CS7140: Advanced Machine Learning **Spring 2020**

Lecture 6: Parameter Learning

Lecturer: Rose Yu *Scribes: Nihang Fu, Yuxuan Cai*

6.1 Maximum Likelihood Estimation

Rather than enumerate all hypotheses, which may be exponential in number, we can save a lot of time by homing in on one good hypothesis that fits the data well. This is the philosophy behind the maximum likelihood method, which identifies the setting of the parameter vector θ that maximizes the likelihood, $P(Data|\theta;\eta)$, that is, the goal of ML is to find a hypothesis that fits the data well

$$\theta^{\star} = \text{argmax}_{\theta} \log P(D|\theta, H) \tag{6.1}$$

The reasons why we work with the logarithm of the likelihood are the products of probabilities tends too be small and can turn the likelihood multiples to log likelihood adds.

MLE is equivalent to minimize the relative entropy

$$KL\left(P\left(x|\theta^{\star}\right) \| P(x|\theta)\right) = \mathbb{E}\left[\log P\left(x|\theta^{\star}\right)\right] - H\left[P\left(x|\theta^{\star}\right)\right] \tag{6.2}$$

Proof:

$$\begin{aligned} D_{KL}\left[P\left(x|\theta^{*}\right) \| P(x|\theta)\right] &= \mathbb{E}_{x\sim P(x|\theta^{*})}\left[\log \frac{P\left(x|\theta^{*}\right)}{P(x|\theta)}\right] \\ &= \mathbb{E}_{x\sim P(x|\theta^{*})}\left[\log P\left(x|\theta^{*}\right) - \log P(x|\theta)\right] \\ &= \mathbb{E}_{x\sim P(x|\theta^{*})}\left[\log P\left(x|\theta^{*}\right)\right] - \mathbb{E}_{x\sim P(x|\theta^{*})}[\log P(x|\theta)] \\ &= \mathbb{E}\left[\log P\left(x|\theta^{*}\right)\right] - H\left[P\left(x|\theta^{\star}\right)\right] \end{aligned} \tag{6.3}$$

Example 6.1: MLE for one Gaussian.

Assume we have data $\{x_n\}_{n=1}^{N}$. The log likelihood is:

$$\ln P\left(\{x_n\}_{n=1}^{N} | \mu, \sigma\right) = -N\ln(\sqrt{2\pi}\sigma) - \sum_{n}\left(x_n - \mu\right)^2 / \left(2\sigma^2\right) \tag{6.4}$$

The sample mean is $\bar{x} \equiv \sum_{n=1}^{N} x_n/N$ and the sum of square deviations is $S = \sum_n \left(x_n - \bar{x}\right)^2$. Then, the likelihood can be expressed in terms of two functions of the data

$$\ln P\left(\{x_n\}_{n=1}^{N} | \mu, \sigma\right) = -N\ln(\sqrt{2\pi}\sigma) - \left[N(\mu - \bar{x})^2 + S\right] / \left(2\sigma^2\right) \tag{6.5}$$

Because the likelihood depends on the data only through $\bar{x}$ and S, these two quantities are known as sufficient statistics.

Sufficient Statistics: In statistics, a statistic is sufficient with respect to a statistical model and its associated unknown parameter if ”no other statistic that can be calculated from the same sample provides any additional information as to the value of the parameter.

Then we use log likelihood for the probability of data:

$$\ln P\left(\{x_n\}_{n=1}^{N}|\mu,\sigma\right) = -N\ln(\sqrt{2\pi}\sigma) - \sum_n (x_n-\mu)^2/\left(2\sigma^2\right) \tag{6.6}$$

For the mean, we take the derivative:

$$\frac{\partial \ln P}{\partial \mu} = -\frac{N(\mu-\bar{x})}{\sigma^2} = 0 \tag{6.7}$$

For the standard deviation, we take the derivative:

$$\frac{\partial \ln P}{\partial \ln \sigma} = -N + \frac{S}{\sigma^2} = 0 \tag{6.8}$$

Finally, combining 6.7 and 6.8, we can get the estimated value of the mean and standard deviation:

$$\begin{cases} \mu = \bar{x} \\ \sigma^2 = S/N \end{cases} \tag{6.9}$$

Example 6.2: MLE for the Mixture Gaussian.

Assume we have data $\{x_n\}_{n=1}^{N}$. The likelihood is:

$$P(x|\mu_1,\mu_2,\sigma) = \left[\sum_{k=1}^{2} p_k \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu_k)^2}{2\sigma^2}\right)\right] \tag{6.10}$$

where, the parameters $\theta = \left\{\{\mu_k\}_{k=1}^{2},\sigma\right\}$.
Let L denote the natural log likelihood. We can compute the derivative of the log likelihood with respect to μ_k is given by

$$\frac{\partial}{\partial \mu_k} L = \sum_n p_{k|n}\frac{(x_n-\mu_k)}{\sigma^2} \tag{6.11}$$

where $p_{k|n} \equiv P(k_n = k|x_n,\boldsymbol{\theta})$.
Neglecting terms in $\frac{\partial}{\partial \mu_k}P(k_n = k|x_n,\boldsymbol{\theta})$, the second derivative is approximately given by

$$\frac{\partial^2}{\partial \mu_k^2} L = -\sum_n p_{k|n}\frac{1}{\sigma^2} \tag{6.12}$$

Hence, from an initial state μ_1,μ_2, an approximate Newton-Raphson step updates these parameters to μ_1',μ_2', where

$$\mu_k' = \frac{\sum_n p_{k|n}x_n}{\sum_n p_{k|n}} \tag{6.13}$$

(The Newton-Raphson method for maximizing $L(\mu)$ updates μ to $\mu' = \mu-\ [\frac{\partial L}{\partial \mu}/\frac{\partial^2}{\partial \mu^2}]$).

Notice that the algorithm we have derived for maximizing the likelihood is identical to the soft K-means algorithm.

6.2 Soft K-means

6.2.1 Soft K-means Algorithm

K-means: corresponding to a modelling assumption that each cluster is a spherical Gaussian. We can divide this algorithm to 2 steps: assignment step and update step, like the following.

Assignment step. The responsibilities are

$$r_k^{(n)} = \frac{\pi_k \frac{1}{\left(\sqrt{2\pi}\sigma_k\right)^I} \exp\left(-\frac{1}{\sigma_k^2} d\left(\mathbf{m}^{(k)}, \mathbf{x}^{(n)}\right)\right)}{\sum_{k'} \pi_k \frac{1}{\left(\sqrt{2\pi}\sigma_{k'}\right)^t} \exp\left(-\frac{1}{\sigma_{k'}^2} d\left(\mathbf{m}^{(k')}, \mathbf{x}^{(n)}\right)\right)} \tag{6.14}$$

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

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

$$\mathbf{m}^{(k)} = \frac{\sum_n r_k^{(n)} \mathbf{x}^{(n)}}{R^{(k)}} \tag{6.15}$$

$$\sigma_k^2 = \frac{\sum_n r_k^{(n)} \left(\mathbf{x}^{(n)} - \mathbf{m}^{(k)}\right)^2}{I R^{(k)}} \tag{6.16}$$

$$\pi_k = \frac{R^{(k)}}{\sum_k R^{(k)}} \tag{6.17}$$

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

$$R^{(k)} = \sum_n r_k^{(n)} \tag{6.18}$$

The algorithm updates the length scales σ_k for itself. The algorithm also includes cluster weight parameters, $\pi_1, \pi_2, ..., \pi_K$, which also update themselves, allowing accurate modelling of data from clusters of unequal weights.

It has the flaws:
Variance is the same in all directions.It is no good at modelling the cigar-shaped clusters.
Can take a long time to converge.

We can see these flaws in Figure 6.3 and Figure 6.4.

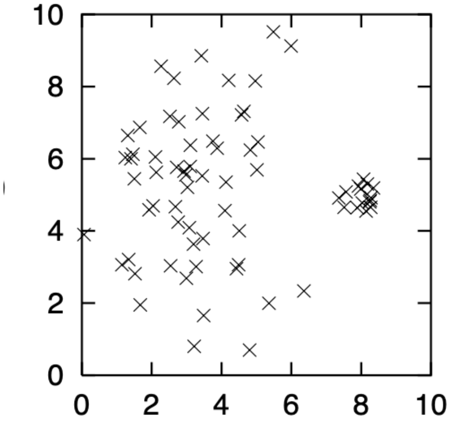

Figure 6.1: The “little ’n’ large” data.

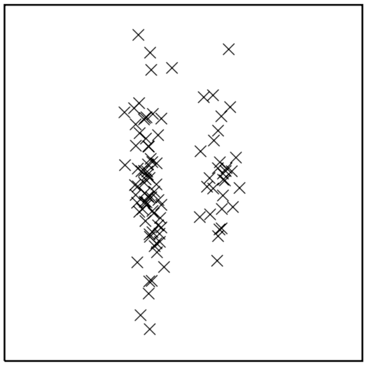

Figure 6.2: Two elongated clusters

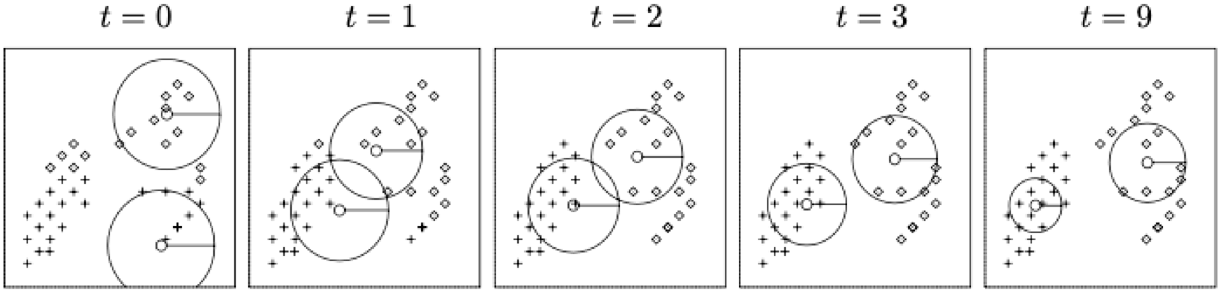

Figure 6.3: Soft K-means algorithm, with K = 2, applied to the 40-point data set

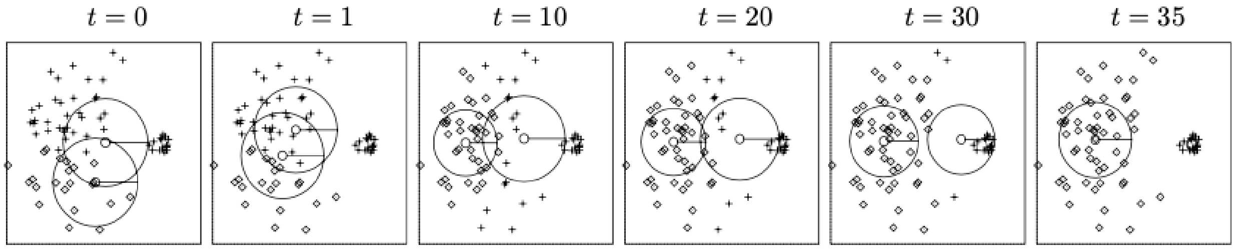

Figure 6.4: Soft K-means algorithm, with K = 2, applied to little 'n' large clusters data set

If we wish to model the clusters by axis-aligned Gaussians with possibly-unequal variances, we replace the assignment rule (6.14) and the variance update rule (6.16) by the rules 6.19 and 6.20 .

$$r_k^{(n)} = \frac{\pi_k \frac{1}{\prod_{i=1}^{I} \sqrt{2\pi}\sigma_i^{(k)}} \exp\left(-\sum_{i=1}^{I}\left(m_i^{(k)} - x_i^{(n)}\right)^2 / 2\left(\sigma_i^{(k)}\right)^2\right)}{\sum_{k'} (\text{ numerator, with } k' \text{ in place of } k)} \tag{6.19}$$

$$\sigma_i^{2(k)} = \frac{\sum_n r_k^{(n)} \left(x_i^{(n)} - m_i^{(k)}\right)^2}{R^{(k)}} \tag{6.20}$$

After redo the enhanced soft-kmeans on the 'two cigars' dataset. We can see in Figure 6.5

After 30 iterations, the algorithm correctly locates the two clusters. Figure 6.6 shows the same algorithm applied to the little 'n' large data set; again, the correct cluster locations are found.

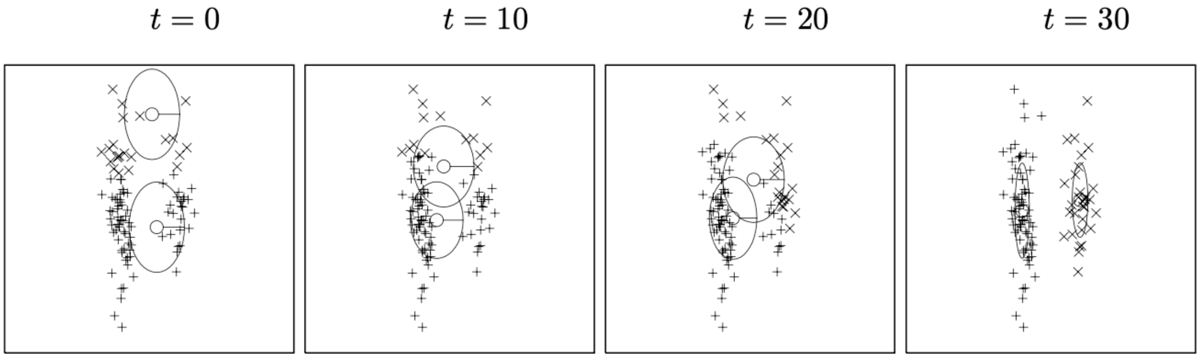

Figure 6.5: Enhanced Soft K-means algorithm, applied to the data consisting of two cigar-shaped clusters. K = 2

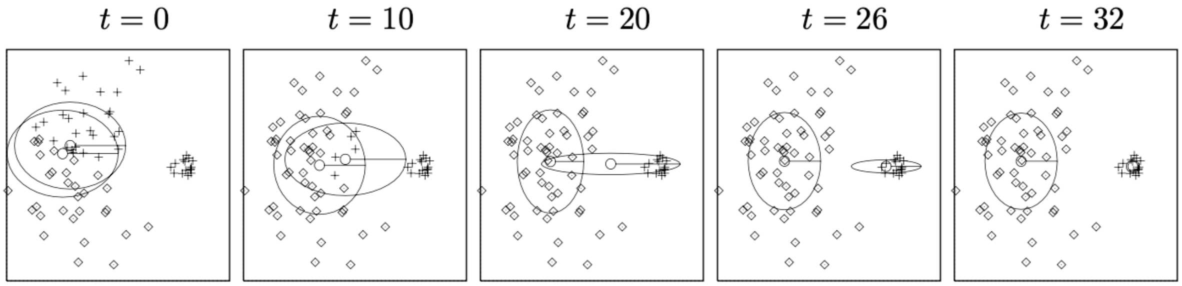

Figure 6.6: Enhanced Soft K-means algorithm, applied to the little 'n' large data set. K = 2

6.2.2 A fatal flaw of MLE

When we fit K = 4 means to our first dataset, we sometimes find that very small clusters form, covering just one or two data points

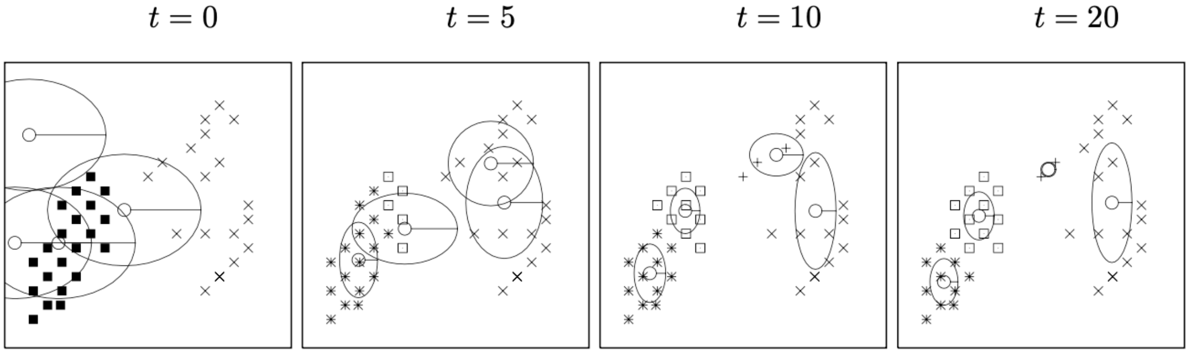

Figure 6.7: Soft K-means algorithm applied to a data set of 40 points. K = 4. Notice that at convergence, one very small cluster has formed between two data points.

Put one cluster exactly on one data point and let its variance go to zero – you can obtain an arbitrarily large likelihood. Maximum likelihood methods can break down by finding highly tuned models that fit part of the data perfectly. This phenomenon is known as overfitting. We conclude that maximum likelihood methods are not a satisfactory general solution to data-modelling problems: the likelihood may be infinitely large at certain parameter settings. Even if the likelihood does not have infinitely-large spikes, the maximum of the likelihood is often unrepresentative, in high-dimensional problems.

Example of Unrepresentative: One Gaussian

In one Gaussian, $\mu = \bar{x} \quad \sigma_N^2 = S/N$. μ is unbiased, because $\mathbb{E}[\mu] = \mu^\star$, but σ_N is biased (σ_{N-1} is unbiased).

Proof:

$$\begin{aligned}
s = \sigma^2 &= \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(x_i-\frac{1}{N}\sum_{i=1}^{N}(x_i)\right)^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}\left[x_i^2-2x_i\frac{1}{N}\sum_{i=1}^{N}(x_i)+\left(\frac{1}{N}\sum_{i=1}^{N}(x_i)\right)^2\right] \\
&= \frac{\sum_{i=1}^{N}x_i^2}{N}-\frac{2\sum_{i=1}^{N}x_i\sum_{i=1}^{N}x_i}{N^2}+\left(\frac{\sum_{i=1}^{N}x_i}{N}\right)^2 \\
&= \frac{\sum_{i=1}^{N}x_i^2}{N}-\left(\frac{\sum_{i=1}^{N}x_i}{N}\right)^2
\end{aligned} \tag{6.21}$$

$$\begin{aligned}
E[s] &= \frac{\sum_{i=1}^{N}E\left[x_i^2\right]}{N}-\frac{E\left[\left(\sum_{i=1}^{N}x_i\right)^2\right]}{N^2} \\
&= s+\mu^2-\frac{E\left[\left(\sum_{i=1}^{N}x_i\right)^2\right]}{N^2} \\
&= s+\mu^2-\frac{E\left[\sum_{i=1}^{N}x_i^2+\sum_{i}^{N}\sum_{j\neq i}^{N}x_ix_j\right]}{N^2} \\
&= s+\mu^2-\frac{E\left[N\left(s+\mu^2\right)+\sum_{i}^{N}\sum_{j\neq i}^{N}E\left[x_i\right]E\left[x_j\right]\right.}{N^2} \\
&= s+\mu^2-\frac{N\left(s+\mu^2\right)+N(N-1)\mu^2}{N^2} \\
&= s+\mu^2-\frac{N\left(s+\mu^2\right)+N^2\mu^2-N\mu^2}{N^2} \\
&= s+\mu^2-\frac{s+\mu^2+N\mu^2-\mu^2}{N} \\
&= s+\mu^2-\frac{s}{N}-\frac{\mu^2}{N}-\mu^2+\frac{\mu^2}{N} \\
&= s-\frac{s}{N} \\
&= s\left(1-\frac{1}{N}\right) \\
&= s\left(\frac{N-1}{N}\right)
\end{aligned} \tag{6.22}$$

6.3 Maximum a Posterior

6.3.1 MAP

A popular replacement for maximizing the likelihood is maximizing the Bayesian posterior probability density of the parameters instead. Given Bayes' rule,

$$P(\boldsymbol{\theta}|\mathcal{D}) = \frac{P(\mathcal{D}|\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(\mathcal{D})} \tag{6.23}$$

where $P(\mathcal{D})$ is the marginal likelihood of data and θ denotes the parameters.
For $P(\mathcal{D})$ is the evidence, we can view it as a constant term, then we have:

$$P(\boldsymbol{\theta}|\mathcal{D}) \propto P(\mathcal{D}|\boldsymbol{\theta})P(\boldsymbol{\theta}) \tag{6.24}$$

We search for parameters that maximize the posterior probability:

$$\tilde{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \log P(\boldsymbol{\theta}|\mathcal{D}) \tag{6.25}$$

More generally, we can view the MAP estimate as a way of regularization using the prior to provide regularization over the likelihood function. And we can also use log posterior to estimate parameters.

$$\begin{aligned} \arg\max_{\boldsymbol{\theta}} \log P(\boldsymbol{\theta}|\mathcal{D}) &= \arg\max_{\boldsymbol{\theta}} \log\left(\frac{P(\boldsymbol{\theta})P(\mathcal{D}|\boldsymbol{\theta})}{P(\mathcal{D})}\right) \\ &= \arg\max_{\boldsymbol{\theta}} (\log P(\boldsymbol{\theta}) + \log P(\mathcal{D}|\boldsymbol{\theta})) \end{aligned} \tag{6.26}$$

That is, $\tilde{\boldsymbol{\theta}}$ is the maximum of a function that sums together the log-likelihood function and $\log P(\boldsymbol{\theta})$.

6.3.2 Conjugate Distributions

If the posterior distributions $P(\boldsymbol{\theta}|\mathcal{D})$ are in the same probability distribution family as the prior probability distribution $P(\boldsymbol{\theta})$, the prior and posterior are then called conjugate distributions, and the prior is called a conjugate prior for the likelihood function. In the (6.22), we say $P(\boldsymbol{\theta}|\mathcal{D})$ and $P(\boldsymbol{\theta})$ are conjugate distribution, since they have the same distribution family.

For example, the Gaussian family is conjugate to itself (or self-conjugate) with respect to a Gaussian likelihood function. If the likelihood function is Gaussian, choosing a Gaussian prior over the mean will ensure that the posterior distribution is also Gaussian. This means that the Gaussian distribution is a conjugate prior for the likelihood that is also Gaussian.

There are some common conjugate distribution. If the likelihood is a multinomial distribution, its conjugate prior distribution is Dirichlet; if the likelihood is a Poison distribution, its conjugate prior distribution is Gamma; if the likelihood is a Bernoulli distribution, its conjugate prior distribution is Beta.

6.3.3 Drawback of MAP: Representation Dependent

The drawback of map is Representation Dependent. Let us look at the Gamma distribution with parameters (s, c).

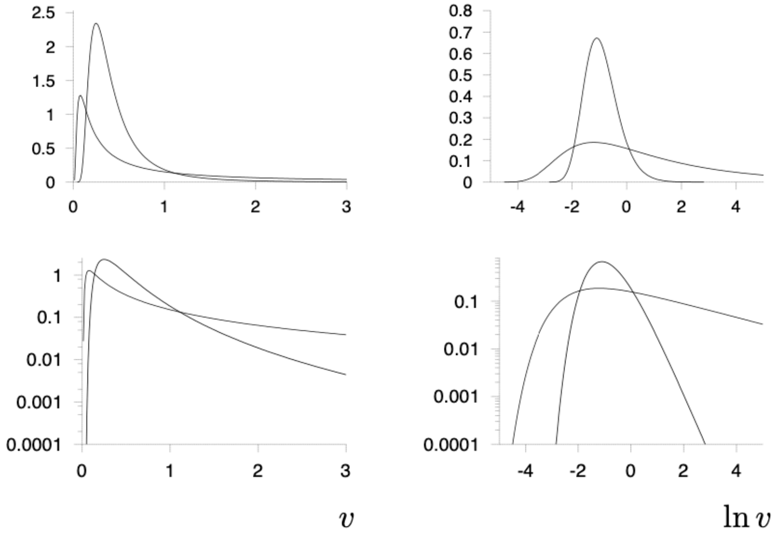

Figure 6.8: Two inverse gamma distributions, with parameters (s, c) = (1, 3) (heavy lines) and 10, 0.3 (light lines), shown on linear vertical scales (top) and logarithmic vertical scales (bottom); and shown as a function of x on the left and l = lnx on the right.

In the left part of Figure 6.8, the gamma distribution is a function of x and right part is a function of $ln(x)$. We use the Equation 6.27 and Equation 6.28 to show denote two different plot. The equation of $P(\ln x)$ can be derived by by $\frac{du^n}{d(log(u))} = du^n$. MAP calculates $\arg\max_{\boldsymbol{\theta}} \log P(\boldsymbol{\theta}|\mathcal{D})$ if we do some non-linear changes to θ the peak of the plot will be different. This is because in some conditions the function of P can decrease but the other function of θ can increase. That's why it is called **Representation Dependent** or **Basis Dependent**.

$$P(x|s,c) = \frac{1}{\Gamma(c)s}\left(\frac{x}{s}\right)^{c-1} \exp\left(-\frac{x}{s}\right) \tag{6.27}$$

$$P(\ln x) = P(x)\left|\frac{\partial x}{\partial \ln x}\right| = \frac{1}{\Gamma(c)}\left(\frac{x}{s}\right)^{c} \exp\left(-\frac{x}{s}\right) \tag{6.28}$$

6.4 Hierarchical Prior

There is another way of enhancing the parameter inference which called Hierarchical Prior. It is build of Bayesian techniques to improve MLE.

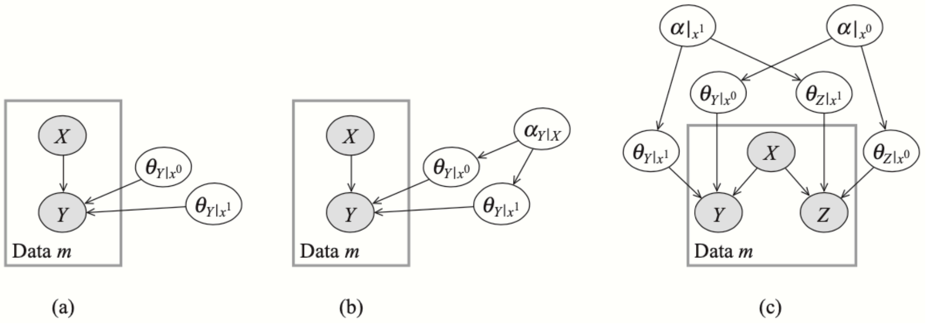

Figure 6.9: Independent and hierarchical priors. (a) A plate model for $P(Y|X)$ under assumption of parameter independence. (b) A plate model for a simple hierarchical prior for the same CPD. (c) A plate model for two CPDs $P(Y|X)$ and $P(Z|X)$ that respond similarly to X.

In Figure 6.9, we can see it is called hierarchical prior because it has multipay layer of prior distributions stack on top of each other. This allows us to perform parameters estimation in a systematic way (deep learning).

It is particular useful for small data like in medical domain. Even if you have very little data in the log-likelihood part, you can encode a strong prior knowledge(doctor knowledge) in the prior distribution using this technique. There is always a trade-off between your prior knowledge and what you see (the log-likelihood).

References

[1] D. Mackay, "Information Theory, Inference, and Learning Algorithms" *Cambridge University press*, 1987, pp. 300–318.

[2] D. Koller and N. Friedman, "Probabilistic Graphical Models: Principles and Techniques" *MIT press*, 2009, pp. 717–781.