For $x = 5$, $\lambda'(p) = \frac{5}{p} > 0$ and $\lambda(p)$ is increasing.

So for $\Theta = [0, 1]$, the MLE is $\frac{X}{5}$.

One context in which the likelihood equations are especially nice is the exponential family - for $X_1, X_2, \ldots, X_n$:

$$f(x | \eta) = \exp \left( -n A(\eta) + \sum_{i=1}^{n} h(x_i) \exp \sum_{j=1}^{k} \eta_j \left( \sum_{i=1}^{n} T_j(x_i) \right) \right)$$

So

$$L(\eta) = -n A(\eta) + \sum_{i=1}^{n} h(x_i) \exp \sum_{j=1}^{k} \eta_j \left( \sum_{i=1}^{n} T_j(x_i) \right)$$

and

$$\frac{\partial}{\partial \eta_j} L(\eta) = - \frac{\partial}{\partial \eta_j} A(\eta) + \sum_{i=1}^{n} T_j(x_i)$$

By corollary 1.5.1, B+D, $A \in E$ has nonempty interior.

Thus, the $j$th of these equations is

$$-n E_\eta T_j(X) + \sum_{i=1}^{n} T_j(x_i) = 0$$

$$E_\eta T_j(X) = \frac{1}{n} \sum_{i=1}^{n} T_j(x_i)$$

The likelihood equation.
So, the ML estimating equation in an exponential family with nonempty interior in $E$ is "set the theoretical mean of $T$ equal to the empirical mean of $T" - a natural exponential family is "MOM".

The question of whether there must be a solution to the exponential family likelihood equations and if any solution must be unique and maximize the likelihood is discussed in B&J Section 2.3 - on simple result from there is

Corollary 2.3.2 If the equations

$$E_y T_j(X) = \frac{1}{n} \sum_{i=1}^{n} T_j(x_i) \quad j = 1, 2, \ldots, k$$

have a solution $\hat{y}$ in the interior of $E$ it is the unique MLE of $y$.

Both for application in exponential families where the "natural" parameterization isn't the "standard" or "usual" parameterization and in other contexts, there is the lemma

**Lemma (Problem 2.2.16a)**

Clearly $\hat{\theta}$ maximizes $f(x|\theta)$ over $\Theta$.

$\Rightarrow \hat{\gamma} = h(\hat{\theta})$ maximizes $f^*(x|\gamma)$ over $h(\Theta)$. 

![Diagram](image-url)
So, e.g., \( \hat{\theta} \) MLE of natural parameter \( \theta \) can be written as \( h(\hat{\theta}) = \theta \) for a 1-1 function \( h \). Then \( \hat{\theta} = h(\hat{\theta}) \) is the MLE of \( \theta \).

Solving the likelihood equations

\[ \nabla l(\theta) = 0 \]

is "just" a numerical analysis problem — various iterative procedures (bisection, Newton–Raphson, etc.) might be applied to do the job — mostly. Their discussion is not a topic for 543 — one procedure for the problem has some probabilistic/statistical content and can/should be touched here is the "EM algorithm" —

Basic idea: Sometimes an observable \( Y \) has a nasty log likelihood \( l_Y(\theta) \) but could be thought of as distributionally equivalent to \( S(X) \) for some (potentially computationally fictitious) \( X \) for which computation and optimization of

\[ E_\theta \left[ l_X(\theta) \mid S(X) = y \right] \]

is feasible — In such cases, I might

(1) pick some starting value \( \theta^{(0)} \)
1. Find
   \[ E_{\Theta^{(0)}} \left[ \ell(\Theta) \mid S(X) = y \right] \]

M-step
2. Optimize this as a function of \( \Theta \) to find \( \Theta^{(1)} \)
3. Replace \( \Theta^{(0)} \) with \( \Theta^{(1)} \)
4. Iterate to convergence

Example

\[ X \sim \exp(\lambda) \]

\[ Y = X I[1 < x < 2] + 1 I[x \leq 1] + 2 I[2 \leq x] \]

(\( Y \) a censored version of \( X \)) — Suppose \( Y_1, Y_2, \ldots, Y_n \) iid with this CDF.

\[ f(y \mid x) = \begin{cases} 
1 - e^{-\lambda} & \text{if } y = 1 \\
\lambda e^{-\lambda y} & 1 < y < 2 \\
e^{-2\lambda} & y = 2
\end{cases} \]

If \( Y_i = 1 \) then \( X_i \) has density \( \alpha x e^{-\lambda x} \) on \((0, 1]\) and if \( Y_i = 2 \) \( X_i \) has density \( \alpha \lambda e^{-\lambda x} \) on \([2, \infty)\).

\[ \ell_X(\lambda) = n \log \lambda - \lambda \sum X_i \]

Then for a particular \( \lambda_0 \) conditioned on \( Y = y \)
\[ \sum_{i=1}^n \text{has mean} \]

\[ \sum_{y_i \in \{1, 2\}} y_i + \# \left[ y_i = 1 \right] E_{\lambda_0} \left[ X | X \leq 1 \right] \\
+ \# \left[ y_i = 2 \right] E_{\lambda_0} \left[ X | X > 2 \right] \]

\[ E_{\lambda_0} \left[ X | X \leq 1 \right] = \frac{1}{1 - e^{-\lambda_0}} \int_0^1 x e^{-\lambda_0 x} \, dx \]

\[ = \frac{1}{1 - e^{-\lambda_0}} \left[ -e^{-\lambda_0 x} \right]_0^1 + \int_0^1 e^{-\lambda_0 x} \, dx \]

\[ = \frac{1}{1 - e^{-\lambda_0}} \left[ -e^{-\lambda_0} + \frac{1}{\lambda_0} \right] \]

\[ = \frac{1}{\lambda_0} - \frac{e^{-\lambda_0}}{(1 - e^{-\lambda_0})} \]

\[ E_{\lambda_0} \left[ X | X > 2 \right] = \frac{1}{e^{-2\lambda_0}} \int_2^\infty x e^{-\lambda_0 x} \, dx \]

\[ = e^{2\lambda_0} \left[ -xe^{-\lambda_0 x} \right]_2^\infty + \int_2^\infty e^{-\lambda_0 x} \, dx \]

\[ = e^{2\lambda_0} \left[ 2e^{-2\lambda_0} + \frac{1}{\lambda_0} e^{-2\lambda_0} \right] \]

\[ = 2 + \frac{1}{\lambda_0} \]

i.e. the \( \lambda_0 \) conditional mean of \( l_X(\lambda) \) given \( Y = y \) is

\[ n \log \lambda - \lambda \left[ \frac{\sum_{y_i \in \{1, 2\}} y_i}{\text{y belongs to } \{1, 2\}} \right] \left( \frac{\lambda_0}{1 - e^{-\lambda_0}} \right) \]

\[ + \# \left[ y_i = 2 \right] \left( 2 + \frac{1}{\lambda_0} \right) \]

\[ \text{E-step} \]
which is maximized at \( \lambda'' = \frac{n}{n} \) [above]

So ultimately, an iterative G-M algorithm sets

\[
\lambda^{(i+1)} = \frac{n}{\sum_{y_i \in \{1,2\}} y_i + \# \{ y_i = 1 \} \left( \frac{1}{\lambda'} - \frac{e^{-\lambda'}}{1 - e^{-\lambda'}} \right) + \# \{ y_i = 2 \} \left( 2 + \frac{1}{\lambda'} \right)}
\]

and iterates to convergence as opposed to straight-up optimization of

\[
\ell_{\lambda}(\lambda) = \# \{ y_i = 1 \} \log (1 - e^{-\lambda}) + \# \{ y_i < 2 \} \log \lambda - \lambda \sum_{y_i \in \{1,2\}} y_i + \# \{ y_i = 2 \} \epsilon \lambda
\]
Example - k possible outcomes, n independent identical trials

\[ P_j = \text{probability that any trial produces outcome } j \]

\[ X_{ij} = I \left[ \text{trial } i \text{ produces outcome } j \right] \]

\[ n_j = \sum_{i=1}^{n} X_{ij} = \# \text{ of trials producing outcome } j \]

Suppose that \( p \) has each \( p_i \geq 0 \) with \( \sum p_i = 1 \)

\[ f(x \mid p) = \frac{k}{n} \sum_{j=1}^{n} n_j \]

\( f(x \mid p) = \left( \sum_{j=1}^{n} x_{ij} = 1, \, n_j \geq 0 \text{ integers} \right) \)

MLE for \( p \) is (as it turns out)

\[ \hat{p} = \left( \frac{n_1}{n}, \frac{n_2}{n}, \ldots, \frac{n_k}{n} \right) \]

But now, what if, e.g. \( k = 4 \) and all we observe is this:

<table>
<thead>
<tr>
<th>Trial</th>
<th>Information Available</th>
<th>In Terms of X's</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>outcome is 1</td>
<td>( X_{11} = 1 )</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>( X_{23} = 1 )</td>
</tr>
<tr>
<td>3</td>
<td>2 or 4</td>
<td>( X_{32} + X_{34} = 1 )</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>( X_{42} = 1 )</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>( X_{53} = 1 )</td>
</tr>
<tr>
<td>6</td>
<td>2 or 3</td>
<td>( X_{62} + X_{63} = 1 )</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>( X_{71} = 1 )</td>
</tr>
<tr>
<td>8</td>
<td>1 or 2</td>
<td>( X_{81} + X_{82} = 1 )</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>( X_{92} = 1 )</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>( X_{104} = 1 )</td>
</tr>
</tbody>
</table>
That is, I don't get the 10 \( X_i = (X_{i1}, X_{i2}, \ldots, X_{i4}) \) but, e.g.:

\[
Y_3 = (X_{31}, X_{32}, X_{32}+X_{34})
\]

The likelihood function based on \( Y_i \)'s (information available) not \( X_i \)'s is

\[
L_Y(p) = p_1^2 p_2^2 p_3^2 (1 - p_1 - p_2 - p_3) \times \frac{(1 - p_1 - p_2)(p_2 + p_3)(p_1 + p_2)}{p_2 + p_4}
\]

and I could probably optimize \( L_Y(p) \) or 
\[
\ell_Y(p) = \log L_Y(p)
\]
via some numerical method.

Another possibility is to use the EM algorithm.

\[
\ell_X(p) = n_1 \log p_1 + n_2 \log p_2 + n_3 \log p_3 + n_4 (1 - p_1 - p_2 - p_3)
\]

For the data in hand this is

\[
= (2 + X_{81}) \log p_1 + (2 + X_{82} + X_{62} + X_{82}) \log p_2 \\
+ (2 + X_{63}) \log p_3 + (1 + X_{84}) \log (1 - p_1 - p_2 - p_3)
\]

For any particular \( p_0 = (p_01, p_02, p_03 + p_04) \)

\[
E_{p_0} \left[ X_{81} \mid Y = \text{data in hand} \right] = \frac{p_01}{p_01 + p_02}
\]

\[
E_{p_0} \left[ X_{82} \mid Y = \text{data in hand} \right] = \frac{p_02}{p_02 + p_04}
\]

\[
E_{p_0} \left[ X_{62} \right] = \frac{p_02}{p_02 + p_03}
\]
\[ E_{P_0} \left[ X_{82} \right] = \frac{P_{o_1}}{P_{o_1} + P_{o_2}} \]
\[ E_{P_0} \left[ X_{83} \right] = \frac{P_{o_3}}{P_{o_2} + P_{o_3}} \]
\[ E_{P_0} \left[ X_{84} \right] = \frac{P_{o_4}}{P_{o_2} + P_{o_4}} \]

Then,
\[ E_{P_0} \left[ \log(P) \mid Y=\text{data in hand} \right] = \frac{b(p_0)}{a(p_0)} \]
\[ = \left( 2 + \frac{P_{o_1}}{P_{o_1} + P_{o_2}} \right) \log p_1 + \left( 2 + \frac{P_{o_2}}{P_{o_2} + P_{o_4}} \right) \log p_2 + \left( 2 + \frac{P_{o_3}}{P_{o_2} + P_{o_3}} \right) \log p_3 + \left( 1 + \frac{P_{o_4}}{P_{o_2} + P_{o_4}} \right) \log \left( 1 - p_1 - p_2 - p_3 \right) \]
\[ c(p_0) + d(p_0) \]

Note
\[ a(p_0) + b(p_0) + c(p_0) + d(p_0) = 10 = n \]
\[ P^{(i+1)} = \left( \frac{a^{(i)} + 1}{10}, \frac{b^{(i)} + 1}{10}, \frac{c^{(i)} + 1}{10}, \frac{d^{(i)} + 1}{10} \right) \]
iterate to a fixed point

BTD give some arguments as to why EM might work — see, in particular, Lemma 2.4.1. That says that at each step \( h_y(\Theta) \) never decreases — the standard complaint with EM is that its convergence is often very slow — you need a good starting point and if other methods are possible they may be faster.
One matter that should be raised regarding EM is that B+D phrase Their version of it concerns not

\[ E_{\theta_0} [l_x(\theta) \mid Y=y] \]

but rather

\[ E_{\theta_0} [l_x(\theta) - l_x(\theta_0) \mid Y=y] \]

sometimes I'll be able to compute the 2nd when I couldn't compute the 1st - and optimization at the first is equivalent to optimization of the 2nd

B+D give some arguments why EM might work - see, in particular, Lemma 2.4! That says that at each step \( l_y(\theta) \) never decreases - the standard complaint about EM is that its convergence is often very slow - you need a good starting point and if other methods are possible, they may be faster -

Bayes Estimators

we've said repeatedly that where \( g(\theta) \) specifies

\[ g(\theta \mid x) \propto L_x(\theta)g(\theta) \]

and that characteristics of the posterior serve as estimators of \( y(\theta) \) - e.g.
under SEL, the Bayes estimator of $Y(\theta)$ is

$$\delta(x) = E \left[ Y(\theta) \mid X = x \right]$$

under AEL, the Bayes estimator of $Y(\theta)$ is

$$\delta(x) = \text{median of the den of } Y(\theta) \mid X = x$$

under WEL with weight function $W(\theta) \geq 0$, the Bayes estimator of $Y(\theta)$ is

$$\delta(x) = \frac{E \left[ Y(\theta)W(\theta) \mid X = x \right]}{E \left[ W(\theta) \mid X = x \right]}$$

In fact, this story can sometimes be generalized by not requiring $g(\theta)$ to specify a probability den.

Example: $X \sim N(\theta, 1)$, $g(\theta) = 1, \int_{-\infty}^{\infty} \infty \theta = \infty$.

$$L_x(\theta)g(\theta) = \frac{1}{\sqrt{2\pi}} \exp \left(-\frac{1}{2}(x-\theta)^2\right) \cdot 1$$

and we can use $L_x(\theta)g(\theta)$ to specify a "posterior" $N(\bar{x}, 1)$ and even though $f(x|\theta)g(\theta)$ doesn't define a (joint) probability den for $(X, \theta)$ — note that

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp \left(-\frac{1}{2}(x-\theta)^2\right) dx d\theta = \int_{-\infty}^{\infty} 1 \ d\theta = \infty$$
What is (Bayes) optimal is in principle "clear":

The problem of actually computing characteristics of a distribution with density proportional to

\[ L(\theta)g(\theta) \]

Note that even finding an exact density for the posterior requires finding

\[ \int L(\theta)g(\theta) \, d\theta \]

the normalizer for the function \( L(\theta)g(\theta) \).

Modern Bayes computation is based on simulation to substitute for calculus - e.g. if I am interested in

\[ Q = \int q(\theta)g(\theta|x) \, d\theta \]

for some \( q: \mathbb{R}^k \to \mathbb{R} \) and can somehow generate \( \theta^*_1, \theta^*_2, \ldots \) iid \( G(\theta|x) \) I might approximate \( Q \) by

\[ \hat{Q}_n = \frac{1}{n} \sum q(\theta^*_i) \]

relying on the LLN to conclude that

\[ \frac{1}{n} \sum q(\theta^*_i) \xrightarrow{P} Q \]

(e.g. I might set \( \theta = (\theta_1, \theta_2) \) have \( q(\theta) = \theta \), or \( q(\theta) = I[\theta < \theta_2] \)

The problem with this idea is that naive ways of doing simulation would require that I know

\[ \int L(\theta)g(\theta) \, d\theta \]

in order to simulate the \( \theta^*_i \).

Happily there is the famous "rejection algorithm"