**EECS E6892: Bayesian Models for Machine Learning**

**Columbia University, Fall 2015**

**Instructor: John Paisley**

**MIDTERM EXAM** (150 points)

# Exam details

- This is a take-home exam. It is open book, but you are not allowed to consult with anyone else on this exam.
- This exam is due by **3:00am** on **Friday, October 23, 2015 through Courseworks**.
- This exam counts 150 points (equivalent of 30%) towards your final grade. Late submissions will have **2.5 points deducted for each minute late**.
- Submission time is non-negotiable. I will only be able to see the time of your last submission to Courseworks. If multiple files are submitted, I will grade the last submitted file. **Under no circumstances will I accept a late test after 4:00am.**
- You must submit your answers in a **single PDF file** that is **no more than 5MB** in size. Failure to do so will result in points being deducted.
- You may answer these questions in any way you like (e.g., Latex, MS Word, scanned handwriting, etc.), so long as the above bullet is satisfied.
- Show your work for full credit. Illegible work won't receive full credit. Photographs of your work that don't show up clearly will not receive full credit. (For example, do not take a picture of your test at a 45 degree angle.)

**Question 1. Bayes rule and predictive distributions (25 + 25 points)**
We have observations $x_1, \ldots, x_n$ with each $x \in \mathbb{R}_+$. We model this as $x_i \overset{iid}{\sim} \text{Gamma}(a, \theta)$. For this problem, use the gamma density function

$$x \sim \text{Gamma}(\tau_1, \tau_2) \quad \Rightarrow \quad p(x) = \frac{\tau_2^{\tau_1}}{\Gamma(\tau_1)} x^{\tau_1 - 1} \mathrm{e}^{-\tau_2 x}$$

We want to learn $\theta$, so we place a gamma prior on it, $\theta \sim \text{Gamma}(b, c)$.

a) Calculate the posterior distribution of $\theta$, $p(\theta | x_1, \ldots, x_n)$.

b) What is the predictive distribution of a new $x$? That is, calculate $p(x_{n+1} | x_1, \ldots, x_n)$ under this modeling assumption.

**Question 2. Expectation-maximization algorithm (50 points)**

You are given a data set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, where $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$. You model this using Bayesian linear regression with the following prior structure,

$$y_i \overset{ind}{\sim} \text{Normal}(x_i^T w, \alpha^{-1}), \quad w \sim \text{Normal}(0, \lambda^{-1} I), \quad \alpha \sim \text{Gamma}(a, b).$$

Derive an Expectation-Maximization algorithm for optimizing $\ln p(y, \alpha | x)$ over $\alpha$, where the vector $w$ functions as the variable being integrated out. Please note the following about what I am looking for:

- It is not necessary to show all work in deriving $q(w)$ using Bayes rule, but it must be clear that you know how Bayes rule is used here, and also what the solution of $q(w)$ is.
- It must be clear that you understand what constitutes the "E" and the "M" steps. Partial credit will be given for correct algorithms without a clear path to the solution.
- You must give pseudo-code for optimizing $\alpha$ including the equations that you would implement in a coding language (similar to the algorithm outlines in the notes).

**Question 3. Variational inference (50 points)**

You are given a data set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, where $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$. You model this using Bayesian linear regression with the following prior structure,

$$y_i \overset{ind}{\sim} \text{Normal}(x_i^T w, \alpha^{-1}), \quad w \sim \text{Normal}(0, \lambda^{-1} I), \quad \lambda \sim \text{Gamma}(a, b).$$

Please note that the Gamma prior is on a different variable from Question 2.

Derive a variational inference algorithm for learning $q(w, \lambda) \approx p(w, \lambda | y, x)$ using the factorization $q(w, \lambda) = q(w)q(\lambda)$. For your $q$ distributions, use

$$q(w) = \text{Normal}(\mu', \Sigma'), \qquad q(\lambda) = \text{Gamma}(a', b').$$

Again, please note the following about what I am looking for:

- You are free to use the "direct method" from the notes, but the "optimal method" will be much simpler. The $q$ distributions above are the optimal ones.
- Therefore, you do not need to explicitly calculate the variational objective function to receive full credit ("$\mathcal{L}$" in the notes). But again, you are free to take this approach.
- What I am looking for is an equation-based algorithm (not a gradient-based algorithm) for learning $(a', b', \mu', \Sigma')$. Since it is technically possible to calculate $\mathcal{L}$ and optimize it using gradient methods, you will be given partial credit if you choose this route.
- For full credit, your work must show a clear path to your answer.