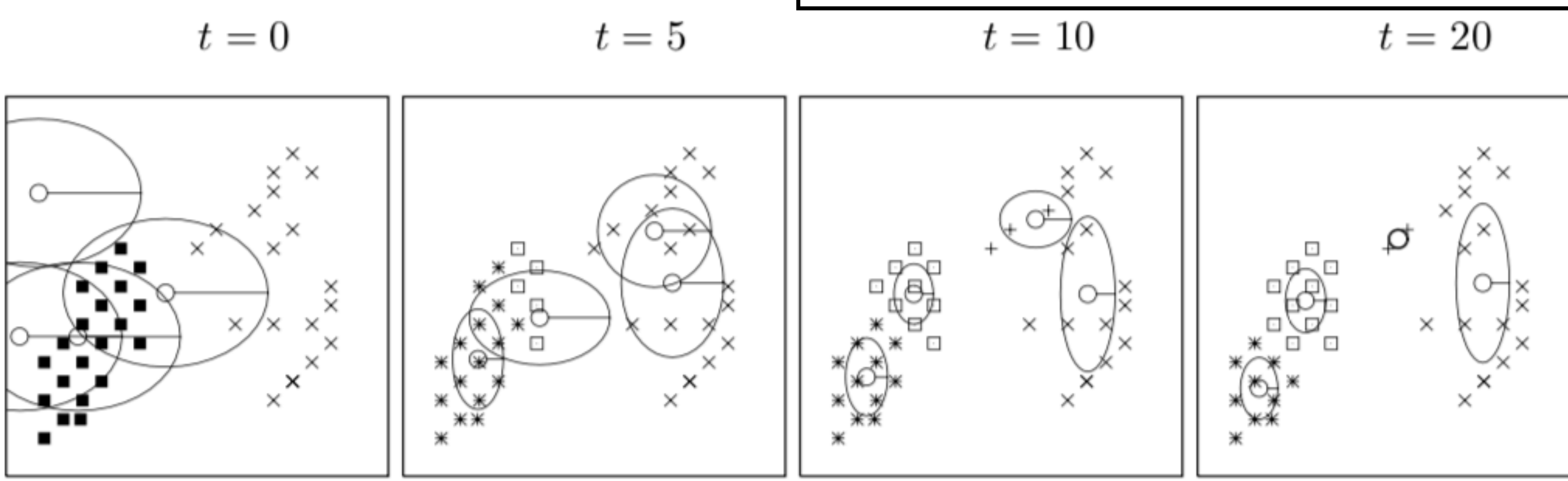
$$\sigma_i^{2\,(k)} = \frac{\sum_n r_k^{(n)} (x_i^{(n)} - m_i^{(k)})^2}{R^{(k)}} \qquad (22.28)$$

Soft K-means



- A fatal flaw of MLE

infinitely large likelihood



Drawback of MLE

- The likelihood may be infinitely large
- Unrepresentative in high-dimensional problems
- Example: one Gaussian: $\mu = \bar{x} \qquad \sigma_N^2 = S/N$
  μ is unbiased $\mathbb{E}[\mu] = \mu^\star$, how about σ_N ?

$$\boxed{\sigma_N \text{ is biased, but } \sigma_{N-1} \text{ is unbiased}}$$

Maximum a Posterior (MAP)

- $P(\theta|D) \propto P(D|\theta)P(\theta)$

  posterior likelihood prior

- Conjugate distributions

  $P(\theta|D) \propto P(D|\theta)P(\theta)$

  same distribution family

- Example: Dirichlet prior is conjugate to multinomial

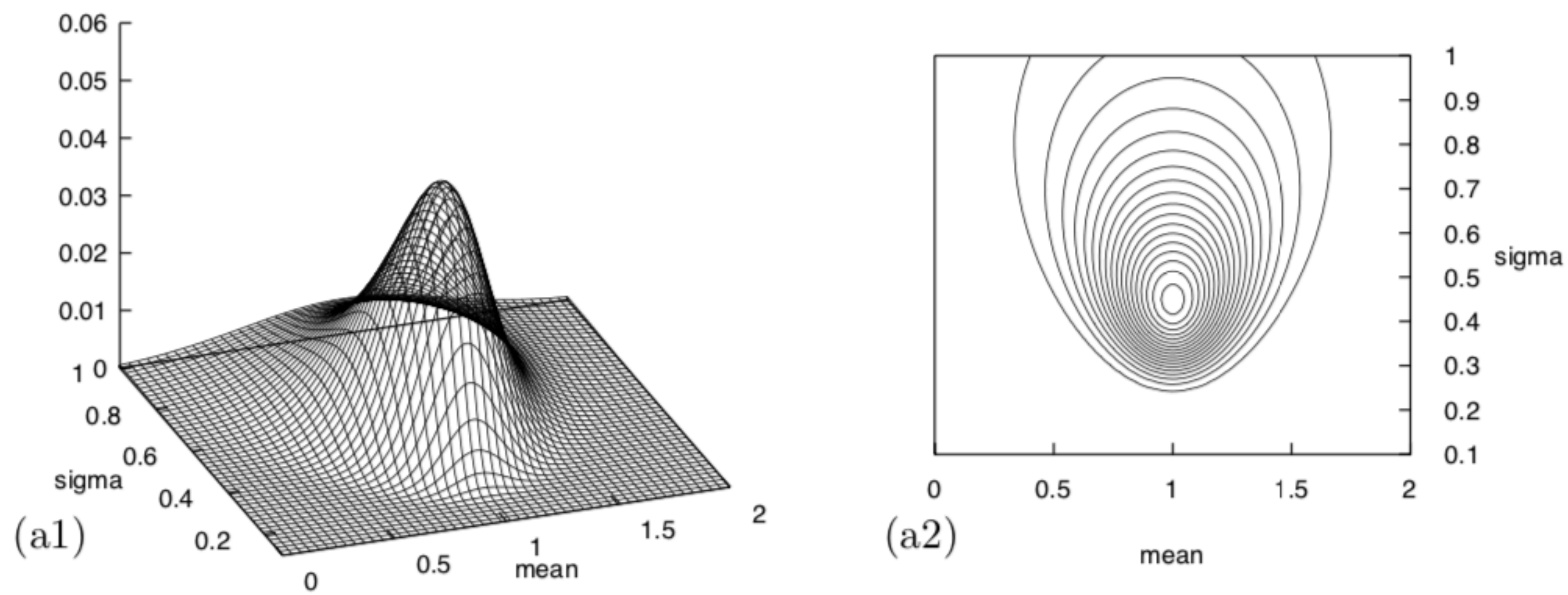
  if $P(\theta) = \text{Dir}(\alpha_1, \cdots, \alpha_K)$ then

  $P(\theta|D) = \text{Dir}(M_1 + \alpha_1, \cdots, M_K + \alpha_K)$

One Gaussian

- Log likelihood
  $$\log P(\{x_n\}_{n=1}^{N} \mid \mu, \sigma) = -N \ln(\sqrt{2\pi}\sigma) - \sum_n (x_n - \mu)^2/(2\sigma^2)$$

- Prior $\dfrac{1}{\sigma_n}$ and $\dfrac{1}{\sigma}$



- Posterior
  $$P(\mu \mid D, \sigma) \propto \exp\big(-N(\mu - \bar{x})^2/(2\sigma^2)\big) = N(\bar{x}, \sigma^2/N)$$

Maximum a Posterior

- Using prior to regularize the likelihood

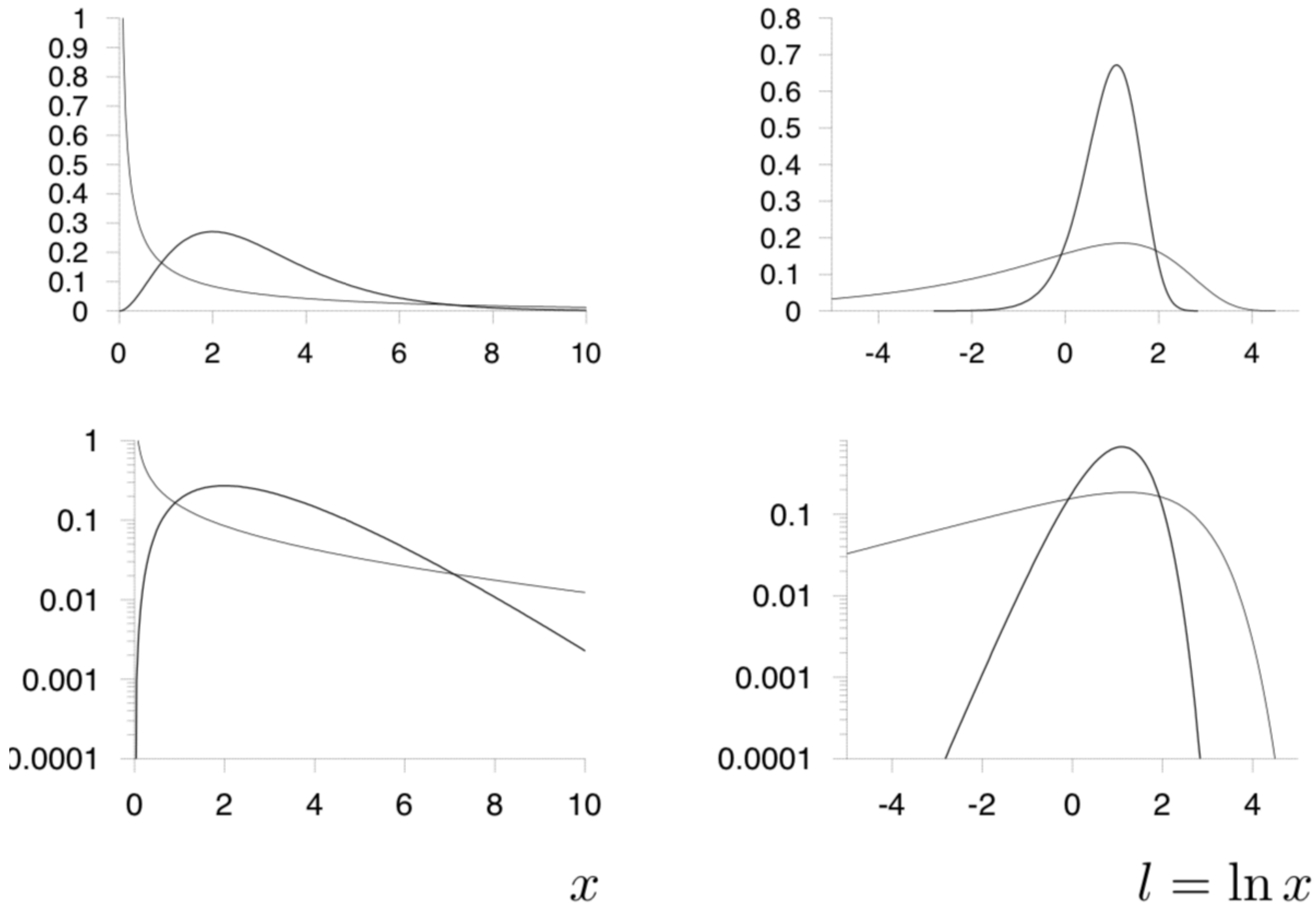
$$\text{argmax}_\theta \log P(\theta \mid D) = \text{argmax}_\theta \log\left(\frac{P(D \mid \theta)P(\theta)}{P(D)}\right)$$
$$= \text{argmax}_\theta \left(\log P(\theta) + \log(P(D \mid \theta)\right)$$

- No harder than MLE estimation

- Draw back: *Representation Dependent*

$$P(u) = P(\theta)\left|\frac{\partial\theta}{\partial u}\right|$$

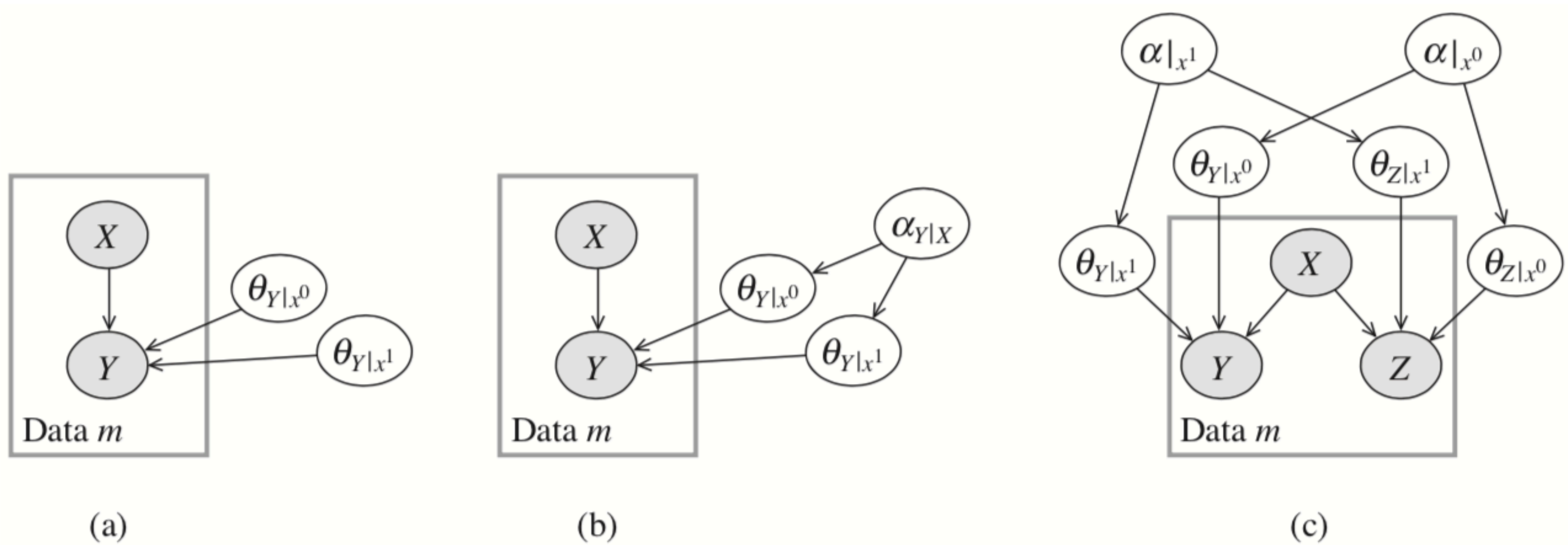
Representation Dependent



Gamma distribution of with parameters (s, c) = (1, 3) (heavy lines) and 10, 0.3 (light lines) shown on linear vertical scales (top) and logarithmic vertical scales (bottom)

- Gamma distribution $P(x\,|\,s,c) = \dfrac{1}{\Gamma(c)s}\left(\dfrac{x}{s}\right)^{c-1}\exp\left(-\dfrac{x}{s}\right)$
- $P(\ln x) = P(x)\left|\dfrac{\partial x}{\partial \ln x}\right| = \dfrac{1}{\Gamma(c)}\left(\dfrac{x}{s}\right)^{c}\exp\left(-\dfrac{x}{s}\right)$

Hierarchical Prior



- Hierarchical Bayesian model: introduce prior over the the parameters of the prior distribution

- Particular useful for small data

Biased Estimator

$$s = \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$
$$= \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \frac{1}{N}\sum_{i=1}^{N}(x_i)\right)^2$$
$$= \frac{1}{N}\sum_{i=1}^{N}\left[x_i^2 - 2x_i\frac{1}{N}\sum_{i=1}^{N}(x_i) + \left(\frac{1}{N}\sum_{i=1}^{N}(x_i)\right)^2\right]$$
$$= \frac{\sum_{i=1}^{N} x_i^2}{N} - \frac{2\sum_{i=1}^{N} x_i \sum_{i=1}^{N} x_i}{N^2} + \left(\frac{\sum_{i=1}^{N} x_i}{N}\right)^2$$
$$= \frac{\sum_{i=1}^{N} x_i^2}{N} - \frac{2\sum_{i=1}^{N} x_i \sum_{i=1}^{N} x_i}{N^2} + \left(\frac{\sum_{i=1}^{N} x_i}{N}\right)^2$$
$$= \frac{\sum_{i=1}^{N} x_i^2}{N} - \left(\frac{\sum_{i=1}^{N} x_i}{N}\right)^2$$

$$E[s] = \frac{\sum_{i=1}^{N} E[x_i^2]}{N} - \frac{E[(\sum_{i=1}^{N} x_i)^2]}{N^2}$$
$$= s + \mu^2 - \frac{E[(\sum_{i=1}^{N} x_i)^2]}{N^2}$$
$$= s + \mu^2 - \frac{E[\sum_{i=1}^{N} x_i^2 + \sum_{i}^{N}\sum_{j\neq i}^{N} x_i x_j]}{N^2}$$
$$= s + \mu^2 - \frac{E[N(s+\mu^2) + \sum_{i}^{N}\sum_{j\neq i}^{N} x_i x_j]}{N^2}$$
$$= s + \mu^2 - \frac{N(s+\mu^2) + \sum_{i}^{N}\sum_{j\neq i}^{N} E[x_i]E[x_j]}{N^2}$$
$$= s + \mu^2 - \frac{N(s+\mu^2) + N(N-1)\mu^2}{N^2}$$
$$= s + \mu^2 - \frac{N(s+\mu^2) + N^2\mu^2 - N\mu^2}{N^2}$$
$$= s + \mu^2 - \frac{s + \mu^2 + N\mu^2 - \mu^2}{N}$$
$$= s + \mu^2 - \frac{s}{N} - \frac{\mu^2}{N} - \mu^2 + \frac{\mu^2}{N}$$
$$= s - \frac{s}{N}$$
$$= s\left(1 - \frac{1}{N}\right)$$
$$= s\left(\frac{N-1}{N}\right)$$