# CMSC 35900-2: A Probabilistic Approach to Machine Learning

## Problem set 2

Due last day of term

# The Probit

The main response ("squash") function we looked at was the logistic response. In this question we will consider the probit response function, given by:

$$g(z) = \int_{t=-\infty}^{z} \tfrac{1}{\sqrt{2\pi}} e^{-t^2/2} dt.$$

Try plotting this function, as well as the logistic response, and verify that they have a similar, but not identical, shape. We will consider a Gaussian Process classification model specified by:

$$\begin{aligned} f &\sim GP(K) \\ f_i &= f(X_i) \\ \pi_i &= g(f_i) \\ y_i &\sim Ber(\pi_i) \end{aligned} \tag{1}$$

### Problem 1

1. Verify that the probit response is symmetric. That is, that $P(y_i|f_i) = g(y_i f_i)$.
2. Prove that the probit model of (1) is equivalent to the following specification:

   $$\begin{aligned} f &\sim GP(K) \\ f_i &= f(X_i) \\ \epsilon_i &\sim \mathcal{N}(0,1) \\ z_i &= f_i + \epsilon_i \\ y_i &= \mathrm{sign}(z_i) \end{aligned} \tag{2}$$

   That is, the joint distribution over $y_i$ is the same under both models.
3. Write down the conditional distribution of $z_1, \ldots, z_N | x_1, \ldots, x_N$ under the model (2) explicitly (marginalizing out $f$ and $f_i$).

# Markov Chain Monte Carlo Sampling

It is often beneficial to combine several different suggestion distributions $Q_i(X'|X)$. A particular case of interest is when each suggestion distribution changes a different component of $X = (X[1], X[2], \ldots, X[N])$, i.e. where $Q_i(X'|X) = 0$ unless $X[j] = X'[j]$ for all $j \neq i$. As discussed in class, Gibbs Sampling is a special case of the Metropolis-Hasting method with multiple such suggestion distribution and $Q_i(X'|X) = P(X'[i]|X[-i])$ when $X[j] = X'[j]$ for all $j \neq i$. We will provide a rigorous basis for combining different suggestion distributions, and so also for Gibbs Sampling.

**Problem 2** We first consider choosing between the different suggestion distributions at random. That is, given suggestion distributions $Q_1, \ldots, Q_N$ and a probability distribution $p = (p_1, \ldots, p_N) \in \triangle_N$, we consider a joint suggestion distribution $Q$ by first picking a random index $i$ according to $p$ and then picking $X'$ according to $Q_i(X'|X)$. That is:

$$Q(X'|X) = \sum_i p_i Q_i(X'|X) \tag{3}$$

To complete a Metropolis-Hasting step with suggestion distribution $Q$, we need to compute the acceptance probability

$$a = \min\left(1, \frac{P(X')Q(X|X')}{P(X)Q(X'|X)}\right). \tag{4}$$

1. Prove that if each suggestion distribution $Q_i$ changes a different component of $X$, then the acceptance probability can be equivalently computed using only $Q_i$:

$$a = \min\left(1, \frac{P(X')Q_i(X|X')}{P(X)Q_i(X'|X)}\right). \tag{5}$$

2. We would also like to consider suggestion distributions for which the above property (changing only different components) does not hold. For example, for $X \in \mathbb{R}^D$, we might consider a low-variance random Gaussian move and a high-variance random Gaussian move. Explain why (4) and (5) are *not* always equivalent in the general case of multiple suggestion distributions, and write down the expression for the correct acceptance probability, in the case of two suggestion distributions $Q_1$ and $Q_2$, in terms of $P, Q_1$ and $Q_2$.

We now turn to combining Metropolis-Hasting moves deterministically, according to some prespecified schedule. Each suggestion distribution $Q_i(X'|X)$ defines a random transition probability $T_i(X_{n+1}|X_n)$ defined by the acceptance procedure:

- Pick $X'$ according to $Q_i(X'|X_n)$
- Calculate the acceptance probability $a$ according to (5)

- Set $X_{n+1} = X'$ with probability $a$ and $X_{n+1} = X_n$ otherwise.

Considering applying $T_1, T_2, \ldots, T_N$ sequentially and cyclically. That is, the following sampling procedure:

- Start with at some initial $X_0$ and $n = 0$.
- Repeat:
  - For $i = 1, \ldots, N$,
    - ∗ Pick $X'$ according to $Q_i(X'|X_n)$
    - ∗ Calculate $a$ according to (5)
    - ∗ Set $X_{n+1} = X'$ with probability $a$ and $X_{n+1} = X_n$ otherwise.
    - ∗ Increase $n \leftarrow n + 1$.

**Problem 3**

1. Prove that if the two transition probabilities $T_1(X'|X)$ and $T_2(X'|X)$ maintain detailed balance (i.e. are reversible) with respect to the same stationary distribution $P(X)$, then the transition obtained by applying them one after the other, $T(X''|X) = \sum_{X'} T_1(X'|X)T_2(X''|X')$, also maintains detailed balance with respect to $P(X)$.
2. Use this to argue that the stationary distribution of the Markov chain $X_i$ described above is in-fact $P(X)$ (assuming the the chain is ergodic).

**Problem 4** Consider a modified procedure where we choose which suggestion distribution to use based on $X_i$:

- Start with at some initial $X_0$ and $n = 0$.
- Repeat:
  - Pick $i$ based on $X_n$ using some pre-specified, perhaps randomized, procedure.
  - Pick $X'$ according to $Q_i(X'|X_n)$
  - Calculate $a$ according to (5)
  - Set $X_{n+1} = X'$ with probability $a$ and $X_{n+1} = X_n$ otherwise.
  - Increase $n \leftarrow n + 1$.

Explain how in the previous Problem we relied on the fact that the order in which we apply the moves $T_i$ does not depend on the sequence $X_i$. Provide a simple example of a distribution $P(X)$, two suggestion distribution $Q_1(X'|X)$ and $Q_2(X'|X)$ and a procedure for picking $i$ based on $X_n$ such that the resulting chain $X_n$ is ergodic, but its stationary distribution is *not* $P(X)$.

# Boltzmann Machines

Consider a Boltzmann Machines involving both observed variables $X = (X[1], \ldots, X[D])$ and latent variables $Z = (Z[1], \ldots, Z[K])$. For convenience we consider the concatenated vector $Y = (X, Z) \in \{\pm 1\}^{D+K}$ containing both observed and latent variables. The joint distribution, parametrized by $W \in \mathbb{R}^{(D+K)\times(D+K)}$ is given by:

$$P(Y|W) \propto e^{\frac{1}{2}Y'WY}. \tag{6}$$

Our goal is to find the maximum likelihood estimator for $W$, given i.i.d. observations $X_1, \ldots, X_N$ drawn from the marginal $P(X|W)$:

$$\hat{W}_{\text{ML}} = \arg\max_{W} P(X_1, \ldots, X_N|W). \tag{7}$$

To do so, we will consider the gradient of the log-likelihood with respect to $W$.

### Problem 5

1. Prove that for each data point $X_n$, the gradient of the log-likelihood of $X_n$ is given by:

$$\frac{\partial \log P(X_n|W)}{\partial W_{ij}} = \mathbb{E}\left[Y[i]Y[j]|Y[1,\ldots,D] = X_n, W\right] - \mathbb{E}\left[Y[i]Y[j]|W\right] \tag{8}$$

2. Describe how you would estimate the gradient of the log-likelihood $\frac{\partial \log P(X_1,\ldots,X_N|W)}{\partial W_{ij}}$. In particular, how many runs of Gibbs sampling are required, and what are the details of each such run.

3. How would your answer to the previous question change if the weight matrix $W$ was constrained such that $W_{ij} = 0$ for all $i, j > D$ (i.e. all weights between hidden units are zero)?