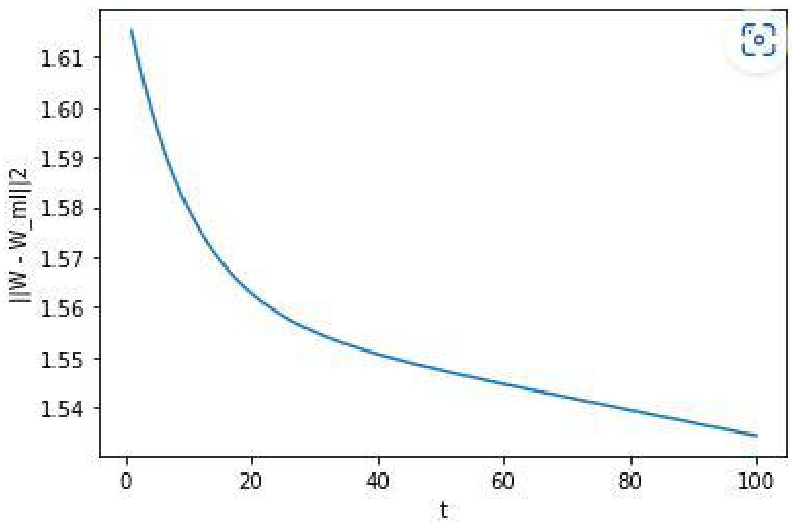
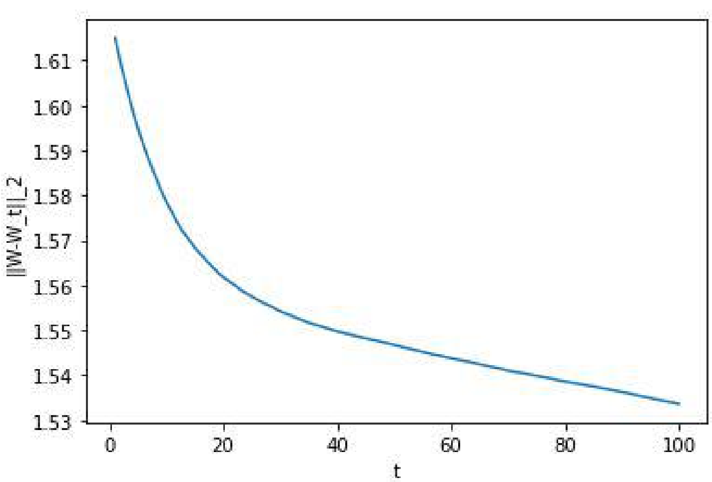
# Report for PRML Assignment 2

Q2.ii) Gradient Descent Algorithm Output




Here we first randomly initialize weight and the weight converges towards the Maximum likelihood value.As we take a small portion of gradient and then update the weights, the convergence is slow as there is quite a large number of very high dimensional datapoints in the given dataset.

Q2.iii) stochastic gradient descent




This plot is identical to Gradient Descent.There is not much noise in the approximated gradient computed from randomly sampled 100 points compared to the gradient obtained from the whole dataset.

2) i) let $\exists$ $x \in R^d$ $y \in R$ $w \in R^d$ $\epsilon \in \mathcal{N}(0, \sigma^2)$

such that $P(y|x) = w^T x + \epsilon$

Each $y$ is the value associated with each $x$ which contains some noise.

Computing $y$ given $x$ can be viewed as an estimation problem, if we assume that the values of noise comes from a Gaussian distribution with mean $0$ & variance $\sigma^2$.

Then we can maximize the likelihood of seeing $y$ given $x$ $\forall x$.

$I^{st}$ compute expectation of $y$ value

$$E[y_i] = E[w^T x_i] + E[\epsilon]$$

$$= w^T x_i + 0$$

$$= w^T x_i$$

Now writing likelihood fn in terms of $w$,

$$\mathcal{L}(w) = \prod_{i=1}^{n} e^{\frac{-(y_i - E[y_i])^2}{2\sigma^2}}$$

$$= \prod_{i=1}^{n} e^{\frac{(-y_i - w^T x_i)^2}{2\sigma^2}}$$

Taking log of likelihood

$$\log(L(w)) = \sum_{i=1}^{n} \frac{-(y_i - w^T x_i)^2}{2\sigma^2}$$

Req. to maximize this fn wrt $w$.

So $\max\limits_{w} \log(L(w))$

$$= \max_{w} \sum_{i=1}^{n} \frac{-(y_i - w^T x_i)}{2\sigma^2}$$

$$\Rightarrow \min_{w} \sum_{i=1}^{n} (w^T x_i - y_i)^2$$

If $X \in$ d-dimension & there are $n$ of them, then

$X \in R^{d \times n}$

Then we can rewrite the minimization problem as:

$$\min_{w \in R^d} \| X^T w - y \|_2^2$$

$$\min_{w \in R^d} (X^T w - y)^T (X^T w - y)$$

Rewriting as fn of $w$:

$$f(w) = (X^T w - y)^T (X^T w - y)$$

$$= w^T X X^T w - w^T X y - y^T X^T w + y^T y$$

Take derivative wrt $w$ & set to '0' To get optimum value for $w$,

$$\nabla f(w) = 2(XX^T) w - 2(Xy) = 0$$

$$\Rightarrow w_{ML} = (XX^T)^{-1} X y$$ } closed form soln for $w$ that maximizes likelihood fn.

Q1. i) As the input data-set consists of discrete data points which can take value 0 or 1, Bernoulli mixture model would have generated this data set.

The probability density function for the Bernoulli distribution is $p^x(1-p)^{1-x}$

where $p$: probability of success.

Probability Mass function:

$$P(X_i = x_i \mid Z_i = 0) = \begin{cases} P(X_i = 1 / Z_i = 0) & x_i = 1 \\ 1 - P(X_i = 1 / Z_i = 0) & x_i = 0 \end{cases}$$

$$= p(X_i = 1 / Z_i = 0)^{x_i} \cdot (1 - p(X_i = 1 / Z_i = 0))^{1-x_i}$$

$$P(X_i = x_i \mid Z_i = 1) = \begin{cases} P(X_i = 1 \mid Z_i = 1), & x_i = 1 \\ 1 - p(X_i = 1 / Z_i = 1) & x_i = 0 \end{cases}$$

$$= p(X_i = 1 \mid Z_i = 1)^{x_i} \cdot (1 - p(X_i = 1 \mid Z_i = 1))^{1-x_i}$$

$$p(Z_i = z_i) = \begin{cases} p(Z_i = 1) & z_i = 1 \\ 1 - p(Z_i = 1) & z_i = 0 \end{cases}$$

$$= p(Z_i = 1)^{z_i} (1 - p(Z_i = 1))^{1-z_i}$$