14 Alternating Minimization

Goal

Alternating minimization, convex function estimation, separability, counterexamples, Nash equilibrium, regularity, convergence condition, coordinate gradient descent

Alert 14.1: Convention

Gray boxes are not required hence can be omitted for unenthusiastic readers.

This note is likely to be updated again soon.

Definition 14.2: Problem

The problem we study in this lecture is the following:

$$\inf_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w}), \quad \text{where} \quad f(\mathbf{w}) = f_0(\mathbf{w}) + \sum_{j=1}^{d} f_j(w_j), \tag{14.1}$$

where we typically have f_0 smooth in mind, while we note that the second component function is separable. More generally, each w_j could itself be a vector, although the alternating minimization algorithm below is more convenient when w_j's are scalars. A special case arises when $f_j(w_j) = \iota_{C_j}(w_j)$, i.e. we minimize a function f_0 over the Cartesian product $C := C_1 \times \cdots \times C_d$.

Example 14.3: Convex function estimation (Hildreth 1954)

Recall in linear regression we assumed the following model:

$$y = f(\mathbf{x}) + \epsilon, \quad \text{where} \quad f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}.$$

Instead of the linear parametric form above, in many applications (e.g. econometrics and more recently optimal transportation) it makes sense to assume the unknown function $f : \mathbb{R}^d \to \mathbb{R}$ to be just convex. In particular, when $d = 1$, this property is known as diminishing return (where x is the price and y is the return):

$$\forall x > v > z, \quad \frac{f(x) - f(v)}{x - v} \geq \frac{f(v) - f(z)}{v - z}.$$

Thus, given a dataset $\{(x_i, y_i) : i = 1, \ldots, n\}$ we are interested in solving the convex estimation problem:

$$\min_{f:\mathbb{R}\to\mathbb{R} \text{ convex}} \sum_{i=1}^{n} \alpha_i (f(x_i) - y_i)^2,$$

where α_i are some given weights (e.g. $\alpha_i \equiv 1/n$). Assuming w.l.o.g. that $x_1 > x_2 > \cdots > x_n$, the above problem can be shown equivalently as:

$$\begin{aligned} \min_{\mathbf{z}\in\mathbb{R}^n} \quad & \sum_{i=1}^{n} \alpha_i (z_i - y_i)^2, \\ \text{s.t.} \quad & \frac{z_i - z_{i+1}}{x_i - x_{i+1}} \geq \frac{z_{i+1} - z_{i+2}}{x_{i+1} - x_{i+2}}, \quad i = 1, \ldots, n-2. \end{aligned}$$

Define the difference matrix

$$B = \begin{bmatrix} -\frac{1}{\Delta_1} & \frac{1}{\Delta_1}+\frac{1}{\Delta_2} & -\frac{1}{\Delta_2} & 0 & \cdots & 0 & 0 & 0 \\ 0 & -\frac{1}{\Delta_2} & \frac{1}{\Delta_2}+\frac{1}{\Delta_3} & -\frac{1}{\Delta_3} & \cdots & 0 & 0 & 0 \\ & & & & \ddots & & & \\ 0 & 0 & 0 & 0 & \cdots & -\frac{1}{\Delta_{n-2}} & \frac{1}{\Delta_{n-2}}+\frac{1}{\Delta_{n-1}} & -\frac{1}{\Delta_{n-1}} \end{bmatrix}, \quad \text{where} \quad \Delta_i = x_i - x_{i+1},$$

and let $D = \operatorname{diag}(\boldsymbol{\alpha})$, we arrive at a simple quadratic program:

$$\min_{B\mathbf{z}\leq \mathbf{0}} \ \tfrac{1}{2}(\mathbf{z}-\mathbf{y})^\top D(\mathbf{z}-\mathbf{y}), \tag{14.2}$$

whose Lagrangian dual (see Definition 0.46) is

$$\min_{\boldsymbol{\lambda}\in\mathbb{R}^{n-2}_+} \ \tfrac{1}{2}\boldsymbol{\lambda}^\top A\boldsymbol{\lambda} + \boldsymbol{\lambda}^\top\mathbf{b}, \quad \text{where} \quad A = BD^{-1}B^\top, \ \mathbf{b} = -B\mathbf{y} \tag{14.3}$$

and $\mathbf{z} = \mathbf{y} - D^{-1}B^\top\boldsymbol{\lambda}$. We may recover an estimate of the convex function f through linearly interpolating $\{(x_i, z_i) : i = 1, \ldots, n\}$.

We have a few choices here, e.g.:

- Solve the primal problem (14.2) using say conditional gradient Algorithm 6.7, each iteration of which requires minimizing a linear function over the cone $B\mathbf{z} \leq \mathbf{0}$: not so trivial.
- Solve the dual problem (14.3), which is a special case of our general problem (14.1). As noted by Hildreth (1957), if we fix all values of $\boldsymbol{\lambda}$ except one, we reduce to a simple univariate constrained quadratic minimization subproblem whose solution is readily available in closed-form. This is the algorithm we detail below.

Estimating a *multivariate* convex function, or more generally functions of certain *shapes*, has regained popularity in recent years, see for instance (Balázs et al. 2015) and the references therein.

Hildreth, C. (1954). “Point Estimates of Ordinates of Concave Functions”. *Journal of the American Statistical Association*, vol. 49, no. 267, pp. 598–619.

— (1957). “A quadratic programming procedure”. *Naval Research Logistics Quarterly*, vol. 4, no. 1, pp. 79–85.

Balázs, G., A. György, and C. Szepesvári (2015). “Near-optimal max-affine estimators for convex regression”. In: *AISTATS*.

Algorithm 14.4: Alternating minimization

```
Algorithm: Alternating Minimization
  Input: $\mathbf{w} \in \operatorname{dom} f$
1 for $t = 1, 2, \ldots$ do
2     choose coordinate $j$                                        // see Remark 15.10 for choices
3     $w_j \leftarrow \operatorname*{argmin}_z f(w_1, \ldots, w_{j-1}, z, w_{j+1}, \ldots, w_d)$  // $\operatorname*{argmin}_z f_0(w_1, \ldots, w_{j-1}, z, w_{j+1}, \ldots, w_d) + f_j(z)$
```

In practice, we may also replace each exact minimization with simply a (proximal) gradient (or descent) step, and the resulting algorithm is usually called coordinate gradient (or alternating descent).

Note that line 3 overwrites the old w_j with the new one in each step, resulting in the so-called Gauss-Seidel update. In contrast, if we overwrite the entire $\mathbf{w}$ only after going through all coordinates, then we obtain a Jacobi update, which is more common in parallel implementations.

Alternating minimization is appealing in practice because of its simplicity, flexibility (could be derivative-free), convenience (could be step size free), lightweight (minimum storage) and surprising efficiency.

Alert 14.5: Notation

To ease later analysis, we denote the t-th iterate of Algorithm 14.4 (with the cyclic rule) as $\mathbf{w}_t$ and let

$$\mathbf{z}_{k,j} = \mathbf{w}_{(k-1)d+j}, \quad \text{where} \quad j = 1, \ldots, d.$$

With the cyclic rule, at iteration $t = (k-1)d + j$, we remind that only the j-th entry is updated while all other entries are held fixed.

Alert 14.6: Why separability?

We remark that if f is completely separable, i.e.

$$f(\mathbf{w}) = \sum_j f_j(w_j),$$

then alternating minimization finds a minimizer in one pass (not surprisingly). Intuitively, this is why we can allow arbitrary (potentially nonsmooth) separable components in our function when applying alternating minimization. Near-separability is also important in improving the analysis of other gradient algorithms.

On the other hand, it is clearly necessary for the domain of f to be separable (i.e. a Cartesian product), for otherwise fixing other entries may significantly restrict any other entry. Consider for instance the "trivial" example:

$$\min_{w+z=0} \ w^2 + z^2.$$

Example 14.7: The difficulty of nonsmoothness

Consider the strongly convex function

$$\min_{w,z} \ w \vee z + \epsilon[(w-2)^2 + (z-2)^2],$$

where $\epsilon > 0$ is arbitrary. Due to symmetry, it is clear that

$$w_\star = z_\star = 2 - \tfrac{1}{2\epsilon}.$$

However, if we start with $w_* = z_* = 2$, then the alternating minimization Algorithm 14.4 immediately gets stuck!

Example 14.8: The difficulty of nonconvexity (Powell 1973)

Consider the following ingenious example due to Powell (1973):

$$\inf_{x,y,z} \ -xy - yz - zx + (x-1)_+^2 + (-x-1)_+^2 + (y-1)_+^2 + (-y-1)_+^2 + (z-1)_+^2 + (-z-1)_+^2,$$

which is continuously differentiable and convex in each coordinate. This function is not bounded from below, as can be seen by taking $x = y = z$, resulting in the objective

$$-3x^2 + 3(x-1)_+^2 + 3(-x-1)_+^2 = \begin{cases} -6x+3, & \text{if } x \geq 1 \\ -3x^2, & \text{if } x \in [-1,1] \\ 6x+3, & \text{if } x \leq -1 \end{cases}.$$

However, stationary points exist at $xyz = 0, x+y+z = 0, x, y, z \in \{0, \pm 2\}$. If one prefers to have a minimizer, we can simply add a box constraint $-a \leq x, y, z \leq a$ so that the minimum is attained at $x_\star = y_\star = z_\star = \pm a$.

Fixing y and z we obtain:

$$\begin{cases} -x(y+z)+(x-1)^2, & \text{if } x \geq 1 \\ -x(y+z), & \text{if } x \in [-1,1] \\ -x(y+z)+(x+1)^2, & \text{if } x \leq -1 \end{cases}, \quad \text{with } x_* = \operatorname{sign}(y+z) + \tfrac{1}{2}(y+z).$$

If we start with $(-1-\epsilon, 1+\frac{1}{2}\epsilon, -1-\frac{1}{4}\epsilon)$, in two passes we obtain

$$\begin{aligned} &(-1-\epsilon, 1+\tfrac{1}{2}\epsilon, -1-\tfrac{1}{4}\epsilon) \to (1+\tfrac{1}{8}\epsilon, 1+\tfrac{1}{2}\epsilon, -1-\tfrac{1}{4}\epsilon) \to (1+\tfrac{1}{8}\epsilon, -1-\tfrac{1}{16}\epsilon, -1-\tfrac{1}{4}\epsilon) \to \\ &\to (1+\tfrac{1}{8}\epsilon, -1-\tfrac{1}{16}\epsilon, 1+\tfrac{1}{32}\epsilon) \to (-1-\tfrac{1}{64}\epsilon, -1-\tfrac{1}{16}\epsilon, 1+\tfrac{1}{32}\epsilon) \to (-1-\tfrac{1}{64}\epsilon, 1+\tfrac{1}{128}\epsilon, 1+\tfrac{1}{32}\epsilon) \to \\ &\to (-1-\tfrac{1}{64}\epsilon, 1+\tfrac{1}{128}\epsilon, -1-\tfrac{1}{256}\epsilon), \end{aligned}$$

which amounts to reducing ϵ by a factor of 64. Thus, alternating minimization Algorithm 14.4 cycles around the 6 limit points:

$$(-1,1,-1) \to (1,1,-1) \to (1,-1,-1) \to (1,-1,1) \to (-1,-1,1) \to (-1,1,1) \to (-1,1,-1),$$

neither of which is optimal or stationary. Note however that 2 of the 3 entries in the gradient at the limit points vanish. This is not a coincidence, as we prove below.

Powell (1973) also constructed similar counterexamples that are robust against rounding errors or infinitely differentiable (but not both, which still seems open).

Powell, M. J. D. (1973). "On search directions for minimization algorithms". *Mathematical Programming*, vol. 4, pp. 193–201.

Exercise 14.9: Convexity

Prove the function in Example 14.8 is *not* (jointly) convex in every two variables.

Remark 14.10: Dykstra's algorithm as alternating minimization in the dual

We can now offer a natural explanation of Dykstra's Algorithm 15.14 due to Han (1988, 1989), Gaffke and Mathar (1989), and Tseng (1993). Recall the problem:

$$\inf_{\mathbf{w} \in \cap_i C_i} f(\mathbf{w}),$$

where each C_i is closed and convex and f is Legendre. Indeed, apply the Fenchel-Rockafellar duality we obtain the dual problem (where σ_i is the support function of C_i):

$$\inf_{\{\mathbf{w}_i^*\}} f^*\big(-\textstyle\sum_i \mathbf{w}_i^*\big) + \textstyle\sum_i \sigma_i(\mathbf{w}_i^*),$$

where the (unique) primal solution $\mathbf{w}$ and dual solution $\{\mathbf{w}_i^*\}$ are connected by:

$$\textstyle\sum_i \mathbf{w}_i^* + \nabla f(\mathbf{w}) = \mathbf{0}. \tag{14.4}$$

Since f is Legendre, f^* is smooth and convex so we have precisely a problem in the format of (14.1). Apply alternating minimization Algorithm 14.4:

$$\mathbf{w}_{i,t+1}^* = \operatorname*{argmin}_{\mathbf{w}_i^*} f^*\Big(-\mathbf{w}_i^* - \sum_{j \neq i} \mathbf{w}_{j,t}^*\Big) + \sigma_i(\mathbf{w}_i^*) \quad \text{or} \quad \mathbf{w}_{t+1} = \operatorname*{argmin}_{\mathbf{w} \in C_i} f(\mathbf{w}) + \Big\langle \mathbf{w}; \sum_{j \neq i} \mathbf{w}_{j,t}^* \Big\rangle,$$

where the equivalence follows from reverting the duality. The primal solution $\mathbf{w}_{t+1}$ and dual solution $\mathbf{w}_{i,t+1}^*$ are now both unique due to the strict convexity in Legendre functions and they are connected by:

$$\nabla f(\mathbf{w}_{t+1}) + \mathbf{w}_{i,t+1}^* + \textstyle\sum_{j \neq i} \mathbf{w}_{j,t}^* = \mathbf{0} = \nabla f(\mathbf{w}_{t+1}) + \textstyle\sum_j \mathbf{w}_{j,t+1}^*, \tag{14.5}$$

since at time t we update $\mathbf{w}^*_{i,t+1}$ and keep $\mathbf{w}^*_{j,t+1} = \mathbf{w}^*_{j,t}$ for all $j \neq i$. Let us define (and maintain)

$$\forall l = 1, \ldots, |I|, \quad \mathbf{a}_{l,t} + \nabla f(\mathbf{w}_t) + \textstyle\sum_{j \neq l} \mathbf{w}^*_{j,t} = \mathbf{0} \overset{(14.5)}{=} \mathbf{a}_{l,t} - \mathbf{w}^*_{l,t},$$

where the last inequality follows from (14.5). Then,

$$\mathbf{a}_{i,t+1} = \mathbf{w}^*_{i,t+1} \overset{(14.5)}{=} -\nabla f(\mathbf{w}_{t+1}) - \sum_{j \neq i} \mathbf{w}^*_{j,t} \overset{(14.5)}{=} -\nabla f(\mathbf{w}_{t+1}) + \mathbf{w}^*_{j,t} + \nabla f(\mathbf{w}_t) = \mathbf{a}_{i,t} + \nabla f(\mathbf{w}_t) - \nabla f(\mathbf{w}_{t+1})$$

while for all $l \neq i$, $\mathbf{a}_{l,t+1} = \mathbf{w}^*_{l,t} = \mathbf{a}_{l,t}$ since $\mathbf{w}^*_{l,t}$ was held fixed. We have thus recovered Dykstra's Algorithm 15.14 and the meaning of $\mathbf{a}_{i,t} = \mathbf{w}^*_{i,t}$ is now clear: they are the dual solutions that are alternatingly minimized! It is also clear that $(\mathbf{w}_{t+1}, \mathbf{w}^*_{i,t+1})$ define a supporting hyperplane for C_i. Moreover, we note that when $\mathbf{w}_t \to \mathbf{w}$ and $\mathbf{w}^*_{i,t} \to \mathbf{w}^*_i$, the optimality condition (14.5) converges to that in (14.4).

Han, S.-P. (1988). "A successive projection method". *Mathematical Programming*, pp. 1–14.

— (1989). "A Decomposition Method and Its Application to Convex Programming". *Mathematics of Operations Research*, no. 2, pp. 237–248.

Gaffke, N. and R. Mathar (1989). "A cyclic projection algorithm via duality". *Metrika*, vol. 36, pp. 29–54.

Tseng, P. (1993). "Dual coordinate ascent methods for non-strictly convex minimization". *Mathematical Programming*, vol. 59, pp. 231–247.

Alert 14.11: When can you go back and forth?

In our interpretation of Dykstra's algorithm in Remark 14.10, we went back and forth between the primal and dual problems, which is where we needed the strict convexity in Legendre functions. In the absence of strict convexity (or more precisely stability), we may run into trouble, which will be discussed in a later lecture.

Definition 14.12: Nash equilibrium and (strictly) regular functions

The above counterexamples motivate us to call $\mathbf{w}$ a (Nash) equilibrium of f if

$$\forall j, \; w_j \in \operatorname*{argmin}_z f(w_1, \ldots, w_{j-1}, z, w_{j+1}, \ldots, w_d).$$

We call a function f strictly regular if any equilibrium is actually a *bona fide* minimizer, and simply regular if any equilibrium is actually stationary (i.e. critical). One may also weaken the notion of equilibrium to alternating stationary, although this is not needed for most settings where Algorithm 14.4 is applied.

It is clear that any minimizer is a equilibrium, while the converse may fail as shown in Example 14.7. Example 14.8 further showed that limit points of the alternating minimization Algorithm 14.4 may not even be an equilibrium.

We call f pairwise (strictly) regular if for all pairs of indices i, j and all $(w_k : k \neq i, k \neq j)$, the bi-variate function $(w_i, w_j) \mapsto f(\mathbf{w})$ is (strictly) regular.

Exercise 14.13: Smooth + separable functions are regular

Prove that functions consisting of a smooth part and a separable part (as in (14.1)) are regular.

Moreover, under convexity we can strengthen the result to strictly regular.

Theorem 14.14: Convergence of alternating minimization for two blocks

Let $d = 2$ and consider any function $f(\mathbf{x}, \mathbf{y})$ that is separately u.s.c. in its product domain. Assume Algorithm 14.4 is well-defined. Then, any limit point (if any) of $\{\mathbf{w}_t\}$ is an equilibrium.

Proof: Let $\mathbf{z}_{k,1} = (\mathbf{x}_{k+1}, \mathbf{y}_k)$ and $\mathbf{z}_{k,2} = (\mathbf{x}_{k+1}, \mathbf{y}_{k+1})$ so that we avoid messy subscripts. Assume w.l.o.g.

$$(\mathbf{x}_{k+1}, \mathbf{y}_k) \to (\mathbf{x}_*, \mathbf{y}_*) \text{ for a subsequence } k \in K.$$

Clearly, the alternating minimization algorithm is descending:

$$f(\mathbf{w}_{t+1}) \leq f(\mathbf{w}_t) \text{ hence } f(\mathbf{w}_t) \downarrow f_* = f(\mathbf{x}_*, \mathbf{y}_*).$$

By definition we have for any $(\mathbf{x}, \mathbf{y}) \in \operatorname{dom} f$:

$$f(\mathbf{x}_{k+1}, \mathbf{y}_k) \leq f(\mathbf{x}, \mathbf{y}_k), \qquad f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}) \leq f(\mathbf{x}_{k+1}, \mathbf{y}).$$

Let k tend to ∞ in K and use upper semicontinuity:

$$f(\mathbf{x}_*, \mathbf{y}_*) = \lim_{k \in K} f(\mathbf{x}_{k+1}, \mathbf{y}_k) \leq \liminf_{k \in K} f(\mathbf{x}, \mathbf{y}_k) \leq f(\mathbf{x}, \mathbf{y}_*),$$
$$f(\mathbf{x}_*, \mathbf{y}_*) = \lim_{k \in K} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}) \leq \liminf_{k \in K} f(\mathbf{x}_{k+1}, \mathbf{y}) \leq f(\mathbf{x}_*, \mathbf{y}),$$

i.e., the limit point $(\mathbf{x}_*, \mathbf{y}_*)$ is alternating minimizing. ■

For $d > 2$, we can similarly prove: Suppose $\mathbf{z}$ is a limit point of $\mathbf{z}_{k,j}$. Then for any w,

$$f(z_1, \ldots, z_{j-1}, z_j, z_{j+1}, \ldots, z_d) \leq f(z_1, \ldots, z_{j-1}, w, z_{j+1}, \ldots, z_d) \wedge f(z_1, \ldots, z_{j-1}, z_j, w, z_{j+2}, \ldots, z_d), \tag{14.6}$$

where of course $d + 1 \equiv 1$. Together, theses results extend Grippof and Sciandrone (2000, Corollary 2, Proposition 3).

Grippof, L. and M. Sciandrone (2000). "On the convergence of the block nonlinear Gauss–Seidel method under convex constraints". *Operations Research Letters*, vol. 26, no. 3, pp. 127–136.

Remark 14.15: Digestion

We can apply Theorem 14.14 to Example 14.7 and confirm that the limit point $(x_*, y_*) = (2, 2)$ is indeed alternating minimizing. However, it is not a minimizer because the function, being nonsmooth, is not strictly alt-reg.

On the other hand, we also see that the counterexample in Example 14.8 is minimal in the sense that it cannot happen with $d = 2$! Moreover, since the function in Example 14.8 is unconstrained, from (14.6) we know that any limit point of the alternating minimization algorithm must have at least 2 consecutive gradient components vanishing, which is indeed the case!

Example 14.16: Convexifying bilinearity

Consider the bilinear problem:

$$\min_{\mathbf{x}, \mathbf{y}} \ \langle \mathbf{x}, Q\mathbf{y} \rangle + f_1(\mathbf{x}) + f_2(\mathbf{y}),$$

which is not convex due to the first bilinear term. Applying alternating minimization Algorithm 14.4 and using Theorem 14.14 we know any limit point is alternating minimizing. When both f_0 and f_1 are convex, each step in Algorithm 14.4 is a convex minimization problem!

More generally, we can apply alternating minimization Algorithm 14.4 to solve any biconvex problems.

Exercise 14.17: Minimizing concave quadratic functions

Consider the following concave quadratic minimization problem:

$$\min_{\mathbf{x}\geq\mathbf{0}} \ \langle \mathbf{x}, Q\mathbf{x}\rangle + \langle \mathbf{x}, \mathbf{c}\rangle, \qquad \text{s.t.} \quad A\mathbf{x} \leq \mathbf{b}, \tag{14.7}$$

where $Q \preceq \mathbf{0}$ is (symmetric) negative semidefinite, and its twin bilinear problem:

$$\min_{\mathbf{x}\geq\mathbf{0},\mathbf{y}\geq\mathbf{0}} \ \langle \mathbf{x}, Q\mathbf{y}\rangle + \langle \mathbf{x}+\mathbf{y}, \mathbf{c}\rangle, \qquad \text{s.t.} \quad A\mathbf{x} \leq \mathbf{b}, A\mathbf{y} \leq \mathbf{b}. \tag{14.8}$$

Prove that

- $(\mathbf{x}_*, \mathbf{y}_*)$ solves (14.8) $\implies$ $\frac{\mathbf{x}_*+\mathbf{y}_*}{2}$ solves (14.7).
- $\mathbf{x}_*$ solves (14.7) $\implies$ $(\mathbf{x}_*, \mathbf{x}_*)$ solves (14.8).

Theorem 14.18: Convergence of alternating minimization for any number of blocks

Let f be continuous on the sublevel set $[\![f \leq f(\mathbf{w}_0)]\!]$ which we assume to be compact. Assume $\operatorname{dom} f$ *to be separable and choose the cyclic rule. If f is pairwise strictly alt-reg, then any limit point of Algorithm 14.4 is an alternating minimizer.*

Proof: Under the compact and continuous assumption, it is clear that Algorithm 14.4 is well-defined and

$$f(\mathbf{w}_{t+1}) \leq f(\mathbf{w}_t) \text{ hence } f(\mathbf{w}_t) \downarrow f_* \in \mathbb{R}.$$

By continuously extracting subsequences we may assume

$$\forall j = 1, \ldots, d, \ \ \mathbf{z}_{k,j} \to \mathbf{x}_j, \quad k \in K, \quad \text{where} \quad f(\mathbf{x}_1) = \cdots = f(\mathbf{x}_d) = f_*. \tag{14.9}$$

We observe that by the consecutiveness of $\{\mathbf{z}_{k,j}\}_j$, their limits satisfy:

$$\forall k \neq j, \ \ x_{k,j} = x_{k,j-1}.$$

Thus, to save subscripts we may write

$$\mathbf{x}_j := (\bar{x}_1, \ldots, \bar{x}_j, x_{j+1}, \ldots, x_d).$$

Using (14.6) we have

$$\begin{aligned} \forall j, \forall w_j, \ f(\mathbf{x}_j) &\leq f(\bar{x}_1, \ldots, \bar{x}_{j-1}, w_j, x_{j+1}, x_{j+2}, \ldots, x_d) \\ \forall j, \forall w_{j+1}, \ f(\mathbf{x}_j) &\leq f(\bar{x}_1, \ldots, \bar{x}_{j-1}, \bar{x}_j, w_{j+1}, x_{j+2}, \ldots, x_d). \end{aligned} \tag{14.10}$$

Since f is $(j, j+1)$ pairwise strict alt-reg, we have

$$\forall j, \forall w_j, \forall w_{j+1}, \quad f(\mathbf{x}_j) \leq f(\bar{x}_1, \ldots, \bar{x}_{j-1}, w_j, w_{j+1}, x_{j+2}, \ldots, x_d),$$

which, together with (14.9), allows us to "telescope" backwards:

$$\begin{aligned} f_* = f(\mathbf{x}_j) = f(\bar{x}_1, \ldots, \bar{x}_{j-1}, \bar{x}_j, x_{j+1}, \ldots, x_d) &\leq f(\bar{x}_1, \ldots, \bar{x}_{j-1}, w_j, w_{j+1}, x_{j+2}, \ldots, x_d) \\ (\text{ setting } w_j = x_j \ ) &= f(\bar{x}_1, \ldots, \bar{x}_{j-1}, x_j, w_{j+1}, x_{j+2}, \ldots, x_d) \\ f_* = f(\mathbf{x}_{j-1}) = f(\bar{x}_1, \ldots, \bar{x}_{j-2}, \bar{x}_{j-1}, x_j, \ldots, x_d) &\leq f(\bar{x}_1, \ldots, \bar{x}_{j-2}, w_{j-1}, x_j, x_{j+1}, \ldots, x_d) \\ (j-1, j+1) \text{ pairwise strictly alt-reg} \implies f_* = f(\mathbf{x}_{j-1}) &\leq f(\bar{x}_1, \ldots, \bar{x}_{j-2}, w_{j-1}, x_j, w_{j+1}, x_{j+2}, \ldots, x_d) \\ (\text{ setting } w_{j-1} = x_{j-1} \ ) &= f(\bar{x}_1, \ldots, \bar{x}_{j-2}, x_{j-1}, x_j, w_{j+1}, x_{j+2}, \ldots, x_d) \\ f_* = f(\mathbf{x}_{j-2}) = f(\bar{x}_1, \ldots, \bar{x}_{j-2}, x_{j-1}, \ldots, x_d) &\leq f(\bar{x}_1, \ldots, \bar{x}_{j-3}, w_{j-2}, x_{j-1}, \ldots, x_d) \end{aligned}$$

$$\begin{aligned}
(j-2, j+1) \text{ pairwise strictly alt-reg} \implies f_* &= f(\mathbf{x}_{j-2}) \le f(\bar x_1, \dots, \bar x_{j-3}, w_{j-2}, x_{j-1}, x_j, w_{j+1}, x_{j+2}, \dots, x_d) \\
&\vdots \\
(2, j+1) \text{ pairwise strictly alt-reg} \implies f_* &= f(\mathbf{x}_2) \le f(\bar x_1, w_2, x_3, \dots, x_j, w_{j+1}, x_{j+2}, \dots, x_d) \\
(\text{ setting } w_2 = x_2\ ) &= f(\bar x_1, x_2, \dots, x_j, w_{j+1}, x_{j+2}, \dots, x_d).
\end{aligned}$$

Since j is arbitrary and $f(\mathbf{x}_1) = f_*$, it follows that $\mathbf{x}_1$ is an alternating minimizer. By a completely similar argument we establish all limit points are alternating minimizing. ■

We point out that if we are only interested in limit points of $\mathbf{z}_{k,j}$, then the pairwise strict alt-reg need *not* involve the j-th or the $(j+1)$-th (if we telescope forwards) coordinate. This observation immediately implies the function in Example 14.8 is not even convex for every pair of variables.

Theorem 14.18 slightly improves Tseng (2001, Theorem 4.1).

Tseng, P. (2001). “Convergence of a Block Coordinate Descent Method for Nondifferentiable Minimization”. *Journal of Optimization Theory and Applications*, vol. 109, pp. 475–494.

Corollary 14.19: Convergence of alternating minimization under uniqueness

Let f be continuous on the sublevel set $[\![f \le f(\mathbf{w}_0)]\!]$ which we assume to be compact. Assume dom f *to be separable and choose the cyclic rule. If for all but one j and any $\mathbf{w}$, the function $z \mapsto f(w_1, \dots, w_{j-1}, z, w_{j+1}, \dots, w_d)$ is attained at a unique minimizer, then any limit point of Algorithm 14.4 is an alternating minimizer.*

Proof: It follows immediately from (14.10) and the uniqueness that $\mathbf{x}_1 = \cdots = \mathbf{x}_d$. ■

Similarly, if we are only interested in the limit points of $\mathbf{z}_{k,j}$, then uniqueness need only hold for all but $(j+1, j+2)$.

Example 14.20: Structured matrix factorization (e.g. Tseng 2001)

Consider the structured matrix factorization problem

$$\min_{A \in \mathbb{R}^{m \times k}, S \in \mathbb{R}^{k \times n}} \quad \|X - AS\|^2 + \sum_{lj} f_{lj}(s_{lj}), \qquad \text{s.t.} \quad \|\mathbf{a}_{:l}\| \le 1,\ l = 1, \dots, k.$$

where the constraint on A is to avoid scaling degeneracy. For small fixed k the problem is not convex (due to the bilinear product AS). However, we may still apply the alternating minimization Algorithm 14.4, with $1 + kn$ blocks $(A, s_{l,j})$. If f_{lj} is convex, continuous and has bounded level sets, then Theorem 14.18 applies to the subsequence $\mathbf{z}_{k,1}$ and any of its limit points is alternating minimizing.

When entries of a low-rank X are partially observed in a random fashion, Jain et al. (2013) and Hardt (2014) and many others have studied when alternating minimization can complete the true matrix X.

One may also add nonnegative constraints to one or both of A and S, leading to the so-called nonnegative matrix factorization problem. Extensions to multi-dimensional matrices (i.e. tensors) are also abundant.

Tseng, P. (2001). “Convergence of a Block Coordinate Descent Method for Nondifferentiable Minimization”. *Journal of Optimization Theory and Applications*, vol. 109, pp. 475–494.

Jain, P., P. Netrapalli, and S. R. Sanghavi (2013). “Low-rank matrix completion using alternating minimization”. In: *Proceedings of the forty-fifth annual ACM symposium on Theory of Computing*, pp. 665–674.

Hardt, M. (2014). “Understanding Alternating Minimization for Matrix Completion”. In: *IEEE 55th Annual Symposium on Foundations of Computer Science*, pp. 651–660.

Example 14.21: The shooting algorithm for lasso (Fu 1998)

Recall the lasso problem for sparse estimation:

$$\min_{\mathbf{w}\in\mathbb{R}^d} \tfrac{1}{2n}\|X\mathbf{w}-\mathbf{y}\|_2^2 + \lambda\|\mathbf{w}\|_1,$$

which fits perfectly into our problem (14.1). Since the objective is strictly alt-reg, any limit point of the alternating minimization Algorithm 14.4 is a *bona fide* minimizer. To update the j-th coordinate, we need to solve the subproblem:

$$\min_{w} \tfrac{1}{2n}\ \|\mathbf{x}_{:j}(w-w_j)+\mathbf{r}\|_2^2 + \lambda|w|, \quad \text{where} \quad \mathbf{r} := X\mathbf{w}_t - \mathbf{y},$$

which, after expanding the norm and completing the squares, reduces to the familiar (univariate) soft-shrinkage operator that is readily available in closed-form. Note that after updating $w_j \leftarrow w_j^+$, we can also update $\mathbf{r}$ by subtracting the old term $\mathbf{x}_{:j}w_j$ and then adding the new term $\mathbf{x}_{:j}w_j^+$. Thus, each update (for 1 coordinate) costs only $O(n)$ while note that a single gradient step (for d coordinates) costs $O(nd)$ due to the matrix vector product $X^\top\mathbf{r}$.

Fu, W. J. (1998). "Penalized Regressions: The Bridge versus the Lasso". *Journal of Computational and Graphical Statistics*, vol. 7, no. 3, pp. 397–416.

Exercise 14.22: k-means clustering

Given a matrix $X = [\mathbf{x}_1,\ldots,\mathbf{x}_n] \in \mathbb{R}^{d\times n}$, we aim to cluster its columns into k centers $M = [\mathbf{m}_1,\ldots,\mathbf{m}_k] \in \mathbb{R}^{d\times k}$ so that each column $\mathbf{x}_i$ is "close" to its assigned center:

$$\min_{M\in\mathbb{R}^{d\times k}} \ \sum_{i=1}^{n} \min_{c=1,\ldots,k} \|\mathbf{x}_i - \mathbf{m}_c\|_2^2.$$

Introducing the assignment matrix $A \in \{0,1\}^{k\times n}$ where $A_{ci} \in \{0,1\}$ indicates whether the i-th column $\mathbf{x}_i$ is assigned to the c-th center $\mathbf{m}_c$. Prove the equivalence:

$$\min_{M\in\mathbb{R}^{d\times k}} \ \min_{A\in\{0,1\}^{k\times n}} \ \|X-MA\|_{\mathsf{F}}^2, \quad \text{s.t.} \quad \operatorname{diag}(A^\top A) = \mathbf{1}, \ \ \operatorname{diag}(AA^\top) \geq \mathbf{d}, \tag{14.11}$$

where $\mathbf{d} \in \mathbb{N}^k$ is the given lower bound on cluster size. Apply the alternating minimization Algorithm 14.4 to solve (14.11).

What if we want to cluster both columns and rows of X simultaneously? The so-called co-clustering problem can be formulated using 3 blocks and solved again by alternating minimization.