MM algorithms for statistical inference and machine learning problems

Hien D. Nguyen[1,2]

[1]DECRA Research Fellow, Australian Research Council. [2]Lecturer, Department of Mathematics and Statistics, La Trobe University, Melbourne Australia. (Contact–Email: h.nguyen5@latrobe.edu.au, Twitter: @tresbienhien, Website: hiendn.github.io)

S4D, Caen, 2018 June 21

Framework

- In machine learning and statistics, we often observe sample data $\{\boldsymbol{z}_i\}$ of $\{\boldsymbol{Z}_i\}$ from some **data generating process** (DGP).
- Inference must be drawn regarding $\{\boldsymbol{z}_i\}$ via some **objective function** of the data $Q_n(\boldsymbol{\theta})$, which is dependent on a parameter vector $\boldsymbol{\theta}$ in a Euclidean space Θ.
- When the sequence is of length $n \in \mathbb{N}$, the parameter of interest can often be estimated from the data, via the **extremum estimator** (cf. Amemiya, 1985, Ch. 4):

$$\hat{\boldsymbol{\theta}}_n \equiv \arg\min_{\boldsymbol{\theta}\in\Theta} \; Q_n(\boldsymbol{\theta}) \text{ or } \arg\max_{\boldsymbol{\theta}\in\Theta} \; Q_n(\boldsymbol{\theta}).$$

A familiar example (1)

- Suppose that we assume the DGP has distribution with marginal **normal mixture model** density

$$f\left(z_i;\boldsymbol{\theta}\right)=\sum_{j=1}^{m}\pi_j\phi\left(z_i;\mu_i,\sigma_i^2\right),$$

where $\boldsymbol{\mu}=(\mu_1,\dots,\mu_m)\in[-L,L]^m$,
$\boldsymbol{\sigma}=\left(\sigma_1^2,\dots,\sigma_m^2\right)\in\left[S^{-1},S\right]^m$, and

$$\boldsymbol{\pi}\in\mathbb{S}_{m-1}=\left\{(\pi_1,\dots,\pi_m):\pi_j\geq 0,\sum_{j=1}^{m}\pi_j=1\right\},$$

for large L and $S>1$. Here $\boldsymbol{\theta}$ contains $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$ and $\boldsymbol{\pi}$, and

$$\Theta=[-L,L]^m\times\left[S^{-1},S\right]^m\times\mathbb{S}_{m-1}.$$

A familiar example (2)

- We wish to obtain a **maximum likelihood estimator** $\hat{\boldsymbol{\theta}}_n$, which we can define as

$$\hat{\boldsymbol{\theta}}_n \in \left\{ \boldsymbol{\theta}_n : Q_n(\boldsymbol{\theta}_n) = \max_{\boldsymbol{\theta} \in \Theta} \; Q_n(\boldsymbol{\theta}),\; Q_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(z_i; \boldsymbol{\theta}) \right\}.$$

- Due to the simplex constraint (i.e. $\boldsymbol{\pi} \in \mathbb{S}_{m-1}$), We must solve for the **first order condition** (FOC)

$$(\nabla \Lambda)(\boldsymbol{\theta}, \lambda) = \mathbf{0},$$

where ∇ is the gradient operator and

$$\Lambda(\boldsymbol{\theta}, \lambda) = Q_n + \lambda \left(\sum_{j=1}^{m} \pi_j - 1 \right),$$

is the Lagrangian (λ is the Lagrange multiplier).

A familiar example (3)

- Recall that the normal **probability density function** (PDF) has form

$$\phi\left(z_i;\mu_j,\sigma_j^2\right) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(-\frac{1}{2}\frac{\left(z_i-\mu_j\right)^2}{\sigma_j^2}\right).$$

- For each j,

$$\frac{\partial \Lambda}{\partial \mu_j} = \sum_{i=1}^{n} \frac{\pi_j}{f\left(z_i;\boldsymbol{\theta}\right)} \times \frac{\partial \phi\left(z_i;\mu_j,\sigma_j^2\right)}{\partial \mu_j},$$

$$\frac{\partial \Lambda}{\partial \sigma_j^2} = \sum_{i=1}^{n} \frac{\pi_j}{f\left(z_i;\boldsymbol{\theta}\right)} \times \frac{\partial \phi\left(z_i;\mu_j,\sigma_j^2\right)}{\partial \sigma_j^2},$$

$$\frac{\partial \Lambda}{\partial \pi_j} = \sum_{i=1}^{n} \frac{\pi_j}{f\left(z_i;\boldsymbol{\theta}\right)} \phi\left(z_i;\mu_j,\sigma_j^2\right) + \lambda.$$

A familiar example (4)

- It is not difficult to see that the system is highly nonlinear and a one-step closed-form solution is not available for the FOC.
- A multi-step iterative algorithm is required to solve the problem.
- There are many available methods for solving the problem (e.g. Newton algorithm, expectation-maximization algorithm, stochastic algorithms, etc.; see for example Berchtold, 2004).
- We will investigate the use of the MM approach of Hunter and Lange (2004) and Lange (2016).

The MM algorithm (1)

- The abbreviation MM can stand for two things:
  - **majorization-minimization**, when the problem is to minimize an objective $Q_n(\boldsymbol{\theta})$.
  - **minorization-maximization**, when the problem is to maximize an objective $Q_n(\boldsymbol{\theta})$.
- Historically, the MM algorithm framework dates back before Hunter and Lange (2004), who first used the terminology "MM algorithm".
  - The basic principle was expressed in Ortega and Rheinboldt (1970, Sec. 8.3).
  - Application to multidimensional scaling was considered by de Leeuw (1977).
  - The quadratic upper-bound principle was analyzed in Bohning and Lindsay (1988).

The MM algorithm (2)

- Although we discuss the minimization problem, the maximization problem is the same, *mutatis mutandis*.
- Suppose we wish to minimize some difficult to manipulate function $g(\boldsymbol{x})$, with respect to $\boldsymbol{x} \in \mathbb{X}$, where $\mathbb{X}$ is a Euclidean space.
  - Here, the difficulty of g may be due to lack of differentiability, awkward FOC, etc.
- Define a function $\bar{g}(\boldsymbol{x}, \boldsymbol{y})$ to be a **majorizer**, if it satisfies the conditions:

(A) For each $\boldsymbol{x} \in \mathbb{X}$, $g(\boldsymbol{x}) = \bar{g}(\boldsymbol{x}, \boldsymbol{x})$.

(B) For each $\boldsymbol{y} \neq \boldsymbol{x}$, $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{X}$, $g(\boldsymbol{x}) \leq \bar{g}(\boldsymbol{x}, \boldsymbol{y})$.

- Define a **minorizer** by flipping the inequality in (B).

The MM algorithm (3)

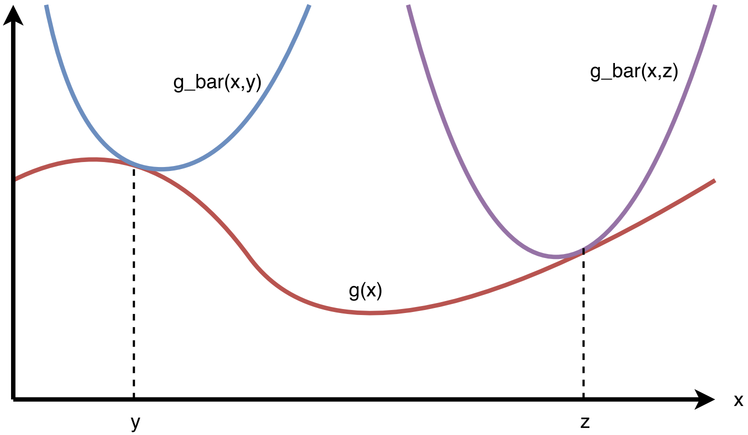

Figure: Majorizers $\bar{g}(x,y)$ and $\bar{g}(x,z)$ of $g(x)$.

The MM algorithm (4)

- Pick some initialization value $\boldsymbol{x}^{(0)} \in \mathbb{X}$.
- We define the majorization-minimization algorithm via the following scheme:

For each $s \in \mathbb{N}$, define

$$\boldsymbol{x}^{(s)} \equiv \arg\min_{\boldsymbol{x} \in \mathbb{X}} \bar{g}\left(\boldsymbol{x}, \boldsymbol{x}^{(s-1)}\right), \tag{1}$$

and stop when some criterion is met.

- A majorization-minimization algorithm is defined by replacing the arg min by arg max, in (1).

The MM algorithm (5)

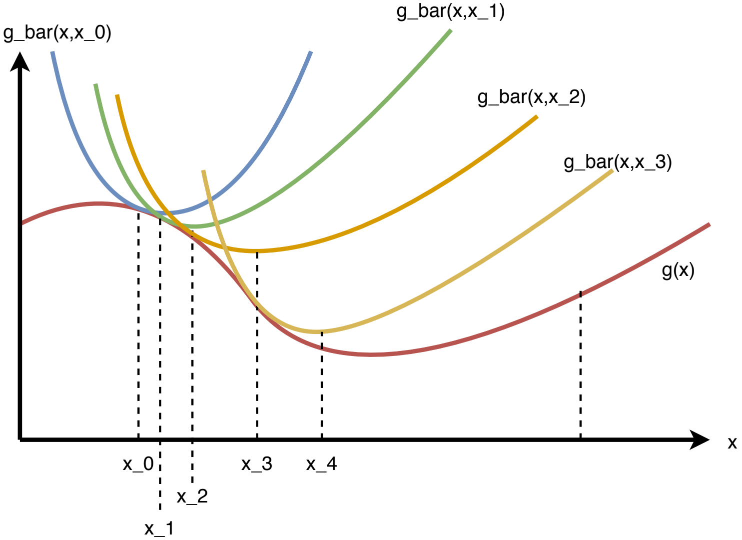

Figure: An MM algorithm that is run for 4 iterations.

The MM algorithm (6)

- From Figure 2, we can observe the **monotonic descent property** of the MM algorithm.
- We can easily prove the descent property as follows:

For any s, (A) implies

$$\bar{g}\left(\boldsymbol{x}^{(s)}, \boldsymbol{x}^{(s)}\right) = g\left(\boldsymbol{x}^{(s)}\right).$$

By (1),

$$\bar{g}\left(\boldsymbol{x}^{(s+1)}, \boldsymbol{x}^{(s)}\right) \le \bar{g}\left(\boldsymbol{x}^{(s)}, \boldsymbol{x}^{(s)}\right).$$

Finally, by (B),

$$g\left(\boldsymbol{x}^{(s+1)}\right) \le \bar{g}\left(\boldsymbol{x}^{(s+1)}, \boldsymbol{x}^{(s)}\right).$$

Thus $g\left(\boldsymbol{x}^{(s+1)}\right) \le g\left(\boldsymbol{x}^{(s)}\right)$.

Some useful majorizers (1)

- Note that all majorizers turn into minorizers when one switches the words convex/concave and positive/negative definite, and the inequality signs.
- All results arise from Lange (2013, Ch. 8) and Lange (2016, Ch. 4).

Suppose that $g(\boldsymbol{x})$ is a concave function, for $\boldsymbol{x} \in \mathbb{X}$ in a Euclidean space. We can majorize g at $\boldsymbol{y}$ via the **supporting hyperplane**

$$\bar{g}(\boldsymbol{x}, \boldsymbol{y}) = g(\boldsymbol{y}) + (\nabla g)(\boldsymbol{y})(\boldsymbol{x} - \boldsymbol{y}).$$

Some useful majorizers (2)

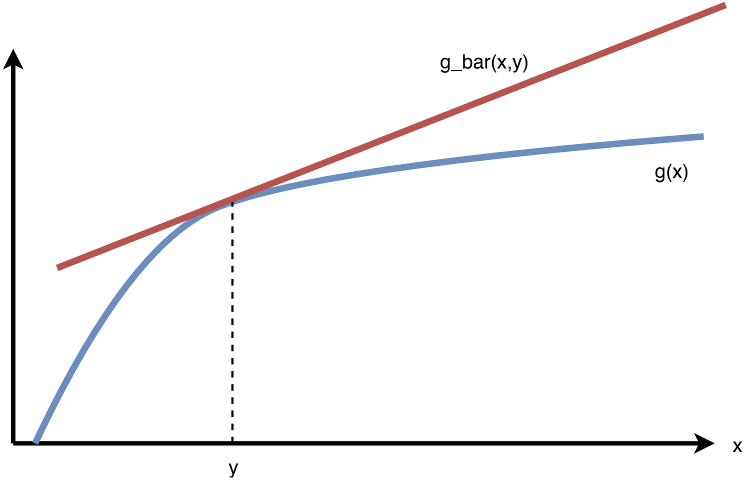

Figure: An example of the supporting hyperplane majorizer.

Some useful majorizers (3)

Suppose that $g(\boldsymbol{x})$ is a convex function with respect to $\boldsymbol{x} \in \mathbb{X} = \mathbb{R}_{+}^{p}$ (the positive cone in $\mathbb{R}^{p}$, $p \in \mathbb{N}$). Also let $\boldsymbol{c} \in \mathbb{R}_{+}^{p}$ be a vector of constants. Via **Jensen's inequality**, we can majorize $g\left(\boldsymbol{c}^{\top}\boldsymbol{x}\right)$ at $\boldsymbol{y}$ by

$$\bar{g}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{j=1}^{m} \frac{c_j y_j}{\boldsymbol{c}^{\top}\boldsymbol{y}} g\left(\frac{\boldsymbol{c}^{\top}\boldsymbol{y}}{y_i} x_i\right),$$

where $\boldsymbol{c} = (c_1, \ldots, c_p)$, $\boldsymbol{x} = (x_1, \ldots, x_p)$, and $\boldsymbol{y} = (y_1, \ldots, y_p)$.

Some useful majorizers (4)

Let $\mathbf{H}$ be the Hessian operator (i.e. $\mathbf{H}g = \partial^2 g/\partial \boldsymbol{x} \partial \boldsymbol{x}^\top$). Let $g(\boldsymbol{x})$ be a function with **bounded curvature**, in the sense that there exists a matrix $\mathbf{C}$, such that $\mathbf{C} - (\mathbf{H}g)(\boldsymbol{x})$ is positive semidefinite for all $\boldsymbol{x} \in \mathbb{X}$. We can majorize g at $\boldsymbol{y}$ by

$$\bar{g}(\boldsymbol{x}, \boldsymbol{y}) = g(\boldsymbol{x}) + (\nabla g)(\boldsymbol{y})(\boldsymbol{x} - \boldsymbol{y}) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{y})^\top \mathbf{C}(\boldsymbol{x} - \boldsymbol{y}).$$

Some useful majorizers (5)

Let $\boldsymbol{x} \in \mathbb{X} = \mathbb{R}_+^p$ and define

$$g(\boldsymbol{x}) = \prod_{j=1}^{p} x_i^{c_i},$$

where $\boldsymbol{c} \in \mathbb{R}_+^p$. Further define $C = \sum_{j=1}^{p} c_j$. Then, via the **arithmetic-geometric mean inequality**, we can majorize g at $\boldsymbol{y}$ by

$$\bar{g}(\boldsymbol{x}; \boldsymbol{y}) = \left(\prod_{j=1}^{p} y_j^{c_j} \right) \sum_{j=1}^{p} \frac{c_j}{C} \left(\frac{x_j}{y_j} \right)^C.$$

Some useful majorizers (6)

- Along with the majorizers that have been presented, we also note that majorization satisfies the following property.
  - **Transitivity** (i.e. if, $\bar{g}$ majorizes g at $\boldsymbol{y}$, and $\bar{\bar{g}}$ majorizes $\bar{g}$ at $\boldsymbol{y}$, then $\bar{\bar{g}}$ majorizes g at $\boldsymbol{y}$).
  - Majorization is **closed under summation** (i.e. if $\bar{g}_1$ majorizes g_1 at $\boldsymbol{y}$ and $\bar{g}_2$ majorizes g_2 at $\boldsymbol{y}$, then $\bar{g}_1 + \bar{g}_2$ majorizes $g_1 + g_2$ at $\boldsymbol{y}$).
  - Majorization is **closed under non-negative multiplication** (i.e. if $\bar{g}_1 > 0$ majorizes $g_1 > 0$ at $\boldsymbol{y}$ and $\bar{g}_2 > 0$ majorizes $g_2 > 0$ at $\boldsymbol{y}$, then $\bar{g}_1\bar{g}_2$ majorizes $g_1 g_2$ at $\boldsymbol{y}$).
  - Majorization is **closed under composition with an increasing function** (i.e. if $\bar{g}$ majorizes g at $\boldsymbol{y}$ and h is an increasing function, then $h \circ \bar{g}$ majorizes $h \circ g$ at $\boldsymbol{y}$).

Normal mixture models (1A)

- Let $g(\boldsymbol{x}) = \log\left(\mathbf{1}^\top \boldsymbol{x}\right)$ (a concave function), where $\boldsymbol{x} \in \mathbb{X} = \mathbb{R}_+^p$. Using the Jensen's inequality majorizer, the following minorizer was proposed by Zhou and Lange (2010):

$$g(\boldsymbol{x}, \boldsymbol{y}) = \sum_{j=1}^{m} \frac{y_j}{\sum_{k=1}^{m} y_k} \log(x_k) - \sum_{j=1}^{m} \frac{y_j}{\sum_{k=1}^{m} y_k} \log\left(\frac{y_j}{\sum_{k=1}^{m} y_k}\right).$$

- Make the substitutions $x_j = \pi_j \phi\left(z_i; \mu_j, \sigma_j^2\right)$ and $y_j = \pi_j^{(s-1)} \phi\left(z_i; \mu_j^{(s-1)}, \sigma_j^{(s-1)2}\right)$.
- Rewrite $g(\boldsymbol{x}) \equiv g_i(\boldsymbol{\theta})$ as

$$g_i(\boldsymbol{\theta}) = \log\left[\sum_{j=1}^{m} \pi_j \phi\left(z_i; \mu_j, \sigma_j^2\right)\right].$$

Normal mixture models (1B)

- We can then write $\bar{g}\left(\boldsymbol{x},\boldsymbol{y}\right)\equiv\bar{g}_i\left(\boldsymbol{\theta},\boldsymbol{\theta}^{(s-1)}\right)$ as

$$\begin{aligned}\bar{g}_i\left(\boldsymbol{\theta},\boldsymbol{\theta}^{(s-1)}\right) &= \sum_{j=1}^{m}\tau_j\left(z_i;\boldsymbol{\theta}^{(s-1)}\right)\left[\log\left(\pi_j\right)+\log\phi\left(z_i;\mu_j,\sigma_j^2\right)\right] \\ &\quad -C_i\left(\boldsymbol{\theta}^{(s-1)}\right),\end{aligned}$$

where $\tau_j\left(z_i;\boldsymbol{\theta}\right)=\pi_j\phi\left(z_i;\mu_j,\sigma_j^2\right)/f\left(z_i;\boldsymbol{\theta}\right)$ and

$$C_i\left(\boldsymbol{\theta}^{(s-1)}\right)=\sum_{j=1}^{m}\tau_j\left(z_i;\boldsymbol{\theta}^{(s-1)}\right)\log\left[\tau_j\left(z_i;\boldsymbol{\theta}^{(s-1)}\right)\right].$$

Normal mixture models (1C)

- Notice that the **log-likelihood** $Q_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(z_i; \boldsymbol{\theta})$ can be written as $Q_n = \sum_{i=1}^{n} g_i$, and thus we can minorize $Q_n(\boldsymbol{\theta})$ by

$$\begin{aligned} Q\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(s-1)}\right) &\equiv \sum_{i=1}^{n} \sum_{j=1}^{m} \tau_j\left(z_i; \boldsymbol{\theta}^{(s-1)}\right) \log(\pi_j) \\ &\quad + \sum_{i=1}^{n} \sum_{j=1}^{m} \tau_j\left(z_i; \boldsymbol{\theta}^{(s-1)}\right) \log \phi\left(z_i; \mu_j, \sigma_j^2\right) \\ &\quad - \sum_{i=1}^{n} C_i\left(\boldsymbol{\theta}^{(s-1)}\right). \end{aligned}$$

Normal mixture models (1D)

- Expand out
  $\phi\left(z_i;\mu_j,\sigma_j^2\right) = \left(2\pi\sigma_i^2\right)^{-1/2}\exp\left[\left(z_i-\mu_j\right)^2/\left(2\sigma_j^2\right)\right]$ to get

$$
\begin{aligned}
Q\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(s-1)}\right) &\equiv \sum_{i=1}^{n}\sum_{j=1}^{m}\tau_j\left(z_i;\boldsymbol{\theta}^{(s-1)}\right)\log\left(\pi_j\right) \\
&\quad -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}\tau_j\left(z_i;\boldsymbol{\theta}^{(s-1)}\right)\log\left(\sigma_j^2\right) \\
&\quad -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}\tau_j\left(z_i;\boldsymbol{\theta}^{(s-1)}\right)\frac{\left(z_i-\mu_j\right)^2}{\sigma_j^2}+C,
\end{aligned}
$$

where C is a constant that is not dependent on $\boldsymbol{\theta}$.

Normal mixture models (1D)

- Construct the Lagrangian

$$\Lambda(\boldsymbol{\theta}, \lambda) = Q\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(s-1)}\right) + \lambda\left(\sum_{j=1}^{m} \pi_j - 1\right),$$

and compute $\nabla\Lambda(\boldsymbol{\theta}, \lambda)$ to obtain the FOC, for each j,

$$\frac{\partial\Lambda(\boldsymbol{\theta}, \lambda)}{\partial \pi_j} = \pi_j^{-1} \sum_{i=1}^{n} \tau_j\left(z_i; \boldsymbol{\theta}^{(s-1)}\right) + \lambda = 0,$$

$$\frac{\partial\Lambda(\boldsymbol{\theta}, \lambda)}{\partial \mu_j} = \sum_{i=1}^{n} \tau_j\left(z_i; \boldsymbol{\theta}^{(s-1)}\right) \frac{z_i - \mu_j}{\sigma_j^2} = 0,$$

$$\frac{\partial\Lambda(\boldsymbol{\theta}, \lambda)}{\partial \sigma_j^2} = \frac{1}{2} \sum_{i=1}^{n} \tau_j\left(z_i; \boldsymbol{\theta}^{(s-1)}\right) \left[-\frac{1}{\sigma_j^2} + \frac{(z_i - \mu_j)^2}{\left(\sigma_j^2\right)^2}\right] = 0,$$

and $\sum_{j=1}^{m} \pi_j - 1 = 0$.

Normal mixture models (1E)

- Solving the FOC yields the MM algorithm update at the sth step, $\boldsymbol{\theta}^{(s)}$, which contains, for each j:

$$\pi_j^{(s)} \equiv n^{-1} \sum_{i=1}^{n} \tau_j \left(z_i; \boldsymbol{\theta}^{(s-1)} \right),$$

$$\mu_j^{(s)} \equiv \frac{\sum_{i=1}^{n} \tau_j \left(z_i; \boldsymbol{\theta}^{(s-1)} \right) z_i}{\sum_{i=1}^{n} \tau_j \left(z_i; \boldsymbol{\theta}^{(s-1)} \right)},$$

and

$$\sigma_j^{(s)2} \equiv \frac{\sum_{i=1}^{n} \tau_j \left(z_i; \boldsymbol{\theta}^{(s-1)} \right) \left(z_i - \mu_j^{(s)} \right)^2}{\sum_{i=1}^{n} \tau_j \left(z_i; \boldsymbol{\theta}^{(s-1)} \right)}.$$

- Note that this is exactly the expectation-maximization algorithm for maximum likelihood estimation for normal mixtures (this is a coincidence; see Meng, 2000).

The EM algorithm (1)

- The MM algorithm (for maximization) is a generalization of the EM algorithm, in the sense that every EM algorithm is an MM algorithm.
- Suppose that $\boldsymbol{Z} = (\boldsymbol{U}, \boldsymbol{V})$ is a random variable, and suppose that we only observe $\boldsymbol{U}$ but not $\boldsymbol{V}$, at $\boldsymbol{u}$.
- Write the PDF of $\boldsymbol{U}$ with respect to some parameter $\boldsymbol{\theta} \in \Theta$ as $f(\boldsymbol{u}; \boldsymbol{\theta})$.
- If we know both $\boldsymbol{U}$ and $\boldsymbol{V}$, then we can write the PDF of $\boldsymbol{Z}$ as $f(\boldsymbol{z}; \boldsymbol{\theta})$ (the complete-data likelihood).
- Starting from some initial parameter $\boldsymbol{\theta}^{(0)}$, the **EM algorithm** proceeds by computing, at the sth step,

$$\boldsymbol{\theta}^{(s)} \equiv \arg\max_{\boldsymbol{\theta} \in \Theta} \mathbb{E}_{\boldsymbol{V}|\boldsymbol{U}=\boldsymbol{u}}^{\boldsymbol{\theta}^{(s-1)}} \left[\log f(\boldsymbol{Z}; \boldsymbol{\theta})\right].$$

The EM algorithm (2)

- Consider the sequence of arguments:

$$
\begin{aligned}
\log(\boldsymbol{u};\boldsymbol{\theta}) &= \log \mathbb{E}_{\boldsymbol{V}}^{\boldsymbol{\theta}}\left[f\left(\boldsymbol{U}|\boldsymbol{V}=\boldsymbol{v};\boldsymbol{\theta}\right)\right] \\
&= \log \mathbb{E}_{\boldsymbol{V}}^{\boldsymbol{\theta}}\left[\frac{f\left(\boldsymbol{V}|\boldsymbol{U}=\boldsymbol{u};\boldsymbol{\theta}^{(s-1)}\right)f\left(\boldsymbol{U}|\boldsymbol{V}=\boldsymbol{v};\boldsymbol{\theta}\right)}{f\left(\boldsymbol{V}|\boldsymbol{U}=\boldsymbol{u};\boldsymbol{\theta}^{(s-1)}\right)}\right] \\
&= \log \mathbb{E}_{\boldsymbol{V}|\boldsymbol{U}=\boldsymbol{u}}^{\boldsymbol{\theta}^{(s-1)}}\left[\frac{f\left(\boldsymbol{V};\boldsymbol{\theta}\right)f\left(\boldsymbol{U}|\boldsymbol{V}=\boldsymbol{v};\boldsymbol{\theta}\right)}{f\left(\boldsymbol{V}|\boldsymbol{U}=\boldsymbol{u};\boldsymbol{\theta}^{(s-1)}\right)}\right] \\
&\geq \mathbb{E}_{\boldsymbol{V}|\boldsymbol{U}=\boldsymbol{u}}^{\boldsymbol{\theta}^{(s-1)}}\left[\log\frac{f\left(\boldsymbol{V};\boldsymbol{\theta}\right)f\left(\boldsymbol{U}|\boldsymbol{V}=\boldsymbol{v};\boldsymbol{\theta}\right)}{f\left(\boldsymbol{V}|\boldsymbol{U}=\boldsymbol{u};\boldsymbol{\theta}^{(s-1)}\right)}\right] \\
&= \mathbb{E}_{\boldsymbol{V}|\boldsymbol{U}=\boldsymbol{u}}^{\boldsymbol{\theta}^{(s-1)}}\left[\log f\left(\boldsymbol{Z};\boldsymbol{\theta}\right)\right] \\
&\quad -\mathbb{E}_{\boldsymbol{V}|\boldsymbol{U}=\boldsymbol{u}}^{\boldsymbol{\theta}^{(s-1)}}\left[\log f\left(\boldsymbol{V}|\boldsymbol{U}=\boldsymbol{u};\boldsymbol{\theta}^{(s-1)}\right)\right]
\end{aligned}
$$

The EM algorithm (3)

- Set

$$g(\boldsymbol{\theta}) \equiv \log f(\boldsymbol{u}; \boldsymbol{\theta}),$$

and

$$\begin{aligned}\bar{g}\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(s-1)}\right) &\equiv \mathbb{E}^{\boldsymbol{\theta}^{(s-1)}}_{\boldsymbol{V}|\boldsymbol{U}=\boldsymbol{u}}\left[\log f(\boldsymbol{Z}; \boldsymbol{\theta})\right] \\ &\quad - \mathbb{E}^{\boldsymbol{\theta}^{(s-1)}}_{\boldsymbol{V}|\boldsymbol{U}=\boldsymbol{u}}\left[\log f\left(\boldsymbol{V}|\boldsymbol{U}=\boldsymbol{u}; \boldsymbol{\theta}^{(s-1)}\right)\right].\end{aligned}$$

- Since the second term of $\bar{g}\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(s-1)}\right)$ does not depend on $\boldsymbol{\theta}$, we can write the EM step as the MM step

$$\boldsymbol{\theta}^{(s)} \equiv \arg\max_{\boldsymbol{\theta}\in\Theta} \; \bar{g}\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(s-1)}\right).$$

A modern example (1)

- Suppose that we observe data $\{\boldsymbol{z}_i\}$, where each $\boldsymbol{z}_i = (\boldsymbol{u}_i, v_i)$, with $\boldsymbol{u}_i \in \mathbb{U} = \mathbb{R}^p$ and $v_i \in \{-1, +1\}$.
- Suppose that we wish to construct a linear classifier that minimizes, with respect to $\boldsymbol{\theta} \in \mathbb{R}^{p+1}$, the average **classification loss**

$$l_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \left[\!\left[v_i \neq \operatorname{sign}\left(\tilde{\boldsymbol{u}}_i^\top \boldsymbol{\theta} \right) \right]\!\right].$$

  - $\operatorname{sign}(a) = -1$ if $a \leq 0$, $+1$ if $a > 0$.
  - $[\![A]\!]$ is the Iverson bracket, which takes value 1 if A is true and 0 otherwise.
  - $\tilde{\boldsymbol{u}}_i = (1, \boldsymbol{u}_i)$.

A modern example (2)

- The problem of obtaining

$$\hat{\boldsymbol{\theta}}_n \equiv \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^{p+1}} \; l_n(\boldsymbol{\theta})$$

  is highly ill-conditioned and combinatorial.

- We can instead replace the average classification loss, by the average **hinge loss**

$$l_n^{\mathrm{h}}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} h(\boldsymbol{z}_i; \boldsymbol{\theta}),$$

  where

$$h(\boldsymbol{z}_i; \boldsymbol{\theta}) = \left[1 - v_i \tilde{\boldsymbol{u}}_i^\top \boldsymbol{\theta}\right]_+,$$

  and $[a]_+ = \max\{0, a\}$.

A modern example (3)

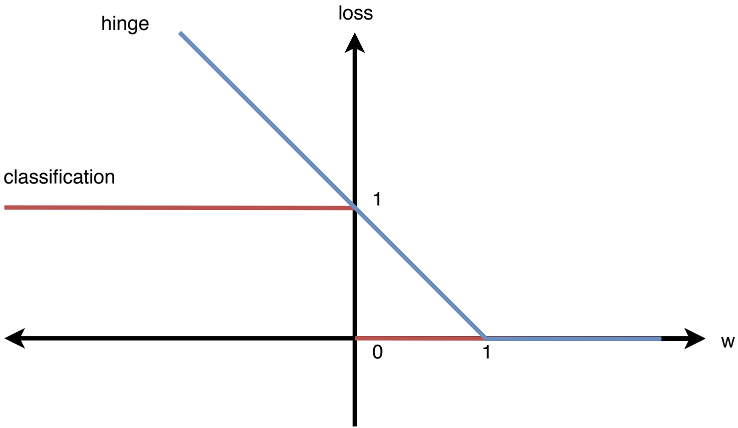

Figure: Example of loss functions, where $w = \tilde{\boldsymbol{u}}_i^\top \boldsymbol{\theta}$ and $v_i = 1$.

A modern example (4)

- Suppose that we penalize the components of $\boldsymbol{\theta}$ that correspond to $\boldsymbol{u}_i$ by

$$\text{pen}(\boldsymbol{\theta}) = \lambda \boldsymbol{\theta}^\top \tilde{\mathbf{I}} \boldsymbol{\theta},$$

where $\lambda \geq 0$

$$\tilde{\mathbf{I}} = \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{I}_p \end{bmatrix}.$$

- The problem:

$$\hat{\boldsymbol{\theta}}_n \equiv \arg \min_{\boldsymbol{\theta} \in \mathbb{R}^{p+1}} g(\boldsymbol{\theta}),$$

where $g(\boldsymbol{\theta}) = l_n^{\text{h}}(\boldsymbol{\theta}) + \text{pen}(\boldsymbol{\theta})$ is the classical **soft-margin support vector machine** (SVM) problem of Cortes and Vapnik (1995).

An MM for the SVM (1)

- The following derivation is from Nguyen and McLachlan (2017).
- Using the supporting hyperplane inequality, we can majorize $g(a) = \sqrt{a}$ ($x \in \mathbb{R}_+$), at b, by

$$\bar{g}(a, b) = \sqrt{b} + \frac{1}{2\sqrt{b}}(a - b).$$

- If we substitute x^2 in for a and y^2 in for b, then we can majorize $g(x) = \sqrt{x^2} = |x|$, at $y \neq 0$, by

$$\begin{aligned} \bar{g}(x, y) &= |y| + \frac{\left(x^2 - y^2\right)}{2|y|} \\ &= \frac{x^2}{2|y|} + \frac{|y|}{2}. \end{aligned}$$

An MM for the SVM (2)

- Consider the identity:

$$\max\{a, b\} = \frac{|a-b|}{2} + \frac{a+b}{2},$$

  which implies that $[a]_+ = |a|/2 + a/2$.

- Using the previous result, we can majorize $g(x) = [x]_+$, at $y \neq 0$, by

$$\begin{aligned}\bar{g}(x, y) &= \frac{x^2}{4|y|} + \frac{|y|}{4} + \frac{x}{2} \\ &= \frac{(x+|y|)^2}{4|y|}.\end{aligned}$$

An MM for the SVM (3)

- Using the Jensen's inequality majorizer, for small $\varepsilon > 0$, we can majorize $g(x) = \sqrt{x^2 + \varepsilon}$, at y, by

$$\bar{g}(x, y) = \sqrt{y^2 + \varepsilon} + \frac{\left(x^2 - y^2\right)}{2\sqrt{y^2 + \varepsilon}}.$$

- Approximate $[x]_+$ by $g(x) = \sqrt{x^2 + \varepsilon}/2 + x/2$, for small $\varepsilon > 0$. We can majorize $g(x)$ by

$$\bar{g}(x, y) = \frac{\left[x + \sqrt{y^2 + \varepsilon}\right]^2}{4\sqrt{y^2 + \varepsilon}}.$$

An MM for the SVM (4)

- For small $\varepsilon > 0$, we can approximate $g(\boldsymbol{\theta}) = l_n^{\mathrm{h}}(\boldsymbol{\theta}) + \mathrm{pen}(\boldsymbol{\theta})$ by

$$g_\varepsilon(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} g_i^\varepsilon(\boldsymbol{\theta}) + \lambda \boldsymbol{\theta}^\top \tilde{\mathbf{I}} \boldsymbol{\theta},$$

where

$$g_i^\varepsilon(\boldsymbol{\theta}) = \frac{\sqrt{\left(1 - v_i \tilde{\boldsymbol{u}}_i^\top \boldsymbol{\theta}\right)^2 + \varepsilon}}{2} + \frac{1 - v_i \tilde{\boldsymbol{u}}_i^\top \boldsymbol{\theta}}{2}.$$

An MM for the SVM (5)

- Using the previously derived majorizer, we can majorize $g_i^{\varepsilon}(\boldsymbol{\theta})$ at some $\boldsymbol{\theta}^{(s-1)}$ by

$$\bar{g}_i\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(s-1)}\right) = \frac{\left[1 - v_i \tilde{\boldsymbol{u}}_i^{\top} \boldsymbol{\theta} + \gamma_i^{\varepsilon}\left(\boldsymbol{\theta}^{(s-1)}\right)\right]^2}{4\gamma_i^{\varepsilon}\left(\boldsymbol{\theta}^{(s-1)}\right)},$$

  where $\gamma_i^{\varepsilon}\left(\boldsymbol{\theta}^{(s-1)}\right) = \sqrt{\left(1 - v_i \tilde{\boldsymbol{u}}_i^{\top} \boldsymbol{\theta}^{(s-1)}\right)^2 + \varepsilon}$.

- Thus, we can majorize $g_{\varepsilon}(\boldsymbol{\theta})$, at $\boldsymbol{\theta}^{(s-1)}$ by

$$\bar{g}_{\varepsilon}\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(s-1)}\right) = \frac{1}{n}\sum_{i=1}^{n} \bar{g}_i\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(s-1)}\right) + \lambda \boldsymbol{\theta}^{\top} \tilde{\mathbf{I}} \boldsymbol{\theta}.$$

An MM for the SVM (6)

- For each s, write

$$\boldsymbol{\gamma}_n^{(s-1)} = \left(1 + \gamma_1^{\varepsilon}\left(\boldsymbol{\theta}^{(s-1)}\right), \ldots, 1 + \gamma_n^{\varepsilon}\left(\boldsymbol{\theta}^{(s-1)}\right)\right)$$

and let $\mathbf{W}_n^{(s-1)}$ be a diagonal matrix with ith element

$$\frac{1}{4\gamma_i^{\varepsilon}\left(\boldsymbol{\theta}^{(s-1)}\right)}.$$

- Let $\mathbf{V}_n$ contain $v_i \tilde{\boldsymbol{u}}_i$ in the ith row.
- We can write $\bar{g}_{\varepsilon}\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(s-1)}\right)$ as

$$\frac{1}{n}\left(\boldsymbol{\gamma}_n^{(s-1)} - \mathbf{V}_n\boldsymbol{\theta}\right)^{\top}\mathbf{W}_n^{(s-1)}\left(\boldsymbol{\gamma}_n^{(s-1)} - \mathbf{V}_n\boldsymbol{\theta}\right) + \lambda\boldsymbol{\theta}^{\top}\tilde{\mathbf{I}}\boldsymbol{\theta}.$$

An MM for the SVM (7)

- The MM update

$$\boldsymbol{\theta}^{(s)} \equiv \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^{p+1}} \bar{g}_{\varepsilon}\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(s-1)}\right)$$

is a **weighted linear ridge regression** problem.

- We can obtain the form of the update by solving the FOC $(\nabla\bar{g}_{\varepsilon})(\boldsymbol{\theta}) = \mathbf{0}$, where

$$\begin{aligned}\nabla\bar{g}_{\varepsilon} &= -\frac{2}{n}\mathbf{V}_n^{\top}\mathbf{W}_n^{(s-1)}\left(\boldsymbol{\gamma}_n^{(s-1)} - \mathbf{V}_n\boldsymbol{\theta}\right) \\ &\quad +2\lambda\tilde{\mathbf{I}}\boldsymbol{\theta}.\end{aligned}$$

- We thus obtain the **iteratively reweighted least-squares** updates

$$\boldsymbol{\theta}^{(s)} = \left(\mathbf{V}_n^{\top}\mathbf{W}_n^{(s-1)}\mathbf{V}_n + n\lambda\tilde{\mathbf{I}}\right)^{\top}\mathbf{V}_n^{\top}\mathbf{W}_n^{(s-1)}\boldsymbol{\gamma}_n^{(s-1)}.$$

A convergence result (1)

- Let $g(\boldsymbol{x})$ be the objective of interest, where $\boldsymbol{x} \in \mathbb{X} \subset \mathbb{R}^p$ and let $\boldsymbol{d} \in \mathbb{X}$. We say that the **directional derivative** of g with respect to $\boldsymbol{d}$ is

$$g'_{\boldsymbol{d}}(\boldsymbol{x}) \equiv \liminf_{t \downarrow 0} \frac{g(\boldsymbol{x} + t\boldsymbol{d}) - g(\boldsymbol{x})}{t}.$$

- We say that a $\boldsymbol{x}^*$ is a **stationary point** of g, if $g'_{\boldsymbol{d}}(\boldsymbol{x}^*) \geq 0$, for all $\boldsymbol{d}$ such that $\boldsymbol{x}^* + \boldsymbol{d} \in \mathbb{X}$.
- In the case where g is a differentiable function, the definition is equivalent to $(\nabla g)(\boldsymbol{x}^*) = \mathbf{0}$.

A convergence result (2)

(A1) Let $\bar{g}(\boldsymbol{x}, \boldsymbol{y})$ majorize the objective $g(\boldsymbol{x})$ ($\boldsymbol{x} \in \mathbb{X}$), at $\boldsymbol{y}$, by satisfying Assumptions (A) and (B).

- Starting from some initialization $\boldsymbol{x}^{(0)}$, we denote the limit point of the MM algorithm by

$$\boldsymbol{x}^{(\infty)} \equiv \lim_{s \to \infty} \boldsymbol{x}^{(s)} < \infty,$$

  where $\boldsymbol{x}^{(s)} \equiv \arg\min_{\boldsymbol{\theta} \in \mathbb{X}} \ \bar{g}\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(s-1)}\right)$.

- Theorem 1 of Razaviyayn et al. (2013) states the following:

Under Assumption (A1), every limit point $\boldsymbol{x}^{(\infty)}$ is a stationary point of the problem

$$\min_{\boldsymbol{x} \in \mathbb{X}} \ g(\boldsymbol{x}).$$

Convergence of the SVM MM (1)

- Assumption (A1) is automatically fulfilled, by construction of the MM algorithm.
- We must therefore show that the updates

$$\boldsymbol{\theta}^{(s)} = \left(\mathbf{V}_n^\top \mathbf{W}_n^{(s-1)} \mathbf{V}_n + n\lambda \tilde{\mathbf{I}}\right)^\top \mathbf{V}_n^\top \mathbf{W}_n^{(s-1)} \boldsymbol{\gamma}_n^{(s-1)}$$

  globally minimizes the majorizer $\bar{g}_\varepsilon\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(s-1)}\right)$.
- In

$$\frac{1}{n}\left(\boldsymbol{\gamma}_n^{(s-1)} - \mathbf{V}_n\boldsymbol{\theta}\right)^\top \mathbf{W}_n^{(s-1)} \left(\boldsymbol{\gamma}_n^{(s-1)} - \mathbf{V}_n\boldsymbol{\theta}\right) + \lambda\boldsymbol{\theta}^\top \tilde{\mathbf{I}}\boldsymbol{\theta},$$

  both $\mathbf{W}_n^{(s-1)}$ and $\tilde{\mathbf{I}}$ are at least positive semidefinite, thus the stationary point of $\bar{g}_\varepsilon$ is also a global minimum (since $\bar{g}_\varepsilon$ is convex).
- Thus, the limit point $\boldsymbol{\theta}^{(\infty)}$ (starting from some $\boldsymbol{\theta}^{(0)}$) converges to a stationary point of g_ε.

Convergence of the SVM MM (2)

- Recall that

$$g_{\varepsilon}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} g_i^{\varepsilon}(\boldsymbol{\theta}) + \lambda \boldsymbol{\theta}^{\top}\tilde{\mathbf{I}}\boldsymbol{\theta},$$

  where

$$g_i^{\varepsilon}(\boldsymbol{\theta}) = \frac{\sqrt{\left(1 - v_i \tilde{\boldsymbol{u}}_i^{\top}\boldsymbol{\theta}\right)^2 + \varepsilon}}{2} + \frac{1 - v_i \tilde{\boldsymbol{u}}_i^{\top}\boldsymbol{\theta}}{2}.$$

- For the function $h(x) = \sqrt{x^2 + \varepsilon}$ $(\varepsilon > 0)$ we can obtain the second derivative

$$h''(x) = \varepsilon / \left(x^2 + \varepsilon\right)^{3/2} > 0.$$

Convergence of the SVM MM (3)

- Since $g_i^{\varepsilon}(\boldsymbol{\theta})$ is a convex composition of an affine function of $\boldsymbol{\theta}$, it is also convex (cf. Boyd and Vandenberghe, 2004, Sec. 3.2.2).
- Thus, every stationary point of g_{ε} is a global minimizer.
- We have the improved result: starting from any $\boldsymbol{\theta}^{(0)}$, the limit point $\boldsymbol{\theta}^{(\infty)}$ converges to a global minimizer of g_{ε}.

Some recent developments

- Stochastic approximation type algorithms have been proposed in Mairal (2013) and Razaviyayn et al. (2016).
  - A stream-data suitable MM algorithm for SVM was proposed in Nguyen et al. (2018).
- Convex analysis and finite-iteration analysis of MM algorithms have been explored in Mairal (2015).
- Block-wise and cyclical MM algorithms have been explored and analyzed in Razaviyayn et al. (2013) and Hong et al. (2016).

Further reading

- Numerous minorizers and majorizers for a variety of problems are presented in Heiser (1995).
- A recent short review and tutorial appears in Nguyen (2017).
- MM algorithms, as applied to signal processing problems are reviewed in Sun et al. (2017).
- Differences between EM and MM algorithms in some contexts are explored Wu and Lange (2010).
- A comprehensive treatment of MM algorithms is presented in the manuscript of Lange (2016).

References I

Amemiya, T. (1985). *Advanced Econometrics*. Harvard University Press, Cambridge.

Berchtold, A. (2004). Optimization of mixture models: comparison of different strategies. *Computational Statistics*, 19:385–406.

Bohning, D. and Lindsay, B. R. (1988). Monotonicity of quadratic-approximation algorithms. *Annals of the Institute of Mathematical Statistics*, 40:641–663.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, Cambridge.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20:273–297.

References II

de Leeuw, J. (1977). *Recent Developments in Statistics*, chapter Applications of convex analysis to multidimensional scaling, pages 133–146. North Holland, Amsterdam.

Heiser, W. J. (1995). *Recent Advances in Descriptive Multivariate Analysis*, chapter Convergent computing by iterative majorization: theory and applications in multidimensional data analysis, pages 157–189. Clarendon Press.

Hong, M., Razaviyayn, M., Luo, Z.-Q., and Pang, J.-S. (2016). A unified algorithmic framework for block-structured optimization involving big data: with applications in machine learning and signal process. *IEEE Signal Processing Magazine*, 33:57–77.

Hunter, D. R. and Lange, K. (2004). A tutorial on MM algorithms. *The American Statistician*, 58:30–37.

Lange, K. (2013). *Optimization*. Springer, New York.

References III

Lange, K. (2016). *MM Optimization Algorithms*. SIAM, Philadelphia.

Mairal, J. (2013). Stochastic majorization-minimization algorithms for large-scale optimization. In *Advances in Neural Information Processing Systems*, pages 2283–2291.

Mairal, J. (2015). Incremental majorization-minimization optimization with application to large-scale machine learning. *SIAM Journal of Optimization*, 25:829–855.

Meng, X.-L. (2000). Discussion of "Optimization transfer using surrogate objective functions" by K. Lange, D. Hunter and I. Yang. *Journal of Computational and Graphical Statistics*, 9:35–43.

Nguyen, H. D. (2017). An introduction to MM algorithms for machine learning and statistical estimation. *WIREs Data Mining and Knowledge Discovery*, 7:e1198.

References IV

Nguyen, H. D., Jones, A. T., and McLachlan, G. J. (2018). Stream-suitable optimization algorithms for some soft-margin support vector machine variants. *Japanese Journal of Statistics and Data Science*, to appear.

Nguyen, H. D. and McLachlan, G. J. (2017). Iteratively-reweighted least-squares fitting of support vector machines: a majorization-minimization algorithm approach. In *Proceedings of the 2017 Future Technologies Conference (FTC)*.

Ortega, J. M. and Rheinboldt, W. C. (1970). *Iterative solution of nonlinear equations in several variables*. Academic Press, San Diego.

Razaviyayn, M., Hong, M., and Luo, Z.-Q. (2013). A unified convergence analysis of block successive minimization methods for nonsmooth optimization. *SIAM Journal of Optimization*, 23:1126–1153.

References V

Razaviyayn, M., Sanjabi, M., and Luo, Z.-Q. (2016). A stochastic successive minimization method for nonsmooth nonconvex optimization with applications to transceiver design in wireless communication networks. *Mathematical Programming Series B*, 157:515–545.

Sun, Y., Babu, P., and Palomar, D. P. (2017). Majorization-minimization algorithms in signal processing, communications, and machine learning. *IEEE Transactions on Signal Processing*, in press.

Wu, T. T. and Lange, K. (2010). The MM alternative to EM. *Statistical Science*, 25:492–505.

Zhou, H. and Lange, K. (2010). MM algorithms for some discrete multivariate distributions. *Journal of Computational and Graphical Statistics*, 19:645–665.

Thank you for your attention!

Email: **h.nguyen5@latrobe.edu.au**

Twitter: **@tresbienhien**

Website: **https://hiendn.github.io**