Optimization in Machine Learning: Majorization Minimization Method

Ziping Zhao

School of Information Science and Technology
ShanghaiTech University, Shanghai, China

CS182: Introduction to Machine Learning (Spring 2023)
http://cs182.sist.shanghaitech.edu.cn

Majorization Minimization

Consider the following problem

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) \quad \text{s.t. } \boldsymbol{x} \in \mathcal{X}, \tag{1}$$

where $\mathcal{X}$ is a closed convex set; $f(\cdot)$ may be non-convex and/or nonsmooth.

- **Challenge**: For a general $f(\cdot)$, problem (1) can be difficult to solve.
- **Majorization Minimization (MM)**: Iteratively generate $\{\boldsymbol{x}^r\}$ as follows:

$$\boldsymbol{x}^r \in \min_{\boldsymbol{x}} u\left(\boldsymbol{x}, \boldsymbol{x}^{r-1}\right) \quad \text{s.t. } \boldsymbol{x} \in \mathcal{X}, \tag{2}$$

  where $u\left(\boldsymbol{x}, \boldsymbol{x}^{r-1}\right)$ is a surrogate function of $f(\boldsymbol{x})$, satisfying

  1. $u(\boldsymbol{x}, \boldsymbol{x}^r) \geq f(\boldsymbol{x}), \quad \forall \boldsymbol{x}^r, \boldsymbol{x} \in \mathcal{X}$;
  2. $u(\boldsymbol{x}^r, \boldsymbol{x}^r) = f(\boldsymbol{x}^r)$;

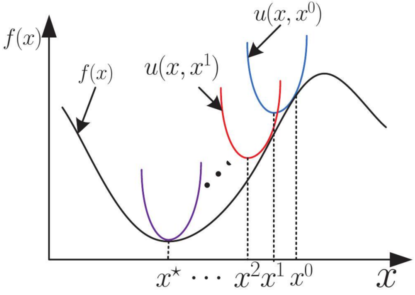


Figure: An pictorial illustration of MM algorithm.

Property 1. $\{f(\boldsymbol{x}^r)\}$ is nonincreasing, i.e., $f(\boldsymbol{x}^r) \leq f(\boldsymbol{x}^{r-1}), \forall r = 1, 2, \ldots$

Proof.

$f(\boldsymbol{x}^r) \leq u(\boldsymbol{x}^r, \boldsymbol{x}^{r-1}) \leq u(\boldsymbol{x}^{r-1}, \boldsymbol{x}^{r-1}) = f(\boldsymbol{x}^{r-1})$. □

- The nonincreasing property of $\{f(\boldsymbol{x}^r)\}$ implies that $f(\boldsymbol{x}^r) \to f^\infty$. But how about the convergence of the iterates $\{\boldsymbol{x}^r\}$?

Technical Preliminaries

- **Limit point**: $\mathbf{x}^\infty$ is a limit point of $\{\mathbf{x}^r\}$ if there exists a subsequence of $\{\mathbf{x}^r\}$ that converges to $\mathbf{x}^\infty$.
  - Note that every bounded sequence in $\mathbb{R}^n$ has a limit point (or convergent subsequence).
- **Directional derivative**: Let $f : \mathcal{D} \to \mathbb{R}$ be a function where $\mathcal{D} \subseteq \mathbb{R}^m$ is a convex set. The directional derivative of f at point x in direction $\boldsymbol{d}$ is defined by

$$f'(\boldsymbol{x}; \boldsymbol{d}) \triangleq \liminf_{\lambda \downarrow 0} \frac{f(\boldsymbol{x} + \lambda \boldsymbol{d}) - f(\boldsymbol{x})}{\lambda}.$$

  - If f is differentiable, then $f'(\boldsymbol{x}; \boldsymbol{d}) = \boldsymbol{d}^\top \nabla f(\boldsymbol{x})$.
- **Stationary point**: $\boldsymbol{x} \in \mathcal{X}$ is a stationary point of $f(\cdot)$ if

$$f'(\boldsymbol{x}; \boldsymbol{d}) \geq 0, \quad \forall \boldsymbol{d} \text{ such that } \boldsymbol{x} + \boldsymbol{d} \in \mathcal{D}. \tag{3}$$

  - A stationary point may be a local min., a local max. or a saddle point;
  - If $\mathcal{D} = \mathbb{R}^n$ and f is differentiable, then (3) $\iff \nabla f(\boldsymbol{x}) = \mathbf{0}$.

Convergence of MM

- **Assumption 1** $u(\cdot,\cdot)$ satisfies the following conditions

$$\begin{cases} u(\boldsymbol{y},\boldsymbol{y}) = f(\boldsymbol{y}), & \forall \boldsymbol{y} \in \mathcal{X}, \quad (4a) \\ u(\boldsymbol{x},\boldsymbol{y}) \geq f(\boldsymbol{x}), & \forall \boldsymbol{x},\boldsymbol{y} \in \mathcal{X}, \quad (4b) \\ u'(\boldsymbol{x},\boldsymbol{y};\boldsymbol{d})\,|_{\boldsymbol{x}=\boldsymbol{y}} = f'(\boldsymbol{y};\boldsymbol{d}) & \forall \boldsymbol{d} \text{ with } \boldsymbol{y}+\boldsymbol{d} \in \mathcal{X}, \quad (4c) \\ u(\boldsymbol{x},\boldsymbol{y}) \text{ is continuous in } (\boldsymbol{x},\boldsymbol{y}) & \quad (4d) \end{cases}$$

- (4c) means the 1st order local behavior of $u\left(\cdot,\boldsymbol{x}^{r-1}\right)$ is the same as $f(\cdot)$.

Theorem (Convergence of MM [1])

Assume that Assumption 1 is satisfied. Then every limit point of the iterates generated by MM algorithm is a stationary point of problem (1).

Convergence of MM

Proof of Theorem 1:
From **Property 1**, we know that $f\left(\boldsymbol{x}^{r+1}\right) \leq u\left(\boldsymbol{x}^{r+1}, \boldsymbol{x}^{r}\right) \leq u\left(\boldsymbol{x}, \boldsymbol{x}^{r}\right), \quad \forall \boldsymbol{x} \in \mathcal{X}$. Now assume that there exists a subsequence $\{\boldsymbol{x}^{r_j}\}$ of $\{\boldsymbol{x}^{r}\}$ converging to a limit point $\boldsymbol{z}$, i.e., $\lim_{j\to\infty} \boldsymbol{x}^{r_j} = \boldsymbol{z}$. Then

$$u\left(\boldsymbol{x}^{r_{j+1}}, \boldsymbol{x}^{r_{j+1}}\right) = f\left(\boldsymbol{x}^{r_{j+1}}\right) \leq f\left(\boldsymbol{x}^{r_j+1}\right) \leq u\left(\boldsymbol{x}^{r_j+1}, \boldsymbol{x}^{r_j}\right) \leq u\left(\boldsymbol{x}, \boldsymbol{x}^{r_j}\right), \quad \forall \boldsymbol{x} \in \mathcal{X}.$$

Letting $j \to \infty$, we obtain $u(\boldsymbol{z}, \boldsymbol{z}) \leq u(\boldsymbol{x}, \boldsymbol{z}), \quad \forall \boldsymbol{x} \in \mathcal{X}$, which implies that

$$u'(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{d})\big|_{\boldsymbol{x}=\boldsymbol{z}} \geq \mathbf{0}, \quad \forall \boldsymbol{z} + \boldsymbol{d} \in \mathcal{X}.$$

Combining the above inequality with (4c) (i.e., $u'(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{d})\big|_{\boldsymbol{x}=\boldsymbol{y}} = f'(\boldsymbol{y}; \boldsymbol{d}), \quad \forall \boldsymbol{d}$ with $\boldsymbol{y} + \boldsymbol{d} \in \mathcal{X}$), we have

$$f'(\boldsymbol{z}; \boldsymbol{d}) \geq \mathbf{0}, \quad \forall \boldsymbol{z} + \boldsymbol{d} \in \mathcal{X}.$$

Applications — Nonnegative Least Squares

In many engineering applications, we encounter the following problem

$$\text{(NLS)} \quad \min_{\boldsymbol{x} \geq \boldsymbol{0}} \|\mathbf{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2 \,, \tag{5}$$

where $\boldsymbol{b} \in \mathbb{R}_+^m$, $\boldsymbol{b} \neq \boldsymbol{0}$, and $\mathbf{A} \in \mathbb{R}_{++}^{m \times n}$.

- It's a least squares (LS) problem with nonnegative constraints, so the conventional LS solution may not be feasible for (5).
- A simple multiplicative updating algorithm:

$$\boldsymbol{x}_l^r = c_l^r \boldsymbol{x}_l^{r-1}, \quad l = 1, \ldots, n, \tag{6}$$

where $\boldsymbol{x}_l^r$ is the l th component of $\boldsymbol{x}^r$, and $c_l^r = \frac{\left[\mathbf{A}^T \boldsymbol{b}\right]_l}{\left[\mathbf{A}^T \mathbf{A} \boldsymbol{x}^{r-1}\right]_l}$.

- Starting with $\boldsymbol{x}^0 > \boldsymbol{0}$, then all $\boldsymbol{x}^r$ generated by (6) are nonnegative.

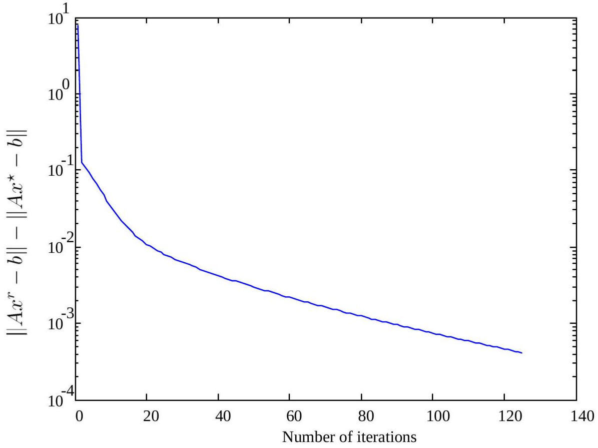


Figure: $\|\mathbf{A}\boldsymbol{x}^r - \boldsymbol{b}\|_2$ vs. the number of iterations.

- Usually the multiplicative update converges within a few tens of iterations.

- **MM interpretation**: Let $f(\boldsymbol{x}) \triangleq \|\mathbf{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2$. The multiplicative update essentially solves the following problem

$$\min_{\boldsymbol{x} \geq \mathbf{0}} \; u\left(\boldsymbol{x}, \boldsymbol{x}^{r-1}\right),$$

where

$$u\left(\boldsymbol{x}, \boldsymbol{x}^{r-1}\right) \triangleq f\left(\boldsymbol{x}^{r-1}\right) + \left(\boldsymbol{x} - \boldsymbol{x}^{r-1}\right)^\top \nabla f\left(\boldsymbol{x}^{r-1}\right) + \frac{1}{2}\left(\boldsymbol{x} - \boldsymbol{x}^{r-1}\right)^\top \boldsymbol{\Phi}\left(\boldsymbol{x}^{r-1}\right)\left(\boldsymbol{x} - \boldsymbol{x}^{r-1}\right),$$

$$\boldsymbol{\Phi}\left(\boldsymbol{x}^{r-1}\right) = \left(\frac{\left[\mathbf{A}^\top \mathbf{A}\boldsymbol{x}^{r-1}\right]_1}{x_1^{r-1}}, \ldots, \frac{\left[\mathbf{A}^\top \mathbf{A}\boldsymbol{x}^{r-1}\right]_n}{x_n^{r-1}}\right).$$

- Observations:

$$\begin{cases} u\left(\boldsymbol{x}, \boldsymbol{x}^{r-1}\right) \text{ is quadratic approx. of } f(\boldsymbol{x}), \\ \boldsymbol{\Phi}\left(\boldsymbol{x}^{r-1}\right) \succeq \mathbf{A}^\top \mathbf{A}, \end{cases} \Rightarrow \begin{cases} u\left(\boldsymbol{x}, \boldsymbol{x}^{r-1}\right) \geq f(\boldsymbol{x}), \; \forall \boldsymbol{x} \in \mathbb{R}^n, \\ u\left(\boldsymbol{x}^{r-1}, \boldsymbol{x}^{r-1}\right) = f\left(\boldsymbol{x}^{r-1}\right). \end{cases}$$

  - The multiplicative update converges to an optimal solution of NLS (by the MM convergence in **Theorem 1** and convexity of NLS).

Applications — Convex-Concave Procedure / Difference-of-Convex (DC) Programming

- Suppose that $f(\boldsymbol{x})$ has the following form $f(\boldsymbol{x}) = g(\boldsymbol{x}) - h(\boldsymbol{x})$, where $g(\boldsymbol{x})$ and $h(\boldsymbol{x})$ are convex and differentiable. Thus, $f(\boldsymbol{x})$ is in general nonconvex.
- **DC Programming:** Construct $u(\cdot,\cdot)$ as

$$u\left(\boldsymbol{x}, \boldsymbol{x}^r\right) = g(\boldsymbol{x}) - \underbrace{\left(h\left(\boldsymbol{x}^r\right) + \nabla_{\boldsymbol{x}} h\left(\boldsymbol{x}^r\right)^\top \left(\boldsymbol{x} - \boldsymbol{x}^r\right)\right)}_{\text{linearization of } h \text{ at } x^r}.$$

- By the 1st order condition of $h(\boldsymbol{x})$, it's easy to show that

$$u\left(\boldsymbol{x}, \boldsymbol{x}^r\right) \geq f(\boldsymbol{x}), \forall \boldsymbol{x} \in \mathcal{X}, \qquad \left(\boldsymbol{x}^r, \boldsymbol{x}^r\right) = f\left(\boldsymbol{x}^r\right).$$

- **Sparse Signal Recovery by DC Programming**

$$\min_{\boldsymbol{x}} \ \|\boldsymbol{x}\|_0 \,, \qquad \text{s.t. } \boldsymbol{y} = \mathbf{A}\boldsymbol{x}. \tag{7}$$

- Apart from the popular ℓ_1 approximation, consider the following concave approximation

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \sum_{i=1}^{n} \log\left(1 + |x_i| / \epsilon\right) \qquad \text{s.t. } \boldsymbol{y} = \mathbf{A}\boldsymbol{x},$$

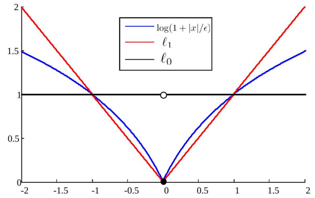


Figure: $\log(1 + |x| / \epsilon)$ promotes more sparsity than ℓ_1

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \sum_{i=1}^{n} \log\left(1 + |x_i| / \epsilon\right) \quad \text{s.t. } \boldsymbol{y} = \mathbf{A}\boldsymbol{x},$$

which can be equivalently written as

$$\min_{\boldsymbol{x},\boldsymbol{z}\in\mathbb{R}^n} \sum_{i=1}^{n} \log\left(z_i + \epsilon\right) \quad \text{s.t. } \boldsymbol{y} = \mathbf{A}\boldsymbol{x}, |x_i| \leq z_i, i = 1, \ldots, n \tag{8}$$

– Problem (8) minimizes a concave objective, so it's a special case of DC programming ($g(\boldsymbol{x}) = 0$). Linearizing the concave function at $(\boldsymbol{x}^r, \boldsymbol{z}^r)$ yields

$$\left(\boldsymbol{x}^{r+1}, \boldsymbol{z}^{r+1}\right) = \arg\min \sum_{i=1}^{n} \frac{z_i}{z_i^r + \epsilon} \quad \text{s.t. } \boldsymbol{y} = \mathbf{A}\boldsymbol{x}, |x_i| \leq z_i, i = 1, \ldots, n$$

– We solve a sequence of reweighted ℓ_1 problems.

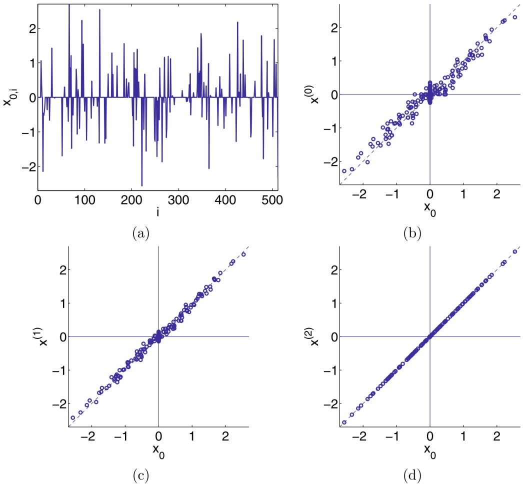


Figure: Sparse signal recovery through reweighted ℓ_1 iterations. (**a**) Original length $n = 512$ signal x_0 with 130 spikes. (**b**) Scatter plot, coefficient-by-coefficient, of x_0 versus its reconstruction $x^{(0)}$ using unweighted ℓ_1 minimization. (**c**) Reconstruction $x^{(1)}$ after the first reweighted iteration. (**d**) Reconstruction $x^{(2)}$ after the second reweighted iteration.

Applications — $\ell_2 - \ell_p$ Optimization

- Many problems involve solving the following problem (e.g., basis-pursuit denoising)

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) \triangleq \frac{1}{2} \|\boldsymbol{y} - \mathbf{A}\boldsymbol{x}\|_2^2 + \mu \|\boldsymbol{x}\|_p , \tag{9}$$

  where $p \geq 1$.
- If $\mathbf{A} = \mathbf{I}$ or $\mathbf{A}$ is unitary, optimal $\boldsymbol{x}^\star$ is computed in closed-form as

$$\boldsymbol{x}^\star = \mathbf{A}^\top \boldsymbol{y} - \mathrm{Proj}_C \left(\mathbf{A}^\top \boldsymbol{y}\right)$$

  where $C \triangleq \left\{\boldsymbol{x} : \|\boldsymbol{x}\|_{p*} \leq \mu\right\}$, $\|\cdot\|_{p*}$ is the dual norm of $\|\cdot\|_p$ and Proj_C denotes the projection operator. In particular, for $p = 1$,

$$x_i^\star = \mathrm{soft}\left(y_i, \mu\right), \quad i = 1, \ldots, n,$$

  where $\mathrm{soft}\left(u, a\right) \triangleq \mathrm{sign}(u) \max\{|u| - a, 0\}$ denotes a *soft-thresholding* operation.
- For general $\mathbf{A}$, there is no simple closed-form solution for (9).

- **MM for $\ell_2 - \ell_p$ Problem**: Consider a modified $\ell_2 - \ell_p$ problem

$$\min_{\boldsymbol{x}} \; u\left(\boldsymbol{x}, \boldsymbol{x}^r\right) \triangleq f(\boldsymbol{x}) + \operatorname{dist}\left(\boldsymbol{x}, \boldsymbol{x}^r\right), \tag{10}$$

where $\operatorname{dist}\left(\boldsymbol{x}, \boldsymbol{x}^r\right) \triangleq \frac{c}{2}\left\|\boldsymbol{x} - \boldsymbol{x}^r\right\|_2^2 - \frac{1}{2}\left\|\mathbf{A}\boldsymbol{x} - \mathbf{A}\boldsymbol{x}^r\right\|_2^2$ and $c > \lambda_{\max}\left(\mathbf{A}^\top\mathbf{A}\right)$.

  - $\operatorname{dist}\left(\boldsymbol{x}, \boldsymbol{x}^r\right) \geq 0 \; \forall \boldsymbol{x} \Longrightarrow u\left(\boldsymbol{x}, \boldsymbol{x}^r\right)$ majorizes $f(\boldsymbol{x})$.
  - $u\left(\boldsymbol{x}, \boldsymbol{x}^r\right)$ can be reexpressed as

$$u\left(\boldsymbol{x}, \boldsymbol{x}^r\right) = \frac{c}{2}\left\|\boldsymbol{x} - \overline{\boldsymbol{x}}^r\right\|_2^2 + \mu\left\|\boldsymbol{x}\right\|_p + \text{ const.},$$

where

$$\overline{\boldsymbol{x}}^r = \frac{1}{c}\mathbf{A}^\top\left(\boldsymbol{y} - \mathbf{A}\boldsymbol{x}^r\right) + \boldsymbol{x}^r.$$

  - The modified $\ell_2 - \ell_p$ problem (10) has a simple soft-thresholding solution.
  - Repeatedly solving problem (10) leads to an optimal solution of the $\ell_2 - \ell_p$ problem (by the MM convergence in **Theorem 1**).

Applications — Gradient Descent (GD)

- GD:

$$
\begin{aligned}
\mathbf{x}^{r+1} &= \mathbf{x}^r - \gamma \nabla f(\mathbf{x}^r) \\
&= \arg\min_{\mathbf{x}} \underbrace{\left\{ f(\mathbf{x}^r) + \nabla f(\mathbf{x}^r)^\top (\mathbf{x} - \mathbf{x}^r) + \frac{1}{2\gamma} \|\mathbf{x} - \mathbf{x}^r\|^2 \right\}}_{L(\mathbf{x} \,|\, \mathbf{x}^r)} \quad \text{[verify it]}
\end{aligned}
$$

- Choose $\gamma \leq 1/L$: $L(\mathbf{x} \,|\, \mathbf{x}^r) \geq f(\mathbf{x})$

Proof of descent

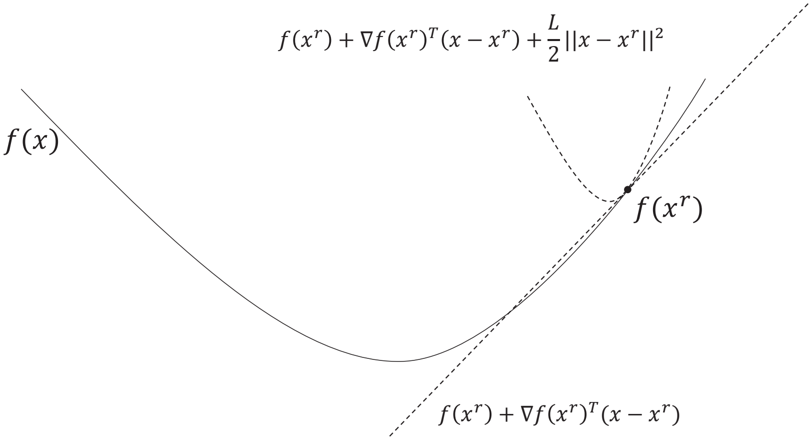
Majorize: by descent lemma and $\gamma \leq 1/L$

$$
\begin{aligned}
f(\mathbf{x}) &\leq f(\mathbf{x}^r) + \nabla f(\mathbf{x}^r)^\top (\mathbf{x} - \mathbf{x}^r) + \frac{L}{2}\|\mathbf{x} - \mathbf{x}^r\|^2 \\
&\leq f(\mathbf{x}^r) + \nabla f(\mathbf{x}^r)^\top (\mathbf{x} - \mathbf{x}^r) + \frac{1}{2\gamma}\|\mathbf{x} - \mathbf{x}^r\|^2
\end{aligned}
$$

Minimize: let $\mathbf{x} = \mathbf{x}^r - \gamma \nabla f(\mathbf{x}^r)$

$$
\begin{aligned}
f(\mathbf{x}^{r+1}) &\leq f(\mathbf{x}^r) - \gamma \|\nabla f(\mathbf{x}^r)\|^2 + \frac{\gamma}{2}\|\nabla f(\mathbf{x}^r)\|^2 \\
&= f(\mathbf{x}^r) - \frac{\gamma}{2}\|\nabla f(\mathbf{x}^r)\|^2.
\end{aligned}
$$

Quadratic upperbound - a majorization minimization perspective



Applications — Expectation Maximization (EM)

- Consider an ML estimate of θ, given the random observation w
$$\hat{\theta}_{\mathrm{ML}} = \arg\min_{\theta} - \ln p(w \mid \theta).$$
- Suppose that there are some missing data or hidden variables z in the model. Then, EM algorithm iteratively compute an ML estimate $\hat{\theta}$ as follows:
  - E-step:
$$g(\theta, \theta^r) \triangleq \mathbb{E}_{z|w,\theta^r} \{\ln p(w, z \mid \theta)\}.$$
  - M-step:
$$\theta^{r+1} = \arg\max_{\theta} g(\theta, \theta^r).$$
  - repeat the above two steps until convergence.
- EM algorithm generates a nonincreasing sequence of $\{-\ln p(w \mid \theta^r)\}$.
- EM algorithm can be interpreted by MM.

- MM interpretation of EM algorithm:

$$
\begin{aligned}
& -\ln p(w \mid \theta) \\
=& -\ln \mathbb{E}_{z|\theta} p(w \mid z, \theta) \\
=& -\ln \mathbb{E}_{z|\theta} \left[\frac{p(z \mid w, \theta^r)\, p(w \mid z, \theta)}{p(z \mid w, \theta^r)} \right] \\
=& -\ln \mathbb{E}_{z|w,\theta^r} \left[\frac{p(z \mid \theta) p(w \mid z, \theta)}{p(z \mid w, \theta^r)} \right] \qquad \text{(interchange the integrations)} \\
\leq& -\mathbb{E}_{z|w,\theta^r} \ln \left[\frac{p(z \mid \theta) p(w \mid z, \theta)}{p(z \mid w, \theta^r)} \right] \qquad \text{(Jensen's inequality)} \\
=& -\mathbb{E}_{z|w,\theta^r} \ln p(w, z \mid \theta) + \mathbb{E}_{z|w,\theta^r} \ln p(z \mid w, \theta^r) \\
\triangleq& u(\theta, \theta^r)
\end{aligned} \tag{11}
$$

  - $u(\theta, \theta^r)$ majorizes $-\ln p(w \mid \theta)$, and $-\ln p(w \mid \theta^r) = u(\theta^r, \theta^r)$;
  - E-step essentially constructs $u(\theta, \theta^r)$;
  - M-step minimizes $u(\theta, \theta^r)$ (note θ appears in the 1st term (11) only).

Meisam Razaviyayn, Mingyi Hong, and Zhi-Quan Luo.
A unified convergence analysis of block successive minimization methods for nonsmooth optimization.
SIAM Journal on Optimization, 23(2):1126–1153, 2013.