# (1)Exponential Family Distributions (Contd)
# (2) Probabilistic Models for Classification

CS772A: Probabilistic Machine Learning

Piyush Rai

# 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 $\{0,1\}$)
- $\boldsymbol{\theta} \in \mathbb{R}^d$ : Natural parameters or canonical parameters defining the distribution
- $\boldsymbol{\phi}(\boldsymbol{x}) \in \mathbb{R}^d$ : 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)
- $\boldsymbol{h}(\boldsymbol{x})$: 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 :))

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

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

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

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

# 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)

# Log-Partition Function

- The log-partition function is $A(\theta) = \log Z(\theta) = \log \int h(\boldsymbol{x}) \exp[\theta^\top \phi(\boldsymbol{x})] d\boldsymbol{x}$
- $A(\boldsymbol{\theta})$ is also called the cumulant function
- Derivatives of $A(\boldsymbol{\theta})$ can be used to generate the cumulants of the sufficient statistics
- Exercise: Assume $\boldsymbol{\theta}$ to be a scalar (thus $\boldsymbol{\phi}(x)$ is also scalar). Show that the first and the second derivatives of $A(\boldsymbol{\theta})$ are

$$
\begin{aligned}
\frac{dA}{d\theta} &= \mathbb{E}_{p(\boldsymbol{x}|\theta)}[\phi(\boldsymbol{x})] \\
\frac{d^2 A}{d\theta^2} &= \mathbb{E}_{p(\boldsymbol{x}|\theta)}[\phi^2(\boldsymbol{x})] - \left[\mathbb{E}_{p(\boldsymbol{x}|\theta)}[\phi(\boldsymbol{x})]\right]^2 = \text{var}[\phi(\boldsymbol{x})]
\end{aligned}
$$

- Above result also holds when $\boldsymbol{\theta}$ and $\boldsymbol{\phi}(x)$ are vector-valued (the "var" will be "covar")
- Important: $A(\boldsymbol{\theta})$ is a convex function of $\boldsymbol{\theta}$. Why?

# MLE for Exponential Family Distributions

- Assume data $\mathcal{D} = \{x_1, \ldots, x_N\}$ drawn i.i.d. from an exp. family distribution

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

- To do MLE, we need the overall likelihood -- a product of the individual likelihoods

$$p(\mathcal{D}|\theta) = \prod_{i=1}^{N} p(\mathbf{x}_i|\theta) = \left[\prod_{i=1}^{N} h(\mathbf{x}_i)\right] \exp\left[\theta^\top \sum_{i=1}^{N} \phi(\mathbf{x}_i) - NA(\theta)\right] = \left[\prod_{i=1}^{N} h(\mathbf{x}_i)\right] \exp\left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right]$$

- To estimate $\theta$ (as we'll see shortly), we only need $\phi(\mathcal{D}) = \sum_{i=1}^{N} \phi(\mathbf{x}_i)$ and $N$
- Size of $\phi(\mathcal{D}) = \sum_{i=1}^{N} \phi(x_i)$ 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 update the sufficient statistics as data comes
  - Useful in probabilistic inference with large-scale data sets and "online" parameter estimation

# MLE and Moment Matching

- The likelihood is of the form $p(\mathcal{D}|\theta) = \left[\prod_{i=1}^{N} h(\boldsymbol{x}_i)\right] \exp\left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right]$
- The log-likelihood is (ignoring constant w.r.t. $\boldsymbol{\theta}$): $\log p(\mathcal{D}|\theta) = \theta^\top \phi(\mathcal{D}) - NA(\theta)$
- This is concave in $\boldsymbol{\theta}$ (since $-A(\boldsymbol{\theta})$ is concave)
  - Maximization (MLE solution) will yield a global maxima of $\boldsymbol{\theta}$
- MLE for exp-fam distributions can also be seen as doing moment-matching

$$\nabla_\theta \left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right] = \phi(\mathcal{D}) - N\nabla_\theta[A(\theta)] = \phi(\mathcal{D}) - N\mathbb{E}_{p(\boldsymbol{x}|\theta)}[\phi(\boldsymbol{x})]$$

$$= \sum_{i=1}^{N} \phi(\boldsymbol{x}_i) - N\mathbb{E}_{p(\boldsymbol{x}|\theta)}[\phi(\boldsymbol{x})]$$

- Therefore, at the "optimal" (i.e., MLE) $\hat{\boldsymbol{\theta}}$, we must have

$$\mathbb{E}_{p(\boldsymbol{x}|\theta)}[\phi(\boldsymbol{x})] = \frac{1}{N}\sum_{i=1}^{N} \phi(\boldsymbol{x}_i)$$

Expected moment

Empirical moment (computed using data)

Can thus solve for the MLE $\boldsymbol{\theta}$ also by matching the expected and empirical moments

# Moment Matching: An Example



- Given data $\mathcal{D} = \{x_1, \ldots, x_N\}$ drawn i.i.d. from a univar Gaussian $p(x) = \mathcal{N}(x|\mu, \sigma^2)$

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

Moment matching

- Since the "true", i.e., expected moments: $\mathbb{E}[\phi(x)] = \mathbb{E}\begin{bmatrix} x \\ x^2 \end{bmatrix}$

$$\mathbb{E}\begin{bmatrix} x \\ x^2 \end{bmatrix} = \begin{bmatrix} \frac{1}{N}\sum_{i=1}^{N} x_i \\ \frac{1}{N}\sum_{i=1}^{N} x_i^2 \end{bmatrix}$$

- For a univariate Gaussian, note that

$$\mathbb{E}[x] = \mu$$
$$\mathbb{E}[x^2] = \text{var}[x] + \mathbb{E}[x]^2 = \sigma^2 + \mu^2$$

Two equations, two unknowns ($\mu$ and $\sigma^2$)

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$
$$\sigma^2 = \mathbb{E}[x^2] - \mu^2$$
$$= \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2$$
$$= \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

Same solution that we get by doing MLE

# Bayesian Inference for Expon. Family Distributions

- Already saw that the total likelihood given $N$ i.i.d. observations $\mathcal{D} = \{x_1, \dots, x_N\}$

$$p(\mathcal{D}|\theta) \propto \exp\left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right] \quad \text{where} \quad \phi(\mathcal{D}) = \sum_{i=1}^{N} \phi(\boldsymbol{x}_i)$$

- Let's choose the following prior (note: looks similar in terms of $\boldsymbol{\theta}$ within exp)

$$p(\theta|\nu_0, \boldsymbol{\tau}_0) = h(\theta) \exp\left[\theta^\top \tau_0 - \nu_0 A(\theta) - A_c(\nu_0, \boldsymbol{\tau}_0)\right]$$

- Ignoring the prior's log-partition function $A_c(\nu_0, \boldsymbol{\tau}_0) = \log \int_\theta h(\theta) \exp\left[\theta^\top \tau_0 - \nu_0 A(\theta)\right] d\theta$

$$p(\theta|\nu_0, \boldsymbol{\tau}_0) \propto h(\theta) \exp\left[\theta^\top \tau_0 - \nu_0 A(\theta)\right]$$

- Comparing the prior's form with the likelihood, note that
  - $\nu_0$ is like the number of "pseudo-observations" coming from the prior
  - $\tau_0$ is the total sufficient statistics of the pseudo-observations ($\tau_0 / \nu_0$ per pseudo-obs)

# The Posterior

- The likelihood and prior were

$$p(\mathcal{D}|\theta) \propto \exp\left[\theta^\top \phi(\mathcal{D}) - NA(\theta)\right] \quad \text{where} \quad \phi(\mathcal{D}) = \sum_{i=1}^{N} \phi(\boldsymbol{x}_i)$$

$$p(\theta|\nu_0, \boldsymbol{\tau}_0) \propto h(\theta)\exp\left[\theta^\top \tau_0 - \nu_0 A(\theta)\right]$$

Assume its log partition function denoted as $A_c(\nu_0, \tau_0)$

- The posterior $p(\theta|\mathcal{D}) \propto p(\theta)p(\mathcal{D}|\theta)$ therefore will be

$$p(\theta|\mathcal{D}) \propto h(\theta)\exp\left[\theta^\top(\tau_0 + \phi(\mathcal{D})) - (\nu_0 + N)A(\theta)\right]$$

Its log partition function will be $A_c(\nu_0 + N, \tau_0 + \phi(\mathcal{D}))$

Posterior is also from the same family as the prior

Happens when the prior is conjugate to the likelihood

- Every exp family likelihood has a conjugate prior having the form above
- Posterior's hyperparams $\tau_0', \nu_0'$ obtained by adding "stuff" to prior's hyperparams

$$\nu_0' \leftarrow \nu_0 + N$$

$$\tau_0' \leftarrow \tau_0 + \phi(\mathcal{D})$$

Number of pseudo-observations plus number of actual observations

Suff-stats of pseudo-obervations plus suff-stats of actual observations

**Another equivalent form**

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

$\bar{\tau}_0 = \tau_0/\nu_0$

$$\nu_0' \leftarrow \nu_0 + N$$

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

$\bar{\phi} = \frac{\phi(D)}{N}$

Convex comb of avg suff-stats of pseudo obs and actual obs

# Posterior Predictive Distribution

- Assume some training data $\mathcal{D} = \{x_1, \ldots, x_N\}$ from some exp-fam distribution
- Assume some test data $\mathcal{D}' = \{\tilde{x}_1, \ldots, \tilde{x}_{N'}\}$ from the same distribution
- The posterior pred. distr. of $\mathcal{D}'$

Exp. Fam. likelihood w.r.t. test data

Posterior (same form as the prior due to conjugacy)

$$p(\mathcal{D}'|\mathcal{D}) = \int p(\mathcal{D}'|\theta)p(\theta|\mathcal{D})d\theta$$

$$= \int \underbrace{\left[\prod_{i=1}^{N'} h(\tilde{\mathbf{x}}_i)\right]}_{\text{constant w.r.t. } \theta} \exp\left[\theta^\top \phi(\mathcal{D}') - N'A(\theta)\right] h(\theta) \exp\left[\theta^\top(\tau_0 + \phi(\mathcal{D})) - (\nu_0 + N)A(\theta) - \underbrace{A_c(\nu_0 + N, \boldsymbol{\tau}_0 + \phi(\mathcal{D}))}_{\text{constant w.r.t. } \theta}\right] d\theta$$

- This gets further simplified into

$$p(\mathcal{D}'|\mathcal{D}) = \left[\prod_{i=1}^{N'} h(\tilde{\mathbf{x}}_i)\right] \frac{\int h(\theta) \exp\left[\theta^\top(\tau_0 + \phi(\mathcal{D}) + \phi(\mathcal{D}')) - (\nu_0 + N + N')A(\theta)\right] d\theta}{\exp\left[A_c(\nu_0 + N, \boldsymbol{\tau}_0 + \phi(\mathcal{D}))\right]}$$

$$= \left[\prod_{i=1}^{N'} h(\tilde{\mathbf{x}}_i)\right] \frac{Z_c(\nu_0 + N + N', \boldsymbol{\tau}_0 + \phi(\mathcal{D}) + \phi(\mathcal{D}'))}{\exp\left[A_c(\nu_0 + N, \boldsymbol{\tau}_0 + \phi(\mathcal{D}))\right]}$$

# Posterior Predictive Distribution

Thus PPD as well as marginal likelihood has closed form expression when working with exp-family distributions

- Since $A_c = \log Z_c$ or $Z_c = \exp(A_c)$, we can write the PPD as

$$p(\mathcal{D}'|\mathcal{D}) = \left[\prod_{i=1}^{N'} h(\tilde{\mathbf{x}}_i)\right] \frac{Z_c(\nu_0 + N + N', \boldsymbol{\tau}_0 + \phi(\mathcal{D}) + \phi(\mathcal{D}'))}{Z_c(\nu_0 + N, \boldsymbol{\tau}_0 + \phi(\mathcal{D}))}$$

$$= \left[\prod_{i=1}^{N'} h(\tilde{\mathbf{x}}_i)\right] \exp\left[A_c(\nu_0 + N + N', \boldsymbol{\tau}_0 + \phi(\mathcal{D}) + \phi(\mathcal{D}')) - A_c(\nu_0 + N, \boldsymbol{\tau}_0 + \phi(\mathcal{D}))\right]$$

- Therefore the posterior predictive is proportional to
  - Ratio of two partition functions of two "posterior distributions" (one with $N + N'$ examples and the other with $N$ examples)
  - 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 this case $p(\mathcal{D}') = \int p(\mathcal{D}'|\theta)p(\theta)d\theta$ is simply the marginal likelihood of test data $\mathcal{D}'$

# Summary

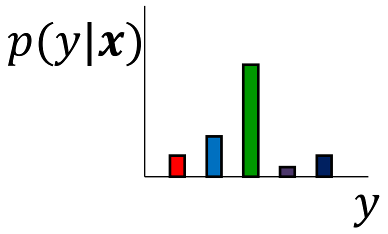
- Exp. family distributions are very useful for modeling diverse types of data/parameters
- Conjugate priors to exp. family distributions make parameter updates very simple
- Other quantities such as posterior predictive can be computed in closed form
- Useful in designing generative classification models.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

# Probabilistic Models for Classification

- Goal: Learn the conditional distribution (PMF) of discrete label $\boldsymbol{y}$ given input $\boldsymbol{x}$

Will depend on some parameters (not shown here for brevity)

$$p(y|\boldsymbol{x})$$




A discrete distribution, e.g., Bernoulli or multinoulli whose parameters will depend on the inputs $\boldsymbol{x}$

- Two ways to learn this conditional distribution
- Discriminative Approach: Don't model the inputs $\boldsymbol{x}$ and directly define $p(y|\boldsymbol{x})$
- Generative Approach: Also model the inputs $\boldsymbol{x}$ and define $p(y|\boldsymbol{x})$ as

Note: Can also use it for regression if we can define the joint $p(\boldsymbol{x}, y)$ and can obtain $p(y|\boldsymbol{x})$ from that joint (usually easy if the joint is Gaussian)

Prior probability of class $y$

a.k.a. "class-prior"

Distribution of inputs from class $y$

a.k.a. "class-conditional" distribution

$$p(y|\boldsymbol{x}) = \frac{p(\boldsymbol{x}, y)}{p(\boldsymbol{x})} = \frac{p(y)p(\boldsymbol{x}|y)}{p(\boldsymbol{x})} = \frac{p(y)p(\boldsymbol{x}|y)}{\sum_y p(y)p(\boldsymbol{x}|y)}$$

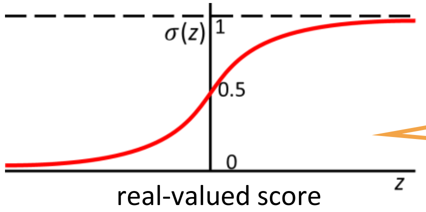
- Both discriminative and generative approaches can be learned via point estimation or by using fully Bayesian inference

# Logistic Regression

There are other ways too that can convert the score into a probability, such as a CDF: $p(y=1|\boldsymbol{x},\boldsymbol{w}) = \mu = \Phi(\boldsymbol{w}^\top \boldsymbol{x})$ where $\Phi$ is the CDF of $\mathcal{N}(0,1)$. This model is known as "Probit Regression".

- A discriminative model for binary classification ($y \in \{0,1\}$)
- A linear model with parameters $\boldsymbol{w} \in \mathbb{R}^D$ computes a score $\boldsymbol{w}^\top \boldsymbol{x}$ for input $\boldsymbol{x}$
- A sigmoid function maps this real-valued score into probability of label being $1$

$$p(y=1|\boldsymbol{x},\boldsymbol{w}) = \mu = \sigma(\boldsymbol{w}^\top \boldsymbol{x})$$




$$\sigma(z) = \frac{1}{1+\exp(-z)} = \frac{\exp(z)}{1+\exp(z)}$$

Large positive score $\boldsymbol{w}^\top \boldsymbol{x}$ means large prob of label being 1, and large negative score means low prob

- Thus conditional distribution of label $y \in \{0,1\}$ given $\boldsymbol{x}$ is the following Bernoulli

Likelihood

$$p(y|\boldsymbol{x},\boldsymbol{w}) = \text{Bernoulli}[y|\mu] = \mu^y(1-\mu)^{1-y} = \left[\frac{\exp(\boldsymbol{w}^\top \boldsymbol{x})}{1+\exp(\boldsymbol{w}^\top \boldsymbol{x})}\right]^y \left[\frac{1}{1+\exp(\boldsymbol{w}^\top \boldsymbol{x})}\right]^{1-y}$$

- Can use a Gaussian prior on $\boldsymbol{w}$: $p(\boldsymbol{w}|\lambda) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \lambda^{-1}\boldsymbol{I})$

Can also use a sparsity-inducing prior, such as spike-and-slab or a scale-mixture of Gaussians

- Point estimation (MLE/MAP) for LR gives global optima (NLL is convex in $\boldsymbol{w}$)
- We will mainly focus on fully Bayesian inference (computing the posterior)

# Logistic Regression: The Posterior

- The posterior will be

$$p(\boldsymbol{w}|\boldsymbol{X},\boldsymbol{y}) = \frac{p(\boldsymbol{w})p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{w})}{p(\boldsymbol{y}|\boldsymbol{X})} = \frac{p(\boldsymbol{w})\prod_{n=1}^{N} p(y_n|\boldsymbol{w},\boldsymbol{x}_n)}{\int p(\boldsymbol{w})\prod_{n=1}^{N} p(y_n|\boldsymbol{w},\boldsymbol{x}_n)\,d\boldsymbol{w}}$$

Gaussian

Bernoulli

Hyperparam $\lambda$ not shown

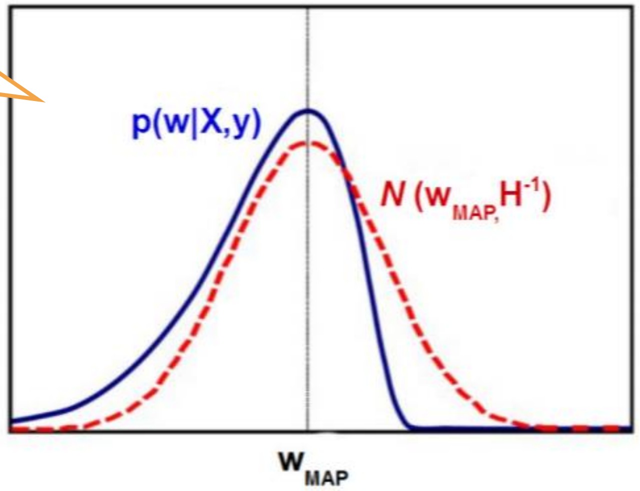
Unfortunately, Gaussian and Bernoulli are not conjugate with each other, so analytic expression for the posterior can't be obtained unlike prob. linear regression

- Need to approximate the posterior in this case
- For now, we will use a simple approximation called Laplace approximation

Other approx. inference methods, such as MCMC and VI later

Laplace approx: Approximates the intractable posterior by a Gaussian whose mean is the MAP solution of the LR model

.. and the covariance matrix of this Gaussian is set to the inverse of the Hessian matrix (second derivative) of the model's negative log-joint of params and data, evaluated at the MAP solution




$$\begin{aligned}\boldsymbol{w}_{MAP} &= \arg\max_{\boldsymbol{w}} \log p(\boldsymbol{w}|\boldsymbol{y},\mathbf{X}) \\ &= \arg\max_{\boldsymbol{w}} \log p(\boldsymbol{y},\boldsymbol{w}|\mathbf{X}) \\ &= \arg\min_{\boldsymbol{w}}[-\log p(\boldsymbol{y},\boldsymbol{w}|\mathbf{X})]\end{aligned}$$

First or second-order optimization methods can be used

$$\mathbf{H} = \nabla^2[-\log p(\boldsymbol{y},\boldsymbol{w}|\mathbf{X})]\big|_{\boldsymbol{w}=\boldsymbol{w}_{MAP}}$$

# Next Class

- Laplace approximation (contd)
- Bayesian logistic regression (contd)