Rules:

1. Homework is due on the due date in the class on April 29. Please see course website for policy on late submission.

2. We recommend that you typeset your homework using appropriate software such as \LaTeX. If you are writing please make sure your homework is cleanly written and legible. The TAs will not invest undue effort to decrypt bad handwriting.

3. You must hand in a hard copy of the homework. The only exception is if you are out of town in which case you can email your homeworks to 10708-instructor@cs.cmu.edu. If this is the case, your homework must be typeset using proper software. Please do not email written and scanned copies. Your email must be sent by the beginning of the class on the due date.

4. The submission procedure for the programming component is described along with the corresponding question.

5. You are allowed to collaborate on the homework, but you should write up your own solution and code. Please indicate your collaborators in your submission.

6. Please staple your homeworks.
1 Dirichlet Process Gaussian Mixture Model (25 Points) (Peng-tao)

The Dirichlet Process Gaussian Mixture Model can be described as follows

- Draw $v_k \sim \text{Beta}(1, \alpha)$, $k = 1, \ldots, \infty$
- Draw $\mu_k \sim \mathcal{N}(\mu_0, \Sigma)$, $k = 1, \ldots, \infty$
- For the $n$th data point:
  - Draw $z_n \sim \text{Multinomial}(\pi(v))$
  - Draw $x_n \sim \text{Gaussian}(\mu_{z_n}, \hat{\Sigma})$

$\alpha$ is a concentration parameter. $\pi(v)$ is an infinite multinomial vector where $\pi(v) = v_k \prod_{j=1}^{k-1}(1 - v_j)$.

1. (10 pts) Prove that
   $$P(\sum_{k=1}^{\infty} \pi_k(v) = 1) = 1$$

2. (15 pts) Given the following distribution $q(\{v_k\}_{k=1}^{\infty}, z_n) = \prod_{k=1}^{\infty} q(v_k)q(z_n)$ where (1) $q(z_n|\phi)$ defined over $z_n$ is a multinomial distribution parametrized by $\phi$; (2) for $k < T$, $q(v_k|\gamma_k)$ defined over $v_k$ is a beta distribution parametrized by $\gamma_k$ which is a two dimensional vector; (3) for $k \geq T$, $q(v_k = 1) = 1$. $T$ is a fixed large integer. Prove that
   $$E_{q(\{v_k\}_{k=1}^{\infty}, z_n)}[\log p(z_n|\{v_k\}_{k=1}^{\infty})] = T - \sum_{i=1}^{T} \left\{ \sum_{j=i+1}^{T} \phi_j(\Psi(\gamma_{i,2}) - \Psi(\gamma_{i,1} + \gamma_{i,2})) + \phi_i(\Psi(\gamma_{i,1}) - \Psi(\gamma_{i,1} + \gamma_{i,2})) \right\}$$

where $\Psi(\cdot)$ is a digamma function.

2 Regularized Bayes (50 Points) (Xun)

In this problem, we will see how to combine RegBayes principal with an existing model.

1. Variational view of Bayes’ theorem. Let $\mathcal{M}$ be the model, $\mathcal{D}$ be the data, $p_0(\mathcal{M})$ be the prior over of the model, and $\mathcal{P}$ be the space of all density functions.

   Show that the posterior $p(\mathcal{M} | \mathcal{D})$ is the unique solution of the following variational problem:

   $$\inf_{q(M) \in \mathcal{P}} \mathcal{L}(q) = \text{KL} [q(\mathcal{M}) || p_0(\mathcal{M})] - E_q [\log p(\mathcal{D} | \mathcal{M})].$$

2. A Bayesian topic model. Let $\mathcal{V} = \{1, \ldots, V\}$ be the set of words and $\mathcal{D} = \{w_d\}_{d=1}^{D}$ be the set of documents, where $w_d = \{w_{di}\}_{i=1}^{N_d}$ is the set of tokens appearing in document $d$, with each $w_{di} \in \mathcal{V}$. Each document is viewed as an admixture of $K$ topics, where each topic $\phi_k$ is a multinomial distribution over $\mathcal{V}$. In a fully Bayesian treatment, topics themselves are considered as random variables and assumed to be generated from the conjugate prior, i.e., for all $k$, $\phi_k \sim \text{Dir}(\beta)$.

   The generative process of document $d$ can be written as:
   - Draw topic mixing coefficients $\theta_d \sim \text{Dir}(\alpha)$;
   - For each position $i = 1, \ldots, N_d$, in the document:
1. Draw topic assignment $z_{di} \sim \text{Mult}(\theta_d)$;  
2. Draw token $w_{di} \sim \text{Mult}(\phi_{z_{di}})$;

Let $\Theta = \{\theta_d\}_{d=1}^D$ be the set of topic proportions, $\Phi = \{\phi_k\}_{k=1}^K$ be the set of topics, and $Z = \{z_d\}_{d=1}^D$ be the set of topic assignments, where $z_d = \{z_{di}\}_{i=1}^N$ represents the topic assignments in document $d$. Identify the set of model variables $M$ and write down the variational form of the posterior $p(M|D)$ as in (3).

3. RegBayes. Now suppose the documents are labeled, i.e., $D = \{(w_d, y_d)\}_{d=1}^D$ with each $y_d \in \{+1, -1\}$.

We can make use of the supervision signals by introducing margin constraints:

$$\begin{align*}
\text{minimize} & \quad \mathcal{L}(q) + U(\xi) \\
\text{subject to} & \quad q(M') \in P_{\text{post}}(\xi),
\end{align*}$$

where

$$\begin{align*}
\mathcal{M}' &= \mathcal{M} \cup \{\eta\}, \\
\mathcal{L}(q) &= \text{KL}[q(M')||p_0(\eta)p_0(M)] - \mathbb{E}_q[\log p(D|M)] \\
U(\xi) &= c \sum_{d=1}^D \xi_d, \\
P_{\text{post}}(\xi) &= \left\{ q(M') : y_d \cdot \mathbb{E}_q[f(\eta, \bar{z}_d)] \geq 1 - \xi_d, \xi_d \geq 0, \forall d \right\}.
\end{align*}$$

The tuning parameter $c$ controls trade-off between posterior divergence and margin constraints; $f(\eta, \bar{z}_d) = \eta^\top \bar{z}_d$ is the linear discriminant function, where $\bar{z} = \frac{1}{N} \sum_{i=1}^N z_i$ is the empirical average of topic assignments in a document.

Does problem (4) have a unique solution? Explain why it is hard to solve exactly.

4. Mean-field variational inference. We need to resort to approximation methods to make the inference tractable. Consider the following mean-field assumption:

$$q(M') = q(\eta)q(M).$$

Show that the optimal solution $q^*(\eta)$ given $q(M)$ has the form

$$q^*(\eta) = \frac{1}{Z} p_0(\eta) \exp \left\{ \eta^\top \sum_{d=1}^D \lambda^*_d x_d \right\},$$

where $x_d = y_d \mathbb{E}[\bar{z}_d]$ and $\lambda^* = \{\lambda^*_d\}_{d=1}^D$ is the solution of the following dual problem:

$$\begin{align*}
\text{maximize} & \quad - \log Z + 1^\top \lambda \\
\text{subject to} & \quad 0 \leq \lambda \leq c.
\end{align*}$$

Note: please refer to the solution$^2$ of homework 1 for notes on the calculus of variations.

---

$^1$ A $K$-dimensional binary vector with only one entry being nonzero.

$^2$ http://www.cs.cmu.edu/~epxing/Class/10708-15/homeworks/homework1_solution.pdf
5. Choice of prior. From (11) we can see that a natural choice of \( p_0(\eta) \) can be the exponential family. In particular, show that if \( p_0(\eta) = \mathcal{N}(\eta; 0, I) \), the optimal solution \( q^*(\eta) \) is also a Gaussian with mean \( \mu \), which can be obtained by solving a standard linear SVM:

\[
\begin{align*}
\min_{\mu, \xi} & \quad \frac{1}{2} \| \mu \|_2^2 + c \cdot \sum_{d=1}^{D} \xi_d \\
\text{subject to} & \quad \forall d: \begin{cases} 
\mu^\top x_d \geq 1 - \xi_d, \\
\xi_d \geq 0.
\end{cases}
\end{align*}
\]

(14)

(15)

6. Surrogate loss. Although solving \( q(\eta) \) is easy under the Gaussian prior, solving for \( q(\mathcal{M}) \) turns out to be nontrivial and often requires stronger assumptions. Now let’s take another approach which does not make mean-field assumptions. But first we need to relax the problem a little bit. Show that the following relationship between two versions of expected hinge loss holds:

\[
\mathcal{R}'(q) = \mathbb{E}_q \left[ (1 - \eta^\top x_d)_{+} \right] \geq (1 - \mathbb{E}_q [\eta^\top x_d])_{+} = \mathcal{R}(q).
\]

(With slight abusing of the notation: \( x_d = y_d z_d \).)

7. Augmented posterior. Let \( \phi(y_d|x_d, \eta) = \exp \{-2c \cdot (1 - \eta^\top x_d)_{+}\} \) be the conditional “likelihood” for the label \( y_d \). Derive the new posterior by first replacing \( \mathcal{R}(q) \) with \( \mathcal{R}'(q) \) in (4) and then solving it. Augment the posterior with the auxiliary variable \( \gamma \), as we have done in homework 3.

(As noted in the previous homework, multiplicative factors can be absorbed into the regularization parameter, so it is fine to replace \( c \) with \( 2c \).)

8.Collapsed Posterior. In the recitation we covered Gibbs sampling for LDA where \( (\Theta, \Phi) \) are integrated out. Following similar derivation, show that the collapsed posterior is

\[
q^*(\eta, \gamma, Z|D) \propto p_0(\eta) \prod_{d=1}^{D} \frac{\mathcal{B}(n_d, \alpha)}{\mathcal{B}(\alpha)} \prod_{k=1}^{K} \frac{\mathcal{B}(n_k, \beta)}{\mathcal{B}(\beta)} \prod_{d=1}^{D} \frac{1}{\sqrt{2\pi\gamma_d}} \exp \left( -\frac{(\gamma_d + c\zeta_d)^2}{2\gamma_d} \right),
\]

(17)

where \( \zeta_d = 1 - \eta^\top x_d \), \( \mathcal{B}(\cdot) \) is the multivariate Beta function, \( n_{kd} \) is the number of tokens in document \( d \) assigned to topic \( k \), \( n_d = \{n_{kd}\}_{k=1}^{K} \) is the topic counts of document \( d \), \( n_{kw} \) is the number of word \( w \) assigned to topic \( k \), \( n_k = \{n_{kw}\}_{w=1}^{V} \) is the word counts of topic \( k \).

9. Gibbs sampling. Derive the conditionals from the collapsed posterior, using isotropic Gaussian prior \( p_0(\eta) = \prod_{k=1}^{K} \mathcal{N}(\eta_k; 0, \nu^{-1}) \) on \( \eta \). Be as specific as possible at each step, especially for \( q(z_{di} = 1 | \text{rest}) \).

3 Structural SVMs (25 Points) (Mrinmaya)

Note: These problems are related to two classic structured prediction papers [1, 2]. If you are confused or are finding it hard to work out the solutions, you can read the two papers which will help you work out the solution.

3.1 Max-Margin Markov Networks (10 points)

Let \( S = \{(x_i, y_i)\}_{i=1}^{N} \) be a labeled training set. Suppose we have a mapping from an input \( x \) to the corresponding Markov network graph \( G(x) = (V, E) \), where the nodes \( V \) correspond to the variables in \( y \), and let \( G_i = G(x_i) \). Suppose the graph is a log-linear conditional random field that represents the conditional distribution \( p(y|x) \). In particular, suppose the CRF is defined via:

\[
\log p_w(y|x) = w^T \phi(x, y) - \log Z_w(x)
\]
Moreover, assume that the graph is tree-structured. We now consider margin-based training approaches for this model. Maximum-margin estimation of this model can be formulated as solving the following optimization problem:

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{2}||w||^2 + C \sum_{i=1}^{N} \xi_i \\
\text{subject to} & \quad w^T \delta \phi_i(y) \geq \Delta(y_i, y) - \xi_i \quad \forall i, y \in \mathcal{Y} \\
& \quad \xi_i \geq 0 \quad \forall i
\end{align*}
\]

where, \( \delta \phi_i(y) = \phi(x_i, y_i) - \phi(x_i, y) \) This is the primal structural SVM with margin rescaling. The dual is given by:

\[
\begin{align*}
\text{maximize} & \quad \sum_{i=1}^{N} \sum_{y} \alpha_i(y) \Delta(y_i, y) - \frac{1}{2} \sum_{i=1}^{N} \sum_{y} \alpha_i(y) \delta \phi_i(y) \\
\text{subject to} & \quad \sum_{y} \alpha_i(y) = C \quad \forall i \\
& \quad \alpha_i(y) \geq 0 \quad \forall i, y
\end{align*}
\]

Here the dual variables are \( \alpha_i(y) \). The primal formulation has exponentially many constraints in the number of labels, so the dual has exponentially many variables. Assume that the loss function is decomposable, i.e., that:

\[
\Delta(y_i, y) = \sum_{c \in C(G_i)} \Delta(y_{ic}, y_c),
\]

where \( C(G_i) \) denotes the cliques of \( G_i \) (e.g. Hamming loss). Assume that \( \delta \phi_i \) decomposes similarly. Show how to use the structure of the model to reparameterize the dual as an equivalent problem with only polynomially many variables.

### 3.2 Latent structural SVMs (15 points)

The standard structural SVM objective can equivalently be written without explicit use of slack variables, as in [2]:

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{2}||w||^2 + C \sum_{i=1}^{N} \left( \max_{\hat{y}} \left( \Delta(y_i, \hat{y}) + w^T \phi(x_i, \hat{y}) \right) - w^T \phi(x_i, y_i) \right) \\
\end{align*}
\]

The objective of this problem is convex because it is the difference between a convex term and a linear term. Note that the second term can be shown to be a convex upper bound on the desired loss function \( \Delta(y_i, \hat{y}_i) \).

The latent structural SVM introduces latent variables \( h \), giving features \( \phi(x, y, h) \) and the prediction rule:

\[
\begin{align*}
f_w(x) &= \arg\max_{y, h} w^T \phi(x, y, h)
\end{align*}
\]

The loss function \( \Delta \) can also depend on the latent variables.

1. The latent structural SVM objective is:

\[
\begin{align*}
\frac{1}{2}||w||^2 + C \sum_{i=1}^{N} \left( \max_{\hat{y}, h} \left( w^T \phi(x_i, \hat{y}, \hat{h}) + \Delta(y_i, \hat{y}, \hat{h}) \right) \right) - C \sum_{i=1}^{N} \left( \max_{h} w^T \phi(x_i, y_i, h) \right)
\end{align*}
\]

a difference of convex functions. Show that:

\[
\begin{align*}
\left( \max_{\hat{y}, h} \left( w^T \phi(x_i, \hat{y}, \hat{h}) + \Delta(y_i, \hat{y}, \hat{h}) \right) \right) - \left( \max_{h} w^T \phi(x_i, y_i, h) \right)
\end{align*}
\]
is a valid upper bound on $\Delta(y_i, \hat{y}_i(w), \hat{h}_i(w))$.

2. Algorithm 1 of [2] describes a convex-concave procedure for optimizing the latent structural SVM objective, which is the difference of two convex functions. At each iteration, the algorithm requires finding a hyperplane $v_t$ such that:

$$-g(w) \leq -g(w_t) - v_t^T (w - w_t)$$

Using subdifferentials, characterize the set of vectors $v_t$ that satisfy this inequality. Show that the particular choice:

$$v_t = C \sum_{i=1}^N \phi(x_i, y_i, h_i^*)$$

satisfies the desired inequality, where

$$h_i^* = \arg\max_h w_t^T \phi(x_i, y_i, h)$$

for each $i$.

References
