15 Fejér-type Algorithms

Goal

Fejér's theorem and sequence, reducing optimization to feasibility, alternating projection, Dykstra's algorithm

Alert 15.1: Convention

See Bauschke and Borwein (1996) for a nice survey on projection algorithms and Escalante and Raydan (2011) for an enjoyable short book on alternating projections. Combettes and Vũ (2013) extended Fejér monotonicity to variable metrics whilte Bauschke et al. (2003) extended to Bregman divergences.

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

This note is likely to be updated again soon.

Bauschke, H. H. and J. M. Borwein (1996). "On Projection Algorithms for Solving Convex Feasibility Problems". *SIAM Review*, vol. 38, no. 3, pp. 367–426.

Escalante, R. and M. Raydan (2011). "Alternating Projection Methods". SIAM.

Combettes, P. L. and B. C. Vũ (2013). "Variable metric quasi-Fejér monotonicity". *Nonlinear Analysis: Theory, Methods & Applications*, vol. 78, no. 384, pp. 17–31.

Bauschke, H. H., J. M. Borwein, and P. L. Combettes (2003). "Bregman Monotone Optimization Algorithms". *SIAM Journal on Control and Optimization*, vol. 42, no. 2, pp. 596–636.

Definition 15.2: Problem

In this lecture we aim to solve the following problem:

$$\begin{aligned} \inf_{\mathbf{w}\in\mathbb{R}^d} \ & f(\mathbf{w}) \\ \text{s.t.} \ & \mathbf{w} \in \bigcap_{i\in I} C_i, \end{aligned}$$

where each $C_i \subseteq \mathbb{R}^d$ is closed ~~and convex~~, and the function $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is convex. We assume each set C_i is simple, in the sense that its projector $\mathrm{P}_i = \mathrm{P}_{C_i}$ can be easily computed. However, projecting to the intersection C is usually much harder.

Example 15.3: Perceptron and SVM revisited

Recall the perceptron problem:

$$\begin{aligned} \min_{\mathbf{w}\in\mathbb{R}^d} \ & f(\mathbf{w}) \equiv 0 \\ \text{s.t.} \ & \mathbf{w} \in \bigcap_{i=1}^n C_i, \end{aligned}$$

where $C_i := \{\mathbf{w} : \langle y_i\mathbf{x}_i, \mathbf{w}\rangle \geq 1\}$. Similarly, we may rewrite the hard-margin SVM problem as:

$$\begin{aligned} \min_{\mathbf{w}\in\mathbb{R}^d} \ & \tfrac{1}{2}\|\mathbf{w}\|_2^2 \\ \text{s.t.} \ & \mathbf{w} \in \bigcap_{i=1}^n C_i. \end{aligned}$$

We note that the projector P_{C_i} is available in closed-form:

$$\mathrm{P}_{C_i}(\mathbf{z}) := \left[\operatorname*{argmin}_{\mathbf{w}\in C_i} \|\mathbf{w}-\mathbf{z}\|_2\right] = \mathbf{z} + \frac{(1-\langle y_i\mathbf{x}_i, \mathbf{z}\rangle)_+}{\|\mathbf{x}_i\|_2^2} y_i\mathbf{x}_i.$$

However, projecting onto the intersection set C is not easy. In fact, the *entire* hard-margin SVM problem is "just" projecting the origin to the intersection set C.

Example 15.4: Linear and quadratic programming

We can reduce the canonical linear program

$$\min_{\mathbf{x}\geq\mathbf{0},A\mathbf{x}=\mathbf{b}} \langle \mathbf{x},\mathbf{c}\rangle$$

into a feasibility problem, through duality:

$$\begin{aligned}\mathbf{x}\geq\mathbf{0}, A\mathbf{x}&=\mathbf{b}\\ A^\top\mathbf{y}&\leq\mathbf{c}\\ \mathbf{c}^\top\mathbf{x}&=\mathbf{b}^\top\mathbf{y}.\end{aligned}$$

In other words, minimizing a linear function over a polyhedron is nothing but solving a linear inequality system! Projecting onto each of the above linear constraints can be done similarly as in Example 15.3.

Similarly, for the canonical quadratic program:

$$\min_{\mathbf{x}\geq\mathbf{0},A\mathbf{x}=\mathbf{b}} \ \tfrac{1}{2}\langle \mathbf{x},Q\mathbf{x}\rangle + \langle \mathbf{x},\mathbf{c}\rangle,$$

where for simplicity $Q\succ\mathbf{0}$ is (symmetric) positive definite. Using duality we may derive the following system:

$$\begin{aligned}\mathbf{x}\geq\mathbf{0}, A\mathbf{x}&=\mathbf{b}\\ Q\mathbf{x}-A^\top\mathbf{y}+\mathbf{c}&\geq\mathbf{0}\\ \langle \mathbf{x},Q\mathbf{x}+\mathbf{c}\rangle - \langle \mathbf{b},\mathbf{y}\rangle &\leq 0,\end{aligned}\tag{15.1}$$

where the last nonlinear quadratic inequality (15.1) can be rewritten as an infinite (uncountable) intersection of closed halfspaces (indexed by $\mathbf{z}$; recall Theorem 0.22):

$$\forall \mathbf{z},\quad \langle \mathbf{x},\mathbf{z}+\mathbf{c}\rangle - \tfrac{1}{2}\langle \mathbf{z},Q^{-1}\mathbf{z}\rangle - \langle \mathbf{b},\mathbf{y}\rangle \leq 0.$$

One may continue to rewrite QCQP, SOCP, and SDP as a similar feasibility problem with infinitely many linear constraints. In fact, for SOCP (and QCQP and QP) it is possible to significantly reduce the number of linear constraints if an approximate solution is sought, which is now routinely used in dealing with nonlinear integer programs.

Example 15.5: Sudoku (Bailey et al. 2008) – A nonconvex example

	2			3		9		7
	1							
4		7				2		8
		5	2				9	
			1	8		7		
	4				3			
				6			7	1
	7							
9		3		2		6		5

See also Elser et al. (2007) and Erlich et al. (2009).

Bailey, R. A., P. J. Cameron, and R. Connelly (2008). "Sudoku, Gerechte Designs, Resolutions, Affine Space, Spreads, Reguli, and Hamming Codes". *The American Mathematical Monthly*, vol. 115, no. 5, pp. 383–404.

Elser, V., I. Rankenburg, and P. Thibault (2007). "Searching with iterated maps". *Proceedings of the National Academy of Sciences of the USA*, vol. 104, no. 2, pp. 418–423.

Erlich, Y., K. Chang, A. Gordon, R. Ronen, O. Navon, M. Rooks, and G. J. Hannon (2009). "DNA Sudoku—harnessing high-throughput sequencing for multiplexed specimen analysis". *Genome Research*, vol. 19, no. 7, pp. 1243–1253.

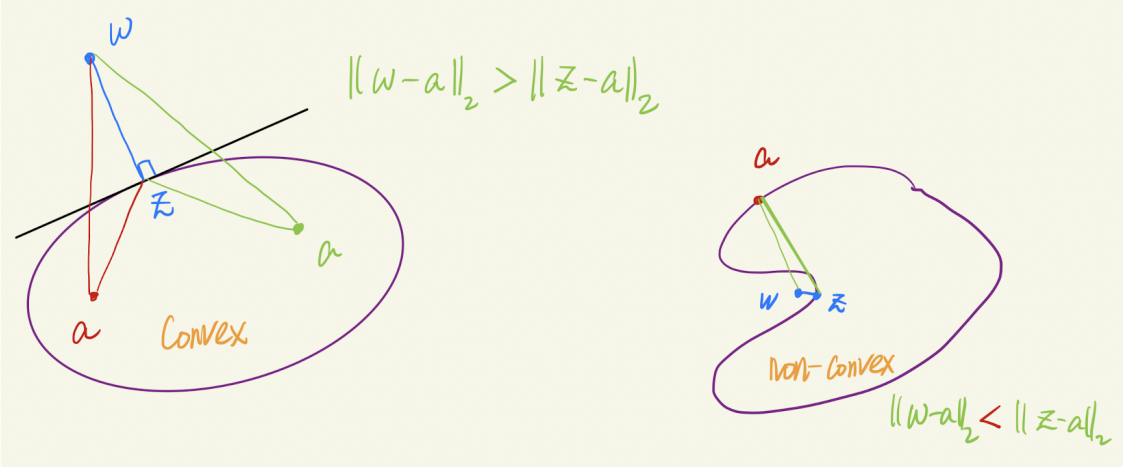
Theorem 15.6: Fejér's characterization of the closed convex hull (Fejér 1922)

Let $A \subseteq \mathbb{R}^d$. Then, $\mathbf{w} \notin \overline{\text{conv}}A$ iff there exists $\mathbf{z} \in \mathbb{R}^d$ such that for all $\mathbf{a} \in A$ (hence all $\mathbf{a} \in \overline{\text{conv}}A$) we have $\|\mathbf{w}-\mathbf{a}\|_2 > \|\mathbf{z}-\mathbf{a}\|_2$.

Proof: If $\mathbf{w} \notin \overline{\text{conv}}A$, then we can find a hyperplane H that separates $\mathbf{w}$ from $\overline{\text{conv}}A$ (in particular, $\mathbf{w} \notin \mathsf{H}$). Let $\mathbf{z} = \mathrm{P}_{\mathsf{H}}(\mathbf{w})$ be the orthogonal projection of $\mathbf{w}$ onto H. For any point $\mathbf{a} \in \overline{\text{conv}}A$, the triangle $\mathbf{azw}$ is obtuse at the point $\mathbf{z}$, thus $\|\mathbf{w}-\mathbf{a}\|_2 > \|\mathbf{z}-\mathbf{a}\|_2$.

Conversely, let $\mathbf{w} \in \overline{\text{conv}}A$ and suppose there exists such a point $\mathbf{z}$ with $\|\mathbf{z}-\mathbf{a}\|_2 < \|\mathbf{w}-\mathbf{a}\|_2$ for all $\mathbf{a} \in A$. Consider the line segment $\mathbf{wz}$ and its orthogonal hyperplane H that passes through the middle point $\mathbf{o} = (\mathbf{w}+\mathbf{z})/2$. We claim that any $\mathbf{a} \in A$ lies on the same side of the half-space H as $\mathbf{z}$, for otherwise the line segment $\mathbf{za}$ intersects H at some $\mathbf{s}$: $\|\mathbf{w}-\mathbf{a}\|_2 \le \|\mathbf{w}-\mathbf{s}\|_2 + \|\mathbf{s}-\mathbf{a}\|_2 = \|\mathbf{z}-\mathbf{s}\|_2 + \|\mathbf{s}-\mathbf{a}\|_2 = \|\mathbf{z}-\mathbf{a}\|_2$, contradiction. But, $\mathbf{w}$ is in the interior of the opposite half-space, contradicting $\mathbf{w} \in \overline{\text{conv}}A$. ■

The proof crucially relies on the fact that the norm $\|\cdot\|_2$ is induced by an inner product (so that we can talk about orthogonal projections meaningfully).



Fejér, L. (1922). "Über die Lage der Nullstellen von Polynomen, die aus Minimumforderungen gewisser Art entspringen". *Mathematische Annalen*, vol. 85, no. 1, pp. 41–48.

Remark 15.7: Significance

Fejér's result is algorithmically significant because it can be used to solve the convex feasibility problem:

$$\text{find} \quad \mathbf{w} \in C,$$

where the closed (and convex) set $C \subseteq \mathbb{R}^d$ represents the solutions set that we are seeking. Indeed, starting from an arbitrary point $\mathbf{w}_0$, if it is in C then we are done; if not then according to Fejér's Theorem 15.6 there exists some $\mathbf{w}_1$ such that $\|\mathbf{w}_1 - \mathbf{w}\| < \|\mathbf{w}_0 - \mathbf{w}\|$ for all $\mathbf{w} \in C$. Of course, this idea by itself is not quite an algorithm yet:

- We need to be able to certify if $\mathbf{w}_0 \in C$, which may be trivial when the set C is defined by *explicit* inequalities, such as $C = \{\mathbf{w} : g(\mathbf{w}) \leq 0\}$.
- If $\mathbf{w}_0 \notin C$, we need to be able to *explicitly and efficiently* find $\mathbf{w}_1$.
- We also need sufficient decrease so that $\operatorname{dist}(\mathbf{w}_t, C) \to 0$.
- We may also want to prove the convergence (rate) of the whole sequence $\mathbf{w}_t$.

Definition 15.8: Fejér monotone sequence

We say that a sequence $\{\mathbf{w}_t\}$ is Fejér monotone w.r.t. a closed and convex set C if

$$\forall t, \ \forall \mathbf{w} \in C, \quad \|\mathbf{w}_{t+1} - \mathbf{w}\|_2 \leq \|\mathbf{w}_t - \mathbf{w}\|_2.$$

Immediate consequences include (more can be found in e.g. Bauschke and Borwein (1996)):

- $\{\mathbf{w}_t\}$ is bounded (hence have limit points);
- $\operatorname{dist}_C(\mathbf{w}_t) = \operatorname{dist}(\mathbf{w}_t, C) := \min_{\mathbf{w} \in C} \|\mathbf{w}_t - \mathbf{w}\|_2$ monotonically decreases;
- $\{\mathbf{w}_t\}$ has at most one limit point in C hence if all limit points are in C, then $\mathbf{w}_t$ actually converges.

Indeed, for the last claim, note that for any limit point $\mathbf{w} \in C$, we know $\|\mathbf{w}_t - \mathbf{w}\|_2^2$ hence $\frac{1}{2}\|\mathbf{w}_t\|_2^2 - \langle \mathbf{w}_t, \mathbf{w} \rangle$ converges. Thus, for limit points $\mathbf{w}, \mathbf{z} \in C$ we know $\langle \mathbf{w}_t, \mathbf{w} - \mathbf{z} \rangle \to \langle \mathbf{w}, \mathbf{w} - \mathbf{z} \rangle = \langle \mathbf{z}, \mathbf{w} - \mathbf{z} \rangle$, i.e. $\mathbf{w} = \mathbf{z}$.

Bauschke, H. H. and J. M. Borwein (1996). "On Projection Algorithms for Solving Convex Feasibility Problems". *SIAM Review*, vol. 38, no. 3, pp. 367–426.

Algorithm 15.9: Method of Alternating Projection (e.g. Bregman 1965)

Now let $C = \cap_{i\in I} C_i \neq \emptyset$. Suppose $\mathbf{w}_0 \notin C$ (otherwise we are done). Then there exists some $C_i \not\ni \mathbf{w}_0$. Apply the constructive part of Fejér's Theorem 15.6 by letting

$$\mathbf{w}_1 = \mathrm{P}_{C_i}(\mathbf{w}_0),$$

we immediately have

$$\forall \mathbf{w} \in C_i \supseteq C,\ \|\mathbf{w} - \mathbf{w}_1\|_2 < \|\mathbf{w} - \mathbf{w}_0\|_2.$$

Iterating the above idea leads to the method of alternating projections:

Algorithm: Method of alternating projections

Input: $\mathbf{w}_0$

for $t = 0, 1, \ldots$ **do**

  choose set C_{i_t} `// see Remark 15.10 for choices`

  $\mathbf{w}_{t+1} \leftarrow (1-\eta_t)\mathbf{w}_t + \eta_t \mathrm{P}_{C_{i_t}}(\mathbf{w}_t)$ `//` $\eta_t \in [0, 2]$

This algorithm has a long history, see Agmon (1954) and Motzkin and Schoenberg (1954) for early analysis when each C_i is a halfspace. See also Goffin (1980, 1982), Mandel (1984), Spingarn (1985, 1987), and Garcoä-Palomares (1993).

Clearly, we have for any $\mathbf{w} \in C$:

$$\begin{aligned}
\|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2 &= \|\mathbf{w}_t - \mathbf{w} - \eta_t(\mathbf{w}_t - \mathrm{P}_{C_{i_t}}(\mathbf{w}_t))\|_2^2 \\
&= \|\mathbf{w}_t - \mathbf{w}\|_2^2 + (\eta_t^2 - 2\eta_t)\|\mathbf{w}_t - \mathrm{P}_{C_{i_t}}(\mathbf{w}_t)\|_2^2 + 2\eta_t \left\langle \mathbf{w} - \mathrm{P}_{C_{i_t}}(\mathbf{w}_t), \mathbf{w}_t - \mathrm{P}_{C_{i_t}}(\mathbf{w}_t)\right\rangle \\
\text{(Theorem 3.12)}\quad &\leq \|\mathbf{w}_t - \mathbf{w}\|_2^2 + (\eta_t^2 - 2\eta_t)\|\mathbf{w}_t - \mathrm{P}_{C_{i_t}}(\mathbf{w}_t)\|_2^2 \qquad (15.2) \\
(\ \eta_t \in [0,2]\ )\quad &\leq \|\mathbf{w}_t - \mathbf{w}\|_2^2,
\end{aligned}$$

i.e. the generated sequence is Fejér monotone, hence explaining the restriction $\eta_t \in [0, 2]$.

Bregman, L. M. (1965). "The method of successive projection for finding a common point of convex sets". *Soviet Mathematics Doklady*, vol. 162, no. 3, pp. 688–692.

Agmon, S. (1954). "The Relaxation Method for Linear Inequalities". *Canadian Journal of Mathematics*, vol. 6, pp. 382–392.

Motzkin, T. S. and I. J. Schoenberg (1954). "The Relaxation Method for Linear Inequalities". *Canadian Journal of Mathematics*, vol. 6, pp. 393–404.

Goffin, J. L. (1980). "The relaxation method for solving systems of linear inequalities". *Mathematis of Operations Research*, vol. 5, no. 3, pp. 388–414.

— (1982). "On the non-polynomiality of the relaxation method for systems of linear inequalities". *Mathematical Programming*, vol. 22, pp. 93–103.

Mandel, J. (1984). "Convergence of the cyclical relaxation method for linear inequalities". *Mathematical Programming*, vol. 30, pp. 218–228.

Spingarn, J. E. (1985). "A primal-dual projection method for solving systems of linear inequalities". *Linear Algebra and its Applications*, vol. 65, pp. 45–62.

— (1987). "A projection method for least-squares solutions to overdetermined systems of linear inequalities". *Linear Algebra and its Applications*, vol. 86, pp. 211–236.

Garcoä-Palomares, U. (1993). "Parallel Projected Aggregation Methods for Solving the Convex Feasibility Problem". *SIAM Journal on Optimization*, vol. 3, no. 4, pp. 882–900.

Remark 15.10: Update order

The following choices for the update order are often used:

- Cyclic: when $|I| < \infty$, we simply set $i_t = t \mod |I|$, i.e., project to each set C_i cyclically.
- Almost cyclic: $\exists B \geq |I|$, so that for all t, $I \subseteq \{i_t, i_{t+1}, \ldots, i_{t+B-1}\}$, i.e. each set is chosen at least once every B iterations.

- Greedy: we can instead choose the furthest set:

$$i_t = \operatorname*{argmax}_{i \in I} \operatorname{dist}(\mathbf{w}_t, C_i),$$

  where ties are broken arbitrarily. In fact, a multiplicative approximation suffices. This choice is particularly useful when the index set $|I| = \infty$.

- Random: when $|I| < \infty$ choose $i_t \in I$ randomly.
- Permutation: in each epoch, randomly permute the sets and then go cyclic.
- Infinite often: make sure each $i \in I$ is chosen infinitely often (which is clearly necessary).

Theorem 15.11: Convergence of alternating projections (Bregman 1965; Gubin et al. 1967)

Let $C = \cap_{i \in I} C_i \neq \emptyset$ where each C_i is closed and convex and $|I| < \infty$. If $0 < \alpha \leq \eta_t \leq 2 - \beta < 2$ for some $\alpha, \beta > 0$, then with the cyclic update order we have

$$\mathbf{w}_t \to \mathbf{w}_\star \in C.$$

Proof: We only prove the case for $\eta_t \equiv 1$.

Let $\mathbf{z}_{k,i} = \mathbf{w}_{k|I|+i}$. Consider any converging subsequence of $\mathbf{w}_t$. Since $|I| < \infty$, we may assume w.l.o.g. the subsequence is contained in $\mathbf{z}_{k,1}$ and has a limit point $\mathbf{w}_\star$. Clearly $\mathbf{w}_\star \in C_1$ since $\mathbf{z}_{k,1} \in C_1$ and C_1 is closed. From (15.2) we know for any $\mathbf{w} \in C$:

$$\|\mathbf{z}_{k,1} - \mathbf{z}_{k,2}\|_2 \leq \|\mathbf{z}_{k,1} - \mathbf{w}\|_2^2 - \|\mathbf{z}_{k,2} - \mathbf{w}\|_2^2 \to 0.$$

Thus, $\mathbf{w}_\star \leftarrow \mathbf{z}_{k,2} \in C_2$. Continuing the same argument we conclude $\mathbf{w}_\star \in \cap_i C_i = C$. Since any limit point of the Fejér monotone sequence $\{\mathbf{w}_t\}$ is in C we know $\mathbf{w}_t \to \mathbf{w}_\star \in C$. ■

Bregman, L. M. (1965). "The method of successive projection for finding a common point of convex sets". *Soviet Mathematics Doklady*, vol. 162, no. 3, pp. 688–692.

Gubin, L. G., B. T. Polyak, and E. V. Raik (1967). "The Method of Projections for Finding the Common Point of Convex Sets". *USSR Computational Mathematics and Mathematical Physics*, vol. 7, no. 6, pp. 1–24. [English translation of paper in *Zh. Vȳchisl. Mat. mat. Fiz.* vol. 7, no. 6, pp. 1211–1228, 1967].

Algorithm 15.12: Alternating Bregman Projection (e.g. Bregman 1966)

Instead of the Euclidean projection, let us now consider the Bregman projection

$$\mathbb{P}_C(\mathbf{z}) = \mathbb{P}_{C,h}(\mathbf{z}) = \operatorname*{argmin}_{\mathbf{w} \in C} \mathsf{D}_h(\mathbf{w}, \mathbf{z}),$$

where $h : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is a Legendre function (see Definition 8.7).

Algorithm: Alternating Bregman projection

Input: $\mathbf{w}_0$, $\operatorname{dom} h \supseteq C$

for $t = 0, 1, \ldots$ **do**

  choose set C_{i_t} // see Remark 15.10 for choices

  $\mathbf{w}_{t+1} \leftarrow (1 - \eta_t)\mathbf{w}_t + \eta_t \mathbb{P}_{C_{i_t}}(\mathbf{w}_t)$ // $\eta_t \in [0, 2]$

Convergence for $\eta_t \equiv 1$ was shown in Bregman (1966).

Bregman, L. M. (1966). "A relaxation method of finding a common point of convex sets and its application to problems of optimization". *Soviet Mathematics Doklady*, vol. 171, no. 5, pp. 1578–1581.

Remark 15.13: A primal-dual view (Bregman 1967)

So far, the alternating projection algorithms allow us to converge to an arbitrary point in C. Quite remarkably, Bregman (1967) observed that a primal-dual modification actually allows us to solve:

$$\begin{aligned} \min_{\mathbf{w}\in\mathbb{R}^d} \ & f(\mathbf{w}) \\ \text{s.t.} \ & A\mathbf{w} \le \mathbf{b}, \end{aligned}$$

where f is a Legendre function (see Definition 8.7) and we define $C_i := \{\mathbf{w} : \langle \mathbf{w}, \mathbf{a}_i\rangle \le b_i\}, i = 1, \dots, n,$. The idea is to introduce the KKT set

$$\mathbb{K} = \{(\mathbf{w}, \mathbf{u}) \in \operatorname{dom} f \times \mathbb{R}^n_+ : \nabla f(\mathbf{w}) + A^\top \mathbf{u} = \mathbf{0}\}$$

and maintain $(\mathbf{w}_t, \mathbf{u}_t) \in \mathbb{K}$ explicitly. Indeed, we start with $\mathbf{w}_0 \in \operatorname{argmin} f$ so that $\nabla f(\mathbf{w}_0) = \mathbf{0} =: \mathbf{u}_0$. Upon choosing C_i, we conduct one of the following updates:

- If $\langle \mathbf{w}_t, \mathbf{a}_i\rangle > b_i$, we compute

$$\begin{aligned} &\mathbf{w}_{t+1} \leftarrow \mathbb{P}_{C_i,f}(\mathbf{w}_t) := \operatorname*{argmin}_{\mathbf{w}\in C_i} \mathsf{D}_f(\mathbf{w}, \mathbf{w}_t), \ \textit{i.e.}, \\ &\exists u_{t+1,i} \ge u_{t,i} \quad \text{s.t.} \quad \begin{cases} \nabla f(\mathbf{w}_{t+1}) + u_{t+1,i}\mathbf{a}_i = \nabla f(\mathbf{w}_t) + u_{t,i}\mathbf{a}_i \\ \langle \mathbf{w}_{t+1}, \mathbf{a}_i\rangle = b_i \end{cases}. \end{aligned}$$

- If $\langle \mathbf{w}_t, \mathbf{a}_i\rangle = b_i$, or $\langle \mathbf{w}_t, \mathbf{a}_i\rangle < b_i$ with $u_{t,i} = 0$, then continue.
- If $\langle \mathbf{w}_t, \mathbf{a}_i\rangle < b_i$ with $u_{t,i} > 0$, we compute

$$\begin{aligned} &\mathbf{w}_{t+1} \leftarrow \operatorname*{argmin}_{\mathbf{w}\in C_i} f(\mathbf{w}) - \langle \mathbf{w}, \nabla f(\mathbf{w}_t) + u_{t,i}\mathbf{a}_i\rangle, \ \textit{i.e.}, \\ &\exists u_{t+1,i} \in [0, u_{t,i}] \quad \text{s.t.} \quad \begin{cases} \nabla f(\mathbf{w}_{t+1}) + u_{t+1,i}\mathbf{a}_i = \nabla f(\mathbf{w}_t) + u_{t,i}\mathbf{a}_i \\ u_{t+1,i}(\langle \mathbf{w}_{t+1}, \mathbf{a}_i\rangle - b_i) = 0 \end{cases}. \end{aligned}$$

It is clear that in all cases we maintain $(\mathbf{w}_t, \mathbf{u}_t) \in \mathbb{K}$ if we start so. In a later lecture we will see how these updates can be derived naturally.

Bregman, L. M. (1967). "The Relaxation Method of Finding the Common Point of Convex Sets and Its Application to the Solution of Problems in Convex Programming". *USSR Computational Mathematics and Mathematical Physics*, vol. 7, no. 3, pp. 200–217. [English translation in *Zh. Vȳchisl. Mat. mat. Fiz.* vol. 7, no. 3, pp. 620–631, 1967].

Algorithm 15.14: Dykstra's algorithm (Dykstra 1983)

We now present a beautiful algorithm for solving:

$$\begin{aligned} \min_{\mathbf{w}} \ & f(\mathbf{w}) \\ \text{s.t.} \ & \mathbf{w} \in C \ne \emptyset, \quad C := \cap_{i\in I} C_i, \end{aligned}$$

where f is Legendre and each C_i is closed and convex. We have seen an algorithm in Remark 15.13 for the case where each C_i is a half-space. On the other hand, the case with $f = \mathbf{q}$ (quadratic) but general C_i was dealt with by Dykstra (1983) and later rediscovered by Han (1988, 1989) and Gaffke and Mathar (1989). We present a unification due to Bregman et al. (1999).

The idea is extremely simple: we simply linearize each convex set C_i by including it in a supporting half-space and then apply Remark 15.13.

Algorithm: Dykstra's algorithm

Input: $\mathbf{w}_0 = \operatorname{argmin} f$, $\mathbf{a}_i = \mathbf{0}, b_i = 0$ for all $i \in I$

for $t = 0, 1, \dots$ **do**

- choose set C_{i_t} `// see Remark 15.10 for choices`
- $\mathbf{w}_{t+1} \leftarrow \operatorname{argmin}_{\mathbf{w} \in C_{i_t}} f(\mathbf{w}) - \langle \mathbf{w}, \nabla f(\mathbf{w}_t) + \mathbf{a}_{i_t} \rangle$ `// Bregman projection`
- $\mathbf{a}_{i_t} \leftarrow \mathbf{a}_{i_t} + \nabla f(\mathbf{w}_t) - \nabla f(\mathbf{w}_{t+1})$
- $b_{i_t} \leftarrow \langle \mathbf{a}_{i_t, t+1}, \mathbf{w}_{t+1} \rangle$ `// needed only for proof`

Indeed, from the optimality condition of $\mathbf{w}_{t+1}$ we obtain:

$$\forall \mathbf{w} \in C_{i_t},\ \langle \nabla f(\mathbf{w}_{t+1}) - \nabla f(\mathbf{w}_t) - \mathbf{a}_{i_t}, \mathbf{w} - \mathbf{w}_{t+1} \rangle \geq 0,\ \textit{i.e.}\ \langle \mathbf{a}_{i_t, t+1}, \mathbf{w} \rangle \leq b_{i_t},$$

and hence the convex set C_{i_t} is contained in the half-space $\mathsf{H}_{i_t} := \{\mathbf{w} : \langle \mathbf{a}_{i_t}, \mathbf{w} \rangle \leq b_{i_t}\}$.

Dykstra, R. L. (1983). "An Algorithm for Restricted Least Squares Regression". *Journal of the American Statistical Association*, vol. 78, no. 384, pp. 837–842.

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.

Bregman, L. M., Y. Censor, and S. Reich (1999). "Dykstra's Algorithm as the Nonlinear Extension of Bregman's Optimization Method". *Journal of Convex Analysis*, vol. 6, no. 2, pp. 319–333.

Exercise 15.15: Entropy-regularized optimal transport

Let $\mathsf{p} \in \Delta_m$ and $\mathsf{q} \in \Delta_n$ be two probability vectors, and we seek a joint coupling (distribution) $\Pi \in \mathbb{R}_+^{m \times n}$ with p and q as marginals such that the transportation cost is minimized:

$$\begin{aligned} \min_{\Pi \in \mathbb{R}_+^{m \times n}} \quad & \langle C, \Pi \rangle \\ \text{s.t.} \quad & \Pi \mathbf{1} = \mathsf{p}, \quad \Pi^\top \mathbf{1} = \mathsf{q}. \end{aligned}$$

While the (discrete) optimal transport problem above can be solved using linear program, we gain a much more scalable algorithm if we add a small entropy regularization:

$$\begin{aligned} \min_{\Pi \in \mathbb{R}_+^{m \times n}} \quad & \langle C, \Pi \rangle + \lambda \sum_{ij} \pi_{ij} \log \pi_{ij} \\ \text{s.t.} \quad & \Pi \mathbf{1} = \mathsf{p}, \quad \Pi^\top \mathbf{1} = \mathsf{q}. \end{aligned}$$

W.l.o.g. we may assume $\Pi_0 \propto \exp(-C/\lambda) \geq \mathbf{0}$ and $\mathbf{1}^\top \Pi_0 \mathbf{1} = 1$. Prove that we have the equivalent problem:

$$\begin{aligned} \min_{\Pi \in \mathbb{R}_+^{m \times n}} \quad & \mathsf{KL}(\Pi \| \Pi_0) \\ \text{s.t.} \quad & \Pi \mathbf{1} = \mathsf{p}, \quad \Pi^\top \mathbf{1} = \mathsf{q}. \end{aligned} \tag{15.3}$$

Can you adapt Dykstra's Algorithm 15.14 to solve (15.3)?