9 Alternating Minimization

Goal

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

Alert 9.1: Convention

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

This note is likely to be updated again soon.

Definition 9.2: Problem

The problem we study in this lecture is the following:

$$\inf_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w}), \quad \text{where} \quad f(\mathbf{w}) = f_0(\mathbf{w}) + \sum_{j=1}^{d} f_j(w_j), \tag{9.1}$$

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 $C := C_1 \times \cdots \times C_d$.

Algorithm 9.3: 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 10.10 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 9.4: Notation

To ease later analysis, we denote the t-th iterate of Algorithm 9.3 (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 9.5: 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:

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

Definition 9.6: 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 9.3 is applied.

It is clear that any minimizer is a equilibrium, while the converse may fail as shown in **??**. **??** further showed that limit points of the alternating minimization Algorithm 9.3 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 9.7: Smooth + separable functions are regular

Prove that functions consisting of a smooth part and a separable part (as in (9.1)) are regular.

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

Theorem 9.8: 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 9.3 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:

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

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

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,

$$f(z_1, \ldots, z_{j-1}, z_j, z_{j+1}, \ldots, z_d) \le f(z_1, \ldots, z_{j-1}, w, z_{j+1}, \ldots, z_d) \wedge f(z_1, \ldots, z_{j-1}, z_j, w, z_{j+2}, \ldots, z_d), \tag{9.2}$$

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.

Theorem 9.9: Convergence of alternating minimization for any number of blocks

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 f is pairwise strictly alt-reg, then any limit point of Algorithm 9.3 is an alternating minimizer.*

Proof: Under the compact and continuous assumption, it is clear that Algorithm 9.3 is well-defined and

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

By continuously extracting subsequences we may assume

$$\forall j = 1, \ldots, d, \quad \mathbf{z}_{k,j} \to \mathbf{x}_j, \quad k \in K, \quad \text{where} \quad f(\mathbf{x}_1) = \cdots = f(\mathbf{x}_d) = f_*. \tag{9.3}$$

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

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

Thus, to save subscripts we may write

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

Using (9.2) we have

$$\begin{aligned} \forall j, \forall w_j, \ f(\mathbf{x}_j) &\le f(\bar{x}_1, \ldots, \bar{x}_{j-1}, w_j, x_{j+1}, x_{j+2}, \ldots, x_d) \\ \forall j, \forall w_{j+1}, \ f(\mathbf{x}_j) &\le f(\bar{x}_1, \ldots, \bar{x}_{j-1}, \bar{x}_j, w_{j+1}, x_{j+2}, \ldots, x_d). \end{aligned} \tag{9.4}$$

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

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

which, together with (9.3), allows us to "telescope" backwards:

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

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

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 **??** is not even convex for every pair of variables.

Theorem 9.9 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 9.10: 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 $\operatorname{dom} f$ *to be separable and choose the cyclic rule. If for all* but one *j and any $\mathbf{w}$, the function $z \mapsto f(w_1, \ldots, w_{j-1}, z, w_{j+1}, \ldots, w_d)$ is attained at a* unique *minimizer, then any limit point of Algorithm 9.3 is an alternating minimizer.*

Proof: It follows immediately from (9.4) 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 $(j+1, j+2)$.