(1) Probabilistic Linear Regression(Contd)
(2) Exponential Family Distributions

CS772A: Probabilistic Machine Learning

Piyush Rai

Probabilistic Linear Regression

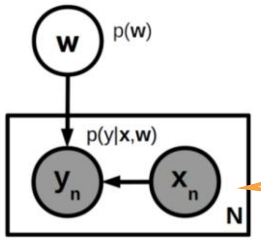


Plate diagram. Hyperparams (λ, β) are fixed and not shown for brevity

- The linear model with Gaussian noise corresponds to a Gaussian likelihood

$$p(y_n|\mathbf{x}_n, \mathbf{w}, \beta) = \mathcal{N}(y_n|\mathbf{w}^\top \mathbf{x}_n, \beta^{-1})$$

NLL corresponds to squared loss prop. to $(y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2$

- Assuming responses to be i.i.d. given features and weights

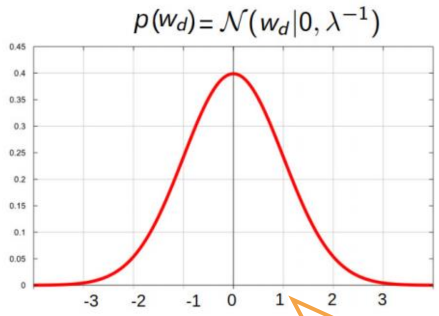
$N \times D$ feature matrix

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^N \mathcal{N}(y_n|\mathbf{w}^\top \mathbf{x}_n, \beta^{-1}) = \mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{w}, \beta^{-1}\mathbf{I}_N)$$

$N \times 1$ response vector

- The above is equivalent to the following

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{\epsilon} \quad \text{where } \boldsymbol{\epsilon} \sim \mathcal{N}(0, \beta^{-1}\mathbf{I}_\mathbf{N})$$



The precision λ of the Gaussian prior controls how aggressively the prior pushes the elements towards mean (0)

- Assume the following Gaussian prior on $\mathbf{w}$,

Neg. log-prior corresponds to ℓ_2 regularizer with λ being the reg. constant

$$p(\mathbf{w}) = \prod_{d=1}^D p(w_d) = \prod_{d=1}^D \mathcal{N}(w_d|0, \lambda^{-1}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \lambda^{-1}\mathbf{I}_D) = \left(\frac{\lambda}{2\pi}\right)^{\frac{D}{2}} \exp\left[-\frac{\lambda}{2}\mathbf{w}^\top \mathbf{w}\right]$$

Can even use different λ's for different w_d's. Useful in sparse modeling (later)

- Then $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{\epsilon}$ is simply a linear Gaussian model
    - Can use all the rules of linear Gaussian models to perform inference/predictions ☺

The Posterior

Will only look at fully Bayesian inference. For MLE/MAP, refer to CS771 slides or book

- The posterior over $\boldsymbol{w}$ (for now, assume hyperparams $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ to be known)

$$p(\boldsymbol{w}|\boldsymbol{y},\mathbf{X},\beta,\lambda) = \frac{p(\boldsymbol{w}|\lambda)p(\boldsymbol{y}|\boldsymbol{w},\mathbf{X},\beta)}{p(\boldsymbol{y}|\mathbf{X},\beta,\lambda)} \propto p(\boldsymbol{w}|\lambda)p(\boldsymbol{y}|\boldsymbol{w},\mathbf{X},\beta)$$

Marginal likelihood for this regression model. Note that it is conditioned on $\mathbf{X}$ too which is assumed given and not being modeled

Must be a Gaussian due to conjugacy

$$p(\boldsymbol{w}|\boldsymbol{y},\mathbf{X},\beta,\lambda) \propto \mathcal{N}(\boldsymbol{w}|\mathbf{0},\lambda^{-1}\mathbf{I}_D) \times \mathcal{N}(\boldsymbol{y}|\mathbf{X}\boldsymbol{w},\beta^{-1}\mathbf{I}_N)$$

- Using the "completing the squares" trick (or linear Gaussian model results)

Note that $\boldsymbol{\lambda}$ and $\boldsymbol{\beta}$ can be learned under the probabilistic set-up (though assumed fixed as of now)

$$p(\boldsymbol{w}|\boldsymbol{y},\mathbf{X},\beta,\lambda) = \mathcal{N}(\boldsymbol{\mu}_N,\boldsymbol{\Sigma}_N)$$

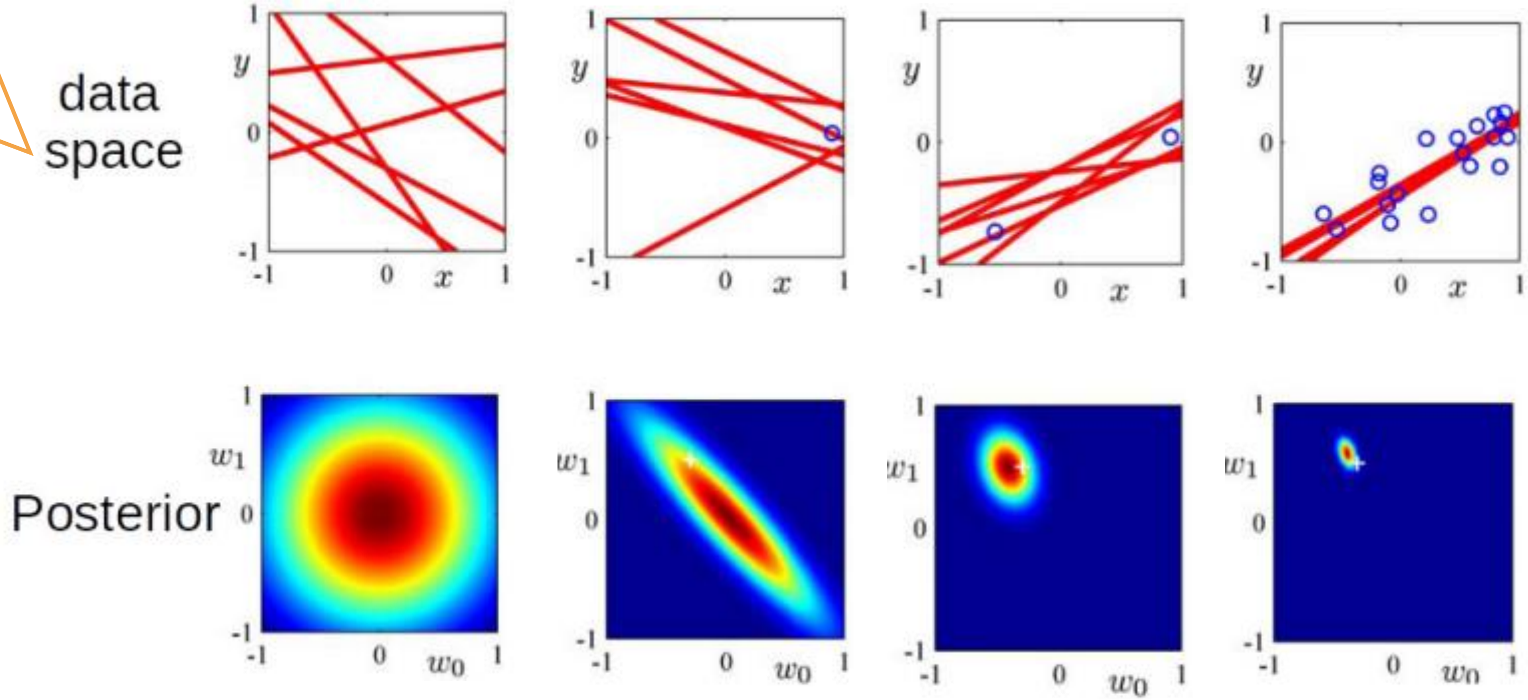
where $\boldsymbol{\Sigma}_N = (\beta\sum_{n=1}^{N}\boldsymbol{x}_n\boldsymbol{x}_n^\top + \lambda\mathbf{I}_D)^{-1} = (\beta\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_D)^{-1}$ (posterior's covariance matrix)

$$\boldsymbol{\mu}_N = \boldsymbol{\Sigma}_N\left[\beta\sum_{n=1}^{N}y_n\boldsymbol{x}_n\right] = \boldsymbol{\Sigma}_N\left[\beta\mathbf{X}^\top\boldsymbol{y}\right] = (\mathbf{X}^\top\mathbf{X} + \frac{\lambda}{\beta}\mathbf{I}_D)^{-1}\mathbf{X}^\top\boldsymbol{y} \quad \text{(posterior's mean)}$$

The Posterior: A Visualization

- Assume a lin. reg. problem with true $\boldsymbol{w} = [w_0, w_1], w_0 = -0.3, w_1 = 0.5$
- Assume data generated by a linear regression model $y = w_0 + w_1 x +$ "noise"
  - Note: It's actually 1-D regression (w_0 is just a bias term), or 2-D reg. with feature $[1, x]$
- Figures below show the "data space" and posterior of $\boldsymbol{w}$ for different number of observations (note: with no observations, the posterior = prior)

Each red line represents the "data" generated for a randomly drawn $\boldsymbol{w}$ from the current posterior



Posterior Predictive Distribution

- To get the prediction y_* for a new input x_*, we can compute its PPD

$$p(y_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}, \beta, \lambda) = \int p(y_*|\mathbf{x}_*, \mathbf{w}, \beta)p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \beta, \lambda)d\mathbf{w}$$

$\mathcal{N}(y_*|\boldsymbol{w}^\top \boldsymbol{x}_*, \beta^{-1})$

$\mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N)$

Only $\boldsymbol{w}$ is unknown with a posterior distribution so only $\boldsymbol{w}$ has to be integrated out

- The above is the marginalization of $\boldsymbol{w}$ from $\mathcal{N}(y_*|\boldsymbol{w}^\top \boldsymbol{x}_*, \beta^{-1})$. Using Gaussian results

$$p(y_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}, \beta, \lambda) = \mathcal{N}(\boldsymbol{\mu}_N^\top \mathbf{x}_*, \beta^{-1} + \mathbf{x}_*^\top \boldsymbol{\Sigma}_N \mathbf{x}_*)$$

Can also derive it by writing $y_* = \boldsymbol{w}^\top \boldsymbol{x}_* + \epsilon$ where $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N)$ and $\epsilon \sim \mathcal{N}(0, \beta^{-1})$

- So we have a predictive mean $\boldsymbol{\mu}_N^\top \boldsymbol{x}_*$ as well as an input-specific predictive variance
- In contrast, MLE and MAP make "plug-in" predictions (using the point estimate of $\boldsymbol{w}$)

$$p(y_*|\mathbf{x}_*, \mathbf{w}_{MLE}) = \mathcal{N}(\mathbf{w}_{MLE}^\top \mathbf{x}_*, \beta^{-1}) \quad \text{- MLE prediction}$$

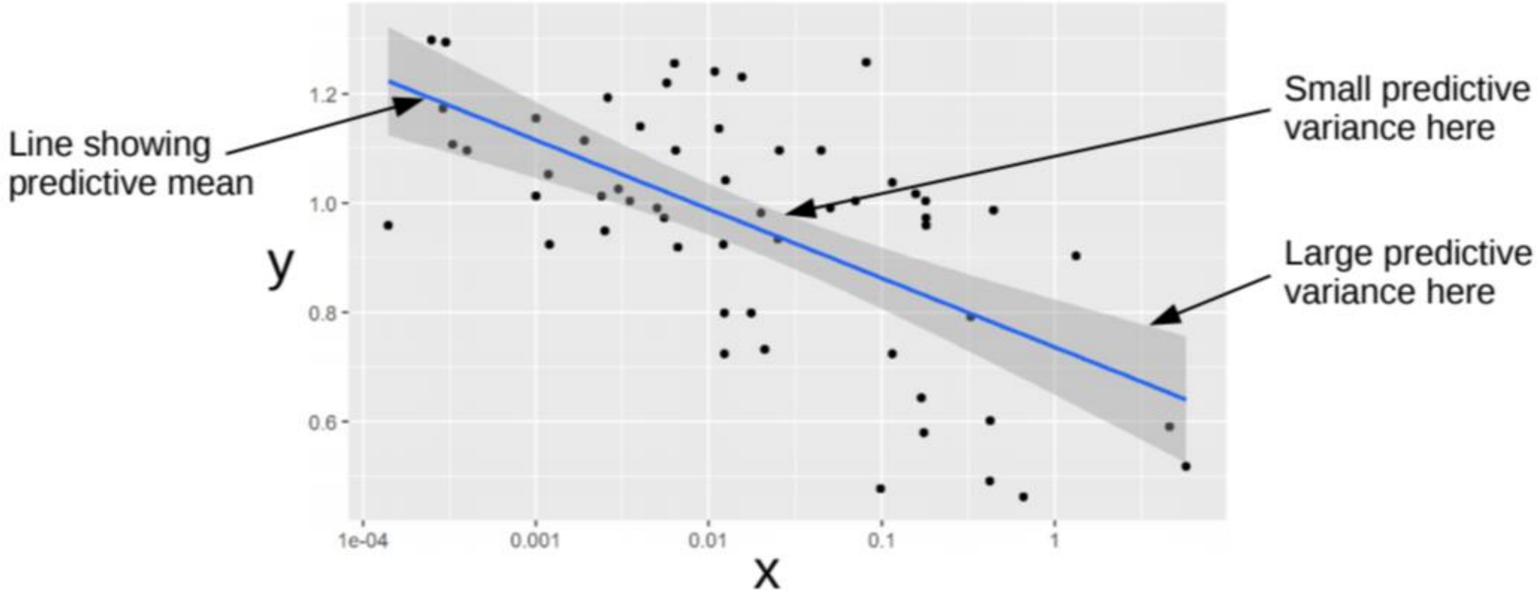
$$p(y_*|\mathbf{x}_*, \mathbf{w}_{MAP}) = \mathcal{N}(\mathbf{w}_{MAP}^\top \mathbf{x}_*, \beta^{-1}) \quad \text{- MAP prediction}$$

Since PPD also takes into account the uncertainty in $\boldsymbol{w}$, the predictive variance is larger

- Unlike MLE/MAP, variance of y_* also depends on the input $\boldsymbol{x}_*$ (this, as we will see later, will be very useful in sequential decision-making problems such as active learning)

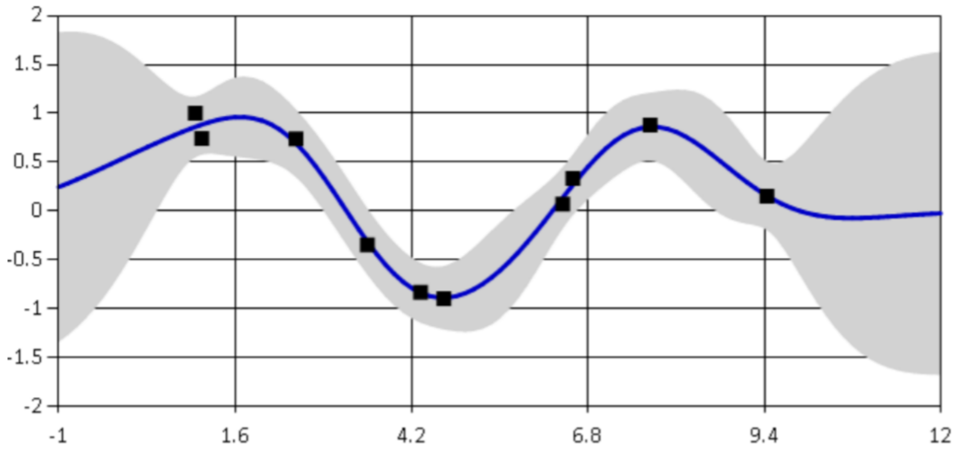
Posterior Predictive Distribution: An Illustration

- Black dots are training examples



- Width of the shaded region at any x denotes the predictive uncertainty at that x (+/- one std-dev)
- Regions with more training examples have smaller predictive variance

Nonlinear Regression



- Can extend the linear regression model to handle nonlinear regression problems
- One way is to replace the feature vectors $\boldsymbol{x}$ by a nonlinear mapping $\boldsymbol{\phi}(\boldsymbol{x})$

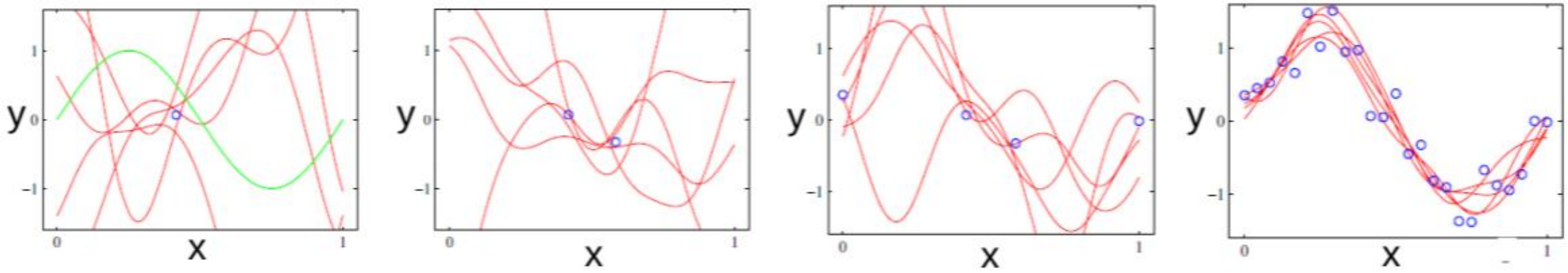
$$p(y|\mathbf{x}, \mathbf{w}) = \mathcal{N}(\mathbf{w}^\top \phi(\mathbf{x}), \beta^{-1})$$

Can be pre-defined or extracted by a pretrained deep neural net

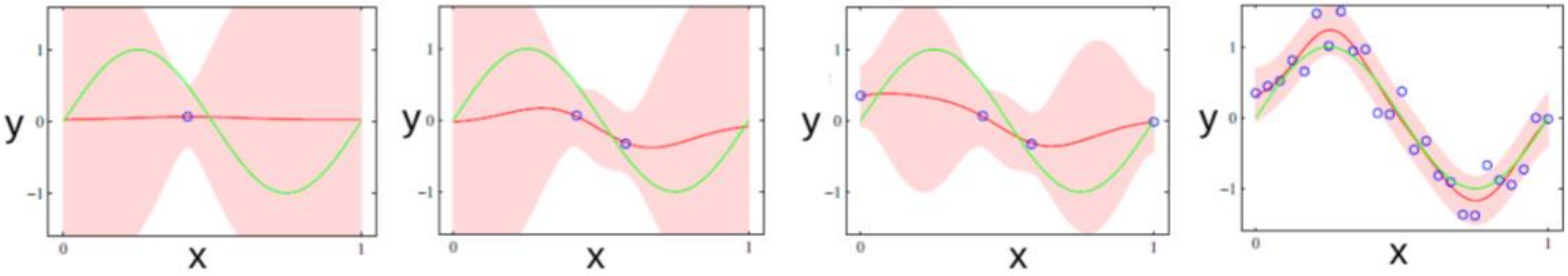
- Alternatively, a kernel function can be used to implicitly define the nonlinear mapping
- More on nonlinear regression when we discuss Gaussian Processes

More on Visualization of Uncertainty

- Figures below: Green curve is the true function and blue circles are observations



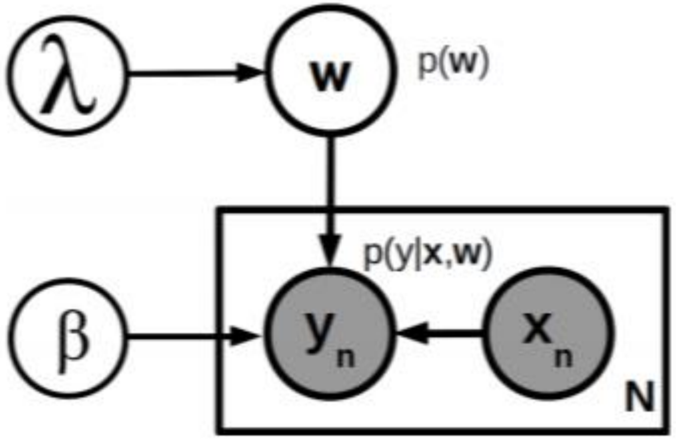
- Posterior of the nonlinear regression model: Some curves drawn from the posterior



- PPD: Red curve is predictive mean, shaded region denotes predictive uncertainty

Hyperparameters

- The probabilistic linear reg. model we saw had two hyperparams $(\boldsymbol{\beta}, \boldsymbol{\lambda})$
  - Thus total three unknowns $(\boldsymbol{w}, \boldsymbol{\beta}, \boldsymbol{\lambda})$



Need posterior over all the 3 unknowns

$$p(\mathbf{w}, \beta, \lambda | \mathbf{X}, \mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \beta, \lambda) p(\mathbf{w}, \lambda, \beta)}{p(\mathbf{y}|\mathbf{X})}$$

$$= \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \beta, \lambda) p(\mathbf{w}|\lambda) p(\beta) p(\lambda)}{\int p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \beta) p(\mathbf{w}|\lambda) p(\beta) p(\lambda) \, d\mathbf{w} \, d\lambda d\beta}$$

PPD would require integrating out all 3 unknowns

$$p(y_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}) = \int p(y_*|\mathbf{x}_*, \mathbf{w}, \beta) p(\mathbf{w}, \beta, \lambda | \mathbf{X}, \mathbf{y}) \, d\mathbf{w} \, d\beta \, d\lambda$$

- Posterior and PPD computation is intractable. Several ways to address this
  - MLE-II for $(\boldsymbol{\beta}, \boldsymbol{\lambda})$: $\hat{\beta}, \hat{\lambda} = \arg\max_{\beta,\lambda} p(\mathbf{y}|\mathbf{X}, \beta, \lambda)$. Use them to infer the posterior of $\boldsymbol{\theta}$ and PPD
  - Use alternating estimation like EM (e.g., E step computes $\boldsymbol{w}$, M step computes $(\boldsymbol{\beta}, \boldsymbol{\lambda})$)
  - Use MCMC or Variational Inference to approximate the above posterior and PPD

MLE-II

For any model where hyperparams are estimated by MLE-II, the posterior and PPD is approximated in a similar fashion

- For the probabilistic linear regression model, the overall posterior over unknowns

$$p(\mathbf{w}, \beta, \lambda|\mathbf{X}, \mathbf{y}) = p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \beta, \lambda)p(\beta, \lambda|\mathbf{X}, \mathbf{y})$$

- With MLE-II approx of $(\boldsymbol{\beta}, \boldsymbol{\lambda})$, $p(\beta, \lambda|\mathbf{X}, \mathbf{y}) \approx \delta(\hat{\beta}, \hat{\lambda})$, a point mass at $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\lambda}}$

$$p(\mathbf{w}, \beta, \lambda|\mathbf{X}, \mathbf{y}) \approx p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \hat{\beta}, \hat{\lambda})$$

Same as the posterior of $\boldsymbol{w}$ with the hyperparameters fixed

- Likewise, the PPD will be approximated as follows

$$\begin{aligned} p(y_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}) &= \int p(y_*|\mathbf{x}_*, \mathbf{w}, \beta)p(\mathbf{w}, \beta, \lambda|\mathbf{X}, \mathbf{y})\, d\mathbf{w}\, d\beta\, d\lambda \\ &= \int p(y_*|\mathbf{x}_*, \mathbf{w}, \beta)p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \beta, \lambda)p(\beta, \lambda|\mathbf{X}, \mathbf{y})d\beta\, d\lambda\, d\mathbf{w} \\ &\approx \int p(y_*|\mathbf{x}_*, \mathbf{w}, \beta)p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \hat{\beta}, \hat{\lambda})\, d\mathbf{w} \end{aligned}$$

Same form for the PPD as in the case of fixed hyperparams

Only need to integrate over $\boldsymbol{w}$, since other two are fixed at their MLE-II solutions

Exp. Family (Pitman, Darmois, Koopman, 1930s)

- Defines a class of distributions. An Exponential Family distribution is of the form

$$p(\mathbf{x}|\theta) = \frac{1}{Z(\theta)} h(\mathbf{x}) \exp[\theta^\top \phi(\mathbf{x})] \quad = \quad h(\mathbf{x}) \exp[\theta^\top \phi(\mathbf{x}) - A(\theta)]$$

- $\boldsymbol{x} \in \mathcal{X}^m$ is the r.v. being modeled ($\mathcal{X}$ denotes some space, e.g., $\mathbb{R}$ or $\{0,1\}$)
- $\theta \in \mathbb{R}^d$: Natural parameters or canonical parameters defining the distribution
- $\phi(\boldsymbol{x}) \in \mathbb{R}^d$: Sufficient statistics (another random variable)
  - Why "sufficient": $p(x|\theta)$ as a function of $\boldsymbol{\theta}$ depends on $\boldsymbol{x}$ only via $\boldsymbol{\phi}(\boldsymbol{x})$
- $Z(\theta) = \int h(\boldsymbol{x}) \exp[\theta^\top \boldsymbol{\phi}(\boldsymbol{x})] d\boldsymbol{x}$: Partition Function
- $A(\theta) = \log Z(\theta)$: Log-partition function (also called cumulant function)
- $h(\boldsymbol{x})$: A constant (doesn't depend on $\boldsymbol{\theta}$)

Expressing a Distribution in Exp. Family Form

- Recall the form of exp-fam distribution $p(x|\theta) = h(x)\exp[\theta^\top \phi(x) - A(\theta)]$
- To write any exp-fam dist $p()$ in the above form, write it as $\exp(\log p())$

$$
\begin{aligned}
\exp\left(\log \text{Binomial}(x|N,\mu)\right) &= \exp\left(\log \binom{N}{x}\mu^x(1-\mu)^{N-x}\right) \\
&= \exp\left(\log \binom{N}{x} + x\log\mu + (N-x)\log(1-\mu)\right) \\
&= \binom{N}{x}\exp\left(x\log\frac{\mu}{1-\mu} - N\log(1-\mu)\right)
\end{aligned}
$$

- Now compare the resulting expression with the exponential family form

$$p(x|\theta) = h(x)\exp[\theta^\top \phi(x) - A(\theta)]$$

.. to identify the natural parameters, sufficient statistics, log-partition function, etc.

(Univariate) Gaussian as Exponential Family

- Let's try to write a univariate Gaussian in the exponential family form

$$p(\mathbf{x}|\theta) = h(\mathbf{x}) \exp[\theta^\top \phi(\mathbf{x}) - A(\theta)]$$

- Recall the PDF of a univar Gaussian (already has exp, so less work needed :))

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] = \frac{1}{\sqrt{2\pi}} \exp\left[\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \frac{\mu^2}{2\sigma^2} - \log\sigma\right]$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left[\begin{bmatrix}\frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2}\end{bmatrix}^\top \begin{bmatrix} x \\ x^2 \end{bmatrix} - \left(\frac{\mu^2}{2\sigma^2} + \log\sigma\right)\right]$$

$$\theta = \begin{bmatrix}\frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2}\end{bmatrix} = \begin{bmatrix}\theta_1 \\ \theta_2\end{bmatrix} \qquad \phi(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix} \quad \text{, and} \begin{bmatrix}\mu \\ \sigma^2\end{bmatrix} = \begin{bmatrix}-\frac{\theta_1}{2\theta_2} \\ -\frac{1}{2\theta_2}\end{bmatrix}$$

$$h(x) = \frac{1}{\sqrt{2\pi}} \qquad A(\theta) = \frac{\mu^2}{2\sigma^2} + \log\sigma = \frac{-\theta_1^2}{4\theta_2} - \frac{1}{2}\log(-2\theta_2) - \frac{1}{2}\log(2\pi)$$

Other Examples

- Many other distribution belong to the exponential family
  - Bernoulli
  - Beta
  - Gamma
  - Multinoulli/Multinomial
  - Dirichlet
  - Multivariate Gaussian
  - .. and many more (https://en.wikipedia.org/wiki/Exponential_family)
- Note: Not all distributions belong to the exponential family, e.g.,
  - Uniform distribution ($x \sim \text{Unif}(a, b)$)
  - Student-t distribution
  - Mixture distributions (e.g., mixture of Gaussians)

Next class

- Continue and wrap up the discussion on exp. family distributions