Question 1 (Mixtures are as good as point distributions): Let $P$ be a Laplace($\lambda$) distribution on $\mathbb{R}$, meaning that $X \sim P$ has density
\[ p(x) = \frac{\lambda}{2} \exp(-\lambda|x|). \]
Assume that $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P$, and let $P^n$ denote the $n$-fold product of $P$. In this problem, we compare the predictive performance of distributions from the normal location family $P = \{N(\theta, \sigma^2) : \theta \in \mathbb{R}\}$ with the mixture distribution $Q^\pi$ over $P$ defined by the normal prior distribution $N(\mu, \tau^2)$, that is, $\pi(\theta) = \frac{1}{(2\pi\tau^2)^{1/2}} \exp(-\frac{(\theta - \mu)^2}{2\tau^2})$.

(a) Let $P_{\theta, \Sigma}$ be the multivariate normal distribution with mean $\theta \in \mathbb{R}^n$ and covariance $\Sigma \in \mathbb{R}^{n \times n}$. What is $D_{\text{KL}}(P^n \| P_{\theta, \Sigma})$?

(b) Show that $\inf_{\theta \in \mathbb{R}^n} D_{\text{KL}}(P^n \| P_{\theta, \Sigma}) = D_{\text{KL}}(P^n \| P_{0, \Sigma})$, that is, the mean-zero normal distribution has the smallest KL-divergence from the Laplace distribution.

(c) Let $Q^\pi_n$ be the mixture of the $n$-fold products in $P$, that is, $Q^\pi_n$ has density
\[ q^\pi_n(x_n) = \int_{-\infty}^{\infty} \pi(\theta)p_\theta(x_1) \cdots p_\theta(x_n) d\theta, \]
where $\pi$ is $N(0, \tau^2)$. What is $D_{\text{KL}}(P^n \| Q^\pi_n)$?

(d) Show that the redundancy of $Q^\pi_n$ under the distribution $P$ is asymptotically nearly as good as the redundancy of any $P_\theta \in P$, the normal location family (so $P_\theta$ has density $p_\theta(x) = (2\pi\sigma^2)^{-1/2} \exp(-\frac{(x - \theta)^2}{2\sigma^2})$). That is, show that
\[ \sup_{\theta \in \mathbb{R}} \mathbb{E}_P \left[ \log \frac{1}{q^\pi_n(X_1^n)} - \log \frac{1}{p_\theta(X_1^n)} \right] = O(\log n) \]
for any prior variance $\tau^2 > 0$ and any prior mean $\mu \in \mathbb{R}$, where the big-Oh hides terms dependent on $\tau^2, \sigma^2, \mu^2$.

(e) **Extra credit:** Can you give an interesting condition under which such redundancy guarantees hold more generally? That is, using Proposition 9.7 in the notes, give a general condition under which
\[ \mathbb{E}_P \left[ \log \frac{1}{q^\pi(X_1^n)} - \log \frac{1}{p_\theta(X_1^n)} \right] = o(n) \]
as $n \to \infty$, for all $\theta \in \Theta$. 

**Question 2:** Find the suboptimality function $H_\phi$ and $\psi$-transform for the binary classification problem with the following losses.

(a) Logistic loss. That is,
$$\phi(\alpha) = \log(1 + e^{-\alpha})$$

(b) Squared error (ordinary regression). The surrogate loss in this case for the pair $(x, y)$ is $\frac{1}{2}(f(x) - y)^2$. Show that for $y \in \{-1, 1\}$, this can be written as a margin-based loss, and compute the associated suboptimality function $H_\phi$ and $\psi$-transform. Is the squared error classification calibrated?

**Question 3:** Suppose we have a regression problem with data (independent variables) $x \in \mathcal{X}$ and $y \in \mathbb{R}$. We wish to find a predictor $f : \mathcal{X} \rightarrow \mathbb{R}$ minimizing the probability of being far away from the true $y$, that is, for some $c > 0$, our loss is of the form
$$L(f(x), y) = 1 \{ |y - f(x)| \geq c \}.$$  
Show that no loss of the form $\phi(\alpha, y) = |\alpha - y|^p$, where $p \geq 1$, is Fisher consistent for the loss $L$, even if the distribution of $Y$ conditioned on $X = x$ is symmetric about its mean $\mathbb{E}[Y \mid X]$. That is, show there exists a distribution on pairs $X,Y$ such that the set of minimizers of the surrogate
$$R_\phi(f) := \mathbb{E}[\phi(f(X), Y)]$$
is not included in the set of minimizers of the true risk, $R(f) = \mathbb{P}(|Y - f(X)| \geq c)$, even if the distribution of $Y$ (conditional on $X$) is symmetric.

**Question 4** (Bayes risk gaps): Consider a general binary classification problem with $(X, Y) \in \mathcal{X} \times \{-1, 1\}$. Let $\phi(\alpha) = \log(1 + e^{-\alpha})$, so that we use the logistic loss. Show that the surrogate risk gap
$$R^*_{\phi, \text{prior}} - R^*_\phi = I(X; Y),$$
where $I$ is the mutual information.

**Question 5** (Empirics of classification calibration): In this problem you will compare the performance of hinge loss minimization and an ordinary linear regression in terms of classification performance. Specifically, we compare the performance of the hinge surrogate loss with the regression surrogate when the data is generated according to the model
$$y = \text{sign}((\theta^*, x) + \sigma Z), \quad Z \sim \mathcal{N}(0, 1)$$
where $\theta^* \in \mathbb{R}^d$ is a fixed vector, $\sigma \geq 0$ is an error magnitude, and $Z$ is a standard normal random variable. We investigate the model (1) with a simulation study.
Specifically, we consider the following set of steps:

(i) Generate two collections of $n$ datapoints in $d$ dimensions according to the model (1), where $\theta \in \mathbb{R}^d$ is chosen (ahead of time) uniformly at random from the sphere $\{\theta \in \mathbb{R}^d : \|\theta\|_2 = R\}$, and where each $x_i \in \mathbb{R}^d$ is chosen as $N(0, I_{d \times d})$. Let $(x_i, y_i)$ denote pairs from the first collection and $(x^\text{test}_i, y^\text{test}_i)$ pairs from the second.
(ii) Set
\[
\hat{\theta}_{\text{hinge}} = \arg\min_{\theta : \|\theta\|_2 \leq R} \frac{1}{n} \sum_{i=1}^{n} [1 - y_i \langle x_i, \theta \rangle]_+
\]
and
\[
\hat{\theta}_{\text{reg}} = \arg\min_{\theta} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \langle x_i, \theta \rangle)^2 = \arg\min_{\theta} \|X\theta - y\|_2^2.
\]

(iii) Evaluate the 0-1 error rate of the vectors \(\hat{\theta}_{\text{hinge}}\) and \(\hat{\theta}_{\text{reg}}\) on the held-out data points \(\{(x_i^{\text{test}}, y_i^{\text{test}})\}_{i=1}^{n}\).

Perform the preceding steps (i)–(iii), using any \(n \geq 100\) and \(d \geq 10\) and a radius \(R = 5\), for different standard deviations \(\sigma = \{0, 1, \ldots, 10\}\); perform the experiment a number of times. Give a plot or table exhibiting the performance of the classifiers learned on the held-out data. How do the two compare? Given that for the hinge loss we know \(H_{\phi}(\delta) = \delta\) (as presented in class), what would you expect based on the answer to Question 2?

I have implemented (in the julia language; see http://julialang.org/) methods for solving the hinge loss minimization problem with stochastic gradient descent so that you do not need to. The file is available at this link. The code should (hopefully) be interpretable enough that if julia is not your language of choice, you can re-implement the method in an alternative language.

**Question 6:** In this question, we generalize our results on classification calibration and surrogate risk consistency to a much broader supervised learning setting. Consider the following general supervised learning problem, where we assume that we have data in pairs \((X, Y) \in \mathcal{X} \times \mathcal{Y}\), where \(\mathcal{X}\) and \(\mathcal{Y}\) are general spaces.

Let \(L : \mathbb{R}^m \times \mathcal{Y} \rightarrow \mathbb{R}_+\) be a loss function we wish to minimize, so that the loss of a prediction function \(f : \mathcal{X} \rightarrow \mathbb{R}^m\) for the pair \((x, y)\) is \(L(f(x), y)\). Let \(\varphi : \mathbb{R}^m \times \mathcal{Y} \rightarrow \mathbb{R}\) be an arbitrary surrogate, where \(\varphi(f(x), y)\) is the surrogate loss. Define the risk and \(\varphi\)-risk
\[
R(f) := \mathbb{E}[L(f(X), Y)] \quad \text{and} \quad R_{\varphi}(f) := \mathbb{E}[\varphi(f(X), Y)].
\]
Let \(\mathcal{P}_Y\) denote the space of all probability distributions on \(\mathcal{Y}\), and define the conditional (pointwise) risks \(\ell : \mathbb{R}^m \times \mathcal{P}_Y \rightarrow \mathbb{R}\) and \(\ell_{\varphi} : \mathbb{R}^m \times \mathcal{P}_Y \rightarrow \mathbb{R}\) by
\[
\ell(\alpha, P) = \int_{\mathcal{Y}} L(\alpha, y)p(y)dy \quad \text{and} \quad \ell_{\varphi}(\alpha, P) = \int_{\mathcal{Y}} \ell(\alpha, y)p(y)dy.
\]
(Here for simplicity we simply write integration against \(dy\); you may make this fully general if you wish.) Let \(\ell^*(P) = \inf_{\alpha} \ell(\alpha, P)\) denote the minimal conditional risk, and similarly for \(\ell_{\varphi}^*(P)\), when \(Y\) has distribution \(P\). If \(P_x\) denotes the distribution of \(Y\) conditioned on \(X = x\), then we may rewrite the risk functionals as
\[
R(f) = \mathbb{E}[\ell(f(X), P_X)] \quad \text{and} \quad R_{\varphi}(f) = \mathbb{E}[\ell_{\varphi}(f(X), P_X)].
\]
We will show that the same machinery we developed for classification calibration extends to this general supervised learning setting.

For \(\epsilon \geq 0\), define the suboptimality gap function
\[
\Delta_{\varphi}(\epsilon, P) := \inf_{\alpha \in \mathbb{R}^m} \left\{ \ell_{\varphi}(\alpha, P) - \ell_{\varphi}^*(P) : \ell(\alpha, P) - \ell^*(P) \geq \epsilon \right\},
\]
which measures the gap between achievable (pointwise) risk and the best surrogate risk when we enforce that the true loss is not minimized. Also define the uniform suboptimality function
\[
\Delta_{\varphi}(\epsilon) := \inf_{\alpha \in \mathbb{R}^m, P \in \mathcal{P}_Y} \{ \ell_{\varphi}(\alpha, P) - \ell_{\varphi}^*(P) : \ell(\alpha, P) - \ell^*(P) \geq \epsilon \}.
\]
(Compare this with the definition of $\Delta$ for the classification case to gain intuition.)

(a) A uniform result: let $\Delta_{\varphi}^{**}(\epsilon)$ be the biconjugate of $\Delta_{\varphi}$ (that is, $\Delta_{\varphi}^{**}$ is the largest convex function below $\Delta_{\varphi}$). Show that
\[
\Delta_{\varphi}^{**}(R(f) - R^*) \leq R_{\varphi}(f) - R_{\varphi}^*.
\]
Prove that $\Delta_{\varphi}(\epsilon) > 0$ for all $\epsilon > 0$ implies if $R_{\varphi}(f_n) \to R_{\varphi}^*$, then $R(f_n) \to R^*$.

(b) We say that the loss $\varphi$ is **uniformly calibrated** if $\Delta_{\varphi}(\epsilon) > 0$ for all $\epsilon > 0$. Show that, in the margin-based binary classification case with loss $\phi : \mathbb{R} \to \mathbb{R}$, uniform calibration as defined here is equivalent to classification-calibration as defined in class. You may assume that the margin-based loss $\phi$ is continuous.

(c) **Extra credit:** A non-uniform result: assume that for all distributions $P \in \mathcal{P}_Y$ on the set $\mathcal{Y}$, we have
\[
\Delta_{\varphi}(\epsilon, P) > 0
\]
if $\epsilon > 0$. (We call this calibration.) Assume that there exists an upper bound function $B : \mathcal{X} \to \mathbb{R}_+$ such that $\mathbb{E}[B(X)] < \infty$ and $\ell(\alpha, P_x) \leq \ell^*(P_x) + B(x)$ for all $x$ and $\alpha \in \mathbb{R}^m$. For example, if the loss $L$ is bounded, this holds. Show that if the sequence of functions $f_n : \mathcal{X} \to \mathbb{R}^m$ satisfies
\[
R_{\varphi}(f_n) \to R_{\varphi}^* \text{ then } R(f_n) \to R^*.
\]
Equivalently, show that for any distribution $P$ on $\mathcal{X} \times \mathcal{Y}$, for all $\epsilon > 0$ there exists a $\delta > 0$ such that
\[
R_{\varphi}(f) \leq R_{\varphi}^* + \delta \text{ implies } R(f) \leq R^* + \epsilon.
\]
(You may ignore any measurability issues that come up.)