Chapter 2: Alternating Minimization

Ernest K. Ryu

MATH 164: Optimization
University of California, Los Angeles
Department of Mathematics

Last edited: February 10, 2025

Minimize one block at a time

Simple algorithm: Minimize each block one at a time.

Commonly used when block updates have closed-form formulae.

Convergence can be shown under appropriate assumptions. Convergence rates are not particularly impressive.

Problem setup

Consider

$$\underset{x \in \mathbb{R}^{n_1+n_2+\cdots+n_p}}{\text{minimize}} \quad f(x_{(1)}, x_{(2)}, \ldots, x_{(p)}),$$

where $x = (x_{(1)}, x_{(2)}, \ldots, x_{(p)})$ and

$$x_{(i)} = \begin{bmatrix} x_{(i),1} \\ x_{(i),2} \\ \vdots \\ x_{(i),n_i} \end{bmatrix} \in \mathbb{R}^{n_i}, \qquad \text{for } i = 1, \ldots, p.$$

This is an unconstrained optimization problem with the x-variable partitioned into p blocks.

Alternating minimization method

Use the notation

$$x^k = \big(x^k_{(1)}, \dots, x^k_{(p)}\big).$$

Alternating minimization (AM) method updates $x^k \mapsto x^{k+1}$ via

$$x^{k+1}_{(i)} \in \underset{z \in \mathbb{R}^{n_i}}{\operatorname{argmin}}\, f\big(x^{k+1}_{(1)}, \dots, x^{k+1}_{(i-1)}, z, x^k_{(i+1)}, \dots, , x^k_{(p)}\big)$$

for $i = 1, \dots, p$. Stars from initialization $x^0 = \big(x^0_{(1)}, \dots, x^0_{(p)}\big)$.
(Actually, no need to initialize $x^0_{(1)}$.)

There are no stepsizes!

Other names of AM

Coordinate minimization: When all of the blocks have size 1, i.e., if $n_1 = n_2 = \cdots = n_p = 1$.

Gauss–Seidel: When there are 2 blocks, i.e., if $p = 2$.

Block coordinate descent: A variant of AM where instead of finding the coordinate-wise minimizer at each step, one performs a coordinate-wise gradient update.

Minimizer vs. coordinate-wise minimizer

We say $x = (x_{(1)}, \dots, x_{(p)})$ is a *minimizer* of f if

$$f(x_{(1)}, \dots, x_{(p)}) \le f(z_{(1)}, \dots, z_{(p)}) \qquad \forall\, (z_{(1)}, \dots, z_{(p)}) \in \mathbb{R}^{n_1 + \cdots + n_p}.$$

I.e., deviating from x in any way cannot reduce f.

We say $x = (x_{(1)}, \dots, x_{(p)})$ is a *coordinate-wise* minimizer of f if

$$\begin{aligned} f(x_{(1)}, \dots, x_{(i-1)}, x_{(i)}, x_{(i+1)}, \dots, x_{(p)}) & \\ \le f(x_{(1)}, \dots, x_{(i-1)}, z, x_{(i+1)}, \dots, x_{(p)}), & \qquad \forall\, z \in \mathbb{R}^{n_i} \end{aligned}$$

for all $i = 1, \dots, p$. I.e., unilaterally changing any individual block of x cannot reduce f.

Convergence of AM without differentiability

Theorem.

Let $f : \mathbb{R}^{n_1+\cdots+n_p} \to \mathbb{R}$ is continuous. Consider alternating minimization, and assume the iterates $\{x^k\}_k$ are well-defined. If $x^k \to \bar{x}$, then $\bar{x}$ is a coordinate-wise minimizer of f.

Alternating minimization is often used for problems non-differentiable optimization problems. Therefore, it is useful to analyze its convergence properties and its failure modes in the absence of differentiability.

Convergence of AM without differentiability

Theorem.

Let $f : \mathbb{R}^{n_1+\cdots+n_p} \to \mathbb{R}$ *is continuous. Consider alternating minimization, and assume the iterates* $\{x^k\}_k$ *are well-defined. If* $x^k \to \bar{x}$, *then* $\bar{x}$ *is a coordinate-wise minimizer of* f.

We clarify a few points about what the theorem is *not* claiming.

- Without further assumptions, we do not know if

$$x_{(i)}^{k+1} \in \underset{z \in \mathbb{R}^{n_i}}{\operatorname{argmin}} f\big(x_{(1)}^{k+1}, \ldots, x_{(i-1)}^{k+1}, z, x_{(i+1)}^{k}, \ldots, , x_{(p)}^{k}\big)$$

  is well-defined, i.e., a minimizer may not exist. (In practice, however, this is often not a problem.)
- The $\{x^k\}_k$ may or may not converge. (In practice, however, this is often not a problem.)
- The coordinate-wise minimum may not be a minimum.

Convergence of AM without differentiability

Theorem.

Let $f : \mathbb{R}^{n_1+\cdots+n_p} \to \mathbb{R}$ is continuous. Consider alternating minimization, and assume the iterates $\{x^k\}_k$ are well-defined. If $x^k \to \bar{x}$, then $\bar{x}$ is a coordinate-wise minimizer of f.

Proof. Since $x_{(1)}^{k+1}$ is defined as a coordinate-wise minimizer,

$$f\left(x_{(1)}^{k+1}, x_{(2)}^{k}, \ldots, x_{(p)}^{k}\right) \leq f\left(z, x_{(2)}^{k}, \ldots, x_{(p)}^{k}\right), \qquad \forall\, z \in \mathbb{R}^{n_1}.$$

Taking the limit $k \to \infty$ on both sides,

$$f\left(\bar{x}_{(1)}, \bar{x}_{(2)}, \ldots, \bar{x}_{(p)}\right) \leq f\left(z, \bar{x}_{(2)}, \ldots, \bar{x}_{(p)}\right), \qquad \forall\, z \in \mathbb{R}^{n_1}.$$

This shows that $\bar{x}$ is a coordinate-wise minimizer with respect to the first block. Repeating the argument for blocks $i = 2, \ldots, p$, we conclude the statement. □

Coordinate-wise minimizer is not a minimizer

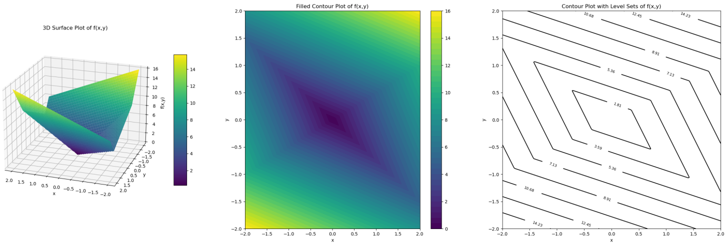
Let $f\colon \mathbb{R}^2 \to \mathbb{R}$ be

$$f(x,y) = |3x + 4y| + |x - 2y|.$$

The global minimizer is $(0,0)$, but $(-4\alpha, 3\alpha)$ for any $\alpha \in \mathbb{R}$ is a coordinate-wise minimizer. (Note that f is convex, so non-convexity is not the cause of any trouble.)

For most starting points, AM will get stuck at $(-4\alpha, 3\alpha)$ with some α

This is a mode of failure of AM. When AM converges the limit may not be a global or local minimum.



Coordinate-wise minimizer is a stationary point under differentiability

Under differentiability, a coordinate-wise minimizer is a stationary point.

Lemma.

If $f : \mathbb{R}^{n_1+\cdots+n_p} \to \mathbb{R}$ is differentiable, then a coordinate-wise minimizer is a stationary point (i.e., $\nabla f(x) = 0$.)

Proof. Let x be a coordinate-wise minimizer. Then,

$$\begin{aligned} f(x+\varepsilon d) &= f(x) + \varepsilon \langle \nabla f(x), d \rangle + o(\varepsilon) \\ &= f(x) + \varepsilon \|\nabla_{x_i} f(x)\|^2 + o(\varepsilon) \end{aligned} \qquad d = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \nabla_{x_i} f(x) \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

for any $i = 1, \ldots, p$. Since x is a coordinate-wise minimizer, $f(x+\varepsilon d) \geq f(x)$ for any ε, so $\|\nabla_{x_i} f(x)\|^2 = 0$ for any $i = 1, \ldots, p$, and we conclude $\nabla f(x) = 0$. $\square$

Convergence of AM with differentiability

Theorem.
Let $f\colon \mathbb{R}^{n_1+\cdots+n_p} \to \mathbb{R}$ *be* **differentiable**. *Consider alternating minimization, and assume the iterates* $\{x^k\}_k$ *are well-defined. If* $x^k \to \bar{x}$, *then* $\bar{x}$ *is a* **coordinate-wise minimizer and a stationary point** *of* f.

In practice, a point that is [a coordinate-wise minimizer and a stationary point] is often a local minimizer. So we can understand this result as essentially a guarantee to converge to a local minimum.

However, although AM finds the coordinate-wise global minimizer at each update, the limit $\bar{x}$ is often *not* a global minimum.

Convergence of AM with differentiability and convexity

Theorem.
Let $f : \mathbb{R}^{n_1+\cdots+n_p} \to \mathbb{R}$ *be* **convex and differentiable**. *Consider alternating minimization, and assume the iterates* $\{x^k\}_k$ *are well-defined. If* $x^k \to \bar{x}$, *then* $\bar{x}$ *is a* **(global) minimizer** *of* f.

Recall, the counterexample showed that with a convex but non-differentiable f, AM may get stuck at a point that is not a minimizer.

Example: Low-rank matrix completion

Given a matrix M, if we observe some of the entries, can we reconstruct the entire matrix?

$$\begin{pmatrix} 1 & ? & ? & 4 & ? \\ ? & 2 & 5 & ? & ? \\ ? & ? & 4 & 5 & ? \\ 5 & ? & ? & ? & 4 \end{pmatrix}$$

Example: Low-rank matrix completion

In the Netflix Competition (Netflix Prize) of 2006–2009, the goal is to recommend movies well to the users.

Specifically, there are m users and n movies. Each user has watched some movies and have provided ratings. Let

$$\Omega = \{(i,j) \,|\, \text{user } i \text{ has rated movie } j\} \subseteq \{1,\dots,m\} \times \{1,\dots,n\}.$$

Let M_{ij} for $(i,j) \in \Omega$ be the score given by user i to movie j.

Can we predict all of $M \in \mathbb{R}^{m\times n}$? Then, if M_{ij} is big for some $(i,j) \notin \Omega$, Netflix can recommend movie j to user i.

Of course, M is not a completely unstructured collection of numbers, and any solution *must* utilize some structure of M. It turns out that assuming M has **low rank** leads to good results.

Example: Low-rank matrix completion

Assume a matrix $M \in \mathbb{R}^{m \times n}$ has rank r. This implies that M can be written as a low-rank product of the form

$$M = \underbrace{\begin{bmatrix} & & & \\ \end{bmatrix}}_{m \times r} \underbrace{\begin{bmatrix} \\ \\ \\ \end{bmatrix}}_{r \times n} \in \mathbb{R}^{m \times n}$$

M has entries M_{ij}, and there are mn such entries. Assume we observe a subset of the entries. Let

$$\Omega = \{(i,j) \,|\, \text{we know the the value of } M_{ij}\} \subseteq \{1, \ldots, m\} \times \{1, \ldots, n\}.$$

be the set of observation indices.

Goal: Reconstruct all of $M \in \mathbb{R}^{m \times n}$.

Example: Low-rank matrix completion

We form an explicit factorization $M = LR$ and fit L and R on the observed entries.

$$\underset{L\in\mathbb{R}^{m\times r},\, R\in\mathbb{R}^{r\times n}}{\text{minimize}} \quad \sum_{(i,j)\in\Omega} \frac{1}{2}\big(M - LR\big)_{ij}^2 = \sum_{(i,j)\in\Omega} \frac{1}{2}\big(M_{ij} - L_i R_j\big)^2,$$

where

$$L = \begin{bmatrix} — L_1 — \\ — L_2 — \\ \vdots \\ — L_m — \end{bmatrix}, \qquad R = \begin{bmatrix} | & | & & | \\ R_1 & R_2 & \cdots & R_n \\ | & | & & | \end{bmatrix}.$$

To clarify, $L_i R_j$ is an inner product between the row vector $L_i \in \mathbb{R}^{1\times r}$ and the column vector $R_j \in \mathbb{R}^{r\times 1}$.

Let us use alternating minimization, minimizing with respect to L and then R, to solve this problem.

Example: Low-rank matrix completion

Let

$$\Omega_i^I = \{j \,|\, (i,j) \in \Omega\} = (\text{list of movies } j \text{ that user } i \text{ rated})$$

for $i = 1, \ldots, m$. Then, we can write

$$\sum_{(i,j)\in\Omega} \square = \sum_{i=1}^{m} \sum_{j\in\Omega_i^I} \square$$

Likewise, let

$$\Omega_j^J = \{i \,|\, (i,j) \in \Omega\} = (\text{list users } i \text{ who have rated movie } j)$$

for $j = 1, \ldots, n$. Then, we can write

$$\sum_{(i,j)\in\Omega} \square = \sum_{j=1}^{n} \sum_{i\in\Omega_j^J} \square$$

Example: Low-rank matrix completion

Next, compute the alternating updates in closed forms. Let

$$\mathcal{J} = \sum_{(i,j)\in\Omega} \frac{1}{2}\big(M_{ij} - L_i R_j\big)^2 = \sum_{i=1}^{m} \sum_{j\in\Omega_i^I} \frac{1}{2}\big(M_{ij} - L_i R_j\big)^2.$$

Then

$$\frac{\partial \mathcal{J}}{\partial (L_i)_k} = \sum_{j\in\Omega_i^I} \big(M_{ij} - L_i R_j\big)(R_j)_k$$

and vectorizing this, we get

$$\begin{aligned}
\nabla_{L_i}\mathcal{J} &= \begin{bmatrix} \frac{\partial \mathcal{J}}{\partial (L_i)_1} & \cdots & \frac{\partial \mathcal{J}}{\partial (L_i)_r} \end{bmatrix} \\
&= \sum_{j\in\Omega_i^I} \big(M_{ij} - L_i R_j\big) \begin{bmatrix} (R_j)_1 & \cdots & (R_j)_r \end{bmatrix} \\
&= \sum_{j\in\Omega_i^I} \big(M_{ij} - L_i R_j\big) R_j^\intercal \\
&= \sum_{j\in\Omega_i^I} M_{ij} R_j^\intercal - L_i \sum_{j\in\Omega_i^I} R_j R_j^\intercal = 0.
\end{aligned}$$

Example: Low-rank matrix completion

Right-multiply $(\sum R_j R_j^\intercal)^{-1}$ on both sides of

$$\sum_{j \in \Omega_i^I} M_{ij} R_j^\intercal = L_i \sum_{j \in \Omega_i^I} R_j R_j^\intercal$$

to get

$$L_i = \Big(\sum_{j \in \Omega_i^I} M_{ij} R_j^\intercal \Big) \Big(\sum_{j \in \Omega_i^I} R_j R_j^\intercal \Big)^{-1} \in \mathbb{R}^{1 \times r}.$$

To get column vectors, we transpose both sides to get

$$L_i^\intercal = \Big(\sum_{j \in \Omega_i^I} R_j R_j^\intercal \Big)^{-1} \Big(\sum_{j \in \Omega_i^I} R_j M_{ij} \Big) \in \mathbb{R}^{r \times 1}.$$

Example: Low-rank matrix completion

We vectorize

$$L_i^\intercal = \Big(\sum_{j \in \Omega_i^I} \underbrace{R_j R_j^\intercal}_{(r\times 1) \text{ by } (1\times r)} \Big)^{-1} \Big(\sum_{j \in \Omega_i^I} \underbrace{R_j M_{ij}}_{(r\times 1) \text{ by } (1\times 1)} \Big)$$

a bit further to get

$$L_i^\intercal = \Big(\underbrace{R_{\Omega_i^I} R_{\Omega_i^I}^\intercal}_{(r\times|\Omega_i^I|) \text{ by } (|\Omega_i^I|\times r)} \Big)^{-1} \Big(\underbrace{R_{\Omega_i^I} M_{i,\Omega_i^I}^\intercal}_{(r\times|\Omega_i^I|) \text{ by } (|\Omega_i^I|\times 1)} \Big),$$

where

$$\Omega_i^I = \{j_1, j_2, \ldots, j_{|\Omega_i^I|}\},$$

$$R_{\Omega_i^I} = \begin{bmatrix} R_{j_1} & R_{j_2} & \cdots & R_{j_{|\Omega_i^I|}} \end{bmatrix} \in \mathbb{R}^{r\times|\Omega_i^I|},$$

$$M_{i,\Omega_i^I} = \begin{bmatrix} M_{i,j_1} & M_{i,j_2} & \cdots & M_{i,j_{|\Omega_i^I|}} \end{bmatrix} \in \mathbb{R}^{1\times|\Omega_i^I|}.$$

Sub-indexing arrays is well-supported in Python.

Example: Low-rank matrix completion

We have arrived at the update

$$L_i^{\mathsf{T}} = \left(R_{\Omega_i^I} R_{\Omega_i^I}^{\mathsf{T}}\right)^{-1} \left(R_{\Omega_i^I} M_{i,\Omega_i^I}^{\mathsf{T}}\right),$$

With an analogous argument, we have

$$R_j = \left(L_{\Omega_j^J}^{\mathsf{T}} L_{\Omega_j^J}\right)^{-1} \left(L_{\Omega_j^J}^{\mathsf{T}} M_{\Omega_j^J,j}\right),$$

We have now derived an alternating minimization algorithm

$$\begin{aligned} L_i^{k+1,\mathsf{T}} &= \left(R_{\Omega_i^I}^k R_{\Omega_i^I}^{k,\mathsf{T}}\right)^{-1} \left(R_{\Omega_i^I}^k M_{i,\Omega_i^I}^{\mathsf{T}}\right), && \text{for } i = 1, \dots, m \\ R_j^{k+1} &= \left(L_{\Omega_j^J}^{k+1,\mathsf{T}} L_{\Omega_j^J}^{k+1}\right)^{-1} \left(L_{\Omega_j^J}^{k+1,\mathsf{T}} M_{\Omega_j^J,j}\right), && \text{for } i = 1, \dots, n \end{aligned}$$

for $k = 0, 1, \dots$.