# Assignment 1

## 1 The Airline Question (40 points)

### 1.

Let $X_i$ equal 1 if customer $i$ arrives and 0 if the customer does not arrive. Then $X_i$ are i.i.d. bernoulli with bias $p = 0.95$. We can explicitly compute the probability that the flight is overbooked. This happens with probability,

$$\mathbb{P}(X_1 = X_2 = \cdots = X_{100} = 1) = \Pi_{i=1}^{100}\mathbb{P}(X_i = 1) = p^{100} = 0.00592052922$$

Where we have used independence.

### 2.

### a)

Let $Y = \sum_{i=1}^{10000} X_i$ denote the amount of show-ups for the 10000 flight reservations, and $Y' = \sum_{i=1}^{100} X_i'$ denote the number of show-ups for the 100 person flight. Consider the events, $A$: in the 10000 sample everyone shows up independently with probability $p$, and we observe 95% show-ups and $B$: Everybody shows up for the 99-seat flight. $A$ and $B$ are independent, and can be written as $\mathbb{P}(A) = \mathbb{P}(Y = 9500)$ and $\mathbb{P}(B) = \mathbb{P}(Y' = 100)$. We want to bound the probability of observing, $A$ and $B$ simoultaneosly. Assume that $p$ is known. Notice that $Y \sim Bin(10000, p)$ and $Y' \sim Bin(100, p)$. By independence we have that

$$\mathbb{P}(A \cap B) = \mathbb{P}(Y = 9500)\mathbb{P}(Y' = 100) = \binom{10000}{9500} p^{9500}(1-p)^{500}p^{100} = \binom{10000}{9500} p^{9600}(1-p)^{500}$$

The 'worst case' value for $p$ is the $p$, that maximizes this probability. Define $f(p) = \log(\mathbb{P}(A \cap B))$. Maximizing this is equivalent to maximizing $\mathbb{P}(A \cap B)$ as the logarithm is monotone, non-decreasing transformation. We have that

$$f'(p) = \frac{9600}{p} - \frac{500}{1-p} = 0 \Leftrightarrow p = \frac{96}{101}$$

And this is indeed a maximum, as

$$f''(p) = -\frac{9600}{p^2} - \frac{500}{(1-p)^2} < 0$$

We therefore have the worst case $p$ being $p^* = \frac{96}{101}$. Let $\mathbb{P}_{p^*}(\cdot)$ denote the probability measure that $p^*$ induces on $Y$ and $Y'$. .We can now bound $\mathbb{P}(A \cap B)$ by using independence and by using the fact that probabilities are bounded by 1,

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B) \leq \max_p \mathbb{P}(A)\mathbb{P}(B) \leq \mathbb{P}_{p^*}(Y' = 100) = \left(\frac{96}{101}\right)^{100} = 0.006237$$

Which is the desired bound.

### b)

Assume that we have sampled 10100 passengers. That is, reusing the same notation as in the previous exercises, we have the sample $\mathcal{X} = \{X_i : i \in \{1, \cdots, 10100\}\}$. Let $(Z_i)_{i\in\{1,\cdots,100\}}$ be sampled without replacement, but uniformly from $\mathcal{X}$. That is, $Z_i$ represents the event, that passenger $i$ shows up for the 99 seat flight. We are interested in bounding the probability

$$\mathbb{P}(A) = \mathbb{P}(\sum_{i=1}^{100} Z_i \quad \textbf{and} \quad \sum_{i=100}^{10100} X_i = 9600)$$

By the definition of conditional expectation,

$$\mathbb{P}\left(\sum_{i=1}^{100} Z_i \quad \textbf{and} \quad \sum_{i=100}^{10100} X_i = 9600\right) = \mathbb{P}\left(\sum_{i=1}^{100} Z_i = 100 \quad \Big| \quad \sum_{i=1}^{10100} X_i = 9600\right)\mathbb{P}\left(\sum_{i=1}^{10100} X_i = 9600\right)$$

Again, using the fact that probabilities are less than 1,

$$\mathbb{P}(A) \leq \mathbb{P}\left(\sum_{i=1}^{100} Z_i = 100 \quad \Big| \quad \sum_{i=1}^{10100} X_i = 9600\right)$$

We can actually compute this probability explicitly, by a combinatorial argument. Without loss of generality, assume that the passengers arrive (or does not arrive) to the flight according to their index. That is $Z_i$ represents (in sequence) the $i$'th passenger arriving or not arriving. Notice that for $\sum_{i=1}^{100} Z_i = 100$, we have to have all $Z_i = 1$ for all $i$. When we have that $\sum_{i=1}^{10100} X_i = 9600$, we have that the probability that $Z_1 = 1$ is $\frac{9600}{10100}$. Since we are sampling without replacement, this means that the probability for $Z_2 = 1$ is $\frac{9599}{10099}$, and so on. This gives us the that,

$$\mathbb{P}\left(\sum_{i=1}^{100} Z_i = 100 \quad \Big| \quad \sum_{i=1}^{10100} X_i = 9600\right) = \prod_{i=0}^{99} \frac{9600 - i}{10100 - i} = 0.006078897097$$

Which gives us a bound of around 0.00608.

## 2 The Growth Function (10 points)

### 1.

There are $2^n$ ways to label $n$ points as either 0 or 1. Therefore $m_{\mathcal{H}}(n) \leq 2^n$. When $|\mathcal{H}| = M$, there are at most $M$ different labellings of $n$ points. Therefore $m_{\mathcal{H}}(n) \leq M$. All in all we have that

$$m_{\mathcal{H}}(n) \leq \min\{2^n, M\}$$

### 2.

Assume $|\mathcal{H}| = 2$. We have that $m_{\mathcal{H}}(n) \leq 2$ per 1. Consider the case where we have one data point, and 2 distinct prediction rules for labelling that point as either -1 or 1. Let $h_-$ be the prediciton rule that always assigns $-1$ and $h_+$ the prediction rule that always assigns 1. This produces two dichotomies, and hence $m_{\mathcal{H}}(n) \geq 2$. We conclude that $m_{\mathcal{H}}(n) = 2$.

### 3.

$m_{\mathcal{H}}(n)$ is the maximal number of dichotomies $\mathcal{H}$ can generate on $n$ arbitrary datapoints binarily. Now consider the set of points $\mathcal{X} = \{x_1, x_2, \cdots, x_n\}$ and the set of points $\mathcal{X}' = \{x_{n+1}, x_{n+2}, \cdots, x_{2n}\}$. Now consider that if we pick a labelling pattern produced by $\mathcal{H}$ on $\mathcal{X}$, then we have up to $m_{\mathcal{H}}(n)$ possibilities to pair this labelling with from the set of labelling patterns produced by $\mathcal{H}$ on $\mathcal{X}'$. Note that there can be less, if some labellings exclude each other - e.g. points lying on the same side of a linear seperator getting different labels. This applies to all labelling paterns produced by $\mathcal{H}$ on $\mathcal{X}$, of which there are at most $m_{\mathcal{H}}(n)$. That is, the number of ways to label $2n$ points, must be less than the ways of combining labellings from two sets of all possible labellings of $n$ points. All this says, is that

$$m_{\mathcal{H}}(2n) \leq m_{\mathcal{H}}(n)^2$$

Which is what we wanted to show.

## KKT Conditions (20 points)

### (a)

We can rewrite $P$ in standard form as

$$\begin{aligned} &\min \sum_{i=1}^{d} x_i \log \frac{x_i}{a_i} \\ \text{s.t.} \quad & 1 - \sum_{i=1}^{d} x_i \leq 0 \\ & -x_i \leq 0, \quad \forall i = 1, \cdots, d \\ & x_i - 1 \leq 0, \quad \forall i = 1, \cdots, d \end{aligned}$$

### (b)

We are optimizing over the set $[0,1]^d$ which is clearly convex. We have that sums of convex functions are convex. This follows easily from the definition of convexity. The objective function $f(x) = \sum_{i=1}^{d} x_i \log \frac{x_i}{a_i}$, is a sum of convex functions, $f_i(x_i) = x_i \log \frac{x_i}{a_i}$. We obtain the convexity of $f_i$ by noting that $f_i''(x_i) = \frac{1}{x_i} > 0$, when $x_i \in [0,1]$.

The inequality constraints involve affine functions, such that these are also convex. All in all we conclude that $P$ is a convex problem.

### (c)

We write down the Lagrangian

$$\mathcal{L}(x, \lambda, \mu, \gamma) = \sum_{i=1}^{d} x_i \log \frac{x_i}{a_i} - \sum_{i=1}^{d} \lambda_i x_i + \sum_{i=1}^{d} \mu_i (x_i - 1) + \gamma(1 - \sum_{i=1}^{d} x_i)$$

### (d)

We write down the KKT conditions. We first have stationarity, for $x_i$ we have that

$$\frac{\partial \mathcal{L}}{\partial x_i} = \log \frac{x_i}{a_i} + 1 - \lambda_i + \mu_i - \gamma$$

Setting this equal to 0 yields that for $x_i^*$ we have that

$$x_i^*(\lambda, \mu, \gamma) = a_i \exp(\lambda_i + \gamma - \mu_i - 1), \quad \forall i = 1, \cdots, d$$

For primal feasibility we have that

$$\begin{aligned} 1 - \sum_{i=1}^{d} x_i^* &\leq 0 \\ -x_i^* &\leq 0, \quad \forall i = 1, \cdots, d \\ x_i^* - 1 &\leq 0, \quad \forall i = 1, \cdots, d \end{aligned}$$

For dual feasibility we have that $\lambda, \mu, \gamma \geq 0$, where there inequality means coordinatewise greater than 0.

For complementary slackness we have that

$$\begin{aligned} \gamma(1 - \sum_{i=1}^{d} x_i^*) &= 0 \\ -\lambda_i x_i^* &= 0, \quad \forall i = 1, \cdots, d \\ \mu_i(x_i^* - 1) &= 0, \quad \forall i = 1, \cdots, d \end{aligned}$$

### (e)

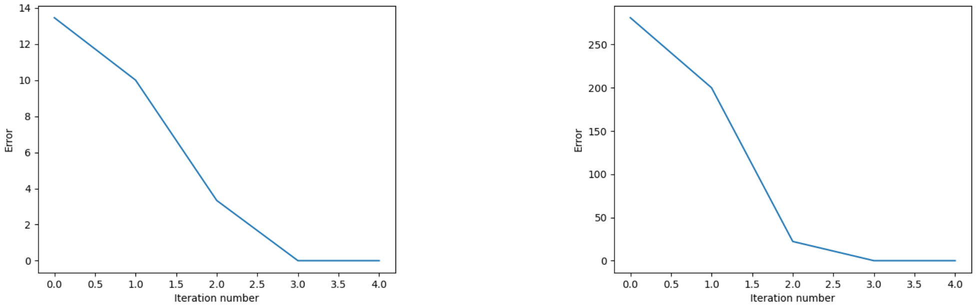
The dual function is given by $D(\lambda, \mu, \gamma) = \min_x L(x, \lambda, \mu, \gamma) = L(x^*(\lambda, \mu, \gamma), \lambda, \mu, \gamma)$. Per (d) we can thus write

$$D(\lambda, \mu, \gamma) = L\left(\left[a_i \exp(\lambda_i + \gamma - \mu_i - 1)\right]_{i=1,\cdots,d}, \lambda, \mu, \gamma\right)$$

## Gradient Descent (30 Points)

### (a)

$f$ takes values from $\mathbb{R}^2$, which is convex. Notice that $f$ is a sum of two convex functions when $\gamma > 0$, as the second derivative of $(x_1 + 1)^2$ w.r.t. $x_1$ is $2 > 0$, and the second derivative of $\gamma x_2^2$ w.r.t. $x_2$ is $2\gamma > 0$. This shows that $P_1$ is convex. Notice also that as we are dealing with a sum of two quadratic functions, $f \geq 0$, and we easily see that $f(x) = 0$ iff $x^* = (-1, 0)$. We plot the convergence error curves below, when $\gamma = 2$. For the implementation we declare convergence of the algorithm, when the distance between $x_t$ and $x_{t-1}$ is less than 0.00000001 in the Euclidean norm. We use the point $x_0 = (-10, 10)$ for initialisation

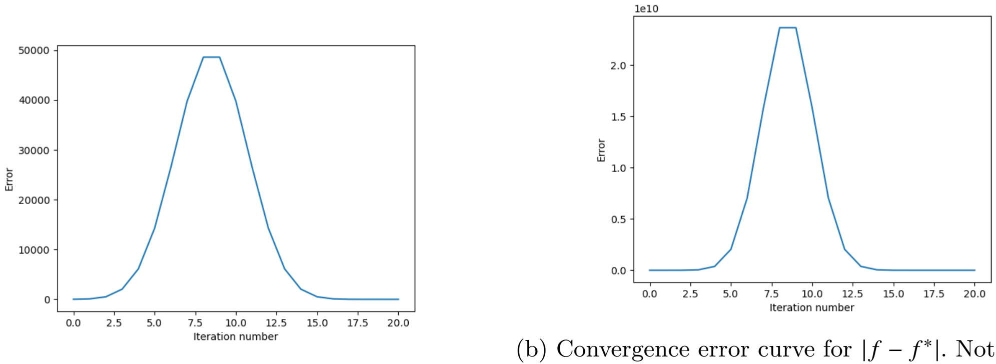


(a) Convergence error curve for $||x_t - x^*||_2$

(b) Convergence error curve for $|f - f^*|$

Figure 1: Plots of convergence error curves when $\gamma = 2$

We do the same for $\gamma = 10$.



(a) Convergence error curve for $||x_t - x^*||_2$

(b) Convergence error curve for $|f - f^*|$. Notice $y$-axis scaled by $10^10$

Figure 2: Plots of convergence error curves when $\gamma = 10$

## (b)

$f$ is once again defined on the convex set $\mathbb{R}^2$. To show the convexity of $f$, a lemma is helpful, to avoid doing too many computations.

**Lemma**: Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be convex and $g : \mathbb{R} \longrightarrow \mathbb{R}$ be convex and non-decreasing. then $g$ composed with $f$ is convex.

**Proof**: By the convexity of $f$ and using the fact that $g$ is non-decreasing,

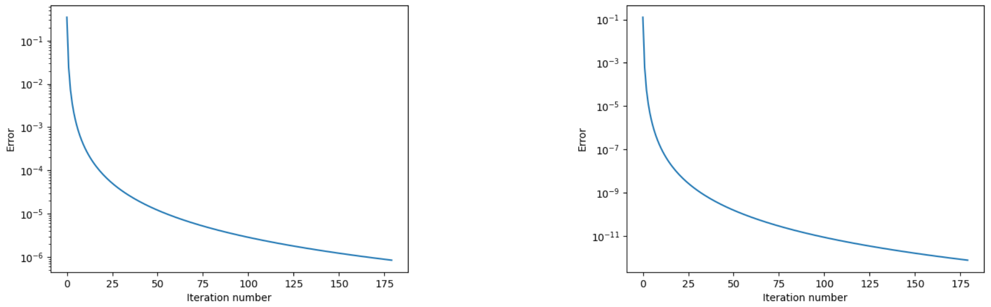
$$g(f(\lambda x) + f((1 - \lambda)y)) \leq g(\lambda f(x) + (1 - \lambda)f(y))$$

And by the convexity of $g$,

$$g(\lambda(f(x) + (1 - \lambda)f(y))) \leq \lambda f(g(x)) + (1 - \lambda)f(g(y))$$

Which is what we wanted to show. Now notice that the exponential function is convex, $(e^x)'' = e^x > 0$. Thus $\exp(\sum_{i=1}^n x_i a_i + b)$ is convex, as $\sum_{i=1}^n x_i a_i + b$ is affine, and hence convex. Since $f$ is a sum of exponential functions composed with affine functions, $f$ is convex, and $P_2$ is a convex problem.

Below we plot convergence error curves for gradient descent. We are a bit careful when choosing our initialisation, as the gradients can get very large and hence numerically unstable when dealing with exponential functions. By looking at $f$ we see that the gradient will probably be relatively small when $x_0 = (0, 0)$, and therefore use this for our initialisation. We plot on a semi-log scale.



(a) Convergence error curve for $||x_t - x^*||_2$

(b) Convergence error curve for $|f - f^*|$

Figure 3: Plots of convergence error curves for $f$ in (b)