Exponential Family Distributions and Conditional Models

Piyush Rai

Topics in Probabilistic Modeling and Inference (CS698X)

Jan 16, 2019
Plan for today

- Exponential family distributions (a very important class of distributions)

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

- Conditional models and parameter estimation for them (our example: Prob. Linear Regression)

\[ p(y_n | w, x_n, \beta) = \mathcal{N}(y_n | w^\top x_n, \beta^{-1}) \]
Exponential Family (Pitman, Darmois, Koopman, Late 1930s)

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

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

- \(x \in \mathcal{X}^m\) is the random variable being modeled (where \(\mathcal{X}\) denotes some space, e.g., \(\mathbb{R}\) or \(\{0,1\}\))
- \(\theta \in \mathbb{R}^d\): **Natural parameters** or **canonical parameters** defining the distribution
- \(\phi(x) \in \mathbb{R}^d\): **Sufficient statistics** (another random variable)
  - **Why “sufficient”**: \(p(x|\theta)\) as a function of \(\theta\) depends on \(x\) only via \(\phi(x)\)
- \(Z(\theta) = \int h(x) \exp[\theta^T \phi(x)] dx\): **Partition function**
- \(A(\theta) = \log Z(\theta)\): **Log-partition function** (also called the **cumulant function**)
- \(h(x)\): A constant (doesn’t depend on \(\theta\))
Expressing a Distribution in Exponential Family Form

- Recall the form of exp-fam distribution: \( h(x) \exp[\theta^T \phi(x) - A(\theta)] \)

- To write any exp-fam dist \( p() \) in the above form, write it as \( \exp(\log p()) \), e.g., for Binomial

\[
\exp(\log \text{Binomial}(x|N, \mu)) = \exp \left( \log \binom{N}{x} \mu^x (1 - \mu)^{N-x} \right) \\
= \exp \left( \log \binom{N}{x} + x \log \mu + (N - x) \log(1 - \mu) \right) \\
= \binom{N}{x} \exp \left( x \log \frac{\mu}{1 - \mu} - N \log(1 - \mu) \right)
\]

- Now compare the resulting expression with the exponential family form

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

.. to identify the natural parameters, sufficient statistics, log-partition function, etc.
Let’s try to write a univariate Gaussian in the exponential family form

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

Recall the standard definition of a univariate Gaussian (already has exp in it, so less work :))

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

- \( h(x) = \frac{1}{\sqrt{2\pi}} \)
- \( \theta = \begin{bmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{bmatrix} = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} \), and \( \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix} = \begin{bmatrix} -\frac{\theta_1}{2\theta_2} \\ \frac{1}{2\theta_2} \end{bmatrix} \)
- \( \phi(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix} \)
- \( A(\theta) = \frac{\mu^2}{2\sigma^2} + \log \sigma = -\frac{\theta_1^2}{4\theta_2} - \frac{1}{2} \log(-2\theta_2) - \frac{1}{2} \log(2\pi) \)
Other Examples

- Many other distributions 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)
Log-Partition Function

- \( A(\theta) = \log Z(\theta) = \log \int h(x) \exp[\theta^\top \phi(x)] dx \) is the log-partition function
- \( A(\theta) \) is also called the cumulant function
- Derivatives of \( A(\theta) \) can be used to generate the cumulants of the sufficient statistics \( \phi(x) \)
- Exercise: Assume \( \theta \) to be a scalar (thus \( \phi(x) \) is also scalar). Show that the first and the second derivatives of \( A(\theta) \) are
  \[
  \frac{dA}{d\theta} = \mathbb{E}_{p(x|\theta)}[\phi(x)] \\
  \frac{d^2A}{d\theta^2} = \mathbb{E}_{p(x|\theta)}[\phi^2(x)] - \left( \mathbb{E}_{p(x|\theta)}[\phi(x)] \right)^2 = \text{var}[\phi(x)]
  \]
- Note: The above result also holds when \( \theta \) and \( \phi(x) \) are vector-valued (the “var” will be “covar”)
- Important: \( A(\theta) \) is a convex function of \( \theta \). Why?
MLE for Exponential Family Distributions

- Suppose we have data $D = \{x_1, \ldots, x_N\}$ drawn i.i.d. from an exponential family distribution

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

- To do MLE, we need the overall likelihood. This is simply a product of the individual likelihoods

$$p(D|\theta) = \prod_{i=1}^{N} p(x_i|\theta) = \left[ \prod_{i=1}^{N} h(x_i) \right] \exp \left[ \theta^T \sum_{i=1}^{N} \phi(x_i) - NA(\theta) \right] = \left[ \prod_{i=1}^{N} h(x_i) \right] \exp \left[ \theta^T \phi(D) - NA(\theta) \right]$$

- To estimate $\theta$ (as we'll see shortly), we only need $\phi(D) = \sum_{i=1}^{N} \phi(x_i)$ and $N$

- Size of $\phi(D) = \sum_{i=1}^{N} \phi(x_i)$ does not grow with $N$ (same as the size of each $\phi(x_i)$)

- Only exponential family distributions have finite-sized sufficient statistics

  - No need to store all the data; can simply store and recursively update the sufficient statistics with more and more data

  - Very useful when doing probabilistic/Bayesian inference with large-scale data sets. Also useful in online parameter estimation problems.
MLE and Moment Matching

- The likelihood is of the form
  \[ p(D|\theta) = \left[ \prod_{i=1}^{N} h(x_i) \right] \exp \left[ \theta^\top \phi(D) - NA(\theta) \right] \]

- The log-likelihood is (ignoring constant w.r.t. \( \theta \)):
  \[ \log p(D|\theta) = \theta^\top \phi(D) - NA(\theta) \]

- Note: This is concave in \( \theta \) (since \(-A(\theta)\) is concave). Maximization will yield a global maxima of \( \theta \)

- MLE for exp-fam distributions can also be seen as doing moment-matching. To see this, note that
  \[ \nabla_\theta \left[ \theta^\top \phi(D) - NA(\theta) \right] = \phi(D) - N\nabla_\theta [A(\theta)] = \phi(D) - NE_{p(x|\theta)}[\phi(x)] = \sum_{i=1}^{N} \phi(x_i) - NE_{p(x|\theta)}[\phi(x)] \]

- Therefore, at the “optimal” (i.e., MLE) \( \hat{\theta} \), where the derivative is 0, the following must hold
  \[ \mathbb{E}_{p(x|\theta)}[\phi(x)] = \frac{1}{N} \sum_{i=1}^{N} \phi(x_i) \]

- This is basically matching the expected moments of the distribution with empirical moments ("empirical" here means what we compute using the observed data)
Moment Matching: An Example

- Given $N$ observations $x_1, \ldots, x_N$ from a univariate Gaussian $N(x|\mu, \sigma^2)$, doing moment-matching

\[ \mathbb{E}[\phi(x)] = \frac{1}{N} \sum_{i=1}^{N} \phi(x_i) \]

- The “true”, i.e., expected moments: $\mathbb{E}[\phi(x)] = \mathbb{E}\left[\frac{x}{x^2}\right]$. Therefore

\[ \mathbb{E}\left[\frac{x}{x^2}\right] = \left[ \frac{1}{N} \sum_{i=1}^{N} x_i \right] \]

- For a univariate Gaussian, note that $\mathbb{E}[x] = \mu$ and $\mathbb{E}[x^2] = \text{var}[x] + \mathbb{E}[x]^2 = \sigma^2 + \mu^2$

- Thus we have two equations and two unknowns

- From the first equation, we immediately get $\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$

- From the second equation, we get $\sigma^2 = \mathbb{E}[x^2] - \mu^2 = \frac{1}{N} \sum_{i=1}^{N} x_i^2 - \mu^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$
Bayesian Inference for Exponential Family Distributions

- We saw that the total likelihood given $N$ i.i.d. observations $D\{x_1, \ldots, x_N\}$

$$p(D|\theta) \propto \exp \left[ \theta^\top \phi(D) - NA(\theta) \right]$$

where $\phi(D) = \sum_{i=1}^{N} \phi(x_i)$

- Let’s choose the following prior (note: it looks similar in terms of $\theta$ within the exponent)

$$p(\theta|\nu_0, \tau_0) = h(\theta) \exp \left[ \theta^\top \tau_0 - \nu_0 A(\theta) - A_c(\nu_0, \tau_0) \right]$$

- Ignoring the prior’s log-partition function $A_c(\nu_0, \tau_0) = \log \int_\theta h(\theta) \exp \left[ \theta^\top \tau_0 - \nu_0 A(\theta) \right] d\theta$

$$p(\theta|\nu_0, \tau_0) \propto h(\theta) \exp \left[ \theta^\top \tau_0 - \nu_0 A(\theta) \right]$$

- Comparing the prior’s form with the likelihood, we notice that

  - $\nu_0$ is like the number of “pseudo-observations” coming from the prior
  - $\tau_0$ is the total sufficient statistics of these $\nu_0$ pseudo-observations
The Posterior Distribution

- As we saw, the likelihood is
  \[ p(D|\theta) \propto \exp \left[ \theta^\top \phi(D) - NA(\theta) \right] \text{ where } \phi(D) = \sum_{i=1}^{N} \phi(x_i) \]

- And the prior we chose is
  \[ p(\theta|\nu_0, \tau_0) \propto h(\theta) \exp \left[ \theta^\top \tau_0 - \nu_0 A(\theta) \right] \]

- For this form of the prior, the posterior \( p(\theta|D) \propto p(\theta)p(D|\theta) \) will be
  \[ p(\theta|D) \propto h(\theta) \exp \left[ \theta^\top (\tau_0 + \phi(D)) - (\nu_0 + N)A(\theta) \right] \]

- Note that the posterior has the same form as the prior; such a prior is called a **conjugate prior** (note: all exponential family distributions have a conjugate prior having a form shown as above)

- Thus posterior hyperparams \( \nu'_0, \tau'_0 \) are obtained by simply adding "stuff" to prior’s hyperparams
  \[ \nu'_0 \leftarrow \nu_0 + N \quad \text{(no. of pseudo-obs + no. of actual obs)} \]
  \[ \tau'_0 \leftarrow \tau_0 + \phi(D) \quad \text{(total suff-stats from pseudo-obs + total suff-stats from actual obs)} \]

- Note: Prior’s log-partition function \( A_c(\nu_0, \tau_0) \) updates to posterior’s: \( A_c(\nu_0 + N, \tau_0 + \phi(D)) \)
The Posterior Distribution

Assuming the prior \( p(\theta|\nu_0, \tau_0) \propto h(\theta) \exp \left[ \theta^T \tau_0 - \nu_0 A(\theta) \right] \), the posterior was

\[
p(\theta|D) \propto h(\theta) \exp \left[ \theta^T (\tau_0 + \phi(D)) - (\nu_0 + N)A(\theta) \right]
\]

Assuming \( \tau_0 = \nu_0 \bar{\tau}_0 \), we can also write the prior as \( p(\theta|\nu_0, \bar{\tau}_0) \propto \exp \left[ \theta^T \nu_0 \bar{\tau}_0 - \nu_0 A(\theta) \right] \)

Can think of \( \bar{\tau}_0 = \tau_0 / \nu_0 \) as the average sufficient statistics per pseudo-observation

The posterior can be written as

\[
p(\theta|D) \propto h(\theta) \exp \left[ \theta^T (\nu_0 + N) \frac{\nu_0 \bar{\tau}_0 + \phi(D)}{\nu_0 + N} - (\nu_0 + N)A(\theta) \right]
\]

Denoting \( \bar{\phi} = \frac{\phi(D)}{N} \) as the average suff-stats per real observation, the posterior updates are

\[
\nu_0' \leftarrow \nu_0 + N
\]

\[
\bar{\tau}_0' \leftarrow \frac{\nu_0 \bar{\tau}_0 + N\bar{\phi}}{\nu_0 + N}
\]

Note that the posterior hyperparam \( \bar{\tau}_0' \) is a convex combination of the average suff-stats \( \bar{\tau}_0 \) of the \( \nu_0 \) pseudo-observations and the average suff-stats \( \bar{\phi} \) of the \( N \) actual observations
Posterior Predictive Distribution

- Assume some past (training) data \( D = \{x_1, \ldots, x_N\} \) generated from an exp. family distribution
- Assume some test data \( D' = \{\tilde{x}_1, \ldots, \tilde{x}_{N'}\} \) from the same distribution \( (N' \geq 1) \)
- The posterior predictive distribution of \( D' \) (probability distribution of new data given old data)

\[
p(D'|D) = \int p(D'|\theta)p(\theta|D)d\theta
\]

- We've already seen some specific examples of computing the posterior predictive dist., e.g.,
  - Beta-Bernoulli case: Posterior predictive distribution of next coin toss
  - Dirichlet-Multinoulli case: Posterior predictive distribution of next dice roll
  - Gaussian-Gaussian, Gaussian-IG, Gaussian-Gamma, Gaussian-NIG, Gaussian-NG case: Posterior predictive distribution of the next observation

- **Nice Property**: If the likelihood is an exponential family distribution, prior is conjugate (and thus is the posterior), the posterior predictive always has a closed form expression (shown next)
Posterior Predictive Distribution

- Recall the form of the likelihood $p(D|\theta)$ for exp. family dist.

$$p(D|\theta) = \left[ \prod_{i=1}^{N} h(x_i) \right] \exp \left[ \theta^T \phi(D) - NA(\theta) \right]$$

- The conjugate prior was

$$p(\theta|\nu_0, \tau_0) = h(\theta) \exp \left[ \theta^T \tau_0 - \nu_0 A(\theta) - A_c(\nu_0, \tau_0) \right]$$

- For this choice of the conjugate prior, the posterior was shown to be

$$p(\theta|D) = h(\theta) \exp \left[ \theta^T (\tau_0 + \phi(D)) - (\nu_0 + N)A(\theta) - A_c(\nu_0 + N, \tau_0 + \phi(D)) \right]$$

- For the test data $D'$, the likelihood will be

$$p(D'|\theta) = \left[ \prod_{i=1}^{N'} h(\tilde{x}_i) \right] \exp \left[ \theta^T \phi(D') - N'A(\theta) \right] \quad \text{where} \quad \phi(D') = \sum_{i=1}^{N'} \phi(\tilde{x}_i)$$
Therefore the posterior predictive distribution will be

\[
p(D'|D) = \int p(D'|\theta) p(\theta|D) d\theta
\]

\[
= \int \left[ \prod_{i=1}^{N'} h(\tilde{x}_i) \right] \exp \left[ \theta^T \phi(D') - N' A(\theta) \right] h(\theta) \exp \left[ \theta^T (\tau_0 + \phi(D)) - (v_0 + N)A(\theta) - A_c(v_0 + N, \tau_0 + \phi(D)) \right] d\theta
\]

where \( Z_c(v_0 + N + N', \tau_0 + \phi(D) + \phi(D')) = \int h(\theta) \exp \left[ \theta^T (\tau_0 + \phi(D) + \phi(D')) - (v_0 + N + N')A(\theta) \right] d\theta \)
Since $A_c = \log Z_c$ or $Z_c = \exp(A_c)$, we can write the posterior predictive distribution as

$$p(D' | D) = \left[ \prod_{i=1}^{N'} h(\tilde{x}_i) \right] \frac{Z_c(\nu_0 + N + N', \tau_0 + \phi(D) + \phi(D'))}{Z_c(\nu_0 + N, \tau_0 + \phi(D))}$$

$$= \left[ \prod_{i=1}^{N'} h(\tilde{x}_i) \right] \exp \left[ A_c(\nu_0 + N + N', \tau_0 + \phi(D) + \phi(D')) - A_c(\nu_0 + N, \tau_0 + \phi(D)) \right]$$

Therefore the posterior predictive is proportional to ..

.. the ratio of two partition functions of two “posterior distributions” (one with $N + N'$ examples and the other with $N$ examples)

.. or exponential of the difference of the corresponding log-partition functions

Note that the form of $Z_c$ (and $A_c$) will simply depend on the chosen conjugate prior

Very useful result. Also holds for $N = 0$

In the $N = 0$ case, $p(D') = \int p(D'|\theta)p(\theta)d\theta$ is simply the marginal likelihood of $D'$
Summary

- Exp. family distributions are very useful for modeling diverse types of data/parameters
- Conjugate priors to exp. family distributions make parameter updates very simple
- Other quantities such as posterior predictive can be computed in closed form
- Useful in designing generative classification models. Choosing class-conditional from exponential family with conjugate priors helps in parameter estimation
- Useful in designing generative models for unsupervised learning
- Uses in designing Generalized Linear Models (GLM): Model $p(y|x)$ using exp. family distribution
  - Linear regression (with Gaussian likelihood) and logistic regression are GLMs
- We will see several use cases when we discuss approximate inference algorithms (e.g., Gibbs sampling, and especially variational inference)
Estimating Conditional Models, e.g., $p(y|x)$

Our Example: Probabilistic/Bayesian Linear Regression
Estimating Conditional Models

- Conditional models of the form $p(y|x)$ are commonly used in supervised learning problems.
- But more broadly applicable (basically any problem where data $y$ depends on another quantity $x$).
- Conditional models can be estimated using one of the following two ways:
  1. Estimate the joint distribution $p(x, y)$ and then use Bayes rule to get $p(y|x)$
     \[
     p(y|x, \theta) = \frac{p(x, y|\theta)}{p(x|\theta)}
     \]
  2. Estimate the conditional $p(y|x)$ directly (used when we don’t care about modeling $x$), e.g.
     \[
     p(y|x) = \mathcal{N}(y|f_\mu(x), f_\sigma^2(x)) \quad \text{(params of $p(y|x)$ will be functions of $x$)}
     \]
- Approach 1 is called generative approach, approach 2 is called discriminative approach.
- For pros/cons, refer to CS771 lecture slides and readings.
- For now, we will focus on learning (2) using fully Bayesian inference.
- Today’s focus will be on regression problems ($y$ is real-valued response for the input $x$).
Linear Regression: A Probabilistic Setup

- Given: \( N \) training examples \( \{ x_n, y_n \}_{n=1}^{N} \), features: \( x_n \in \mathbb{R}^D \), response \( y_n \in \mathbb{R} \)
- Assume a “noisy” linear model with regression weight vector \( w = [w_1, w_2, \ldots, w_D] \in \mathbb{R}^D \)

\[
y_n = w^\top x_n + \epsilon_n
\]

where \( \epsilon_n \sim \mathcal{N}(0, \beta^{-1}) \), \( \beta \): precision (inverse variance) of Gaussian (assumed known)
- Therefore \( p(y_n|x_n, w, \beta) = \mathcal{N}(y_n|w^\top x_n, \beta^{-1}) \)

Note: Some books (e.g., PRML) use \( \phi(x_n) \) to denote the features where \( \phi \) is some transformation of the original features \( x_n \) (we will only use this notation when talking about nonlinear regression)
The Likelihood Model

- Notation: \( X = [x_1 \ldots x_N]^{\top} \): \( N \times D \) feature matrix, \( y = [y_1 \ldots y_N]^{\top} \): \( N \times 1 \) response vector

- Assuming independent observations, the likelihood model

\[
p(y|w, X, \beta) = \prod_{n=1}^{N} p(y_n|w, x_n, \beta) = \prod_{n=1}^{N} \mathcal{N}(y_n|w^{\top}x_n, \beta^{-1})
\]

\[
= \prod_{n=1}^{N} \sqrt{\frac{\beta}{2\pi}} \exp \left[ -\frac{\beta}{2} (y_n - w^{\top}x_n)^2 \right]
\]

\[
= \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} \exp \left[ -\frac{\beta}{2} \sum_{n=1}^{N} (y_n - w^{\top}x_n)^2 \right]
\]

- Note that NLL = sum of squared errors! Minimizing w.r.t. \( w \) will give MLE/least squares solution!

- For brevity, can also write the likelihood \( p(y|w, X) \) as an \( N \)-dim multivariate Gaussian

\[
p(y|X, w, \beta) = \mathcal{N}(y|Xw, \beta^{-1}I_N) = \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} \exp \left[ -\frac{\beta}{2} (y - Xw)^{\top} (y - Xw) \right]
\]
The Prior

- Assume the entries in $w$ are i.i.d. with zero mean Gaussian priors. Therefore
  \[ p(w) = \prod_{d=1}^{D} p(w_d) = \prod_{d=1}^{D} \mathcal{N}(w_d | 0, \lambda^{-1}) = \mathcal{N}(w | 0, \lambda^{-1} I_D) = \left( \frac{\lambda}{2\pi} \right)^{\frac{D}{2}} \exp \left[ -\frac{\lambda}{2} w^\top w \right] \]

- This prior promotes the entries in $w$ to be small (close to zero)
  - Also, the negative of log-prior is the same as an $\ell_2$ regularizer on $w$

- This prior is conjugate to the likelihood (Gaussian) which makes posterior inference easy
The role of the precision hyperparam $\lambda$ in the prior is important.

- Large values of $\lambda$ would more aggressively encourage $w_d$ to be close to zero.
- Can think of $\lambda$ as the regularization hyperparam for the weights.
- **Important**: Can infer $\lambda$ as well (will see later how to do this).
- Can even have different $\lambda$ for each $w_d$, i.e.,
  \[
p(w \mid \{\lambda_d\}_{d=1}^D) = \prod_{d=1}^D \mathcal{N}(w_d \mid 0, \lambda_d^{-1})
  \]
  Useful in **sparse regression/classification** models in which very few features are relevant which can be identified by inferring $\{\lambda_d\}_{d=1}^D$. Popularly known as **sparse Bayesian learning** (more on this later).
Inference Tasks for Bayesian Linear Regression

(Hyperparameters $\lambda, \beta$ not shown as they are fixed/known)

- Want to infer the posterior distribution over $w$ (for now, assume $\beta$ and $\lambda$ to be known)

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

- Want to infer the posterior predictive distribution

$$p(y_*|x_*, X, y, \beta, \lambda) = \int p(y_*|w, x_*, \beta)p(w|X, y, \beta, \lambda)dw$$

- Likelihood $p(y|w, x, \beta)$ and prior $p(w|\lambda)$ are Gaussians, so above computations are easy!

- Also note that it’s also like a noisy linear Gaussian model: $y = Xw + \epsilon$ with noise $\epsilon = [\epsilon_1, \ldots, \epsilon_N]$

  - $D \times 1$ Gaussian r.v. $w$ transformed via $N \times D$ matrix $X$ to produce $N \times 1$ vector $y$
Bayesian Linear Regression: The Posterior

• The posterior over \( w \) (for now, assume hyperparams \( \beta \) and \( \lambda \) to be known)

\[
p(w|y, X, \beta, \lambda) = \frac{p(w|\lambda)p(y|w, X, \beta)}{p(y|X, \beta, \lambda)} \propto p(w|\lambda)p(y|w, X, \beta)
\]

• Computing \( p(w|X, y, \beta, \lambda) \)

\[
p(w|y, X, \beta, \lambda) \propto \mathcal{N}(w|0, \lambda^{-1}I_D) \times \mathcal{N}(y|Xw, \beta^{-1}I_N)
\]

• Using the “completing the squares” trick (or directly using Gaussian conditioning formula)

\[
p(w|y, X, \beta, \lambda) = \mathcal{N}(\mu_N, \Sigma_N)
\]

where \( \Sigma_N = (\beta \sum_{n=1}^{N} x_n x_n^\top + \lambda I_D)^{-1} = (\beta X^\top X + \lambda I_D)^{-1} \) (posterior’s covariance matrix)

\[
\mu_N = \Sigma_N \left[ \beta \sum_{n=1}^{N} y_n x_n \right] = \Sigma_N \left[ \beta X^\top y \right] = (X^\top X + \frac{\lambda}{\beta} I_D)^{-1} X^\top y \) (posterior’s mean)
Assume a linear regression problem with ground truth $\mathbf{w} = [w_0, w_1]$ with $w_0 = -0.3, w_1 = 0.5$.

Assume data generated by a linear regression model $y = w_0 + w_1 x + \text{"noise"}$

- 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 $\mathbf{w}$ for different number of observations (note: with no observations, the posterior = prior).

The “data space” (red lines) shown above denotes various possible linear regression datasets with data of the form $y = w_0 + w_1 x$ generated using $\mathbf{w}$ drawn from the current posterior of $\mathbf{w}$. 
Bayesian Linear Regression: Posterior Predictive Distribution

- Given the posterior \( p(w|y, X, \beta, \lambda) = \mathcal{N}(\mu_N, \Sigma_N) \), how to make prediction \( y_* \) for a new input \( x_* \)?
- The posterior predictive distribution will be
  \[
  p(y_*|x_*, X, y, \beta, \lambda) = \int p(y_*|x_*, w, \beta)p(w|X, y, \beta, \lambda)dw
  \]
- Using Gaussian predictive/marginal formula, the posterior predictive will be another Gaussian
  \[
  p(y_*|x_*, X, y, \beta, \lambda) = \mathcal{N}(\mu_N^T x_*, \beta^{-1} + x_*^T \Sigma_N x_*)
  \]
- So we get a predictive mean \( \mu_N^T x_* \) and an input-specific predictive variance \( \beta^{-1} + x_*^T \Sigma_N x_* \)
- In contrast, MLE and MAP make “plug-in” predictions (using the point estimate of \( w \))
  \[
  p(y_*|x_*, w_{MLE}) = \mathcal{N}(w_{MLE}^T x_*, \beta^{-1}) \quad \text{- MLE prediction}
  \]
  \[
  p(y_*|x_*, w_{MAP}) = \mathcal{N}(w_{MAP}^T x_*, \beta^{-1}) \quad \text{- MAP prediction}
  \]
- Important: Unlike MLE/MAP, the variance of \( y_* \) also depends on the input \( x_* \) (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 $x$ ($\pm$ 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 $\mathbf{x}$ by a nonlinear mapping $\phi(\mathbf{x})$

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

- The nonlinear mapping can be defined directly, e.g., for a one-dimensional feature $x$

$$\phi(x) = [1, x, x^2]$$

- Alternatively, a kernel function can be used to implicitly define the nonlinear mapping

- More on nonlinear regression when we discuss Gaussian Processes
What about the hyperparameters of the regression model?

- If hyperparameters are to be estimated, we will have a hierarchical/multiparameter model.
- Posterior inference in slightly more involved in this case.
- Iterative methods required to learn the weight vector and the hyperparameters, e.g.,
  - Marginal likelihood maximization for hyperparameter estimation.
  - Expectation maximization (EM).
  - MCMC or variational inference.
- We will discuss more when we talk about inference in hierarchical/multiparameter models.
Summary and What Lies Ahead..

- Seen Bayesian inference for several models with a single unknown parameter (and another simple case where we had two unknown parameters - Gaussian with unknown mean and precision)
- Focused on the cases where the likelihood and prior are conjugate
- Both posterior as well as posterior predictive are computable easily in such cases
- Saw various nice properties of exponential family distributions and parameter estimation for such distributions. Also saw estimation in a conditional model (linear regression)
- Things become more challenging/interesting for more complex models, e.g.,
  - Multiple unknown parameters (e.g., hyperparameters, latent variables, hierarchical models etc)
  - Likelihood and prior are not conjugate
- The basic ideas we have seen will turn out to be useful in more complex models as well
  - Conditionally-conjugate models
  - Approximate inference methods (e.g., EM, Gibbs sampling, etc) that resemble alternating optimization techniques