STATS 406 F15: Lab 11
Optimization and EM algorithm

1 Preliminary: Likelihood and MLE

1.1 Likelihood function

- (My own definition, just to help you understand) The likelihood function is the probability/-density that the observed data is generated.
- To formulate the likelihood function for an observed dataset:
  1. Write down the PDF.
  2. Replace the random variable(s) with observed data.
- Log-likelihood is more frequently used than likelihood.

1.2 Maximum Likelihood Estimation (MLE)

- Suppose we have a likelihood function $L(\Theta; X)$, where Θ is the parameter and X is the data.
- We estimate the true parameter by maximizing $L(\Theta; X)$ over Θ.
- Intuitively, this amounts to find the parameter that "best explains" the observed data.
- We usually equivalently maximize $l(\Theta; X) = \log L(\Theta; X)$ over Θ.

1.3 Examples

- Example 1: Model: Exponential(λ); observed data: $X_1, \ldots, X_n$. Likelihood function:

$$L(\lambda; \boldsymbol{X}) = \prod_{i=1}^{n} \left\{ \lambda e^{-\lambda X_i} \mathbb{1}[X_i > 0] \right\}$$

  Log-likelihood (when all $X_i > 0$):

$$l(\lambda; \boldsymbol{X}) = n \log \lambda - \left(\sum_{i=1}^{n} X_i \right) \cdot \lambda$$

  Take the derivative and set it to 0:

$$\hat{\lambda} = 1/\bar{X}$$

- Example 2: Model: Poisson(λ); observed data: $X_1, \ldots, X_n$. Likelihood function:

$$L(\lambda; \boldsymbol{X}) = \prod_{i=1}^{n} \left(\frac{\lambda^{X_i} e^{-\lambda}}{X_i!} \right)$$

  Log-likelihood:

$$l(\lambda; \boldsymbol{X}) = -n\lambda + \left(\sum_{i=1}^{n} X_i \right) \cdot \log \lambda - \sum_{i=1}^{n} \log\left(X_i!\right)$$

  Take the derivative and set it to 0:

$$\hat{\lambda} = \bar{X}$$

- Examples 1 and 2 are closely related. You will see their connection upon taking the Stochastic Processes course.

- Example 3: Model: $N(\mu, \sigma^2)$; observed data: $X_1, \ldots, X_n$. Likelihood function:

$$L(\mu, \sigma; \boldsymbol{X}) = \prod_{i=1}^{n} \left\{ \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(X_i - \mu)^2}{2\sigma^2} \right) \right\}$$

  Log-likelihood:

$$l(\mu, \sigma; \boldsymbol{X}) = -n \cdot \log \sigma - \left\{ \sum_{i=1}^{n} (X_i - \mu)^2 \right\} \Big/ \left(2\sigma^2\right) - \frac{n}{2} \cdot \log(2\pi)$$

  First, take the partial derivative over μ and set it to 0:

$$\hat{\mu} = \bar{X}$$

  Then plug in $\mu = \hat{\mu}$, take the partial derivative over σ and set it to 0:

$$\widehat{\sigma^2} = \left\{ \sum_{i=1}^{n} (X_i - \bar{X})^2 \right\} \Big/ n$$

2 Optimization

Why do we need optimization tools in statistics? One reason is to maximize likelihoods.

What have we learned in class?

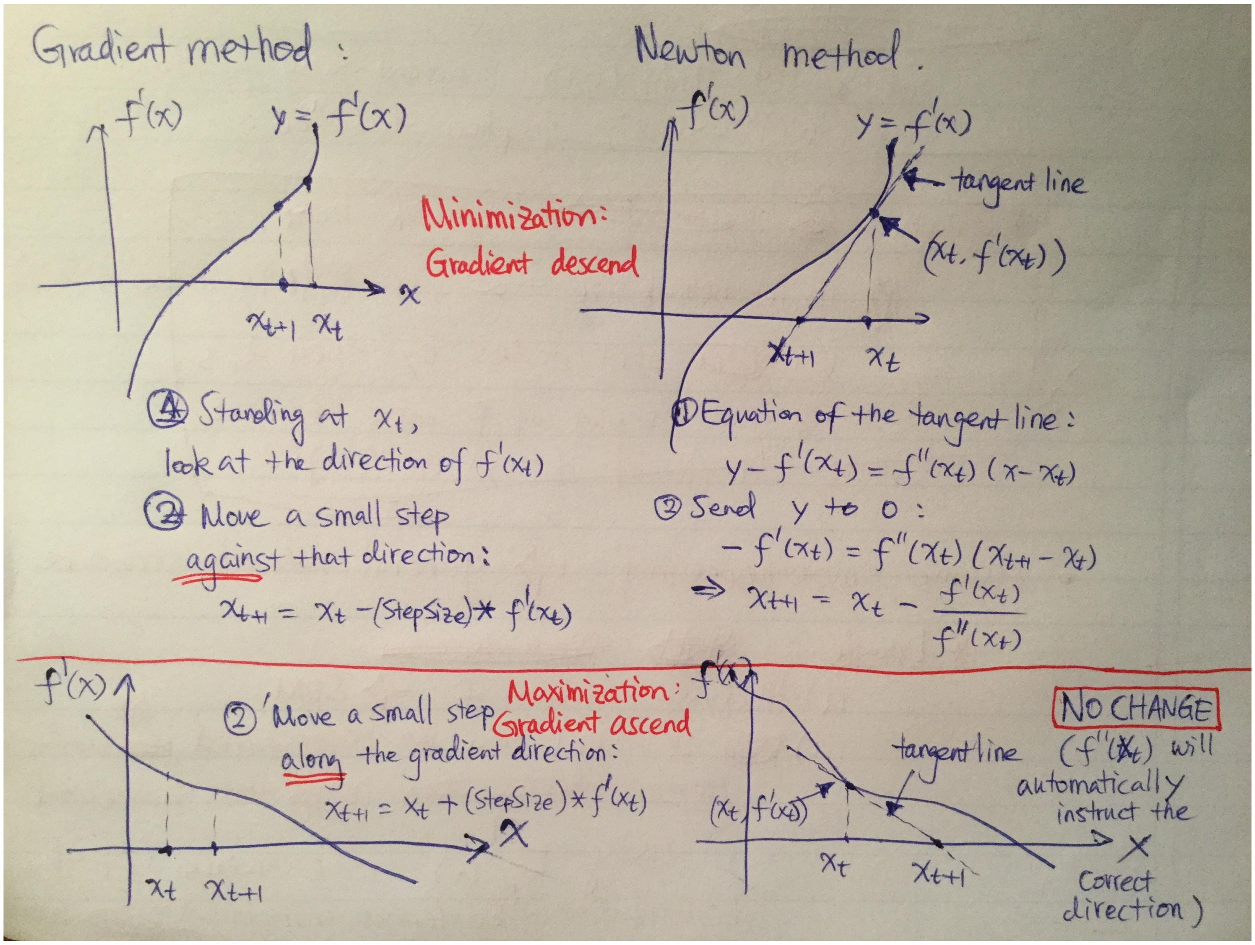
1. Use the “black box” in R: **optim()**
2. Gradient methods
3. Newton’s method

2.1 The optimization function in R

- Syntax: **optim(**_par_, _fn_, _gr=NULL_, _method_**)**
  - * _par_: initial guess
  - * _fn_: the function to be optimized, usually a likelihood
  - * _gr_: (optional) the gradient function if the chosen method requires
  - * _method_: the optimization method to use
- The **optim()** function starts from the initial guess, and explore the shape of the function (when applicable, with the help of the gradient function), to minimize the function values.

2.2 Gradient methods and Newton's method: a route map

1. Our goal can be summarized as maximizing/minimizing a function $f(x)$[*].
2. This is equivalent to finding the zero-point of $f'(x)$[†], i.e. solving $f'(x) = 0$.



① Standing at x_t, look at the direction of $f'(x_t)$

② Move a small step against that direction:

$$x_{t+1} = x_t - (\text{StepSize}) * f'(x_t)$$

① Equation of the tangent line:

$$y - f'(x_t) = f''(x_t)(x - x_t)$$

② Send y to 0:

$$-f'(x_t) = f''(x_t)(x_{t+1} - x_t)$$

$$\Rightarrow x_{t+1} = x_t - \frac{f'(x_t)}{f''(x_t)}$$

②' Move a small step along the gradient direction:

$$x_{t+1} = x_t + (\text{StepSize}) * f'(x_t)$$

[*] Recall that the likelihood function L (or its log l) is just a function of the parameter Θ.

[†] In this course we always assume the smoothness of $f(x)$.

Additional remarks:

- Whether we do a minimization or a maximization depends on the context.
    - * In Examples 1 and 2, it is more convenient to do maximizations; while in Example 3, we usually take a negative sign and do minimization.
    - * The "max-or-min" context determines the (local) shape of the gradient function $f'(x)$.
- (Optional) The step size in gradient methods is often chosen adaptively. One simple choice is to compare $f(x_{t+1})$ to $f(x_t)$. If $f(x_{t+1})$ is improved over $f(x_t)$ (i.e. smaller in minimization, larger in maximization), then inflate the step size and take x_{t+1}; otherwise shrink the step size and retry based on x_t.
- Newton's method usually converges much faster than gradient methods, but it also fails under some cases where gradient methods work.

Optimization you learned in this course is very simple. We will not spend time on specific examples.

3 EM algorithm

When do we need the EM algorithm? When the model contains latent random variables.

3.1 Implementation: a step-by-step implementation guidance

Notation:

- Θ: the model parameter we want to estimate
- Θ_t: the current estimation of Θ at the t-th iteration.
- X: the observable random variable
- Z: the latent random variable
- $l_c(\Theta; X, Z)$: the complete log-likelihood function – **it is the log-likelihood l_c, NOT the likelihood L_c, that is used here! Using L_c is mathematically wrong.**
- $l(\Theta; X)$: the incomplete log-likelihood function – **NOTICE: we DO NOT work with l in the EM algorithm. We only use l_c.**

A step-by-step implementation guidance:

1. Write down the formulation of l_c, the complete log-likelihood function.
2. **E-step:** Evaluate $\mathbb{E}\left[l_c(\Theta; X, Z) | \Theta_t; X\right]$.
   - This expectation is taken over Z. Here, only Z is random; Θ_t is fixed and X is known.
   - **A common conceptual mistake:** evaluate $\mathbb{E}\left[Z | \Theta_t; X\right]$ and plug it into $l_c(\Theta; X, Z)$.
   - **This wrong E-step coincides with the correct E-step if $l_c(\Theta; X, Z)$ is linear in Z.**
   - Otherwise: for example, if $l_c(\Theta; X, Z)$ is linear in $Z_i Z_j$'s, then in the E-step, one should evaluate $\mathbb{E}\left[Z_i Z_j | \Theta_t; X\right]$ to replace $Z_i Z_j$, **NOT** $\mathbb{E}\left[Z_i | \Theta_t; X\right] \cdot \mathbb{E}\left[Z_j | \Theta_t; X\right]$.
3. **M-step:** Update Θ with
   $$\Theta_{t+1} = \arg\max_{\Theta} \mathbb{E}\left[l_c(\Theta; X, Z) | \Theta_t; X\right]$$
   - Recall that the E-step "sketches" $\mathbb{E}\left[l_c(\Theta; X, Z) | \Theta_t; X\right]$ as a function of Θ.
4. Iterate the **E-step** and the **M-step** until convergence. Report Θ_T at convergence.

3.2 Example: 1-dimensional Gaussian mixture

Model set up:

1. Suppose we have two Gaussian distributions $\mathcal{F}_1 := N(\mu_1, \sigma_1^2)$ and $\mathcal{F}_2 := N(\mu_2, \sigma_2^2)$.
2. Toss a coin to decide whether to sample from $\mathcal{F}_1$ or $\mathcal{F}_2$. For simplicity, let the coin $Z \sim \text{Bernoulli}(p)$. Sample from $\mathcal{F}_1$ if $Z = 1$, otherwise sample from $\mathcal{F}_2$.
3. Then sample from the decided $\mathcal{F}$ to obtain X.

Data: $X = \{X_1, \dots, X_n\}$

Parameter to be estimated: $\Theta := (\mu_1, \sigma_1, \mu_2, \sigma_2, p)$. Let $\Theta_t := \left(\mu_1^{(t)}, \sigma_1^{(t)}, \mu_2^{(t)}, \sigma_2^{(t)}, p^{(t)}\right)$ denote the current estimation of Θ at the t-th iteration.

We now follow the guideline and derive an EM algorithm to estimate Θ.

1. **Write down the complete log-likelihood** l_c

$$\begin{aligned}
l_c(\Theta|X, Z) = &\sum_{i=1}^{n} \{Z_i \cdot \log \text{PDF}(N(\mu_1, \sigma_1^2); X_i) + (1 - Z_i) \cdot \log \text{PDF}(N(\mu_2, \sigma_2^2); X_i)\} \\
&+ \sum_{i=1}^{n} (Z_i \log p + (1 - Z_i) \log(1 - p)) \\
&= \sum_{i=1}^{n} \left\{ Z_i \cdot \left(-\log \sigma_1 - \frac{(X_i - \mu_1)^2}{2\sigma_1^2} \right) + (1 - Z_i) \cdot \left(-\log \sigma_2 - \frac{(X_i - \mu_2)^2}{2\sigma_2^2} \right) - \log(2\pi) \right\} \\
&+ \left(\sum_{i=1}^{n} Z_i \right) \log p + \left(n - \sum_{i=1}^{n} Z_i \right) \log(1 - p) \\
&= - \left(\sum_{i=1}^{n} Z_i \right) \log \sigma_1 - \frac{\sum_{i=1}^{n} Z_i (X_i - \mu_1)^2}{2\sigma_1^2} - \left(n - \sum_{i=1}^{n} Z_i \right) \log \sigma_2 - \frac{\sum_{i=1}^{n} (1 - Z_i)(X_i - \mu_2)^2}{2\sigma_2^2} \\
&+ \left(\sum_{i=1}^{n} Z_i \right) \log p + \left(n - \sum_{i=1}^{n} Z_i \right) \log(1 - p) + \text{Constant}
\end{aligned}$$

2. **E-step: Compute** $\mathbb{E}\left[l_c(\Theta; X, Z)|\Theta_t; X\right]$, where the conditional expectation is over Z.

   * Notice that l_c is linear in each Z_i, we can evaluate $\mathbb{E}\left[Z_i|\Theta_t; X\right]$.
   * Since in the model, given X_i, Z_i is conditionally independent of all other X_i's, we have $\mathbb{E}\left[Z_i|\Theta_t; X\right] = \mathbb{E}\left[Z_i|\Theta_t; X_i\right]$.
   * We need the **Bayes Formula**, for two events A and B, we have:

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c)} \tag{1}$$

   * Applying (1) in our context, we can roughly think that $Z_i = 1$ is A and "$X = X_i$" is B.

* Rigorously, we have

$$\begin{aligned}
&\mathbb{E}\left[Z_i|\Theta_t;X_i\right]=\mathbb{P}(Z_i=1|\Theta_t;X_i)\\
=&\frac{\text{PDF}(X_i|\Theta_t;Z_i=1)\cdot\mathbb{P}(Z_i=1|\Theta_t)}{\text{PDF}\left(X_i|\Theta_t;Z_i=1\right)\cdot\mathbb{P}(Z_i=1|\Theta_t)+\text{PDF}(X_i|\Theta_t;Z_i=0)\cdot\mathbb{P}(Z_i=0|\Theta_t)}\\
=&\frac{\text{PDF}\left(N\left(\mu_1^{(t)},\left(\sigma_1^{(t)}\right)^2\right);X_i\right)\cdot p^{(t)}}{\text{PDF}\left(N\left(\mu_1^{(t)},\left(\sigma_1^{(t)}\right)^2\right);X_i\right)\cdot p^{(t)}+\text{PDF}\left(N\left(\mu_2^{(t)},\left(\sigma_2^{(t)}\right)^2\right);X_i\right)\cdot(1-p^{(t)})}\\
=&\frac{\frac{1}{\sqrt{2\pi}\sigma_1^{(t)}}\exp\left(-\frac{\left(X_i-\mu_1^{(t)}\right)^2}{2\left(\sigma_1^{(t)}\right)^2}\right)\cdot p^{(t)}}{\frac{1}{\sqrt{2\pi}\sigma_1^{(t)}}\exp\left(-\frac{\left(X_i-\mu_1^{(t)}\right)^2}{2\left(\sigma_1^{(t)}\right)^2}\right)\cdot p^{(t)}+\frac{1}{\sqrt{2\pi}\sigma_2^{(t)}}\exp\left(-\frac{\left(X_i-\mu_2^{(t)}\right)^2}{2\left(\sigma_2^{(t)}\right)^2}\right)\cdot(1-p^{(t)})}\\
=&\frac{\exp\left(-\frac{\left(X_i-\mu_1^{(t)}\right)^2}{2\left(\sigma_1^{(t)}\right)^2}\right)\cdot p^{(t)}/\sigma_1^{(t)}}{\exp\left(-\frac{\left(X_i-\mu_1^{(t)}\right)^2}{2\left(\sigma_1^{(t)}\right)^2}\right)\cdot p^{(t)}/\sigma_1^{(t)}+\exp\left(-\frac{\left(X_i-\mu_2^{(t)}\right)^2}{2\left(\sigma_2^{(t)}\right)^2}\right)\cdot\left(1-p^{(t)}\right)/\sigma_2^{(t)}}=:\langle Z_i\rangle
\end{aligned}$$

3. **M-step: update the estimation Θ_{t+1} by maximizing $\mathbb{E}\left[l_c(\Theta;X,Z)|\Theta_t;X\right]$ over Θ.**

   * In order to do this, we first write down the formulation of $\mathbb{E}\left[l_c(\Theta;X,Z)|\Theta_t;X\right]$.
   * For simplicity, we write the conditional expectation $\mathbb{E}\left[Z_i|\Theta_t;X_i\right]$ that we computed in the E-step as $\langle Z_i\rangle$.
   * Ignoring the constant, we have:

$$\begin{aligned}
l_c=&-\left(\sum_{i=1}^n\langle Z_i\rangle\right)\log\sigma_1-\frac{\sum_{i=1}^n\langle Z_i\rangle(X_i-\mu_1)^2}{2\sigma_1^2}-\left(n-\sum_{i=1}^n\langle Z_i\rangle\right)\log\sigma_2-\frac{\sum_{i=1}^n(1-\langle Z_i\rangle)(X_i-\mu_2)^2}{2\sigma_2^2}\\
&+\left(\sum_{i=1}^n\langle Z_i\rangle\right)\log p+\left(n-\sum_{i=1}^n\langle Z_i\rangle\right)\log(1-p)
\end{aligned}$$

   * Take derivatives and set them to 0:

$$\begin{aligned}
\frac{\partial l_c}{\partial\mu_1}&=0\Rightarrow 2\sum_{i=1}^n\langle Z_i\rangle(\mu_1-X_i)=0 &&\Rightarrow\mu_1^{(t+1)}=\frac{\sum_{i=1}^n\langle Z_i\rangle X_i}{\sum_{i=1}^n\langle Z_i\rangle}\\
\frac{\partial l_c}{\partial\mu_2}&=0\Rightarrow 2\sum_{i=1}^n(1-\langle Z_i\rangle)(\mu_2-X_i)=0 &&\Rightarrow\mu_2^{(t+1)}=\frac{\sum_{i=1}^n(1-\langle Z_i\rangle)X_i}{\sum_{i=1}^n(1-\langle Z_i\rangle)}\\
\frac{\partial l_c}{\partial p}&=0\Rightarrow\frac{\sum_{i=1}^n\langle Z_i\rangle}{p}-\frac{\sum_{i=1}^n(1-\langle Z_i\rangle)}{1-p}=0\Rightarrow p^{(t+1)}=\frac{\sum_{i=1}^n\langle Z_i\rangle}{n}
\end{aligned}$$

Then we have:

$$
\begin{aligned}
\frac{\partial l_c}{\partial \sigma_1} = 0 &\Rightarrow \frac{\sum_{i=1}^n \langle Z_i \rangle}{\sigma_1} - \frac{\sum_{i=1}^n \langle Z_i \rangle (X_i - \mu_1^{(t+1)})^2}{\sigma_1^3} = 0 \\
&\Rightarrow \left(\sigma_1^{(t+1)}\right)^2 = \frac{\sum_{i=1}^n \langle Z_i \rangle (X_i - \mu_1^{(t+1)})^2}{\sum_{i=1}^n \langle Z_i \rangle} \\
\frac{\partial l_c}{\partial \sigma_2} = 0 &\Rightarrow \frac{\sum_{i=1}^n (1 - \langle Z_i \rangle)}{\sigma_2} - \frac{\sum_{i=1}^n (1 - \langle Z_i \rangle)(X_i - \mu_2^{(t+1)})^2}{\sigma_2^3} = 0 \\
&\Rightarrow \left(\sigma_2^{(t+1)}\right)^2 = \frac{\sum_{i=1}^n (1 - \langle Z_i \rangle)(X_i - \mu_2^{(t+1)})^2}{\sum_{i=1}^n (1 - \langle Z_i \rangle)}
\end{aligned}
$$

4. Iterate the **E-step** and the **M-step** until convergence.

```
#Example: see Lab_11.r
```

3.3 Theory for EM algorithm (Optional)

No theory taught in class? Theory is beyond undergraduate level. If interested (optional):

- Validity: https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm#Proof_of_correctness
- Convergence properties: https://projecteuclid.org/euclid.aos/1176346060