# COMS 4771 Machine Learning 2020
# Problem Set #1

Joseph High & Eliza Mace - `jph2185,emm2314`

April 11, 2020

## Problem 1: Statistical Estimators

**Joe: This is my version of Problem 1. Eliza's is on the following page. Let's revise and possibly combine later.**

**(i)** We are given that $x_1, \ldots, x_n$ are drawn independently from
$p(x|\theta = (a,b)) \propto \begin{cases} 1 & \text{if } \ a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$.

That is, $x_i \overset{\text{iid}}{\sim} \text{unif}[a,b]$ , $\forall i \in \{1, \ldots, n\}$. Then for each $i$, the pdf of $x_i$ is

$p(x_i|\theta) = \begin{cases} \frac{1}{b-a} & \text{if } \ a \leq x_i \leq b \\ 0 & \text{otherwise} \end{cases}$. Therefore, the likelihood function is

$$\mathcal{L}(\theta|X) = \prod_{i=1}^{n} p(x_i|\theta) = \prod_{i=1}^{n} \frac{1}{b-a} = \frac{1}{(b-a)^n}, \ \text{ for } \quad a \leq x_i \leq b \ \ \forall i \in \{1, \ldots, n\}$$

The constraints on $\theta$ can be written equivalently as $a \leq \min_i\{x_i\}$ and $b \geq \max_i\{x_i\}$.

The values of $a$ and $b$ that maximize $\frac{1}{(b-a)^n}$ are equivalent to the values of $a$ and $b$ that minimize $b-a$. Subject to the constraints, $b-a$ is minimized when $a = \min_i\{x_i\}$ and $b = \max_i\{x_i\}$ (both of which are feasible). The MLE estimate of $\theta = (a,b)$, denoted by $\theta_{ML}$, is then

$$\theta_{ML} = \arg\max_{\theta} \mathcal{L}(\theta|X) = \arg\max_{a \leq x_i \leq b} \frac{1}{(b-a)^n} = \arg\min_{a \leq x_i \leq b} (b-a) = (\min_i\{x_i\} \ , \ \max_i\{x_i\})$$

Therefore,

$$\theta_{ML} = (\min\{x_1, \ldots, x_n\}, \ \max\{x_1, \ldots, x_n\})$$

**(ii)** *Proof.* For an arbitrary, differentiable function $g$, let $\Gamma$ be such that $g : \Omega \longrightarrow \Gamma$, where $\Omega$ is the parameter space. That is, $\Gamma := \{\tau : g(\theta) = \tau\}$. For each $\tau \in \Gamma$, define $\Theta_\tau := \{\theta : g(\theta) = \tau\}$, and note that $\Theta_\tau \subseteq \Omega$. Finally, let $\hat{\tau}$ be the $MLE$ of $g(\theta)$. That is,

$$\hat{\tau} = \arg\max_{\tau \in \Gamma} \left( \max_{\theta \in \Theta_\tau} \log \mathcal{L}(\theta|\mathbf{x}) \right)$$

Since $\Theta_\tau \subseteq \Omega$, $\max\limits_{\theta\in\Theta_\tau} \log \mathcal{L}(\theta|\mathbf{x}) \leq \max\limits_{\theta\in\Omega} \log \mathcal{L}(\theta|\mathbf{x}) = \log \mathcal{L}(\theta_{ML}|\mathbf{x})$ , for all $\tau \in \Gamma$.

That is, since $\log \mathcal{L}(\theta|\mathbf{x})$ is maximized by $\theta_{ML}$ over all $\theta \in \Omega$, then it also maximizes $\log \mathcal{L}(\theta|\mathbf{x})$ over $\Theta_\tau \subseteq \Omega$, for all $\tau \in \Gamma$. More specifically,

$$\max_{\tau\in\Gamma}\left(\max_{\theta\in\Theta_\tau} \log \mathcal{L}(\theta|\mathbf{x})\right) = \max_{\theta\in\Omega} \log \mathcal{L}(\theta|\mathbf{x}) = \log \mathcal{L}(\theta_{ML}|\mathbf{x})$$

Then it must be the case that $\theta_{ML} \in \Theta_{\hat{\tau}} = \{\theta : g(\theta) = \hat{\tau}\}$
$\implies g(\theta_{ML}) = \hat{\tau}$. Hence, $g(\theta_{ML})$ is the MLE of $g(\theta)$.

$\square$

**(iii)** • *Consistent and unbiased*: (i) $\hat{\mu} = \dfrac{1}{N}\sum\limits_{i=1}^{N} X_i$ and (ii) a linear combination of the data with unequal weights on each data point:
$\hat{\mu} = \sum\limits_{i=1}^{N} \gamma_i X_i$ where $\gamma_i \neq \gamma_j$ and $\sum_{i=1}^{N} \gamma_i = 1$
For each estimator, we will show why each estimate is consistent and unbiased.

(i) $\boxed{\hat{\mu} = \dfrac{1}{N}\sum\limits_{i=1}^{N} X_i}$

Unbiased: $\mathbb{E}[\hat{\mu}] = \mathbb{E}\left[\dfrac{1}{N}\sum\limits_{i=1}^{N} X_i\right] = \dfrac{1}{N}\sum\limits_{i=1}^{N}\mathbb{E}[X_i] = \dfrac{1}{N}\sum\limits_{i=1}^{N}\mu = \mu$

Consistent:

$$\lim_{N\to\infty}\mathbb{E}[\hat{\mu}] = \lim_{N\to\infty}\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} X_i\right] = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\mu = \lim_{N\to\infty}\mu = \mu$$

and

$$\lim_{N\to\infty}\mathrm{Var}[\hat{\mu}] = \lim_{N\to\infty}\mathrm{Var}\left[\frac{1}{N}\sum_{i=1}^{N} X_i\right] = \lim_{N\to\infty}\frac{1}{N^2}\sum_{i=1}^{N}\mathrm{Var}[X_i] = \lim_{N\to\infty}\frac{N\sigma^2}{N^2} = \lim_{N\to\infty}\frac{\sigma^2}{N} = 0$$

(ii) $\boxed{\hat{\mu} = \mathbb{E}\left[\sum\limits_{i=1}^{N}\gamma_i X_i\right] \text{ where } \gamma_i \neq \gamma_j \text{ and } \sum\limits_{i=1}^{N}\gamma_i = 1}$
Unbiased:

$$\mathbb{E}[\hat{\mu}] = \mathbb{E}\left[\sum_{i=1}^{N}\gamma_i X_i\right] = \sum_{i=1}^{N}\gamma_i\mathbb{E}[X_i] = \sum_{i=1}^{N}\gamma_i\mu = \mu\sum_{i=1}^{N}\gamma_i = \mu\times 1 = \mu$$

<u>Consistent</u>:

$$\lim_{N\to\infty} \mathbb{E}[\hat{\mu}] = \lim_{N\to\infty} \mathbb{E}\left[\sum_{i=1}^{N} \gamma_i X_i\right] = \lim_{N\to\infty} \sum_{i=1}^{N} \gamma_i \mu = \lim_{N\to\infty} \mu = \mu$$

- *Consistent, but not unbiased*: (i) $\hat{\mu} = \dfrac{1}{N-1}\sum_{i=1}^{N} X_i$ and (ii) $\hat{\mu} = \dfrac{1}{N}\sum_{i=1}^{N} X_i + \dfrac{1}{N}$

For each estimator, we will show why each estimate is consistent and biased.

(i) $\boxed{\hat{\mu} = \frac{1}{N-1}\sum_{i=1}^{N} X_i}$

<u>Biased</u>: $\mathbb{E}[\hat{\mu}] = \mathbb{E}\left[\frac{1}{N-1}\sum_{i=1}^{N} X_i\right] = \frac{1}{N-1}\sum_{i=1}^{N} \mathbb{E}[X_i] = \frac{1}{N}\sum_{i=1}^{N} \mu = \frac{N\mu}{N-1} \neq \mu$

<u>Consistent</u>:

$$\lim_{N\to\infty} \mathbb{E}[\hat{\mu}] = \lim_{N\to\infty} \mathbb{E}\left[\frac{1}{N-1}\sum_{i=1}^{N} X_i\right] = \lim_{N\to\infty} \frac{1}{N-1}\sum_{i=1}^{N} \mu = \lim_{N\to\infty} \frac{N\mu}{N-1} = \mu$$

and

$$\begin{aligned}\lim_{N\to\infty} \mathrm{Var}[\hat{\mu}] &= \lim_{N\to\infty} \mathrm{Var}\left[\frac{1}{N-1}\sum_{i=1}^{N} X_i\right] \\ &= \lim_{N\to\infty} \left(\frac{1}{N-1}\right)^2 \sum_{i=1}^{N} \sigma^2 = \lim_{N\to\infty} \frac{N\sigma^2}{(N-1)^2} = \lim_{N\to\infty} \frac{\sigma^2}{N-2+\frac{1}{N}} = 0\end{aligned}$$

(ii) $\boxed{\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} X_i + \frac{1}{N}}$

<u>Biased</u>:

$$\begin{aligned}\mathbb{E}[\hat{\mu}] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} X_i + \frac{1}{N}\right] &= \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} X_i\right] + \mathbb{E}\left[\frac{1}{N}\right] = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[X_i] + \frac{1}{N} \\ &= \frac{1}{N}\sum_{i=1}^{N} \mu + \frac{1}{N} = \mu + \frac{1}{N} \neq \mu\end{aligned}$$

<u>Consistent</u>:

$$\lim_{N\to\infty} \mathbb{E}[\hat{\mu}] = \lim_{N\to\infty} \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} X_i + \frac{1}{N}\right] = \lim_{N\to\infty} \left(\frac{1}{N}\sum_{i=1}^{N} \mu \ + \ \frac{1}{N}\right) = \lim_{N\to\infty} \left(\mu + \frac{1}{N}\right) = \mu$$

$$\lim_{N\to\infty} \text{Var}[\hat{\mu}] = \lim_{N\to\infty} \text{Var}\left[\frac{1}{N}\sum_{i=1}^{N} X_i + \frac{1}{N}\right] = \lim_{N\to\infty} \frac{1}{N^2}\sum_{i=1}^{N} \text{Var}[X_i] = \lim_{N\to\infty} \frac{N\sigma^2}{N^2}$$
$$= \lim_{N\to\infty} \frac{\sigma^2}{N} = 0$$

- *Not consistent, but unbiased*: (i) $\hat{\mu} = X_k \in \{X_1, \dots, X_N\}$ and (ii) $\hat{\mu} = \frac{X_1+X_2}{2}$
  For each estimator, we will show why each estimate is not consistent, but unbiased.

  (i) $\boxed{\hat{\mu} = X_k}$
  Unbiased: $\mathbb{E}[\hat{\mu}] = \mathbb{E}[X_k] = \mu$
  Not consistent: Inconsistent since $X_k$ is fixed and will not change as $N \longrightarrow \infty$. That is,

  $$\lim_{N\to\infty} \text{Var}[\hat{\mu}] = \lim_{N\to\infty} \text{Var}[X_k] = \lim_{N\to\infty} \sigma^2 = \sigma^2 \neq 0$$

  (ii) $\boxed{\hat{\mu} = \frac{X_1+X_2}{2}}$
  Unbiased: $\mathbb{E}[\hat{\mu}] = \mathbb{E}\left[\frac{X_1+X_2}{2}\right] = \frac{1}{2}\mathbb{E}[X_1] + \frac{1}{2}\mathbb{E}[X_2] = \frac{1}{2}\mu + \frac{1}{2}\mu = \mu$
  Not consistent:

  $$\lim_{N\to\infty} \text{Var}[\hat{\mu}] = \lim_{N\to\infty} \text{Var}\left[\frac{X_1+X_2}{2}\right] = \lim_{N\to\infty} \frac{1}{2}\sigma^2 = \frac{\sigma^2}{2} \neq 0$$

- *Neither consistent, nor unbiased*: (i) $X_k + \alpha$ and (ii) $\alpha X_k$, for some $X_k \in \{X_1, \dots, X_N\}$ and a fixed constant $\alpha > 1$
  For each estimator, we will show why each estimate is neither consistent, nor unbiased.

  (i) $\boxed{\hat{\mu} = X_k + \alpha}$
  Biased: $\mathbb{E}[\hat{\mu}] = \mathbb{E}[X_k + \alpha] = \mu + \alpha \neq \mu$
  Not consistent:

  $$\lim_{N\to\infty} \mathbb{E}[\hat{\mu}] = \lim_{N\to\infty} \mathbb{E}[X_k + \alpha] = \lim_{N\to\infty} \mu + \alpha = \mu + \alpha \neq \mu$$

  and

  $$\lim_{N\to\infty} \text{Var}[\hat{\mu}] = \lim_{N\to\infty} \text{Var}[X_k + \alpha] = \lim_{N\to\infty} \text{Var}[X_k] = \lim_{N\to\infty} \sigma^2 = \sigma^2 > 0$$

  (ii) $\boxed{\hat{\mu} = \alpha X_k}$
  Biased: $\mathbb{E}[\hat{\mu}] = \mathbb{E}[\alpha X_k] = \alpha\mu \neq \mu$

Not consistent:

$$\lim_{N\to\infty} \mathbb{E}[\hat{\mu}] = \lim_{N\to\infty} \mathbb{E}[\alpha X_k] = \lim_{N\to\infty} \alpha\mu = \alpha\mu \neq \mu$$

and

$$\lim_{N\to\infty} \mathrm{Var}[\hat{\mu}] = \lim_{N\to\infty} \mathrm{Var}[\alpha X_k] = \lim_{N\to\infty} \alpha^2\sigma^2 = \alpha^2\sigma^2 > 0$$

# Problem 2: On Forecasting Product Demand

1.

$$\begin{aligned}
\pi(D) &= \int_0^{Q-1} [(P-C)\cdot D - C\cdot(Q-D)]\cdot f(D)dD + \int_Q^{\infty}(P-C)\cdot Q\cdot f(D)dD \\
&= \int_0^{Q-1} [(P-C)D - C(Q-D)]\cdot f(D)dD + (P-C)\cdot Q[1-\int_0^{Q-1} f(D)dD] \\
&= \int_0^{Q-1} P\cdot D\cdot f(D)dD - C\cdot Q\int_0^{Q-1} f(D)dD + (P-C)\cdot Q + (P-C)\cdot Q\int_0^{Q-1} f(D)dD \\
&= (P-C)\cdot Q + P\int_0^{Q-1} D\cdot f(D)dD + [(P-C)\cdot Q - C\cdot Q]\int_0^{Q-1} f(D)dD \\
&= (P-C)\cdot Q + P\int_0^{Q-1} D\cdot f(D)dD + [Q\cdot(P-2C)][F(Q-1)-F(0)]
\end{aligned}$$

2.

$$\frac{d\pi}{dQ} = (P-C) + P\cdot(Q-1)\cdot f(Q-1) + \frac{d}{dQ}[(P-2C)\cdot Q\cdot F(Q-1) - (P-2C)\cdot Q\cdot F(0)]$$

# Problem 3: Evaluating Classifiers

(i) We get an error when $x_i > t, y_i = y_2$ or when $x_i \leq t, y_i = y_1$

$$\begin{aligned} P[f_t(X) \neq y] &= P[f_t(\vec{x}) = y_1, Y = y_2|X = \vec{x}] + P[f_t(\vec{x}) = y_2, Y = y_1|X = \vec{x}] \\ &= P[x_i > t, Y = y_2|X = x_i] + P[x_i \leq t, Y = y_1|X = x_i] \end{aligned}$$

$f_t(x)$ is conditionally independent of $y$ given $x$

$$\begin{aligned} P[f_t(x) \neq y] &= P[x_i > t|X = x_i]P[Y = y_2|X = x_i] + P[x_i \leq t|X = x_i]P[Y = y_1|X = x_i] \\ &= \nVdash(x_i > t)P[Y = y_2, X = x_i] + (1 - \nVdash(x_i > t))P[Y = y_i|X = x_i] \end{aligned}$$

(ii) optimal threshold

(iii) bayes error rate

# Problem 4: Analyzing iterative optimization

**(i)** *Proof.* We first show that $M$ is *symmetric.*
Recall that a (square) matrix $M$ is symmetric $\iff M = M^\top$
Clearly, $M = A^\top A$ is a square matrix ($A^\top A$ a $d \times d$ matrix). Consider the following

$$M^\top = (A^\top A)^\top = A^\top (A^\top)^\top = A^\top A = M \implies M^\top = M$$

Thus, $M$ is symmetric.

To prove that $M$ is also *positive semi-definite*, it suffices to show that for any $x \in \mathbb{R}^d$, $x^\top M x \geq 0$. As such, consider an arbitrary vector $\mathbf{x} \in \mathbb{R}^d$ and let $\mathbf{w} = A\mathbf{x}$, where $\mathbf{w} \in \mathbb{R}^n$ and $\mathbf{w} = [w_1,\ w_2,\ \ldots, w_n]^\top$. We then have that,

$$\mathbf{x}^\top M \mathbf{x} = \mathbf{x}^\top A^\top A \mathbf{x} = (A\mathbf{x})^\top A\mathbf{x} = \|A\mathbf{x}\|_2^2 = \|\mathbf{w}\|_2^2 = \sum_{i=1}^{n} w_i^2 \geq 0$$

$\implies M$ is positive semi-definite. □

**(ii)** *Proof.* Proof by induction on $N$.

<u>Base case</u> ($N = 1, 2$)

For $N = 1$ we have

$$\begin{aligned}
\beta^{(1)} &= \beta^{(0)} + \eta A^\top (b - A\beta^{(0)}) && \text{(by definition of the Richardson iteration)}\\
&= \eta A^\top b = \eta v && \text{(since } \beta^{(0)} \text{ is the zero vector and } v = A^\top b)\\
&= \eta I v = \eta \underbrace{(I - \eta M)^0}_{=I} v && \text{(Note: } (I - \eta M) \text{ is a square matrix since } I \text{ and } M \text{ are square)}\\
&= \eta \sum_{k=0}^{0} (I - \eta M)^k v && \text{Thus, it holds for } N = 1.
\end{aligned}$$

For $N = 2$ we have

$$\begin{aligned}
\beta^{(2)} &= \beta^{(1)} + \eta A^\top (b - A\beta^{(1)}) = \eta v + \eta (A^\top b - A^\top A \eta v) && \text{(since } \beta^{(1)} = \eta v \text{ from the above)}\\
&= \eta v + \eta (v - M \eta v) && \text{(since } M = A^\top A \text{ and } v = A^\top b)\\
&= \eta \underbrace{(I - \eta M)^0}_{=I} v + \eta (I - \eta M) v && (\eta \text{ a real number})\\
&= \eta [(I - \eta M)^0 v + (I - \eta M)^1 v]\\
&= \eta \sum_{k=0}^{1} (I - \eta M)^k v && \text{Thus, it holds for } N = 2.
\end{aligned}$$

(Inductive hypothesis) Now assume the result holds for $k = 1, 2, \ldots, N - 1$.
That is, assume the following holds:

$$\beta^{(N-1)} = \eta \sum_{k=0}^{N-2} (I - \eta M)^k v$$

From the definition of the Richardson iteration, the $N^{th}$ iterate is

$$\begin{aligned}
\beta^{(N)} &= \beta^{(N-1)} + \eta A^\top(b - A\beta^{(N-1)}) = \beta^{(N-1)} + \eta(v - M\beta^{(N-1)}) \\
&= \eta\sum_{k=0}^{N-2}(I-\eta M)^k v + \eta\left[v - M\left(\eta\sum_{k=0}^{N-2}(I-\eta M)^k v\right)\right] && \text{(from the induction hypothesis)} \\
&= \eta\sum_{k=0}^{N-2}(I-\eta M)^k v - \eta M\left(\eta\sum_{k=0}^{N-2}(I-\eta M)^k v\right) + \eta v \\
&= (I-\eta M)\left(\eta\sum_{k=0}^{N-2}(I-\eta M)^k v\right) + \eta v \\
&= \eta\sum_{k=0}^{N-2}(I-\eta M)^{k+1} v + \eta v \\
&= \eta\sum_{k=1}^{N-1}(I-\eta M)^k v + \eta v && \text{(rearrange indices)} \\
&= \eta\sum_{k=1}^{N-1}(I-\eta M)^k v + \eta\underbrace{(I-\eta M)^0}_{=I} v && \text{(adding } k=0 \text{ summand)} \\
&= \eta\sum_{k=0}^{N-1}(I-\eta M)^k v
\end{aligned}$$

Hence, $\beta^{(N)} = \eta\sum_{k=0}^{N-1}(I-\eta M)^k v$ □

**(iii)** We are given that the eigenvalues of $M$ are $\lambda_1, \lambda_2, \ldots, \lambda_d$. Then, the eigenvalues of $I - \eta M$ are $1 - \eta\lambda_i$, for all $i = 1, \ldots, d$. Indeed, without loss of generality, let $\mathbf{x}$ be the eigenvector associated with $\lambda_i$, then $M\mathbf{x} = \lambda_i\mathbf{x} \implies (\eta M)\mathbf{x} = (\eta\lambda_i)\mathbf{x} \implies I\mathbf{x} - (\eta M)\mathbf{x} = I\mathbf{x} - \eta\lambda_i\mathbf{x} = 1\mathbf{x} - \eta\lambda_i\mathbf{x} \implies (I - \eta M)\mathbf{x} = (1 - \eta\lambda_i)\mathbf{x}$.

We also claim that since $((1-\eta\lambda_i),\ \mathbf{x})$ is the eigenvalue–eigenvector pair for $(I-\eta M)$, then $((1-\eta\lambda_i)^k,\ \mathbf{x})$ is the eigenvalue–eigenvector pair for $(I-\eta M)^k\ ,\ k \in \mathbb{N}\cup\{0\}$.

*Proof of claim*: For any $i \in \{1, 2, \ldots, d\}$ we have

$$\begin{aligned}
(I-\eta M)\mathbf{x} = (1-\eta\lambda_i)\mathbf{x} &\implies (I-\eta M)^2\mathbf{x} = (1-\eta\lambda_i)(I-\eta M)\mathbf{x} = (1-\eta\lambda_i)^2\mathbf{x} \\
&\implies (I-\eta M)^3\mathbf{x} = (1-\eta\lambda_i)^2(I-\eta M)\mathbf{x} = (1-\eta\lambda_i)^3\mathbf{x} \\
&\ \vdots \\
&\implies (I-\eta M)^k\mathbf{x} = (1-\eta\lambda_i)^{k-1}(I-\eta M)\mathbf{x} = (1-\eta\lambda_i)^k\mathbf{x}
\end{aligned}$$

□

Using the above results, we have the following

$$\begin{aligned}
&\eta I\mathbf{x} + \eta(I-\eta M)\mathbf{x} + \eta(I-\eta M)^2\mathbf{x} + \cdots + \eta(I-\eta M)^{N-1}\mathbf{x} = \\
&\qquad\qquad \eta\mathbf{x} + \eta(1-\eta\lambda_i)\mathbf{x} + \eta(1-\eta\lambda_i)^2\mathbf{x} + \cdots + \eta(1-\eta\lambda_i)^{N-1}\mathbf{x}
\end{aligned}$$

$$\implies \eta(I + (I - \eta M) + (I - \eta M)^2 + \cdots + (I - \eta M)^{N-1})\mathbf{x} = \\ \eta(1 + (1 - \eta\lambda_i) + (1 - \eta\lambda_i)^2 + \cdots + (1 - \eta\lambda_i)^{N-1})\mathbf{x}$$

$$\implies \left(\eta \sum_{k=0}^{N-1} (I - \eta M)^k\right)\mathbf{x} = \left(\eta \sum_{k=0}^{N-1} (1 - \eta\lambda_i)^k\right)\mathbf{x} = \left(\frac{1 - (1 - \eta\lambda_i)^N}{\lambda_i}\right)\mathbf{x}\,,\ \forall i = 1, 2, \ldots, d$$

Thus, the eigenvalues of $\eta \sum_{k=0}^{N-1} (I - \eta M)^k$ are

$$\frac{1 - (1 - \eta\lambda_1)^N}{\lambda_1}\,,\ \frac{1 - (1 - \eta\lambda_2)^N}{\lambda_2}\,,\ \ldots\,,\ \frac{1 - (1 - \eta\lambda_d)^N}{\lambda_d}$$

.

**(iv)** *Proof.*
Note that

$$\begin{aligned}
\hat{\beta} - \beta^{(N)} &= \left(\hat{\beta} + \eta A^\top(b - A\hat{\beta})\right) - \left(\beta^{(N-1)} + \eta A^\top(b - A\beta^{(N-1)})\right) \\
&= \left(\hat{\beta} + \eta v - \eta M\hat{\beta}\right) - \left(\beta^{(N-1)} + \eta v - \eta M\beta^{(N-1)}\right) \\
&= \left((I - \eta M)\hat{\beta} + \cancel{\eta v}\right) - \left((I - \eta M)\beta^{(N-1)} + \cancel{\eta v}\right) \\
&= (I - \eta M)(\hat{\beta} - \beta^{(N-1)}) \\
&= (I - \eta M)^2((\hat{\beta} - \beta^{(N-2)}) \\
&= (I - \eta M)^3(\hat{\beta} - \beta^{(N-3)}) \\
&\vdots \\
&= (I - \eta M)^N(\hat{\beta} - \beta^{(0)}) \\
&= (I - \eta M)^N\hat{\beta}
\end{aligned}$$

Now consider,

$$\begin{aligned}
\|\beta^{(N)} - \hat{\beta}\|_2^2 = \|\hat{\beta} - \beta^{(N)}\|_2^2 &= \|(I - \eta M)^N\hat{\beta}\|_2^2 \\
&\leq \|(I - \eta M)^N\|_2^2\|\hat{\beta}\|_2^2 \\
&\leq (\|(I - \eta M)\|_2^2)^N\|\hat{\beta}\|_2^2 \\
&= \|I - \eta M\|_2^{2N}\|\hat{\beta}\|_2^2 \\
&\leq (1 - 2\eta\lambda_{min})^N\|\hat{\beta}\|_2^2 \\
&\leq e^{-2\eta\lambda_{min}N}\|\hat{\beta}\|_2^2
\end{aligned}$$

□

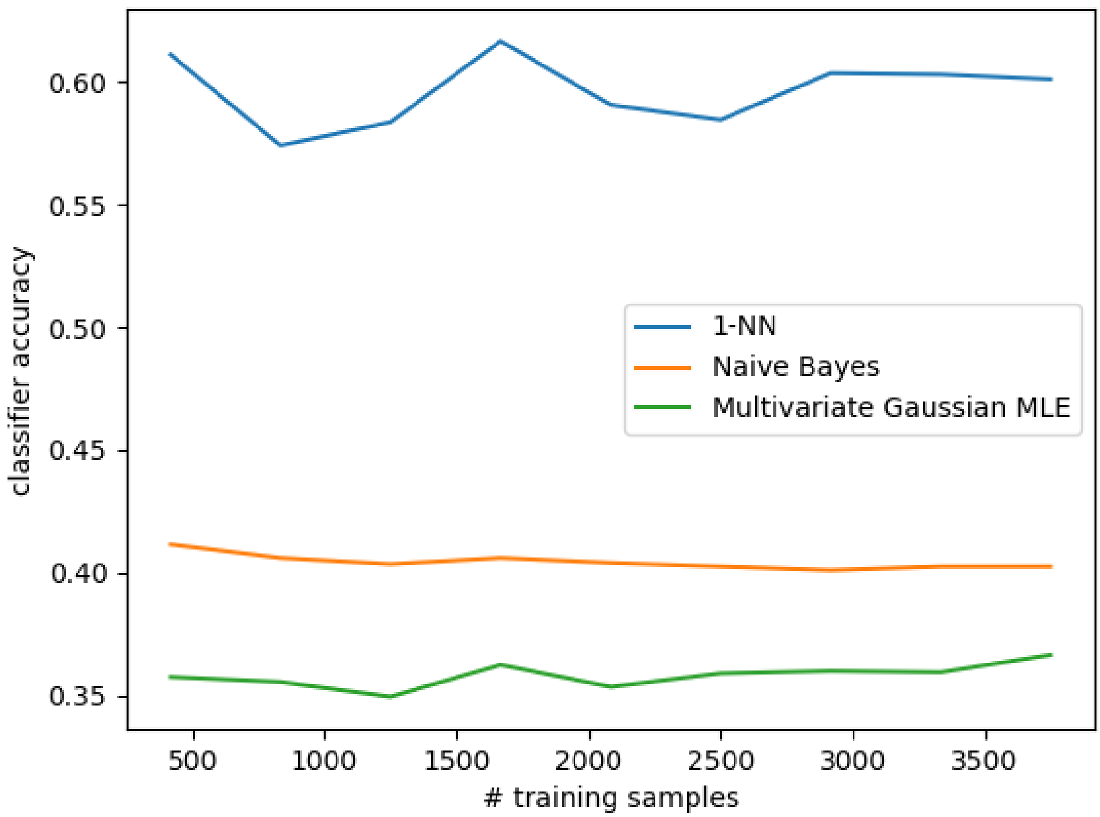
# Problem 5: Designing socially aware classifiers

(i) It is not enough just to remove the sensitive attribute $A$ from the dataset because it is possible that other attributes in the feature vector are correlated with this attribute.

(ii) Demographic parity

(iii) equivalence relationship

(iv) classifiers




(v)

(vi) positive rate across sensitive attribute

(vii) real-world

# Problem 6: Email spam classification case study

(i) Bag-of-words

(ii) classifiers

(iii) Naive bayes is best!